Teaching *with* the Test: Experimental Evidence on Diagnostic Feedback and Capacity Building for Public Schools in Argentina[*]

Rafael de Hoyos[†]
World Bank

Alejandro J. Ganimian[‡]
New York University

Peter A. Holland[§]
World Bank

July 8, 2019

Abstract

We examine the impact of two strategies to use large-scale assessment results to improve school management and classroom instruction in the Province of La Rioja, Argentina. We randomly assigned 104 public primary schools to: a diagnostic-feedback group, in which we administered standardized tests at baseline and two follow-ups and made results available to schools; a capacity-building group, in which we also conducted workshops and school visits; or a control group, in which we administered tests at the second follow-up. After two years, diagnostic-feedback schools outperformed control schools by $.33\sigma$ in math and $.36\sigma$ in reading. In fact, feedback schools still performed $.26\sigma$ better in math and $.22\sigma$ in reading in the national assessment a year after the end of the intervention. Additionally, principals at these schools were more likely to use assessment results for management decisions and students were more likely to report that their teachers used more instructional strategies and rated them more favorably. Combining feedback with capacity building does not seem to lead to additional improvements, but this might be due to schools assigned to receive both components starting from lower learning levels and participating in fewer workshops and visits than expected.

JEL codes: C93, I21, I22, I25.

Keywords: Diagnostic feedback, capacity building, large-scale assessments, Argentina.

[*] We gratefully acknowledge funding provided by the World Bank for this study. We thank Rita Abdala, Walter Flores, Silvia Romero, and Juan Vegas at the Ministry of Education of La Rioja, and Eduardo Cascallar, Jorge Fasce, and Gustavo Iaies at the Center for Studies of Public Policies for making this study possible. María Cortelezzi, Alvina Erman, Bárbara Funtowicz, Andrés Felipe Pérez, Melissa Rofman, María José Vargas, and Guillermo Toral assisted with project management. Finally, we thank Larry Aber, Felipe Barrera-Osorio, Elise Cappella, Joe Cimpian, Sean Corcoran, Andy de Barros, Emmerich Davies, Andrew Ho, Jimmy Kim, Dan Koretz, Isaac Mbiti, Dick Murnane, Jeff Puryear, Jonah Rockoff, Jennifer Steele, Hiro Yoshikawa, and seminar participants at APPAM, FRA, HGSE, NEUDC, NYU, and the World Bank for useful comments. All views expressed are those of the authors and not of any of the institutions with which they are affiliated.

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

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

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

1 Introduction

Over the past decade, developing countries have shifted their attention from expanding access to schooling to ensuring children acquire basic skills at school.[1] This emerging consensus has led a growing number of low- and middle-income countries to administer large-scale student assessments and to participate in international assessments. According to one mapping effort, 85 national school systems have conducted 306 assessments of math, reading, and science since 2004 (Cheng and Gale 2014). A similar effort found that 328 national and sub-national school systems have participated in 37 international assessments of the same subjects from 1963 to 2015; nearly half of them began participating since 1995 (Ganimian and Koretz 2017).[2]

Yet, in spite of the exponential growth of student assessments among the developing world, we still know surprisingly little about whether (and if so, how) governments in these settings can leverage the results of these tests to improve school management and classroom instruction. Most prior studies have either focused on how to use national assessments for accountability purposes, such as nudging parents to send their children to better-performing schools (see, for example, Andrabi et al. 2017; Camargo et al. 2018; Mizala and Urquiola 2013), or on how to use classroom assessments to inform the implementation of differentiated or scripted instruction (see Banerjee et al. 2011; Bassi et al. 2016; Duflo et al. 2015; Piper and Korda 2011). The few studies that evaluate the impact of using large-scale assessments for formative purposes reach conflicting conclusions (see de Hoyos et al. 2017; Muralidharan and Sundararaman 2010).[3]

This paper presents experimental evidence on two strategies to use large-scale assessment results for improvement in a developing country: diagnostic feedback and capacity building. We randomly assigned 104 public primary schools in the Province of La Rioja, Argentina to: (a) a diagnostic-feedback (T1) group, in which we administered standardized tests in math and reading comprehension at baseline and two follow-ups and made results available to schools through user-friendly reports; (b) a capacity-building (T2) group, in which we also conducted professional development workshops and school visits for supervisors, principals, and teachers; or (c) a control group, in which we only administered the tests at the second follow-up. This setup allows us to understand whether disseminating assessment results to schools is enough

[1] In the Millennium Development Goals, adopted by the United Nations General Assembly in 2000, 191 countries pledged to ensure that "by 2015, children everywhere, boys and girls alike will be able to complete a full course of primary schooling" (UNGA 2000). In the Sustainable Development Goals, adopted in 2015, 194 countries set a new target: "by 2030... all girls and boys [should] complete free, equitable, and *quality* primary and secondary education learning to relevant and effective *learning outcomes*" (UNGA 2015, emphasis added).

[2] These figures are likely to increase as testing agencies develop assessments for low-income countries, which have been reluctant to join existing global tests (e.g., the Organization for Economic Cooperation and Development's Program for International Student Assessment for Development and the International Association for the Evaluation of Educational Achievement's Literacy and Numeracy Assessment).

[3] This question has received more attention in developed countries (see, for example, Betts et al. 2017). Yet, it is not clear that the lessons from these studies apply to developing countries, which have both less technically solid assessments and lower capacity to analyze and disseminate their results.

to prompt improvements in management and instruction or whether principals and teachers need additional support to understand and act on this information.

We report three sets of results based on this experiment. First, we find that simply providing schools with information on how their students perform relative to the rest of the province, with minimal capacity building, led to large improvements in student learning after two years: T1 schools outperformed control schools by $.33\sigma$ in math and $.36\sigma$ in reading. Importantly, this intervention led to improvements in nearly all content and cognitive domains assessed in both subjects, not just in those that are easier to improve in the short-term, suggesting that it did not result in a narrowing of the curriculum, or what is known as "teaching to the test." Further, the intervention led to improvements on both items that were common across rounds and on new items, indicating that test-score gains are not driven by familiarity with the test. The bounds on the effects range from moderately negative to very large positive effects. Yet, we also find that students who were exposed to the intervention for two years still outperformed control peers by $.26\sigma$ in math and $.22\sigma$ in reading in the national student assessment a year after the end of the experiment, indicating improvements are not test-dependent or short-lived.

Second, consistent with these effects, we find that diagnostic feedback also led to changes in school management and classroom instruction. Principals in T1 schools were more likely than their control counterparts to report using assessment results to inform management decisions (e.g., making changes to the curriculum, appraising teachers' effectiveness, informing parents about their children's results, and making results public).[4] Students in T1 schools were more likely than their control peers to report that their teachers used more instructional strategies (e.g., using textbooks, assigning homework, writing on the blackboard, and explaining topics).[5] They also rated their teachers more positively on all domains of a widely used student survey (e.g., demonstrating interest in students, managing the classroom, clarifying difficult concepts and tasks, challenging students to do their best, delivering engaging lessons, engaging students in discussions, and summarizing the material at the end of every lesson). Diagnostic feedback did not, however, improve teacher attendance or punctuality.

Finally, we do not find evidence that combining information provision with workshops and school visits leads to additional improvements in school management, instruction, or learning. In fact, for most outcomes and potential mechanisms, we do not find a statistically significant impact of T2 and we cannot discard the possibility that T1 and T2 had the same effect. Yet, these results should be interpreted with caution. First, it is possible that schools randomly assigned to T2 were already at a disadvantage with respect to those assigned to T1 at baseline. T1 and T2 had similar internal efficiency and student characteristics, but T2 schools performed

[4] Interestingly, however, these principals were no more likely to use test scores to assign students to sections, suggesting that the intervention did not increase segregation of educational opportunities within schools.

[5] Yet, these students were no more likely to report that their teachers made them complete practice tests, providing further evidence that the intervention did not lead to test "coaching."

slightly below T1 schools on the assessments. The gap was small and statistically insignificant, but it might explain why T2 did not have an effect that was at least as large as that of T1.[6] Second, participation in workshops and visits among T2 schools was lower than expected, and a few T1 schools participated in these activities even if they were not supposed to do so. The low take-up and contamination may have resulted in most T2 schools not receiving much more support than T1 schools, preserving the initial imbalance between these two groups.[7]

Our results contribute to the impact evaluation literature on the use of large-scale assessments in developing countries in at least two ways. First, they demonstrate that it is possible to use assessment results to improve student learning without having to attach stakes to them. This is important because, even when the high-stakes use of assessments has worked (e.g., Muralidharan 2012; Muralidharan and Sundararaman 2011), it has faced political opposition. Second, they offer a potential explanation for the mixed results of prior feedback interventions. One possibility is that feedback has less scope for impact in settings where the binding constraint is the extensive margin of principal and teacher effort (i.e., improving attendance) (Muralidharan and Sundararaman 2010), but holds promise where the constraint is the intensive effort margin (raising productivity, conditional on attendance) (de Hoyos et al. 2017).

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

2 Experiment

2.1 Context

Schooling in Argentina is compulsory from the age of 4 until the end of secondary school. In 12 of the country's 24 provinces, including the Province of La Rioja, primary school runs from grades 1 to 7 and secondary school runs from grades 8 to 12 (DiNIECE 2013).[8] According to the latest official figures, the Argentine school system serves 11.1 million students: 1.7 million in pre-school, 4.5 million in primary, and 3.9 million in secondary school (DiNIEE 2015).

Argentina achieved near-universal access to primary education before most of Latin America (see Bassi et al. 2013). Yet, the relative performance of Argentina's primary school students

[6] We present two results consistent with this possibility. First, we only find a statistically significant effect of T2 on grade 5, where the initial difference between T1 and T2 schools was smaller. Second, the coefficients on T2 become larger when we account for baseline covariates in grade 3, where the initial difference was larger.

[7] We present data from principal surveys and intervention monitoring that verify that T1 and T2 schools received similar levels of support.

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

in the region has deteriorated. In 1997, on the first regional student assessment in primary school, Argentine third graders had ranked second in math, after their Cuban counterparts. In 2013, on the third regional assessment, they ranked seventh—on par with their peers in Peru and Ecuador, who had ranked near the bottom in 1997 and 2006 (Ganimian 2014).[9]

Education policy in Argentina is shaped by both the national and the sub-national (province) governments. According to the National Education Law of 2006, the federal government is responsible for higher education and for providing technical and financial assistance to the provinces, and the provincial governments for pre-primary, primary, and secondary education. The Ministry of Education, Culture, Science, and Technology at the national level is also tasked with coordinating the national student assessment (formerly known as the *Operativo Nacional de Evaluación* and currently as *Aprender*), which has been in place since 1993, in conjunction with its counterparts in each province. To our knowledge, only the Autonomous City of Buenos Aires conducts its own sub-national student assessment regularly.

Argentina is an interesting setting to evaluate the impact of leveraging student assessments for diagnostic feedback and capacity building for schools. Over the past two decades, the country has taken multiple steps that limited the generation, dissemination, and use of student achievement data: (a) it reduced the frequency of its national assessment from an annual basis (in 1999-2000), to a biennial basis (in 2002-2007), to a triennial basis (in 2008-2013); (b) it prohibited by law the publication of learning outcomes disaggregated at the student-, teacher-, or school-level; and (c) in 2013, it discontinued the publication of assessment results at the province level, publishing them instead at the regional level (Ganimian 2015). These policies have stood in stark contest with those of other Latin American countries (e.g., Brazil, Chile, Colombia, Mexico, and Peru), which have technically robust and long-standing assessments and use them for multiple purposes (Ferrer 2006; Ferrer and Fiszbein 2015).

In recent years, a new government has reversed some of these policies: (a) it adopted a new national assessment, to be administered annually, cover all students at the end of primary and secondary (grades 7 and 12), and a sample halfway through each level (grades 3 and 8), and assess math, reading, and natural and social sciences (SEE-MEDN 2016);[10] and (b) it started sharing school-level assessment results to all principals using reports that resemble the ones evaluated in this paper.[11] Therefore, the questions that we examine in this paper are not only of general interest to developing countries, but also of specific interest to Argentina.

We conducted our study in La Rioja for three reasons. First, it is one of the lowest-performing provinces in Argentina. In the latest national assessment, 41% of its sixth graders performed at

[9] The 1997 and 2006 assessments are not strictly comparable, but no other country participating in both assessments has changed its ranking so radically. Further, the relative standing of Argentina's secondary school students has also deteriorated over the same period (de Hoyos et al. 2015; Ganimian 2013).

[10] In 2017, the national government began alternating the grades and subjects assessed each year.

[11] A Spanish version of the report can be accessed at: `https://bit.ly/2MfSpLu`.

the two lowest levels in math and 53% in reading (SEE-MEDN 2018). Second, it is one of the smallest sub-national school systems in the country, so it is better positioned to implement a quality assurance mechanism. It is the seventh-smallest system in schools (377 primary schools) and the fourth-smallest in students (41,571 primary school students) (DiNIEE 2015). Third, it was one of the few provinces with the political will to experiment with a sub-national assessment. The assessment was endorsed by the governor and the minister of education.

2.2 Sample

The sampling frame for the study included all 126 public primary schools in urban and semi-urban areas of La Rioja.[12] We arrived at this frame as follows. First, of the 394 primary schools in the province, we excluded all 29 private schools because we were interested in the effect of the interventions on public schools. Then, we dropped all 239 schools in rural areas because they are spread across the province, which would have limited our capacity to implement the interventions.[13] We drew a random sample of 104 urban and semi-urban public primary schools from this frame, stratified by enrollment tercile in the 2013 school year.

The schools in our sample are comparable to all public primary schools in the province, and even more so to other urban and semi-urban public primary schools (Table A.1, Appendix A). The average school in our sample enrolls more students, but this is mostly because we excluded rural schools, which are typically smaller. We do not find any statistically significant difference between in- and out-of-sample schools in the internal efficiency indicators collected by the school system, including: passing, failure, dropout, and repetition rates.

We sampled students and teachers to obtain *cross-sectional* information in grades 3 and 5 every year, as well as *longitudinal* information on the students who started grade 3 in 2013. Thus, in 2013, all students and teachers from grades 3 and 5 participated; in 2014, all students and teachers from grades 3 *to* 5 participated; and in 2015, all students and teachers from grades 3 and 5 participated. All principals in selected schools participated in the study.

2.3 Randomization

We randomly assigned the 104 public primary schools in our sample to: (a) a "diagnostic feedback" (T1) group, in which we administered student assessments at baseline and two follow-ups and made results available to schools through user-friendly reports; (b) a "capacity

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

[13] It is worth noting, however, that while rural schools account for a large share of the total number of public schools in La Rioja (65% of the total), they serve a small share of the students (less than 10%).

building" (T2) group, in which we also provided schools with professional development workshops and school visits for supervisors, principals, and teachers; or (c) a control group, in which we only administered the assessments at endline.[14] We stratified our randomization by terciles of enrollment in primary school to increase statistical power.[15]

This setup allows us to estimate the effect of: (a) diagnostic feedback (comparing T1 to control schools in 2015); (b) combining diagnostic feedback with capacity building (comparing T2 to control schools in 2015); and (c) the value-added of capacity building, over and above diagnostic feedback (comparing T1 to T2 schools in 2014 and 2015).

2.4 Interventions

Table 1 shows the timeline for the interventions and rounds of data collection for the study. The school year in Argentina starts in February and ends in December. As the table shows, we administered the student assessments at the end of each year (only in T1 and T2 schools in 2013 and 2014 and in all schools in 2015). We delivered school reports based on the prior-year assessments in T1 and T2 schools at the start of each year (in 2014 and 2015). We conducted workshops and school visits for T2 schools during each school year (in 2014 and 2015).

2.4.1 Diagnostic-feedback (T1) group

The diagnostic feedback intervention provided schools with reliable, timely, and actionable data on student learning outcomes to inform school management and classroom instruction. At the beginning of each year, and for two consecutive years (2014 and 2015), schools randomly assigned to the T1 group received reports that summarized the results of student assessments of math and reading administered at the end of the previous year.[16]

The reports were brief (10 pages) and had four sections: (a) an introduction, which described the assessments and reported the percentage of students at the school who completed them; (b) an overview of the school's average performance, which included the school's average score in each grade and subject, the change in each score from the previous year, and comparisons between the school's scores and those of the average school in the area and in the province;[17] (c) an analysis of the distribution of the school's performance, which includeda box-and-whiskers

[14] We discuss our rationale for this decision in the next section.

[15] We conducted the randomization in June of 2013. T1 and T2 schools were informed that they would be part of the study in August of that year, but control schools were not disclosed until right before endline to minimize John Henry effects (i.e., schools seeking to compensate for not receiving an intervention).

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

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

plot for the school and the province for each grade and subject; and (d) a "traffic light" display of the school's performance on each item of the assessments for each each grade and subject.[18]

As Table 1 reveals, some T1 schools participated in the workshops and visits designed for T2 schools (described below). Therefore, our estimate of the causal effect of the T1 intervention should be interpreted as the effect of diagnostic feedback with minimal capacity building.

2.4.2 Capacity building (T2) group

The capacity building intervention provided supervisors, principals, and teachers with support to understand and make decisions based on the student learning outcomes in the school reports. During the school year, and for two consecutive years (2014 and 2015), schools randomly assigned to the T2 group were offered the same reports as the T1 group, five workshops (three in 2014, two in 2015), and two school visits (one per year).

Two of the workshops explained the assessment results after each round of delivery of the reports, one focused on how to develop school improvement plans, one on how to institute quality assurance mechanisms at the school level, and one on geometry instruction. The first four workshops were offered to supervisors and principals and the last one to teachers. Each school visit entailed a meeting with the principal and his/her leadership team, a classroom observation, and a meeting with the teaching staff. The workshops and visits were conducted by the ministry of education of the province, in collaboration with a local think tank. After each visit, the ministry prepared and shared a report with the school, which included a diagnosis and some recommendations for improvement.[19]

As Table 1 shows, participation in the workshops and school visits was lower than expected. In section 5.2, we use the endline data to discuss variation in dosage across T2 schools.

2.4.3 Control group

Control schools were only assessed in 2015, at the end of the last year of the study. Student assessments were rare in La Rioja,[20] so administering assessments in 2013 and 2014 could have prompted behavioral responses from principals, teachers, and students that would not have accurately represented business-as-usual school management and classroom instruction. Thus, following Muralidharan and Sundararaman (2010), we only administered the endline assessments at these schools to estimate the impact of the interventions after two years.

[18] An English version of the report can be accessed at: `http://bit.ly/2xrRaoc`.

[19] The workshops and school visits were based on the recommendations in Boudett et al. (2005).

[20] The last national student assessment had been conducted in a sample of primary schools in 2010. There had not been any census-based national assessments in primary schools in the province since 2000. The province had never administered sub-national assessments (Ganimian 2015).

3 Data

As Table 1 indicates, we administered student assessments of math and reading and student and teacher surveys in T1 and T2 schools on each year of the study (from 2013 to 2015). We also conducted principal surveys at the end of each intervention year (in 2014 and 2015). We only administered the assessments in control schools at the end of the study (in 2015).[21] Additionally, we obtained internal efficiency data on all schools in our study (for 2013 to 2017) and learning outcomes data from the national student assessment (for 2016).

3.1 Student assessments

We administered student assessments of math and reading before the interventions (in 2013) and after one and two years of the interventions (in 2014 and 2015) in T1 and T2 schools. We only administered the assessments in control schools at the end of the second year (in 2015).[22]

The assessments evaluated what students ought to know and be able to do according to: (a) the national curriculum (*Contenidos Básicos Comunes*); (b) the topics of the curriculum that the national government has identified as priorities (*Núcleos de Aprendizaje Prioritario*); and (c) the curriculum of the province (*Diseño Curricular de La Rioja*).[23] Specifically, the math assessment covered three content domains (numbers, geometry, measurement, and probability and statistics) and four cognitive domains (identifying mathematical concepts, understanding and using symbolic math, performing calculations, and solving abstract and applied problems). The reading assessment covered three content domains (narrative, informative, and short texts) and four cognitive domains (locating information in texts, understanding relationships between parts of texts, identifying the main idea of texts, and interpreting the meaning of words from context). Each assessment included 30 to 35 multiple-choice items.

We used a two-parameter Item Response Theory model to scale the assessment results in a way that accounts for differences between items (their difficulty, capacity to distinguish between students of similar ability, and propensity to be answered correctly by guessing) and to leverage common items across data collection rounds to link the results over time.

[21] The grades assessed on each year are shown in Table 1.

[22] Some students may be missing test scores because they were absent on the day of the assessments, they were present but excused from the tests, they dropped out of school, or they transferred to another school. However, we find no evidence of treatment schools having more students than control schools on such days; in fact, for 2015, we observe the opposite and all differences are below 5 percentage points (see Table A.2). The dropout rate in our sample is extremely low and varies little across experimental groups (Tables A.3 and A.15). We do not have school-level data on transfers or students who were present and excluded on test day, but we have no reason to believe that either occurred frequently or differentially across groups.

[23] Therefore, the school reports, which were based on these assessments, were aligned with the national and sub-national curricular requirements.

Appendix B provides further details on the design, scaling, and linking of the assessments, as well as the distribution of scores for all subjects, grades, and years of the study.[24]

3.2 Student surveys

We also administered surveys of students in the same years and to the same groups as the student assessments. In 2013 (i.e., the year before the interventions), the surveys asked about students' demographic characteristics, home assets, schooling trajectory, and study supports to allow us to describe our study sample. In 2015 (i.e., the second year of the interventions), they asked students about their teachers' effort, as measured by the frequency of attendance, punctuality, and a set of classroom activities, and about their teachers' effectiveness, as measured by an abridged version of the Tripod survey developed by Ron Ferguson at Harvard (see, for example, Ferguson 2010, 2012; Ferguson and Danielson 2014).[25]

3.3 Teacher surveys

We conducted surveys of teachers in the same years and groups as the student assessments. In 2013, we asked about teachers' demographic characteristics, education and experience, professional development, and teaching practices to describe our study sample. In 2014 and 2015, we asked teachers about aspects that could plausibly be influenced by the interventions (e.g., monitoring and evaluation practices at their schools and job satisfaction).[26]

3.4 Principal surveys

We administered surveys of principals in T1 and T2 schools in 2014 and in all schools in 2015. In both years, we asked principals about aspects that could be affected by the interventions (e.g., management practices and resources and materials at their schools).[27]

3.5 National assessments

We obtained the results from the national student assessment (called *Aprender*) from the national ministry of education for grade 6 students one year after the interventions (2016). This is the only primary school grade for which the assessments are census-based (i.e., they

[24] The assessments are available at `https://bit.ly/2MRudPZ` (2013), `https://bit.ly/2Gf6it4` (2014), and `https://bit.ly/2DXnhxD` (2015).

[25] The surveys are available at `http://bit.ly/1qZeYHC` (2013) and `http://bit.ly/1VrPBek` (2014-2015).

[26] The surveys are available at: `http://bit.ly/2OR1ni3` (2013) and `http://bit.ly/1THNgr0` (2014-2015).

[27] The survey is available at: `http://bit.ly/1TUkwyO` (2014-2015).

cover all schools and students). This is also the cohort of students that was in grade 3 (one of the grades targeted by the interventions) in 2013, the first year of our study, and received two years of the interventions. We use these data to verify whether we find evidence of fade-out.

3.6 Internal efficiency

Finally, we also obtained data on schools' internal efficiency (including enrollment and passing, failure, repetition, and dropout rates) from the national ministry of education for the year prior to the interventions (2013), the two years of the interventions (2014 and 2015), and two years after the interventions. We use the 2013 data to check balance across experimental groups, the 2014 and 2015 data to estimate the impact of the interventions on internal efficiency, and the 2016 and 2017 data to verify whether we find evidence of fade-out and/or dormant effects.

4 Empirical strategy

We estimate the effect of the offer (i.e., the intent-to-treat or ITT effect) of diagnostic feedback and capacity building after two years by fitting the following model:

$$Y_{igs} = \alpha_{r(s)} + X_{gs}\gamma + T_s'\beta + \epsilon_{igs} \tag{1}$$

where Y_{igs} is the outcome of interest for student i in grade g and school s after two years of the intervention, $r(s)$ is the randomization stratum of school s and $\alpha_{r(s)}$ is a stratum fixed effect, X_{gs} is the first principal component from a principal component analysis of internal efficiency indicators for grade g and school s before the intervention,[28] and T_s is a vector of intervention indicators.[29] The parameter of interest is β, which measures the effects of each intervention relative to the control group. The null hypotheses are that the elements of β equal zero. We use cluster-robust standard errors to account for within-school correlations across students in outcomes. We also test the sensitivity of our estimates to the inclusion of X_{gs}.

We also fit several variations of this model, including: (a) two nearly identical models, in which outcomes are measured at the principal or teacher level (to estimate the impact of the interventions on school management and classroom instruction, respectively); (b) models that interact the intervention indicators with student-level covariates, such as students' sex (to estimate the heterogeneous effects of the interventions on these sub-groups of students); and (c) models in which the outcomes are measured at the student- and school-level in T1 and T2

[28] These indicators include enrollment, as well as passing, failure, dropout, and repetition rates.

[29] We cannot account for students' outcomes before the intervention because, as discussed in sections 2.3, 2.4, and 3, control schools only participated in the last round of data collection of the study.

schools after the first year of the interventions (to estimate the impact of capacity building, over and above diagnostic feedback, after one year).[30]

5 Results

5.1 Balancing checks

Control, diagnostic feedback (T1), and capacity building (T2) schools were comparable on all indicators of internal efficiency tracked by the school system in 2013 (Table A.3). This is true both when we compare schools with respect to all their primary school students and when we compare them on students in the grades targeted by the intervention (grades 3 and 5 in 2013). The direction of the differences do not systematically favor any group.[31] The magnitudes of these differences are small (less than 3.2 percentage points in all internal efficiency indicators). Only one of these differences is (marginally) statistically significant, which is less than what we would expect given that number of hypotheses that we are testing.

If we compare groups using baseline data (from 2013), T1 schools fare better than T2 schools. T1 students score higher in the math and reading assessments, and the difference is larger in grade 3 (.15σ in math, .16σ in reading) than in grade 5 (.1σ and .12σ, respectively) (Table A.4). Students at T1 schools also have marginally more educated parents, more household assets, and more study supports than their T2 counterparts (Table A.5). None of these differences are statistically significant, but as we discuss below, they may explain why we do not find a statistically significant effect of the capacity building intervention on most outcomes.

5.2 Implementation fidelity

The interventions were implemented mostly as intended (Table 2). In surveys administered right before the endline (in 2015), the principals of nearly all diagnostic feedback (T1) and capacity building (T2) schools report having administered student assessments at their schools that year (Panel A, cols. 2-3).[32] Some control principals also reported administering assessments, but given that there were no national assessments or any other sub-national assessments in 2015, we believe that these principals either thought that they were supposed

[30] In these models, we cannot account for students' outcomes before the intervention because neither students nor teachers were assigned unique identifiers to allow us to track them over time.

[31] For example, T1 schools have lower passing rates than control schools (which suggests that they fare worse), but they also have lower dropout rates (which suggests that they fare better).

[32] The survey asked principals about whether their students had taken national or sub-national assessments. It did not specifically refer to the intervention to reduce social desirability bias (i.e., principals reporting that they had administered the assessments because they believed that is what they were supposed to do).

to administer assessments and claimed to have done so even if they did not, or considered other types of assessments (e.g., classroom tests) (col. 1).[33] Principals in T1 and T2 schools were also more likely than their control counterparts to compare the learning outcomes of their school, both over time and against those of the province and other schools. Given that these comparisons were included in the reports distributed during the study, these results suggest that most principals in T1 and T2 schools used the reports as intended.[34]

According to the intervention monitoring data, the distribution of school reports occurred as planned, but participation in workshops and school visits did not. All treatment schools received three reports (two during the study and one after the endline) and all control schools received one report after the endline (Panel B, cols. 1-3). There were two problems with the implementation of the capacity-building intervention. First, participation in workshops and school visits was lower than expected among T2 schools (col. 3). Principals at these schools were supposed to attend five workshops and two visits, but the average principal in this group attended three workshops and one visit. Second, a few principals from T1 schools attended workshops and visits, even if they were not supposed to do so (col. 2). Therefore, the effect of T1 should be interpreted as that of diagnostic feedback with minimal capacity building.

5.3 Average ITT effects

5.3.1 Student achievement

We find that diagnostic feedback had a positive, large, and statistically significant effect on student achievement, but we see less clear evidence of a positive effect of capacity building. After the first year of the interventions (in 2014), diagnostic-feedback (T1) schools performed better in math and reading than capacity-building (T2) schools, but the differences are small (less than $.07\sigma$ or 2 pp.) and statistically insignificant (Table 3, cols. 1-3). After two years (in 2015), T1 and T2 schools outperformed control schools (cols. 4-8). Only the differences between control and T1 schools are statistically significant, and they are large in both subjects ($.33\sigma$ or 6.3 pp. in math and $.36\sigma$ or 7.7 pp. in reading), but we cannot discard the possibility that T1 and T2 had the same impact on student achievement (col. 9).[35]

[33] A key drawback of relying solely on self-reported data, and the reason why we also collected intervention monitoring data, is that principals sometimes report engaging in practices that are extremely unlikely to have actually occurred (e.g., because they are forbidden by law). For example, many public school principals claim to make decisions over the hiring and firing of teachers in the school survey of the Program for International Student Assessment (PISA) in countries where public schools have no such discretion (see OECD 2016).

[34] We believe that the responses of principals from control schools to these three questions are subject to the same potential problems as those to the previous question.

[35] We do not find any statistically significant heterogeneous ITT effects on learning by school size or location. We have omitted these estimates from the manuscript, but they are available upon request.

The results above are robust to checks for student attrition and multiple hypothesis testing. We observe a similar pattern when we compute Lee (2009) bounds to account for the potential influence of student attrition in 2015.[36] The bounds are very wide: for T1 schools, they range from -.18 to .89σ in math and from -.17 to .92σ in reading; for T2 schools, they range from .31 to .76σ in math and from -.36 to .7σ in reading. Yet, they are mostly positive (Table A.13).[37] Further, the statistical significance of the coefficients on T1 and T2 remains unchanged when we compute False Discovery Rate q-values for 2014 and 2015 (Table A.14).

The positive effect of diagnostic feedback is not limited to only some topics or skills. In fact, T1 schools outperformed control schools in nearly all content and cognitive domains in both grades and subjects (Table A.7 and A.8). This finding is important because it suggests that the intervention did not lead schools to focus on more malleable skills to increase test scores.[38]

The positive effect of diagnostic feedback is not driven by students' familiarity with the test. Quite the opposite, T1 schools outperformed control schools in both items that were repeated from prior assessment rounds (which we call "familiar" items) and in items that were introduced for the first time on each assessment round (which we call "unfamiliar" items) (Table A.9). In fact, the magnitude of the coefficient on T1 is remarkably similar across both types of items.

The statistically insignificant effect of T2 is puzzling because it combines the reports of T1 with workshops and school visits, so its effect should be at least at large as that of T1.[39] Yet, some evidence suggests that it may be due to the initial imbalance between T1 and T2.[40] First, the effect of T2 is commensurate to that imbalance: it is small (.14σ in math and .11σ in reading) and statistically insignificant in grade 3, where the imbalance was larger, and it is large (.29σ and .22σ) and statistically significant in grade 5, where it was smaller (Table A.6). Second, the effect of T2 is slightly larger in grade 3 if we use covariates (from .14 to .15σ in math and from .11 to .12σ in reading), but it does not change in grade 5 (Table A.10, col. 8). Admittedly, however, these patterns are suggestive rather than conclusive.

[36] All Lee (2009) bounds estimated in this paper report analytic standard errors because the Stata command leebounds does not support clustered standard errors. Therefore, they should be interpreted with caution, as they ignore the school-level error component.

[37] We also find our ITT estimates change little when we account for absences on test day (Table A.11). Further, we see no evidence that the positive and statistically significant effects for diagnostic feedback in 2015 are predicted by student absenteeism in 2013. The correlation coefficients for T1 schools is below .1 for all grades and subjects. The coefficients for T2 schools are slightly larger, but this may be due to the baseline imbalance we have already discussed. In fact, the differences in the magnitudes of the coefficients across grades are consistent with these hypotheses: coefficients are larger in grade 3 (where we see greater imbalance between T1 and T2 schools at baseline) and smaller in grade 5 (where we observe less imbalance).

[38] Capacity building did not have a larger effect on geometry. This might be because only 20% of T2 schools participated in the workshop on that topic or because the assessments took place one month later (see Table 1).

[39] It is possible that the additional components may have contradicted the reports and/or diverted resources away from activities that improve student achievement, but this is unlikely due to the little time demanded by these components (see section 2.4) and their uneven implementation (see section 5.2).

[40] See section 5.1 for the discussion of this imbalance.

The statistically insignificant effect of T2 may also result from the problems with compliance.[41] If T2 schools were at a disadvantage with respect to T1 schools on the first year of the study (in 2013), the low take-up of workshops and school visits among T2 principals, and the participation of some T1 principals in those activities may have preserved that disadvantage.[42] We cannot, however, ascertain the role of non-compliance on the effect of capacity building.

5.3.2 National assessments

We find evidence that the effects of the interventions persisted after the end of the experiment. One year after the interventions (in 2016), T1 and T2 schools outperformed control schools in math and reading, even if control schools had received one report at the end of the experiment (see section 5.2) (Table 5, cols. 4-5).[43] The effects small to moderate (.26σ in math and .22σ in reading among T1 schools and .17σ in math and .18σ in reading among T2 schools). The coefficients on T1 are larger than those on T2, but consistent with the results from our assessments, we cannot rule out that both interventions had the same effect (col. 6).

5.3.3 Internal efficiency

We do not find clear evidence that the interventions improved schools' internal efficiency (Table 4), either while they were being implemented (in 2014-2015, columns 1-6), or up to two years after they concluded (in 2016-2017, columns 7-12). In some indicators and years, the sign of the difference between control and treatment schools is as expected (e.g., in 2014, T1 and T2 schools had slightly lower dropout rates). In other indicators and years, the sign seems counterintuitive (e.g., in 2014, T1 and T2 schools had *lower* passing rates). And in yet other indicators and years, the sign differs between T1 and T2 schools (e.g., in 2014, T1 schools had higher dropout rates but T2 schools had lower dropout rates than control schools). Further, only a handful of these differences are (marginally) statistically significant. We observe a similar pattern when we estimate these effects separately by grade (Table A.15).

Importantly, however, internal efficiency indicators are collected at the school (rather than at the student) level. At this level of aggregation, it is possible that we do not have enough statistical power to detect small or moderate effects, even if they occurred.

[41] See section 5.2 for the discussion of these problems.

[42] This possibility is consistent with analyses in which we exploited random assignment to T2 to estimate the effect of take up of capacity building and found it had a positive impact on student achievement. These analyses are available from the authors upon request.

[43] As mentioned in section 3, the national assessments tested grade 6 students, who were in grade 3 at the start of the study and received two years of the interventions.

5.4 Potential mechanisms

5.4.1 School management

One way in which the interventions may have improved student achievement is by leading principals to use assessment results to inform management decisions.[44] Consistent with this expectation, principal surveys administered after two years (in 2015) indicate that principals at T1 and T2 schools used assessment results for several management-related aspects.[45]

First, principals at intervention schools used assessment results for planning purposes. Principals at T1 and T2 schools reported that they were 48 pp. and 28 pp. more likely to set goals for their schools based on assessment results and 30 pp. and 23 pp. more likely to change the curriculum based on the results than their control peers (Table 6, cols. 4-5). All of these differences are statistically significant and we cannot discard the possibility that T1 and T2 had the same effect on either indicator (col. 6).

Second, principals at these schools also used assessment results to make staffing decisions. Principals at T1 and T2 schools reported that they were 28 pp. and 37 pp. more likely to use assessment results to evaluate principals than their control counterparts. Principals at T1 schools also claimed that they were 23 pp. more likely to use these results to evaluate teachers, but this difference is only marginally statistically significant, insignificant in the case of T2 schools, and we cannot reject the null that T1 and T2 had the same effect on this indicator.[46]

Finally, principals made assessment results available to others. Principals at T1 and T2 schools reported that they were 42 and 49 pp. more likely to inform parents of assessment results, and 32 pp. and 31 pp. more likely to make these results public, than those at control schools. All of these differences are statistically significant and we cannot discard the possibility that T1 and T2 had the same effect. This finding suggests that the improvements in learning could have occurred in part due to changes in parental allocation of educational investments.[47] All results in this section are robust to checks for multiple hypothesis testing (Table A.16).

[44] As discussed in a prior footnote, we asked principals whether they used results from national or sub-national student assessments in 2015 for the purposes discussed in this section. Given that there was no national assessment in Argentina in 2015, and no other sub-national assessments in La Rioja that year, the only assessments to which principals could be referring would be those associated with the interventions.

[45] As mentioned on a prior footnote, the results in this section should be interpreted with caution, as principals reports do not always correspond with school practices.

[46] Note that in La Rioja, as in the rest of Argentina, public schools have limited authority over the hiring and firing of principals and teachers. Thus, it is unlikely that they attached stakes to these results. Instead, it is more likely that they made learning outcomes part of the appraisal process for principals and teachers.

[47] As discussed in section 2.1, the National Education Law prohibits the government from disseminating assessment results at the school, teacher, or student level. However, the law does not forbid schools from making their own results available to parents or community members.

5.4.2 Classroom instruction

Another way in which feedback may have improved learning is through classroom instruction. The students surveys administered after two years (in 2015) indicate that teachers at T1 schools devoted more time to instruction, employed more activities during lessons, and were rated more favorably by their students than those at control schools.

First, teachers at T1 schools were no more likely to go to school or arrive on time or to devote more time to instruction conditional on attendance. Students at T1 schools reported that their teachers were less likely to start class late, less likely to end class early, and less likely to leave school early than those at control schools (Table 7, col. 4). Yet, none of these results are statistically significant once we account for multiple hypothesis testing (Table A.18).

Second, teachers at T1 schools also used more activities during lessons. Students at T1 schools reported that their teachers were 6 pp. more likely to use a textbook, 2 pp. more likely to assign homework, 3 pp. more likely to write on the board, 2 pp. more likely to explain topics, and 2 pp. more likely to grade homework than those at control schools (Table 8, col. 4). The statistical significance of these coefficients remains unchanged once we account for multiple hypothesis testing (Table A.19). Yet, teachers were no more likely to assign practice tests, which suggests that the intervention did not lead to teaching to the test. In most of these outcomes, we cannot rule out the possibility that T1 and T2 had a similar effect (col. 6).

Finally, teachers at T1 schools were rated more favorably. Students at T1 schools reported that their teachers were more likely to demonstrate interest in students ($.19\sigma$), manage their classrooms ($.13\sigma$), clarify difficult concepts or tasks ($.17\sigma$), motivate students to perform at their best ($.18\sigma$), deliver captivating lessons ($.16\sigma$), engage students in discussions ($.14\sigma$), and summarize the material at the end of a lesson ($.18\sigma$) than control teachers (Table 9, col. 4). The statistical significance of all coefficients remains unchanged once we account for multiple hypothesis testing (Table A.20). In fact, T1 schools outperformed T2 schools (col. 6).

6 Conclusion

We present experimental evidence on the impact of diagnostic feedback and capacity building for public primary schools in La Rioja, Argentina and find that providing schools with information on their relative performance led to large test-score gains in math and reading. The results presented in Tables 6 to 9 indicate that this improvement was driven by principals using assessment results to inform school management decisions, teachers resorting to more instructional strategies, and an improvement in interactions between teachers and students.

The impact of diagnostic feedback demonstrates the potential of large-scale assessments to improve system performance in developing countries. Specifically, it suggests that, at least in settings where the binding constraint is not the extensive but the intensive margin of principal and teacher effort, informing schools about their relative standing can prompt improvements in school management and classroom instruction, which in turn can raise learning outcomes. Given that many developing nations recently started administering large-scale assessments (see Cheng and Gale 2014; Ganimian and Koretz 2017), that these assessments account for only a small share of their education budgets (see Wolff 2007), and that school reports may be automated and distributed at little to no cost (e.g., if online), diagnostic feedback promises to be a cost-effective way to improve student learning in these settings (see World Bank 2018).

We cannot distinguish between the relative contributions of each components of diagnostic feedback, but it is important to note that the intervention that we evaluated was based on assessments that were informative over a wide range of achievement and comparable over time. There are good reasons to believe that these aspects are fundamental for feedback purposes. Assessments that only cover the material that students are supposed to know in a given grade (i.e., without testing material from lower grades), which are quite prevalent in some settings (e.g., South Asia) often result in most students performing poorly and do not provide enough information to make meaningful distinctions between schools (see Muralidharan et al. 2019). Similarly, assessments that are not comparable over time (e.g., because they include too few common items across rounds and/or do not use Item Response Theory for linking purposes) may convey inconsistent information about the relative and absolute performance of schools, leading its users to make incorrect decisions (see Barrera-Osorio and Ganimian 2016).

In the same vein, it seems equally important to highlight that the reports that we evaluated featured different types of information that may prove useful for principals and teachers, including not only comparisons between the average performance of a school with that of the average school in the school system, but also between sections within a school, as well as the school's overall performance on each content and cognitive domains of each subject assessed. We cannot determine which of these different types of information was most useful for schools, but we would caution against expecting that reports that simply compare a school to the rest of the system would result in the large test-score gains that we have observed in this setting. We see experimentation with the types of information presented in school reports, and with the format in which that information is presented, as a promising area for further research.

Finally, our results on the lower-than-expected take-up of capacity building illustrate the challenges of implementing meaningful professional development in the developing world. They are consistent with those of evaluations of traditional teacher training programs in other developing countries, which have also found low take-up and limited effects on learning (see Angrist and Lavy 2001; Yoshikawa et al. 2015; Yue et al. 2014; Zhang et al. 2013). We see experimentation with more innovative models of professional development (e.g., coaching) to be a more promising area for future research (see, for example, Cilliers et al. 2019).

References

Andrabi, T., J. Das, and A. I. Khwaja (2017). Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review 107*(6), 1535–1563.

Angrist, J. D. and V. Lavy (2001). Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools. *Journal of Labor Economics 19*(2).

Banerjee, A. V., R. Banerji, E. Duflo, and M. Walton (2011). Effective pedagogies and a resistant education system: Experimental evidence on interventions to improve basic skills in rural India. *Unpublished manuscript*. New Delhi, India: Abdul Latif Jameel Poverty Action Lab (J-PAL).

Barrera-Osorio, F. and A. J. Ganimian (2016). The barking dog that bites: Test score volatility and school rankings in Punjab, Pakistan. *International Journal of Educational Development 49*, 31–54.

Bassi, M., M. Busso, and J. S. Muñoz (2013). Is the glass half empty or half full? School enrollment, graduation, and dropout rates in Latin America. (IDB Working Paper No. 462). Washington, DC: Inter-American Development Bank.

Bassi, M., C. Meghir, and A. Reynoso (2016). Education quality and teaching practices. Technical report, National Bureau of Economic Research.

Betts, J. R., Y. Hahn, and A. C. Zau (2017). Can testing improve student learning? An evaluation of the mathematics diagnostic testing project. *Journal of Urban Economics 100*, 54–64.

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

Camargo, B., R. Camelo, S. Firpo, and V. Ponczek (2018). Information, market incentives, and student performance: Evidence from a regression discontinuity design in Brazil. *The Journal of Human Resources 53*(2), 414–444.

Cheng, X. and C. Gale (2014). National assessments mapping metadata. Washington, DC: FHI 360. Retrieved from: `http://bit.ly/2yxBeBd`.

Cilliers, J., B. Fleisch, C. Prinsloo, and S. Taylor (2019). How to improve teaching practice? an experimental comparison of centralized training and in-classroom coaching. *Journal of Human Resources*, 0618–9538R1.

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

de Hoyos, R., P. A. Holland, and S. Troiano (2015). Understanding the trends in learning outcomes in Argentina, 2000 to 2012. (Policy Research Working Paper No. 7518). The World Bank. Washington.

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

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

DiNIECE (2013). Anuario Estadístico 2013. Buenos Aires, Argentina: Dirección Nacional de Información de la Calidad Educativa (DiNIECE).

DiNIEE (2015). Anuario Estadístico 2015. Buenos Aires, Argentina: Dirección Nacional de Información de la Calidad Educativa (DiNIECE).

Duflo, E., J. Berry, S. Mukerji, and M. Shotland (2015). A wide angle view of learning: Evaluation of the CCE and LEP programmes in Haryana, India. (Impact Evaluation Report No. 22). International Initiative for Impact Evaluation (3ie). New Delhi, India.

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

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

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

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

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

Ganimian, A. J. (2013). No logramos mejorar: Informe sobre el desempeño de Argentina en el Programa para la Evaluación Internacional de Alumnos (PISA) 2012. Buenos Aires, Argentina: Proyecto Educar 2050.

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

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

Ganimian, A. J. and D. M. Koretz (2017). Dataset of international large-scale assessments. Last updated: February 8, 2017. Cambridge, MA: Harvard Graduate School of Education.

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

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

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

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies 76*(3), 1071–1102.

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

Muralidharan, K. (2012). Long-term effects of teacher performance pay: Experimental evidence from India. *Unpublished manuscript.* San Diego, CA: University of California, San Diego.

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

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

Muralidharan, K. and V. Sundararaman (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy 119*(1), 39–77.

OECD (2016). PISA 2015 results: Excellence and equity in education. Volume I. Paris, France: Organization for Economic Cooperation and Development (OECD).

Piper, B. and M. Korda (2011). EGRA plus: Liberia. Program evaluation report. *Unpublished manuscript.* RTI International. Research Triangle Park, NC.

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

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

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

UNGA (2000). A/RES/55/2. Resolution adopted by the General Assembly on 18 September 2000. New York, NY: United Nations General Assembly.

UNGA (2015). A/RES/70/1. Resolution adopted by the General Assembly on 25 September 2015. New York, NY: United Nations General Assembly.

Wolff, L. (2007). The costs of student assessments in Latin America. (PREAL Working Paper No. 38). Washington, DC: Partnership for Educational Revitalization in the Americas (PREAL).

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

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

Yoshikawa, H., D. Leyva, C. E. Snow, E. Treviño, M. C. Arbour, M. C. Barata, C. Weiland, C. Gómez, L. Moreno, A. Rolla, and N. D'Sa (2015). Experimental impacts of a teacher professional development program in chile on preschool classroom quality and child outcomes. *Journal of Developmental Psychology 51*, 309–322.

Yue, A., Y. Shi, F. Chang, C. Yang, H. Wang, H. Yi, R. Luo, C. Liu, L. Zhang, J. Yanjey Chu, et al. (2014). Dormitory management and boarding students in china's rural primary schools. *China Agricultural Economic Review 6*(3), 523–550.

Zhang, L., F. Lai, X. Pang, H. Yi, and S. Rozelle (2013). The impact of teacher training on teacher and student outcomes: Evidence from a randomised experiment in Beijing migrant schools. *Journal of Development Effectiveness 5*(3), 339–358.

Table 1: Timeline of the study

(1)	(2)	(3)	(4)	(5)
		School participation rates		
		Control	T1	T2
Month	Event	schools	schools	schools
Panel A. 2013				
February	School year starts			
April	Administrative data (grades 3 and 5)	-	100%	100%
October	Student assessments (grades 3 and 5)	-	100%	100%
	Student surveys (grades 3 and 5)	-	100%	100%
	Teacher surveys (grades 3 and 5)	-	100%	100%
December	School year ends			
Panel B. 2014				
February	School year starts			
March	Reports are delivered to schools	-	100%	100%
	Workshop 1: Assessment results	-	-	53%
April	School visit 1	-	40%	60%
May	Workshop 2: School improvement plans	-	-	90%
September	Workshop 3: Quality assurance	-	-	87%
November	Student assessments (grades 3, 4, and 5)	-	100%	100%
	Student surveys (grades 3, 4, and 5)	-	100%	100%
	Teacher surveys (grades 3, 4, and 5)	-	100%	93%
	Principal surveys	-	100%	93%
December	School year ends			
Panel C. 2015				
February	School year starts			
April	Reports are delivered to schools	-	100%	100%
	Workshop 4: Assessment results	-	-	97%
June	School visit 2	-	33%	87%
September	Workshop 5: Teaching geometry	-	23%	20%
October	Student assessments (grades 3 and 5)	100%	100%	100%
	Student surveys (grades 3 and 5)	-	100%	100%
	Teacher surveys (grades 3 and 5)	-	100%	100%
	Principal surveys	-	100%	100%
December	School year ends			

Notes: (1) The table shows the timeline for the interventions 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 3-5).

Table 2: Administration and use of student assessments (2015)

	(1)	(2)	(3)	(4)	(5)	(6)
	Control	T1	T2	Col. (5)-	Col. (6)-	F-test
	schools	schools	schools	Col. (4)	Col. (4)	$\beta_1 = \beta_2$
Panel A. Principal survey						
Students at my school took national or sub-national assessments	.4	.962	.931	.562***	.529***	.291
	(.497)	(.196)	(.258)	(.093)	(.098)	(.591)
I compared my school's results with those of the province	.25	.679	.75	.426***	.501***	.391
	(.441)	(.476)	(.441)	(.123)	(.119)	(.533)
I compared my school's results with those of other schools	.25	.741	.519	.47***	.276**	2.215
	(.441)	(.447)	(.509)	(.122)	(.129)	(.141)
I compared my school's results over time	.677	1	.966	.323***	.29***	.706
	(.475)	(0)	(.186)	(.086)	(.095)	(.403)
Panel B. Implementation monitoring						
Number of reports received	1	3	3	2	2	.
	(0)	(0)	(0)	(0)	(0)	(.)
Number of workshops attended	0	.233	3.4	.22***	3.401***	334.158
	(0)	(.43)	(.855)	(.075)	(.156)	(0)
Number of visits received	0	.733	1.467	.758***	1.462***	24.782
	(0)	(.583)	(.571)	(.1)	(.105)	(0)

Notes: (1) The table shows, for the 2015 school year, the means and standard deviations for the control group (column 1), diagnostic feedback or T1 group (column 2), and capacity building or T2 group (column 3). It also estimates the ITT effect of T1 and T2 with respect to control schools, using randomization fixed effects (columns 4-5). Finally, it shows the F-statistic and associated p-value for the null hypothesis that the coefficients of T1 and T2 are equal in 2015 (column 6). Panel A uses data from principal surveys from 2015 and Panel B uses implementation monitoring data from 2013-2015. (2) In the surveys, principals were asked to indicate whether their schools used student assessment results for the purposes listed above. The figures indicate the share of principals who reported that their schools used assessments for each purpose, based on the school year of data collection. (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 3: ITT effect of the interventions on student achievement (2014-2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	2014			2015					
	T1 schools	T2 schools	Col. (2)- Col. (1)	Control schools	T1 schools	T2 schools	Col. (5)- Col. (4)	Col. (6)- Col. (4)	F-test $\beta_1 = \beta_2$
Math (percent-correct score)	55.779	54.941	-.635	51.552	57.69	55.46	6.275**	3.939	.538
	(18.059)	(19.614)	(2.432)	(19.417)	(19.599)	(20.761)	(2.422)	(2.757)	(.465)
Math (IRT-scaled score)	.278	.243	-.023	0	.324	.216	.333**	.216	.481
	(.992)	(1.075)	(.139)	(1)	(1.036)	(1.104)	(.131)	(.145)	(.489)
Reading (percent-correct score)	63.181	61.043	-1.431	58.319	66.048	61.674	7.713***	3.627	1.981
	(19.666)	(20.092)	(2.167)	(22.371)	(21.625)	(22.921)	(2.13)	(2.518)	(.162)
Reading (IRT-scaled score)	.312	.212	-.066	0	.357	.153	.356***	.166	1.987
	(.969)	(.966)	(.105)	(1)	(1.018)	(1.035)	(.102)	(.111)	(.162)
N (students)	3950	3588	7538	3446	3950	3588	10984	10984	10984

Notes: (1) The table shows, for 2014 and 2015, the means and standard deviations of all control schools (column 4), diagnostic feedback or T1 schools (columns 1 and 5), and capacity building or T2 schools (columns 2 and 6). It also estimates the ITT effect of T2 with respect to T1 in 2014 (column 3) and of T1 and T2 with respect to control schools in 2015, using randomization fixed effects (columns 7-8). Finally, it shows the F-statistic and associated p-value for the null hypothesis that the coefficients of T1 and T2 are equal in 2015 (column 9). (2) All test scores are shown as percent-correct scores and as IRT-scaled scores, standardized with respect to the control group in 2015. (3) Control schools were only assessed in 2015 (see sections 2.3 and 2.4). (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 4: ITT effect of the interventions on internal efficiency (2014-2017)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
		2014			2015			2016			2017	
		Difference			Difference			Difference			Difference	
	Control	T1	T2	Control	T1	T2	Control	T1	T2	Control	T1	T2
Number of students enrolled	326.567	-9.443	20.38	319.481	-14.888	26.339	311.462	-13.086	32.386	306.692	-5.397	33.576
	(272.272)	(41.564)	(41.351)	(271.878)	(41.103)	(40.893)	(263.714)	(38.986)	(38.787)	(259.112)	(37.872)	(37.679)
Percentage of students who passed the grade	96.28	-.626	-1.69*	97.155	-.167	.556	98.416	.147	.009	98.774	-.358	-.116
	(4.31)	(1.018)	(1.014)	(3.497)	(.842)	(.838)	(2.662)	(.644)	(.641)	(1.904)	(.452)	(.45)
Percentage of students who failed the grade	3.682	.694	1.718*	2.571	-.067	-.67	1.432	-.188	-.202	1.175	.313	.017
	(4.32)	(1.02)	(1.016)	(3.412)	(.824)	(.82)	(2.516)	(.606)	(.603)	(1.869)	(.444)	(.441)
Percentage of students who dropped out of school	.038	-.067*	-.028	.274	.234	.115	.152	.041	.193	.051	.045	.099*
	(.153)	(.037)	(.036)	(.929)	(.22)	(.219)	(.75)	(.18)	(.18)	(.214)	(.051)	(.05)
Percentage of students who repeated the grade	2.649	.581	-.076	1.922	.436	-.107	1.453	-.711	-1.09**	2.535	-.025	-1.552*
	(3.191)	(.771)	(.767)	(2.567)	(.615)	(.612)	(2.279)	(.536)	(.533)	(3.593)	(.851)	(.847)
N (schools)	44	104	104	44	104	104	44	104	104	44	104	104

Notes: (1) The table shows, for each year in the 2014-2017 period, the means and standard deviations for all control schools (columns 1, 4, 7, and 10) and the ITT effect of diagnostic feedback or T1 (columns 2, 5, 8, 11) and capacity building or T2 (columns 3, 6, 9, 12), with respect to control schools, using randomization fixed effects. (2) Dropout rates should be interpreted as a upper-bound estimate, as they actually refer to the percentage of students who leave their schools without asking for a pass to another school. (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 5: ITT effect of the interventions on national student assessment (2016)

	(1) Control schools	(2) T1 schools	(3) T2 schools	(4) Col. (2)- Col. (1)	(5) Col. (3)- Col. (1)	(6) F-test $\beta_1 = \beta_2$
Math (IRT-scaled score)	0	.259	.164	.257**	.173**	.503
	(1)	(1.082)	(1.038)	(.107)	(.079)	(.48)
Reading (IRT-scaled score)	0	.226	.164	.221**	.177*	.139
	(1)	(1.034)	(1.04)	(.099)	(.102)	(.71)
N (students)	1661	1291	1217	3538	3538	3538

Notes: (1) The table shows, for 2016, the means and standard deviations of all control schools (column 1), diagnostic feedback or T1 schools (column 2), and capacity building or T2 schools (column 3). It also estimates the ITT effect of T1 and T2 with respect to control schools in 2015, using randomization fixed effects (columns 4-5). Finally, it shows the F-statistic and associated p-value for the null hypothesis that the coefficients of T1 and T2 are equal (column 6). (2) All test scores are shown as IRT-scaled scores, standardized with respect to the control group in 2016. (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 6: ITT effect of the interventions on principal-reported school management (2015)

	(1)	(2)	(3)	(4)	(5)	(6)
	Control	T1	T2	Col. (2)-	Col. (3)-	F-test
	schools	schools	schools	Col. (1)	Col. (1)	$\beta_1 = \beta_2$
My school set goals based on assessment results	.483	.964	.963	.475***	.484***	.021
	(.509)	(.189)	(.192)	(.103)	(.103)	(.885)
I made changes to the curriculum based on assessment results	.7	1	.931	.295***	.233**	1.665
	(.466)	(0)	(.258)	(.086)	(.099)	(.2)
I am evaluated partly based on assessment results	.333	.615	.704	.284**	.37***	.405
	(.479)	(.496)	(.465)	(.135)	(.127)	(.526)
My teachers are evaluated partly based on assessment results	.452	.69	.654	.233*	.208	.035
	(.506)	(.471)	(.485)	(.128)	(.133)	(.853)
I assign students to sections based on assessment results	.107	.231	.107	.11	0	1.226
	(.315)	(.43)	(.315)	(.099)	(.086)	(.271)
I informed parents about the results of their children	.452	.857	.931	.423***	.485***	.542
	(.506)	(.356)	(.258)	(.11)	(.104)	(.464)
I made my school's assessment results public	.207	.519	.519	.318**	.309**	.004
	(.412)	(.509)	(.509)	(.126)	(.126)	(.952)
N (schools)	42	29	30	101	101	101

Notes: (1) The table shows, for the 2015 school year, the means and standard deviations for the control group (column 1), diagnostic feedback or T1 group (column 2), and capacity building or T2 group (column 3). It also estimates the ITT effect of T1 and T2 with respect to control schools, using randomization fixed effects (columns 4-5). Finally, it shows the F-statistic and associated p-value for the null hypothesis that the coefficients of T1 and T2 are equal in 2015 (column 6). (2) Principals were asked to indicate whether their schools used student assessment results for the purposes listed above. The figures indicate the share of principals who reported that their schools used assessments for each purpose, based on the school year of data collection. (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 7: ITT effect of the interventions on student-reported teacher time use (2015)

	(1) Control schools	(2) T1 schools	(3) T2 schools	(4) Col. (2)- Col. (1)	(5) Col. (3)- Col. (1)	(6) F-test $\beta_1 = \beta_2$
My teacher was absent to school	.578	.554	.6	-.025	.026	2.206
	(.494)	(.497)	(.49)	(.031)	(.03)	(.141)
My teacher arrived late to school	.394	.364	.413	-.032	.021	1.411
	(.489)	(.481)	(.492)	(.033)	(.04)	(.238)
My teacher started class late	.443	.394	.446	-.05*	.005	1.763
	(.497)	(.489)	(.497)	(.029)	(.036)	(.187)
My teacher ended class early	.5	.441	.492	-.06**	-.006	1.699
	(.5)	(.497)	(.5)	(.03)	(.035)	(.195)
My teacher left school early	.449	.389	.433	-.06*	-.014	.982
	(.497)	(.488)	(.496)	(.035)	(.039)	(.324)
N (students)	4034	3014	2854	9902	9902	9902

Notes: (1) The table shows, for the 2015 school year, the means and standard deviations for the control group (column 1), diagnostic feedback or T1 group (column 2), and capacity building or T2 group (column 3). It also estimates the ITT effect of T1 and T2 with respect to control schools, using randomization fixed effects (columns 4-5). (2) Students were asked to indicate how frequently they or their teachers engaged in the activities above. The figures indicate the share of students who reported that these activities occurred two or more times a week, based on the two weeks prior to the round of data collection.
(3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 8: ITT effect of the interventions on student-reported teacher activity (2015)

	(1) Control schools	(2) T1 schools	(3) T2 schools	(4) Col. (2)- Col. (1)	(5) Col. (3)- Col. (1)	(6) F-test $\beta_1 = \beta_2$
I used a textbook	.814	.872	.85	.058***	.036**	1.869
	(.389)	(.334)	(.357)	(.016)	(.016)	(.175)
My teacher assigned me homework	.936	.958	.949	.022***	.014	.888
	(.245)	(.2)	(.22)	(.008)	(.009)	(.348)
I copied from the blackboard	.912	.938	.913	.026**	.001	4.186
	(.283)	(.241)	(.281)	(.012)	(.01)	(.043)
I worked with a group	.916	.936	.921	.021	.004	1.506
	(.278)	(.245)	(.27)	(.014)	(.013)	(.223)
My teacher explained a topic	.96	.977	.969	.018***	.009	1.992
	(.196)	(.149)	(.174)	(.006)	(.006)	(.161)
My teacher asked me to take a practice test	.892	.911	.9	.02	.008	.758
	(.311)	(.285)	(.299)	(.014)	(.011)	(.386)
My teacher graded my homework	.958	.975	.956	.018***	-.004	9.084
	(.2)	(.156)	(.206)	(.006)	(.007)	(.003)
N (students)	4034	3014	2854	9902	9902	9902

Notes: (1) The table shows, for the 2015 school year, the means and standard deviations for the control group (column 1), diagnostic feedback or T1 group (column 2), and capacity building or T2 group (column 3). It also estimates the ITT effect of T1 and T2 with respect to control schools, using randomization fixed effects (columns 4-5). (2) Students were asked to indicate how frequently they or their teachers engaged in the activities above. The figures indicate the share of students who reported that these activities occurred two or more times a week, based on the two weeks prior to the round of data collection.
(3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 9: ITT effect of the interventions on student-reported teacher effectiveness (2015)

	(1) Control schools	(2) T1 schools	(3) T2 schools	(4) Col. (2)- Col. (1)	(5) Col. (3)- Col. (1)	(6) F-test $\beta_1 = \beta_2$
Care (standardized score)	0	.186	.003	.191***	-.006	17.419
	(1)	(.854)	(.967)	(.05)	(.052)	(0)
Control (standardized score)	0	.128	.029	.133***	.02	4.108
	(1)	(.9)	(.949)	(.05)	(.053)	(.045)
Clarify (standardized score)	0	.163	.053	.168***	.048	4.386
	(1)	(.865)	(.951)	(.056)	(.053)	(.039)
Challenge (standardized score)	0	.171	.017	.176***	.01	7.21
	(1)	(.887)	(1.01)	(.06)	(.057)	(.008)
Captivate (standardized score)	0	.152	.022	.157***	.017	8.523
	(1)	(.822)	(.943)	(.042)	(.049)	(.004)
Confer (standardized score)	0	.129	-.001	.137**	-.009	6.244
	(1)	(.874)	(.995)	(.055)	(.056)	(.014)
Consolidate (standardized score)	0	.168	-.004	.177***	-.017	11.054
	(1)	(.881)	(.983)	(.054)	(.054)	(.001)
N (students)	4034	3014	2854	9902	9902	9902

Notes: (1) The table shows, for the 2015 school year, the means and standard deviations for the control group (column 1), diagnostic feedback or T1 group (column 2), and capacity building or T2 group (column 3). It also estimates the ITT effect of T1 and T2 with respect to control schools, using randomization fixed effects (columns 4-5). (2) Students were asked to indicate how frequently their teacher engaged in certain behaviors (e.g., treating them nicely when they ask questions) using a Likert-type scale, from 1 (never) to 5 (always). Their responses were then used to calculate a score for each teacher on seven domains, including: (a) demonstrating interest in their students; (b) managing the classroom; (c) clarifying difficult concepts/tasks; (d) challenging students to perform at their best; (e) capturing students' attention with their lessons; (f) engaging students in discussions; and (g) summarizing the material learned at the end of every lesson. Students' scores were standardized with respect to the control group in 2015. The scores for each domain are expressed in student-level standard deviations. (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix A Additional tables

Table A.1: Comparison between in- and out-of-sample schools on internal efficiency (2013)

	(1)	(2)	(3)	(4)	(5)	(6)
	All	Out-of-sample schools		In-sample	Col.(4)-	Col.(4)-
	schools	All	Non-rural	schools	Col.(2)	Col.(3)
Panel A. Primary school						
Number of students enrolled	122.753	38.019	310.684	332.144	294.125***	24.539
	(212.564)	(92.519)	(176.837)	(272.879)	(19.261)	(64.993)
Percentage of students who passed the grade	95.503	95.59	95.946	95.292	-.298	-.7
	(9.296)	(10.648)	(3.413)	(4.61)	(1.084)	(1.111)
Percentage of students who failed the grade	3.703	3.385	3.882	4.477	1.092	.638
	(7.161)	(7.981)	(3.459)	(4.533)	(.833)	(1.096)
Percentage of students who dropped out of school	.794	1.026	.173	.232	-.794	.061
	(6.198)	(7.332)	(.545)	(.906)	(.722)	(.216)
Percentage of students who repeated the grade	3.816	3.752	4.26	3.975	.223	-.247
	(6.557)	(7.223)	(3.827)	(4.536)	(.763)	(1.109)
N (schools)	361	257	19	104	361	122
Panel B. Grade 3						
Number of students enrolled	21.565	7.193	47.895	49.481	42.288***	2.028
	(33.602)	(15.909)	(28.075)	(40.714)	(3.259)	(9.746)
Percentage of students who passed the grade	95.593	95.755	96.869	95.283	-.473	-1.632
	(14.461)	(17.179)	(3.996)	(6.758)	(1.752)	(1.607)
Percentage of students who failed the grade	3.899	3.729	2.992	4.225	.496	1.274
	(13.306)	(15.767)	(3.915)	(6.406)	(1.612)	(1.526)
Percentage of students who dropped out of school	.508	.516	.138	.492	-.024	.358
	(5.912)	(7.09)	(.603)	(2.414)	(.716)	(.561)
Percentage of students who repeated the grade	4.807	5.284	2.885	3.882	-1.402	1.035
	(14.975)	(18.01)	(3.345)	(5.454)	(1.808)	(1.299)
N (schools)	306	202	19	104	306	122
Panel C. Grade 5						
Number of students enrolled	21.01	7.059	43.316	48.24	41.181***	5.383
	(33.015)	(14.384)	(25.699)	(41.235)	(3.216)	(9.797)
Percentage of students who passed the grade	96.267	96.36	97.931	96.093	-.267	-1.876
	(11.532)	(13.658)	(3.548)	(5.823)	(1.404)	(1.387)
Percentage of students who failed the grade	3.153	2.775	1.919	3.859	1.085	1.978
	(10.659)	(12.501)	(3.573)	(5.835)	(1.296)	(1.39)
Percentage of students who dropped out of school	.58	.865	.151	.048	-.817	-.102
	(4.801)	(5.925)	(.656)	(.49)	(.583)	(.13)
Percentage of students who repeated the grade	2.567	2.243	2.218	3.201	.959	1.014
	(9.078)	(10.245)	(4.528)	(6.194)	(1.095)	(1.496)
N (schools)	307	203	19	104	307	122

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

Table A.2: Differences in student absenteeism across experimental groups (2013-2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	2013			2014			2015				
	T1 schools	T2 schools	Col. (2)- Col. (1)	T1 schools	T2 schools	Col. (5)- Col. (4)	Control schools	T1 schools	T2 schools	Col. (8)- Col. (7)	Col. (9)- Col. (7)
Panel A. Grade 3											
Percentage of absent students in math	17.43	18.27	.847	11.37	12.26	.891	20.61	19.51	15.89	-1.102	-4.717
	(19.86)	(12.92)	(4.378)	(9.75)	(10.62)	(2.652)	(14.82)	(14.48)	(11.45)	(3.46)	(3.062)
Percentage of absent students in reading	20.48	20.5	.025	7.64	11.67	4.023	17.2	15.24	15.14	-1.967	-2.069
	(19.76)	(11.66)	(4.242)	(20.06)	(7.59)	(3.974)	(10.39)	(12.12)	(10.34)	(2.711)	(2.453)
N (schools)	30	30	60	30	30	60	44	30	30	74	74
Panel B. Grade 5											
Percentage of absent students in math	12.58	8.79	-3.787	12.64	11.74	-.901	15.83	11.36	15.72	-4.474*	-.108
	(20.37)	(39.6)	(8.161)	(12.05)	(9.02)	(2.747)	(10.54)	(11.62)	(6.72)	(2.65)	(2.01)
Percentage of absent students in reading	15.64	12.69	-2.945	13.46	11.22	-2.238	14.52	9.99	13.89	-4.538*	-.639
	(18.04)	(42.13)	(8.391)	(14.44)	(8.51)	(3.06)	(11.02)	(8.93)	(8.08)	(2.328)	(2.223)
N (schools)	30	30	60	30	30	60	44	30	30	74	74

Notes: (1) The table shows the mean absenteeism rate on test day by subject, year, and experimental group and its corresponding standard deviation (columns 1-2, 4-5, and 7-9). It also estimates the difference between absenteeism rates across all experimental groups assessed on a given year (columns 3, 6, and 10-11). (2) Control schools were only assessed in 2015 (see sections 2.3 and 2.4). (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.3: Balancing checks between experimental groups on internal efficiency (2013)

	(1) In-sample schools	(2) Control schools	(3) T1 schools	(4) T2 schools	(5) Col.(4)- Col.(2)	(6) Col.(4)- Col.(3)
Panel A. Primary school						
Number of students enrolled	332.144	324.205	356.4	319.533	-17.335	11.59
	(272.879)	(334.744)	(210.542)	(230.495)	(48.034)	(48.296)
Percentage of students who passed the grade	95.292	95.281	95.118	95.482	-.103	.222
	(4.61)	(5.274)	(4.465)	(3.772)	(1.2)	(1.134)
Percentage of students who failed the grade	4.477	4.329	4.668	4.501	.29	.141
	(4.533)	(5.098)	(4.492)	(3.769)	(1.176)	(1.104)
Percentage of students who dropped out of school	.232	.39	.214	.016	-.187	-.362
	(.906)	(1.186)	(.86)	(.052)	(.258)	(.218)
Percentage of students who repeated the grade	3.975	3.487	3.543	5.121	.009	1.589
	(4.536)	(4.264)	(3.381)	(5.725)	(.951)	(1.162)
N (schools)	104	44	30	30	74	74
Panel B. Grade 3						
Number of students enrolled	49.481	48.114	52.9	48.067	-2.732	2.424
	(40.714)	(49.116)	(32.648)	(35.039)	(6.797)	(6.887)
Percentage of students who passed the grade	95.283	95.883	93.211	96.474	-2.596	.555
	(6.758)	(5.743)	(9.064)	(4.968)	(1.757)	(1.302)
Percentage of students who failed the grade	4.225	3.295	6.313	3.501	2.95*	.218
	(6.406)	(4.581)	(9.139)	(4.977)	(1.643)	(1.134)
Percentage of students who dropped out of school	.492	.822	.476	.024	-.354	-.772
	(2.414)	(3.499)	(1.461)	(.133)	(.692)	(.646)
Percentage of students who repeated the grade	3.882	3.646	3.718	4.393	.137	.692
	(5.454)	(5.416)	(4.57)	(6.393)	(1.231)	(1.391)
N (schools)	104	44	30	30	74	74
Panel C. Grade 5						
Number of students enrolled	48.24	47.591	54.067	43.367	-.916	-1.841
	(41.235)	(52.049)	(30.773)	(31.856)	(7.654)	(7.644)
Percentage of students who passed the grade	96.093	95.142	96.08	97.499	1.043	2.309
	(5.823)	(7.094)	(5.16)	(3.973)	(1.542)	(1.443)
Percentage of students who failed the grade	3.859	4.858	3.753	2.501	-1.206	-2.309
	(5.835)	(7.094)	(5.205)	(3.973)	(1.544)	(1.443)
Percentage of students who dropped out of school	.048	0	.167	0	.163	0
	(.49)	(0)	(.913)	(0)	(.138)	(0)
Percentage of students who repeated the grade	3.201	3.077	2.471	4.114	-.673	1.02
	(6.194)	(6.207)	(3.646)	(8.025)	(1.286)	(1.663)
N (schools)	104	44	30	30	74	74

Notes: (1) The table shows the means and standard deviations of all schools in our sample (column 1), control schools (column 2), diagnostic feedback or T1 schools (column 3), and capacity building or T2 schools (column 4). It also tests for differences between control and treatment schools, using randomization fixed effects (columns 5-6). Panel A shows results for all primary school students, Panel B for grade 3 students, and Panel C for grade 5 students. (2) Dropout rates should be interpreted as a upper-bound estimate, as they actually refer to the percentage of students who leave their schools without asking for a pass to another school. (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.4: Balancing checks between treatment groups on student achievement (2013)

	(1)	(2)	(3)	(4)
	Treatment	T1	T2	Col. (3)-
	schools	schools	schools	Col. (2)
Panel A. Primary school				
Math (percent-correct score)	55.459	56.513	54.193	-2.224
	(19.502)	(18.823)	(20.219)	(2.397)
Math (IRT-scaled score)	.038	.096	-.032	-.123
	(1.027)	(.994)	(1.062)	(.135)
Reading (percent-correct score)	57.012	58.541	55.155	-2.918
	(20.378)	(19.97)	(20.717)	(2.68)
Reading (IRT-scaled score)	.072	.145	-.016	-.139
	(.968)	(.961)	(.969)	(.125)
N (students)	4903	2689	2214	4903
Panel B. Grade 3				
Math (percent-correct score)	59.86	61.405	58.075	-3.152
	(20.694)	(19.375)	(21.994)	(2.907)
Math (IRT-scaled score)	.095	.172	.007	-.155
	(.985)	(.926)	(1.044)	(.139)
Reading (percent-correct score)	56.443	58.394	54.178	-3.825
	(21.718)	(21.176)	(22.125)	(3.14)
Reading (IRT-scaled score)	.041	.124	-.054	-.16
	(.943)	(.926)	(.953)	(.134)
N (students)	2457	1320	1137	2457
Panel C. Grade 5				
Math (percent-correct score)	50.977	51.714	50.054	-1.718
	(17.073)	(16.936)	(17.206)	(2.45)
Math (IRT-scaled score)	-.02	.023	-.074	-.098
	(1.065)	(1.051)	(1.08)	(.157)
Reading (percent-correct score)	57.584	58.683	56.187	-1.977
	(18.925)	(18.74)	(19.075)	(2.493)
Reading (IRT-scaled score)	.104	.166	.024	-.117
	(.992)	(.994)	(.985)	(.131)
N (students)	2446	1369	1077	2446

Notes: (1) The table shows the means and standard deviations of all treatment schools (column 1), diagnostic feedback or T1 schools (column 2), and capacity building or T2 schools (column 3). It also tests for differences between T1 and T2 schools, using randomization fixed effects (column 4). Panel A shows results for grade 3 and 5 students, Panel B for grade 3 students, and Panel C for grade 5 students. (2) All test scores are shown as percent-correct scores and as IRT-scaled scores, standardized with respect to the control group in 2015. (3) Control schools were not assessed in 2013 (see sections 2.3 and 2.4). (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.5: Balancing checks between treatment groups on student survey (2013)

	(1) Treatment schools	(2) T1 schools	(3) T2 schools	(4) Col. (3)- Col. (2)
Panel A. Primary school				
Female	.498	.494	.502	.008
	(.5)	(.5)	(.5)	(.013)
Index of parental education	0	.061	-.066	-.27
	(2.195)	(2.173)	(2.279)	(.196)
Index of household assets	0	.014	-.016	.005
	(1.424)	(1.415)	(1.415)	(.146)
Index of study supports	0	.009	-.01	-.064
	(1.1)	(1.115)	(1.077)	(.057)
N (students)	4212	3278	3018	4212
Panel B. Grade 3				
Female	.498	.502	.493	-.008
	(.5)	(.5)	(.5)	(.019)
Index of parental education	0	.106	-.129	-.195
	(2.288)	(2.199)	(2.387)	(.187)
Index of household assets	0	.018	-.022	.015
	(1.397)	(1.387)	(1.409)	(.129)
Index of study supports	0	.049	-.055	-.084
	(1.092)	(1.114)	(1.064)	(.063)
N (students)	2084	1103	981	2084
Panel C. Grade 5				
Female	.499	.487	.513	.024
	(.5)	(.5)	(.5)	(.022)
Index of parental education	0	.157	-.201	-.336
	(2.11)	(1.974)	(2.257)	(.255)
Index of household assets	0	.037	-.048	-.003
	(1.447)	(1.443)	(1.453)	(.179)
Index of study supports	0	.026	-.033	-.046
	(1.108)	(1.155)	(1.045)	(.073)
N (students)	2128	1194	934	2128

Notes: (1) The table shows the mean and standard deviation of all treatment schools (column 1), diagnostic feedback or T1 schools (column 2), and capacity building or T2 schools (column 3). It also tests for differences between T1 and T2 schools, using randomization fixed effects (column 4). Panel A shows results for grade 3 and 5 students, Panel B for grade 3 students, and Panel C for grade 5 students. (2) The index of parental education is the first principal component from a principal component analysis (PCA) of the highest education level attained by the mother and father of each student and their literacy level, as reported by the student. (3) The index of household assets is the first principal component from a PCA of variables indicating whether each student had a fridge, microwave, fan, TV, washroom, computer, and internet at home. (4) The index of study supports is the first principal component from a PCA of variables indicating whether each student was assigned homework every day, attended private tuition, completed homework alone, and the number of books at his/her home. (5) All indices are standardized with respect to all treatment within each grade in 2013. (6) Control schools were not surveyed in 2013 (see sections 2.3 and 2.4). (7) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.6: ITT effect of the interventions on student achievement, by grade (2014-2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		2014					2015		
	T1 schools	T2 schools	Col. (2)- Col. (1)	Control schools	T1 schools	T2 schools	Col. (5)- Col. (4)	Col. (6)- Col. (4)	F-test $\beta_1 = \beta_2$
Panel A. Both grades									
Math (percent-correct score)	55.779	54.941	-.635	51.552	57.69	55.46	6.275**	3.939	.538
	(18.059)	(19.614)	(2.432)	(19.417)	(57.69)	(55.46)	(2.422)	(2.757)	(.465)
Math (IRT-scaled score)	.278	.243	-.023	0	.324	.216	.333**	.216	.481
	(.992)	(1.075)	(.139)	(1)	(.324)	(.216)	(.131)	(.145)	(.489)
Reading (percent-correct score)	63.181	61.043	-1.431	58.319	66.048	61.674	7.713***	3.627	1.981
	(19.666)	(20.092)	(2.167)	(22.371)	(66.048)	(61.674)	(2.13)	(2.518)	(.162)
Reading (IRT-scaled score)	.312	.212	-.066	0	.357	.153	.356***	.166	1.987
	(.969)	(.966)	(.105)	(1)	(.357)	(.153)	(.102)	(.111)	(.162)
N (students)	3950	3588	7538	3446	3950	3588	10984	10984	10984
Panel B. Grade 3									
Math (percent-correct score)	59.92	58.481	-1.418	55.334	61.955	58.084	6.776**	2.853	1.064
	(18.839)	(21.014)	(2.801)	(21.629)	(61.955)	(58.084)	(2.603)	(3.468)	(.305)
Math (IRT-scaled score)	.322	.262	-.062	0	.31	.14	.321***	.143	1.034
	(.944)	(1.043)	(.137)	(1)	(.31)	(.14)	(.119)	(.158)	(.312)
Reading (percent-correct score)	61.91	58.74	-2.511	56.779	64.6	59.238	7.714***	2.781	2.275
	(21.274)	(21.919)	(2.621)	(24.029)	(64.6)	(59.238)	(2.413)	(2.98)	(.135)
Reading (IRT-scaled score)	.332	.192	-.112	0	.327	.098	.322***	.112	2.279
	(.948)	(.96)	(.112)	(1)	(.327)	(.098)	(.104)	(.122)	(.134)
N (students)	1951	1769	3720	1650	1951	1769	5370	5370	5370
Panel C. Grade 5									
Math (percent-correct score)	51.484	51.366	.269	48.094	53.704	52.907	5.746**	4.754*	.106
	(16.133)	(17.382)	(2.586)	(16.407)	(53.704)	(52.907)	(2.537)	(2.556)	(.746)
Math (IRT-scaled score)	.233	.223	.015	0	.336	.29	.344**	.288*	.093
	(1.037)	(1.105)	(.166)	(1)	(.336)	(.29)	(.155)	(.155)	(.761)
Reading (percent-correct score)	64.497	63.358	-.412	59.734	67.418	64.032	7.727***	4.518*	1.223
	(17.762)	(17.781)	(2.069)	(20.636)	(67.418)	(64.032)	(2.264)	(2.407)	(.271)
Reading (IRT-scaled score)	.291	.233	-.019	0	.385	.207	.387***	.218*	1.324
	(.99)	(.973)	(.118)	(1)	(.385)	(.207)	(.118)	(.118)	(.253)
N (students)	1999	1819	3818	1796	1999	1819	5614	5614	5614

Notes: (1) The table shows, for 2014 and 2015, the means and standard deviations of all control schools (column 4), diagnostic feedback or T1 schools (columns 1 and 5), and capacity building or T2 schools (columns 2 and 6). It also estimates the ITT effect of T2 with respect to T1 in 2014 (column 3) and of T1 and T2 with respect to control schools in 2015, using randomization fixed effects (columns 7-8). Finally, it shows the F-statistic and associated p-value for the null hypothesis that the coefficients of T1 and T2 are equal in 2015 (column 9). Panel A shows results for grade 3 and 5 students, Panel B for grade 3 students, and Panel C for grade 5 students. (2) All test scores are shown as percent-correct scores and as IRT-scaled scores, standardized with respect to the control group in 2015. (3) Control schools were only assessed in 2015 (see sections 2.3 and 2.4). (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.7: ITT effect of the interventions on student achievement, by grade and content domain (2014-2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		2014					2015		
	T1 schools	T2 schools	Col. (2)-Col. (1)	Control schools	T1 schools	T2 schools	Col. (5)-Col. (4)	Col. (6)-Col. (4)	F-test $\beta_1 = \beta_2$
Panel A. Both grades									
Math - Number	54.261	54.047	-.106	52.094	57.154	56.276	5.202**	4.22	.08
	(19.536)	(20.62)	(2.423)	(21.783)	(57.154)	(56.276)	(2.594)	(3.035)	(.777)
Math - Geometry	58.993	58.047	-.406	53.732	60.471	57.115	6.771***	3.525	1.239
	(25.062)	(26.88)	(2.538)	(26.413)	(60.471)	(57.115)	(2.297)	(2.558)	(.268)
Math - Measurement	50.143	48.87	-1.349	43.419	51.303	47.836	8.142***	4.297	1.461
	(25.102)	(26.211)	(2.734)	(24.801)	(51.303)	(47.836)	(2.516)	(2.744)	(.23)
Reading - Informative texts	60.669	59.457	-.406	58.52	66.107	62.657	7.569***	4.441*	1.119
	(22.679)	(22.862)	(2.161)	(25.26)	(66.107)	(62.657)	(2.068)	(2.544)	(.293)
Reading - Narrative texts	64.319	61.822	-1.909	58.418	66.298	61.578	7.87***	3.35	2.471
	(23.133)	(23.503)	(2.165)	(25.878)	(66.298)	(61.578)	(2.104)	(2.48)	(.119)
Reading - Short texts	65.459	61.498	-3.183	60.818	68.811	62.482	7.963***	1.974	3.855
	(28.321)	(29.204)	(2.455)	(32.493)	(68.811)	(62.482)	(2.269)	(2.819)	(.052)
N (students)	3950	3588	7538	3446	3950	3588	10984	10984	10984
Panel B. Grade 3									
Math - Number	60.782	59.427	-1.292	58.324	64.411	60.855	6.166**	2.678	.799
	(18.503)	(19.811)	(2.55)	(22.391)	(64.411)	(60.855)	(2.704)	(3.566)	(.374)
Math - Geometry	65.265	64.002	-.976	58.043	64.172	60.199	6.343**	2.241	1.41
	(26.797)	(29.273)	(3.292)	(29.317)	(64.172)	(60.199)	(2.654)	(3.185)	(.238)
Math - Measurement	53.648	51.859	-2.078	45.454	53.845	49.223	8.701***	3.772	1.391
	(27.679)	(29.741)	(3.434)	(28.701)	(53.845)	(49.223)	(2.754)	(3.844)	(.241)
Reading - Informative texts	59.269	57.885	-.687	58.141	65.101	61.116	6.877***	3.252	1.24
	(22.405)	(22.501)	(2.324)	(26.011)	(65.101)	(61.116)	(2.299)	(2.917)	(.268)
Reading - Narrative texts	62.068	58.467	-2.999	56.929	65.108	59.083	8.023***	2.507	2.776
	(25.392)	(25.558)	(2.872)	(28.702)	(65.108)	(59.083)	(2.56)	(3.034)	(.099)
Reading - Short texts	69.396	62.683	-5.994*	56.582	66.027	57.813	9.363***	1.553	3.854
	(31.301)	(33.492)	(3.283)	(34.2)	(66.027)	(57.813)	(2.847)	(3.759)	(.052)
N (students)	1951	1769	3720	1650	1951	1769	5370	5370	5370
Panel C. Grade 5									
Math - Number	47.497	48.614	1.279	46.397	50.372	51.821	4.182	5.323*	.101
	(18.241)	(19.995)	(2.916)	(19.551)	(50.372)	(51.821)	(2.854)	(3.13)	(.751)
Math - Geometry	52.487	52.035	.35	49.789	57.011	54.115	7.138***	4.499*	.828
	(21.263)	(22.703)	(2.428)	(22.749)	(57.011)	(54.115)	(2.435)	(2.4)	(.365)
Math - Measurement	46.508	45.852	-.52	41.559	48.928	46.486	7.594***	4.677*	1
	(21.529)	(21.68)	(2.944)	(20.431)	(48.928)	(46.486)	(2.637)	(2.366)	(.32)
Reading - Informative texts	62.118	61.037	-.184	58.867	67.057	64.149	8.229***	5.61**	.785
	(22.878)	(23.12)	(2.478)	(24.552)	(67.057)	(64.149)	(2.192)	(2.453)	(.378)
Reading - Narrative texts	66.648	65.194	-.928	59.786	67.423	63.993	7.702***	4.226*	1.444
	(20.284)	(20.707)	(1.775)	(22.903)	(67.423)	(63.993)	(2.264)	(2.411)	(.232)
Reading - Short texts	61.385	60.307	-.23	64.71	71.444	67.003	6.741***	2.593	2.266
	(24.211)	(24.094)	(2.273)	(30.335)	(71.444)	(67.003)	(2.305)	(2.456)	(.135)
N (students)	1999	1819	3818	1796	1999	1819	5614	5614	5614

Notes: (1) The table shows, for 2014 and 2015, the means and standard deviations of all control schools (column 4), diagnostic feedback or T1 schools (columns 1 and 5), and capacity building or T2 schools (columns 2 and 6). It also estimates the ITT effect of T2 with respect to T1 in 2014 (column 3) and of T1 and T2 with respect to control schools in 2015, using randomization fixed effects (columns 7-8). Finally, it shows the F-statistic and associated p-value for the null hypothesis that the coefficients of T1 and T2 are equal in 2015 (column 9). Panel A shows results for grade 3 and 5 students, Panel B for grade 3 students, and Panel C for grade 5 students. (2) All test scores are shown as percent-correct scores and as IRT-scaled scores, standardized with respect to the control group in 2015. (3) Control schools were only assessed in 2015 (see sections 2.3 and 2.4). (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.8: ITT effect of the interventions on student achievement, by grade and cognitive domain (2014-2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		2014				2015			
	T1 schools	T2 schools	Col. (2)-Col. (1)	Control schools	T1 schools	T2 schools	Col. (5)-Col. (4)	Col. (6)-Col. (4)	F-test $\beta_1 = \beta_2$
Panel A. Both grades									
Math - Knowing	62.023	60.537	-1.091	55.958	63.023	59.554	7.172***	3.672	1.364
	(21.479)	(23.395)	(2.491)	(22.558)	(63.023)	(59.554)	(2.226)	(2.69)	(.246)
Math - Algorithms	59.915	60.1	.53	56.313	60.798	59.899	4.661	3.648	.067
	(24.136)	(25.175)	(3.237)	(26.94)	(60.798)	(59.899)	(2.894)	(3.552)	(.796)
Math - Reasoning	39.818	39.856	-.102	38.032	44.527	42.592	6.688**	4.492	.423
	(22.82)	(24.074)	(2.288)	(23.21)	(44.527)	(42.592)	(2.719)	(2.79)	(.517)
Reading - Retrieving explicit information	70.121	67.983	-1.502	64.359	71.919	67.6	7.573***	3.375	2.543
	(22.562)	(23.47)	(2.065)	(25.632)	(71.919)	(67.6)	(1.832)	(2.381)	(.114)
Reading - Making straightforward inferences	59.524	56.723	-2.214	57.295	64.483	59.988	7.207***	2.978	1.999
	(23.827)	(23.637)	(2.24)	(26.46)	(64.483)	(59.988)	(2.161)	(2.606)	(.16)
Reading - Interpreting and integrating information	58.967	57.309	-.728	54.533	63.209	58.594	8.586***	4.426*	1.698
	(22.883)	(23.329)	(2.27)	(25.585)	(63.209)	(58.594)	(2.371)	(2.664)	(.195)
Reading - Evaluating and critiquing textual elements	64.032	62.279	-1.066	55.639	63.131	59.226	7.481***	3.929	1.264
	(27.296)	(27.492)	(2.538)	(28.087)	(63.131)	(59.226)	(2.643)	(2.693)	(.264)
N (students)	3950	3588	7538	3446	3950	3588	10984	10984	10984
Panel B. Grade 3									
Math - Knowing	63.867	61.84	-1.922	56.065	63.354	58.881	7.483***	2.93	1.418
	(23.613)	(26.272)	(3.134)	(25.441)	(63.354)	(58.881)	(2.538)	(3.537)	(.236)
Math - Algorithms	61.722	59.766	-1.712	61.108	67.526	63.792	6.464**	2.88	.732
	(21.842)	(23.065)	(2.919)	(26.138)	(67.526)	(63.792)	(2.883)	(3.842)	(.394)
Math - Reasoning	48.49	48.501	-.197	44.853	51.521	47.394	6.939**	2.538	1.169
	(23.223)	(24.331)	(2.785)	(25.972)	(51.521)	(47.394)	(2.891)	(3.67)	(.282)
Reading - Retrieving explicit information	70.122	67.149	-2.294	61.364	68.985	63.911	7.591***	2.686	2.369
	(24.736)	(25.821)	(2.749)	(26.367)	(68.985)	(63.911)	(2.299)	(2.949)	(.127)
Reading - Making straightforward inferences	61.691	57.084	-3.95	59.012	66.707	60.908	7.619***	2.247	2.505
	(25.872)	(26.194)	(2.878)	(28.484)	(66.707)	(60.908)	(2.508)	(3.113)	(.117)
Reading - Interpreting and integrating information	53.877	51.923	-1.305	51.111	59.526	53.696	8.168***	3.061	2.069
	(22.631)	(22.62)	(2.328)	(26.896)	(59.526)	(53.696)	(2.621)	(3.135)	(.153)
Reading - Evaluating and critiquing textual elements	60.635	57.711	-2.284	53.919	61.391	56.981	7.397***	3.406	1.391
	(28.554)	(28.819)	(3.137)	(29.169)	(61.391)	(56.981)	(2.799)	(3.079)	(.241)
N (students)	1951	1769	3720	1650	1951	1769	5370	5370	5370
Panel C. Grade 5									
Math - Knowing	60.111	59.221	-.226	55.861	62.715	60.21	6.898***	4.385*	.898
	(18.834)	(20.001)	(2.312)	(19.561)	(62.715)	(60.21)	(2.238)	(2.262)	(.345)
Math - Algorithms	58.041	60.439	2.825	51.928	54.509	56.112	2.866	4.089	.08
	(26.182)	(27.145)	(4.064)	(26.922)	(54.509)	(56.112)	(3.353)	(3.925)	(.778)
Math - Reasoning	30.823	31.127	.259	31.796	37.99	37.921	6.354**	5.958**	.014
	(18.517)	(20.389)	(2.827)	(18.257)	(37.99)	(37.921)	(2.918)	(2.644)	(.906)
Reading - Retrieving explicit information	70.119	68.822	-.715	67.11	74.694	71.173	7.624***	4.192*	1.799
	(20.074)	(20.821)	(1.838)	(24.625)	(74.694)	(71.173)	(2.042)	(2.175)	(.183)
Reading - Making straightforward inferences	57.282	56.36	-.405	55.718	62.381	59.096	6.748***	3.588	1.132
	(21.285)	(20.759)	(2.063)	(24.355)	(62.381)	(59.096)	(2.276)	(2.48)	(.29)
Reading - Interpreting and integrating information	64.234	62.721	-.366	57.678	66.692	63.337	9.017***	5.91**	.917
	(21.938)	(22.78)	(2.442)	(23.898)	(66.692)	(63.337)	(2.537)	(2.636)	(.34)
Reading - Evaluating and critiquing textual elements	67.548	66.868	.013	57.22	64.777	61.401	7.597**	4.521	.786
	(25.467)	(25.283)	(2.554)	(26.966)	(64.777)	(61.401)	(2.961)	(2.85)	(.377)
N (students)	1999	1819	3818	1796	1999	1819	5614	5614	5614

Notes: (1) The table shows, for 2014 and 2015, the means and standard deviations of all control schools (column 4), diagnostic feedback or T1 schools (columns 1 and 5), and capacity building or T2 schools (columns 2 and 6). It also estimates the ITT effect of T2 with respect to T1 in 2014 (column 3) and of T1 and T2 with respect to control schools in 2015, using randomization fixed effects (columns 7-8). Finally, it shows the F-statistic and associated p-value for the null hypothesis that the coefficients of T1 and T2 are equal in 2015 (column 9). Panel A shows results for grade 3 and 5 students, Panel B for grade 3 students, and Panel C for grade 5 students. (2) All test scores are shown as percent-correct scores and as IRT-scaled scores, standardized with respect to the control group in 2015. (3) Control schools were only assessed in 2015 (see sections 2.3 and 2.4). (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.9: ITT effect of the interventions on student achievement, by grade and familiarity (2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		2014					2015		
	T1 schools	T2 schools	Col. (2)- Col. (1)	Control schools	T1 schools	T2 schools	Col. (5)- Col. (4)	Col. (6)- Col. (4)	F-test $\beta_1 = \beta_2$
Panel A. Both grades									
Math - Familiar items	51.407	50.958	-.253	53.595	59.091	56.953	5.555**	3.49	.462
	(17.985)	(19.365)	(1.767)	(22.79)	(59.091)	(56.953)	(2.405)	(2.666)	(.498)
Math - Non-familiar items	57.224	56.259	-.771	50.777	57.16	54.925	6.549***	4.142	.543
	(20.042)	(21.467)	(2.696)	(20.085)	(57.16)	(54.925)	(2.461)	(2.812)	(.463)
Reading - Familiar items	61.933	58.803	-2.417	61.191	68.519	63.456	7.329***	2.534	3.046
	(21.764)	(21.987)	(2.176)	(26.093)	(68.519)	(63.456)	(2.034)	(2.514)	(.084)
Reading - Non-familiar items	64.453	62.585	-1.165	58.247	66.22	62.093	7.965***	4.106	1.646
	(21.097)	(21.525)	(2.217)	(23.084)	(66.22)	(62.093)	(2.221)	(2.57)	(.202)
N (students)	3950	3588	7538	3446	3950	3588	10984	10984	10984
Panel B. Grade 3									
Math - Familiar items	57.096	55.843	-1.002	58.76	64.243	59.889	5.512**	1.375	1.398
	(18.26)	(20.238)	(2.234)	(23.901)	(64.243)	(59.889)	(2.468)	(3.318)	(.24)
Math - Non-familiar items	61.05	59.536	-1.584	53.963	61.039	57.362	7.281***	3.445	.936
	(20.856)	(22.897)	(3.06)	(22.504)	(61.039)	(57.362)	(2.726)	(3.573)	(.336)
Reading - Familiar items	60.957	56.394	-3.864	58.133	66.667	60.261	8.429***	2.463	2.927
	(24.86)	(25.132)	(2.913)	(27.992)	(66.667)	(60.261)	(2.511)	(3.294)	(.09)
Reading - Non-familiar items	62.291	59.679	-1.97	56.238	63.773	58.829	7.428***	2.908	1.964
	(21.673)	(22.439)	(2.557)	(24.037)	(63.773)	(58.829)	(2.416)	(2.891)	(.164)
N (students)	1951	1769	3720	1650	1951	1769	5370	5370	5370
Panel C. Grade 5									
Math - Familiar items	45.507	46.025	.673	48.872	54.276	54.098	5.502**	5.222**	.008
	(15.651)	(17.085)	(1.88)	(20.628)	(54.276)	(54.098)	(2.734)	(2.591)	(.929)
Math - Non-familiar items	53.255	52.949	.149	47.863	53.534	52.554	5.818**	4.616*	.153
	(18.344)	(19.374)	(2.88)	(17.077)	(53.534)	(52.554)	(2.52)	(2.587)	(.696)
Reading - Familiar items	62.944	61.223	-1.018	64	70.27	66.549	6.337***	2.745	2.013
	(17.962)	(17.98)	(1.795)	(23.883)	(70.27)	(66.549)	(2.015)	(2.208)	(.159)
Reading - Non-familiar items	66.691	65.506	-.453	60.093	68.533	65.253	8.494***	5.363**	.986
	(20.251)	(20.156)	(2.253)	(22.017)	(68.533)	(65.253)	(2.476)	(2.599)	(.323)
N (students)	1999	1819	3818	1796	1999	1819	5614	5614	5614

Notes: (1) The table shows, for 2014 and 2015, the means and standard deviations of all control schools (column 4), diagnostic feedback or T1 schools (columns 1 and 5), and capacity building or T2 schools (columns 2 and 6). It also estimates the ITT effect of T2 with respect to T1 in 2014 (column 3) and of T1 and T2 with respect to control schools in 2015, using randomization fixed effects (columns 7-8). Finally, it shows the F-statistic and associated p-value for the null hypothesis that the coefficients of T1 and T2 are equal in 2015 (column 9). Panel A shows results for grade 3 and 5 students, Panel B for grade 3 students, and Panel C for grade 5 students. (2) All test scores are shown as percent-correct scores and as IRT-scaled scores, standardized with respect to the control group in 2015. (3) Control schools were only assessed in 2015 (see sections 2.3 and 2.4). (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.10: ITT effect of the interventions on student achievement, by grade with covariates (2014-2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		2014				2015			
	T1 schools	T2 schools	Col. (2)- Col. (1)	Control schools	T1 schools	T2 schools	Col. (5)- Col. (4)	Col. (6)- Col. (4)	F-test $\beta_1 = \beta_2$
Panel A. Both grades									
Math (percent-correct score)	55.779	54.941	-.51	51.552	57.69	55.46	6.395***	3.988	.585
	(18.059)	(19.614)	(2.321)	(19.417)	(57.69)	(55.46)	(2.431)	(2.799)	(.446)
Math (IRT-scaled score)	.278	.243	-.016	0	.324	.216	.338**	.218	.516
	(.992)	(1.075)	(.131)	(1)	(.324)	(.216)	(.132)	(.147)	(.474)
Reading (percent-correct score)	63.181	61.043	-1.581	58.319	66.048	61.674	8.174***	3.748	2.447
	(19.666)	(20.092)	(2.121)	(22.371)	(66.048)	(61.674)	(2.105)	(2.509)	(.121)
Reading (IRT-scaled score)	.312	.212	-.073	0	.357	.153	.375***	.171	2.412
	(.969)	(.966)	(.102)	(1)	(.357)	(.153)	(.101)	(.111)	(.123)
N (students)	3950	3588	7538	3446	3950	3588	10984	10984	10984
Panel B. Grade 3									
Math (percent-correct score)	59.92	58.481	-1.498	55.334	61.955	58.084	7.257***	2.995	1.283
	(18.839)	(21.014)	(2.768)	(21.629)	(61.955)	(58.084)	(2.651)	(3.502)	(.26)
Math (IRT-scaled score)	.322	.262	-.065	0	.31	.14	.341***	.149	1.238
	(.944)	(1.043)	(.136)	(1)	(.31)	(.14)	(.122)	(.16)	(.268)
Reading (percent-correct score)	61.91	58.74	-2.665	56.779	64.6	59.238	8.505***	2.893	3.245
	(21.274)	(21.919)	(2.579)	(24.029)	(64.6)	(59.238)	(2.351)	(2.879)	(.075)
Reading (IRT-scaled score)	.332	.192	-.118	0	.327	.098	.353***	.117	3.202
	(.948)	(.96)	(.11)	(1)	(.327)	(.098)	(.101)	(.118)	(.076)
N (students)	1951	1769	3720	1650	1951	1769	5370	5370	5370
Panel C. Grade 5									
Math (percent-correct score)	51.484	51.366	.608	48.094	53.704	52.907	5.609**	4.682*	.095
	(16.133)	(17.382)	(2.391)	(16.407)	(53.704)	(52.907)	(2.514)	(2.57)	(.758)
Math (IRT-scaled score)	.233	.223	.037	0	.336	.29	.338**	.284*	.085
	(1.037)	(1.105)	(.153)	(1)	(.336)	(.29)	(.155)	(.155)	(.771)
Reading (percent-correct score)	64.497	63.358	-.506	59.734	67.418	64.032	7.875***	4.577*	1.309
	(17.762)	(17.781)	(2.05)	(20.636)	(67.418)	(64.032)	(2.297)	(2.441)	(.255)
Reading (IRT-scaled score)	.291	.233	-.026	0	.385	.207	.395***	.222*	1.417
	(.99)	(.973)	(.116)	(1)	(.385)	(.207)	(.12)	(.119)	(.237)
N (students)	1999	1819	3818	1796	1999	1819	5614	5614	5614

Notes: (1) The table shows, for 2014 and 2015, the means and standard deviations of all control schools (column 4), diagnostic feedback or T1 schools (columns 1 and 5), and capacity building or T2 schools (columns 2 and 6). It also estimates the ITT effect of T2 with respect to T1 in 2014 (column 3) and of T1 and T2 with respect to control schools in 2015, using randomization fixed effects and accounting for internal efficiency at baseline as specified in equation 1 (columns 7-8). Finally, it shows the F-statistic and associated p-value for the null hypothesis that the coefficients of T1 and T2 are equal in 2015 (column 9). Panel A shows results for grade 3 and 5 students, Panel B for grade 3 students, and Panel C for grade 5 students. (2) All test scores are shown as percent-correct scores and as IRT-scaled scores, standardized with respect to the control group in 2015. (3) Control schools were only assessed in 2015 (see sections 2.3 and 2.4). (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.11: ITT effect of the interventions on student achievement, by grade accounting by student absenteeism (2014-2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		2014				2015			
	T1 schools	T2 schools	Col. (2)- Col. (1)	Control schools	T1 schools	T2 schools	Col. (5)- Col. (4)	Col. (6)- Col. (4)	F-test $\beta_1 = \beta_2$
Panel A. Grade 3									
Math (percent-correct score)	59.92	58.481	-1.421	55.334	61.955	58.084	6.236**	2.198	1.284
	(18.839)	(21.014)	(2.724)	(21.629)	(61.955)	(58.084)	(2.454)	(3.359)	(.26)
Math (IRT-scaled score)	.322	.262	-.062	0	.31	.14	.297***	.114	1.23
	(.944)	(1.043)	(.134)	(1)	(.31)	(.14)	(.113)	(.154)	(.27)
Reading (percent-correct score)	61.91	58.74	-2.29	56.779	64.6	59.238	6.82***	2.412	2.648
	(21.274)	(21.919)	(2.439)	(24.029)	(64.6)	(59.238)	(1.999)	(2.457)	(.107)
Reading (IRT-scaled score)	.332	.192	-.102	0	.327	.098	.285***	.097	2.665
	(.948)	(.96)	(.104)	(1)	(.327)	(.098)	(.086)	(.101)	(.106)
N (students)	1951	1769	3720	1650	1951	1769	5370	5370	5370
Panel B. Grade 5									
Math (percent-correct score)	51.484	51.366	.206	48.094	53.704	52.907	4.692**	4.91**	.006
	(16.133)	(17.382)	(2.55)	(16.407)	(53.704)	(52.907)	(2.263)	(2.465)	(.936)
Math (IRT-scaled score)	.233	.223	.011	0	.336	.29	.277**	.298**	.017
	(1.037)	(1.105)	(.164)	(1)	(.336)	(.29)	(.136)	(.149)	(.896)
Reading (percent-correct score)	64.497	63.358	-.422	59.734	67.418	64.032	7.422***	5.17**	.799
	(17.762)	(17.781)	(2.05)	(20.636)	(67.418)	(64.032)	(2.184)	(2.1)	(.373)
Reading (IRT-scaled score)	.291	.233	-.02	0	.385	.207	.372***	.251**	.9
	(.99)	(.973)	(.116)	(1)	(.385)	(.207)	(.114)	(.102)	(.345)
N (students)	1999	1819	3818	1796	1999	1819	5614	5614	5614

Notes: (1) The table shows, for 2014 and 2015, the means and standard deviations of all control schools (column 4), diagnostic feedback or T1 schools (columns 1 and 5), and capacity building or T2 schools (columns 2 and 6). It also estimates the ITT effect of T2 with respect to T1 in 2014 (column 3) and of T1 and T2 with respect to control schools in 2015, using randomization fixed effects and accounting for absenteeism on test day as specified in equation 1 (columns 7-8). Finally, it shows the F-statistic and associated p-value for the null hypothesis that the coefficients of T1 and T2 are equal in 2015 (column 9). Panel A shows results for grade 3 and 5 students, Panel B for grade 3 students, and Panel C for grade 5 students. (2) All test scores are shown as percent-correct scores and as IRT-scaled scores, standardized with respect to the control group in 2015. (3) Control schools were only assessed in 2015 (see sections 2.3 and 2.4). (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.12: Correlation between absenteeism in 2013 and achievement in 2015, by grade

(1)	(2) T1 schools	(3) T2 schools
Panel A. Grade 3		
Math (percent-correct score)	.014	-.141***
Math (IRT-scaled score)	.009	-.139***
Reading (percent-correct score)	-.077***	-.132***
Reading (IRT-scaled score)	-.08***	-.138***
Panel B. Grade 5		
Math (percent-correct score)	-.077***	.062**
Math (IRT-scaled score)	-.079***	.069**
Reading (percent-correct score)	-.075***	.107***
Reading (IRT-scaled score)	-.084***	.096***

Notes: (1) The table shows the pairwise correlations between student absenteeism in 2013 and student achievement in 2015 by grade and experimental group. The former is measured at the school level and the latter at the student level. (2) Control schools were only assessed in 2015 (see sections 2.3 and 2.4). (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.13: Lee bounds on ITT effect of the interventions on student achievement (2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Control	Effect of T1			Effect of T2		
	schools	Lower	Upper	95% CI	Lower	Upper	95% CI
Math (percent-correct score)	51.552	-4.006***	17.329***	[-5.005, 18.264]	-6.542***	14.723***	[-7.591, 15.722]
	(19.417)	(.607)	(.568)		(.638)	(.607)	
Math (IRT-scaled score)	0	-.183***	.892***	[-.234, .94]	-.306***	.76***	[-.36, .812]
	(1)	(.031)	(.029)		(.033)	(.031)	
Reading (percent-correct score)	58.319	-4.687***	20.03***	[-5.776, 21.17]	-8.738***	15.147***	[-9.882, 16.32]
	(22.371)	(.662)	(.693)		(.696)	(.714)	
Reading (IRT-scaled score)	0	-.167***	.917***	[-.218, .965]	-.356***	.697***	[-.408, .747]
	(1)	(.031)	(.029)		(.032)	(.03)	
N (students)	3446		7396			7034	

Notes: (1) The table shows, for 2015, the means and standard deviations of all control schools (column 1), and Lee (2009) bounds on the ITT effects on student achievement for that year. It shows the lower (columns 2 and 5) and upper (columns 3 and 6) bounds with their associated analytic standard errors and the 95% confidence interval (columns 4 and 7). (2) We computed these bounds using a regression with endline test scores as the dependent variable to keep our analysis of bounds analogous to the main ITT effects. We tightened the bounds using the randomization fixed effects for school size. (3) All test scores are shown as percent-correct scores and as IRT-scaled scores, standardized with respect to the control group in 2015. (4) Control schools were only assessed in 2015 (see sections 2.3 and 2.4). (5) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.14: ITT effect of the interventions on student achievement with FDR q-values (2014-2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	2014			2015					
	T1 schools	T2 schools	Col. (2)- Col. (1)	Control schools	T1 schools	T2 schools	Col. (5)- Col. (4)	Col. (6)- Col. (4)	F-test $\beta_1 = \beta_2$
Math (percent-correct score)	55.779	54.941	-.635	51.552	57.69	55.46	6.275**	3.939	.538
	(18.059)	(19.614)	(2.432)	(19.417)	(19.599)	(20.761)	(2.422)	(2.757)	(.465)
			[.8659]				[.0373]	[.2341]	
Math (IRT-scaled score)	.278	.243	-.023	0	.324	.216	.333**	.216	.481
	(.992)	(1.075)	(.139)	(1)	(1.036)	(1.104)	(.131)	(.145)	(.489)
			[.8659]				[.0373]	[.2341]	
Reading (percent-correct score)	63.184	61.043	-1.436	58.319	66.048	61.674	7.713***	3.627	1.981
	(19.665)	(20.092)	(2.167)	(22.371)	(21.625)	(22.921)	(2.13)	(2.518)	(.162)
			[.6386]				[.0043]	[.2341]	
Reading (IRT-scaled score)	.312	.213	-.066	0	.357	.153	.356***	.166	1.987
	(.969)	(.966)	(.105)	(1)	(1.018)	(1.035)	(.102)	(.111)	(.162)
			[.6386]				[.0043]	[.2341]	
N (students)	3950	3588	7538	3446	3950	3588	10984	10984	10984

Notes: (1) The table shows, for 2014 and 2015, the means and standard deviations of all control schools (column 4), diagnostic feedback or T1 schools (columns 1 and 5), and capacity building or T2 schools (columns 2 and 6). It also estimates the ITT effect of T2 with respect to T1 in 2014 (column 3) and of T1 and T2 with respect to control schools in 2015, using randomization fixed effects (columns 7-8). Finally, it shows the F-statistic and associated p-value for the null hypothesis that the coefficients of T1 and T2 are equal in 2015 (column 9). (2) All test scores are shown as percent-correct scores and as IRT-scaled scores, standardized with respect to the control group in 2015. (3) Control schools were only assessed in 2015 (see sections 2.3 and 2.4). (4) False discovery rate (FDR) q-values (shown between brackets) were computed using the Benjamini-Hochberg's step-up procedure. (5) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.15: ITT effect of the interventions on internal efficiency (2014-2017)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
		2014			2015			2016			2017	
		Difference			Difference			Difference			Difference	
	Control	T1	T2	Control	T1	T2	Control	T1	T2	Control	T1	T2
Panel A. Primary school												
Number of students enrolled	326.567	-9.443	20.38	319.481	-14.888	26.339	311.462	-13.086	32.386	306.692	-5.397	33.576
	(272.272)	(41.564)	(41.351)	(271.878)	(41.103)	(40.893)	(263.714)	(38.986)	(38.787)	(259.112)	(37.872)	(37.679)
Percentage of students who passed the grade	96.28	-.626	-1.69*	97.155	-.167	.556	98.416	.147	.009	98.774	-.358	-.116
	(4.31)	(1.018)	(1.014)	(3.497)	(.842)	(.838)	(2.662)	(.644)	(.641)	(1.904)	(.452)	(.45)
Percentage of students who failed the grade	3.682	.694	1.718*	2.571	-.067	-.67	1.432	-.188	-.202	1.175	.313	.017
	(4.32)	(1.02)	(1.016)	(3.412)	(.824)	(.82)	(2.516)	(.606)	(.603)	(1.869)	(.444)	(.441)
Percentage of students who dropped out of school	.038	-.067*	-.028	.274	.234	.115	.152	.041	.193	.051	.045	.099*
	(.153)	(.037)	(.036)	(.929)	(.22)	(.219)	(.75)	(.18)	(.18)	(.214)	(.051)	(.05)
Percentage of students who repeated the grade	2.649	.581	-.076	1.922	.436	-.107	1.453	-.711	-1.09**	2.535	-.025	-1.552*
	(3.191)	(.771)	(.767)	(2.567)	(.615)	(.612)	(2.279)	(.536)	(.533)	(3.593)	(.851)	(.847)
N (schools)	44	104	104	44	104	104	44	104	104	44	104	104
Panel B. Grade 3												
Number of students enrolled	47.779	-.946	4.03	46.058	-2.634	3.268	45.462	-1.84	5.741	46	0	6.281
	(40.522)	(6.184)	(6.152)	(40.237)	(6.278)	(6.246)	(39.21)	(5.872)	(5.842)	(39.06)	(5.698)	(5.597)
Percentage of students who passed the grade	96.66	.085	-1.697	97.233	1.188	1.231	97.83	-.884	-1.052	97.667	-1.022	-.078
	(5.528)	(1.291)	(1.286)	(4.331)	(1.001)	(.996)	(4.064)	(.97)	(.965)	(5.085)	(1.24)	(1.218)
Percentage of students who failed the grade	3.31	-.022	1.727	2.684	-1.306	-1.39	2.108	.671	1.058	2.333	1.022	.078
	(5.53)	(1.291)	(1.286)	(4.282)	(.99)	(.985)	(4.071)	(.976)	(.971)	(5.085)	(1.24)	(1.218)
Percentage of students who dropped out of school	.029	-.063	-.03	.083	.118	.159	.062	.213**	-.007	0	0	0
	(.175)	(.041)	(.041)	(.597)	(.141)	(.141)	(.455)	(.105)	(.105)	(0)	(0)	(0)
Percentage of students who repeated the grade	2.809	-.401	-.354	2.261	-.562	-.514	1.645	.175	-.561	2.473	.71	-.216
	(4.493)	(1.09)	(1.085)	(4.082)	(.985)	(.98)	(3.532)	(.861)	(.846)	(4.21)	(1.015)	(.998)
N (schools)	44	104	104	44	104	104	44	104	104	44	104	104
Panel C. Grade 5												
Number of students enrolled	46.904	-1.726	3.082	47.048	-3.682	4.225	45.942	-.633	5.882	44.231	-.799	4.437
	(39.868)	(6.163)	(6.131)	(39.796)	(5.843)	(5.814)	(38.984)	(6.022)	(5.992)	(39.094)	(6.045)	(6.014)
Percentage of students who passed the grade	96.601	-.664	-.069	97.472	.734	.649	98.218	.755	.874	98.619	-.315	.668
	(5.781)	(1.4)	(1.395)	(5.141)	(1.245)	(1.239)	(4.022)	(.969)	(.964)	(2.947)	(.691)	(.688)
Percentage of students who failed the grade	3.384	.701	.102	2.517	-.705	-.624	1.65	-.735	-.77	1.305	.267	-.572
	(5.766)	(1.397)	(1.391)	(5.145)	(1.246)	(1.24)	(3.951)	(.951)	(.946)	(2.877)	(.676)	(.673)
Percentage of students who dropped out of school	.014	-.038	-.032	.011	-.029	-.025	.132	-.021	-.103	.075	.048	-.096
	(.145)	(.035)	(.035)	(.112)	(.027)	(.027)	(.773)	(.186)	(.185)	(.541)	(.13)	(.129)
Percentage of students who repeated the grade	2.452	.687	-.853	1.91	-.616	-1.046	1.556	-.104	.025	3.402	.285	-2.662*
	(4.466)	(1.074)	(1.069)	(4.427)	(1.063)	(1.058)	(4.362)	(1.065)	(1.046)	(6.452)	(1.535)	(1.527)
N (schools)	44	104	104	44	104	104	44	104	104	44	104	104

Notes: (1) The table shows, for each year in the 2014-2017 period, the means and standard deviations for all control schools (columns 1, 4, 7, and 10) and the ITT effect of diagnostic feedback or T1 (columns 2, 5, 8, 11) and capacity building or T2 (columns 3, 6, 9, 12), with respect to control schools, using randomization fixed effects. Panel A shows results for all primary school students, Panel B for grade 3 students, and Panel C for grade 5 students. (2) Dropout rates should be interpreted as a upper-bound estimate, as they actually refer to the percentage of students who leave their schools without asking for a pass to another school. (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.16: ITT effect of the interventions on principal-reported school management with FDR q-values (2015)

	(1)	(2)	(3)	(4)	(5)	(6)
	Control	T1	T2	Col. (2)-	Col. (3)-	F-test
	schools	schools	schools	Col. (1)	Col. (1)	$\beta_1 = \beta_2$
My school set goals based on assessment results	.483	.964	.963	.475***	.484***	.021
	(.509)	(.189)	(.192)	(.103)	(.103)	(.885)
				[.0001]	[.0001]	
I made changes to the curriculum based on assessment results	.7	1	.931	.295***	.233**	1.665
	(.466)	(0)	(.258)	(.086)	(.099)	(.2)
				[.0026]	[.0321]	
I am evaluated partly based on assessment results	.333	.615	.704	.284*	.37**	.405
	(.479)	(.496)	(.465)	(.135)	(.127)	(.526)
				[.0532]	[.0106]	
My teachers are evaluated partly based on assessment results	.452	.69	.654	.233*	.208	.035
	(.506)	(.471)	(.485)	(.128)	(.133)	(.853)
				[.0912]	[.1418]	
I assign students to sections based on assessment results	.107	.231	.107	.11	0	1.226
	(.315)	(.43)	(.315)	(.099)	(.086)	(.271)
				[.2912]	[1]	
I informed parents about the results of their children	.452	.857	.931	.423***	.485***	.542
	(.506)	(.356)	(.258)	(.11)	(.104)	(.464)
				[.0008]	[.0001]	
I made my school's assessment results public	.207	.519	.519	.318**	.309**	.004
	(.412)	(.509)	(.509)	(.126)	(.126)	(.952)
				[.0273]	[.0288]	
N (schools)	42	29	30	101	101	101

Notes: (1) The table shows, for the 2015 school year, the means and standard deviations for the control group (column 1), diagnostic feedback or T1 group (column 2), and capacity building or T2 group (column 3). It also estimates the ITT effect of T1 and T2 with respect to control schools, using randomization fixed effects (columns 4-5). Finally, it shows the F-statistic and associated p-value for the null hypothesis that the coefficients of T1 and T2 are equal in 2015 (column 6). (2) Principals were asked to indicate whether their schools used student assessment results for the purposes listed above. The figures indicate the share of principals who reported that their schools used assessments for each purpose, based on the school year of data collection. (3) False discovery rate (FDR) q-values (shown between brackets) were computed using the Benjamini-Hochberg's step-up procedure. (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.17: Lee bounds on ITT effect of the interventions on principal-reported school management (2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Control		Effect of T1			Effect of T2	
	schools	Lower	Upper	95% CI	Lower	Upper	95% CI
My school set goals based on assessment results	.483	.444***	.517***	[.233, .682]	.467***	.517***	[.276, .685]
	(.509)	(.121)	(.094)		(.108)	(.094)	
I made changes to the curriculum based on assessment results	.7	.3***	.3***	[.133, .467]	.208*	.3***	[.025, .445]
	(.466)	(.085)	(.085)		(.108)	(.085)	
I am evaluated partly based on assessment results	.333	.229	.417**	[-.053, .713]	.292*	.556***	[.047, .844]
	(.479)	(.168)	(.177)		(.148)	(.174)	
My teachers are evaluated partly based on assessment results	.452	.109	.464***	[-.16, .752]	.14	.305*	[-.118, .568]
	(.506)	(.163)	(.175)		(.154)	(.157)	
I assign students to sections based on assessment results	.107	-.048	.161	[-.274, .396]	-.107*	.046	[-.206, .226]
	(.315)	(.137)	(.142)		(.06)	(.109)	
I informed parents about the results of their children	.452	.375***	.548***	[.157, .699]	.459***	.527***	[.264, .724]
	(.506)	(.131)	(.091)		(.112)	(.113)	
I made my school's assessment results public	.207	.193	.522***	[-.088, .829]	.166	.469***	[-.107, .75]
	(.412)	(.17)	(.186)		(.165)	(.17)	
N (schools)	42		71			72	

Notes: (1) The table shows, for 2015, the means and standard deviations of all control schools (column 1), and Lee (2009) bounds on the ITT effects on student achievement for that year. It shows the lower (columns 2 and 5) and upper (columns 3 and 6) bounds with their associated analytic standard errors and the 95% confidence interval (columns 4 and 7). (2) We computed these bounds using a regression with endline test scores as the dependent variable to keep our analysis of bounds analogous to the main ITT effects. We tightened the bounds using the randomization fixed effects for school size. (3) Principals were asked to indicate whether their schools used student assessment results for the purposes listed above. The figures indicate the share of principals who reported that their schools used assessments for each purpose, based on the school year of data collection. (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.18: ITT effect of the interventions on student-reported teacher time use with FDR q-values (2015)

	(1) Control schools	(2) T1 schools	(3) T2 schools	(4) Col. (2)- Col. (1)	(5) Col. (3)- Col. (1)	(6) F-test $\beta_1 = \beta_2$
My teacher was absent to school	.578	.554	.6	-.025	.026	2.206
	(.494)	(.497)	(.49)	(.031)	(.03)	(.141)
				[.699]	[.699]	
My teacher arrived late to school	.394	.364	.413	-.032	.021	1.411
	(.489)	(.481)	(.492)	(.033)	(.04)	(.238)
				[.699]	[.865]	
My teacher started class late	.443	.394	.446	-.05	.005	1.763
	(.497)	(.489)	(.497)	(.029)	(.036)	(.187)
				[.3131]	[.8817]	
My teacher ended class early	.5	.441	.492	-.06	-.006	1.699
	(.5)	(.497)	(.5)	(.03)	(.035)	(.195)
				[.3131]	[.8817]	
My teacher left school early	.449	.389	.433	-.06	-.014	.982
	(.497)	(.488)	(.496)	(.035)	(.039)	(.324)
				[.3131]	[.8817]	
N (students)	4034	3014	2854	9902	9902	9902

Notes: (1) The table shows, for the 2015 school year, the means and standard deviations for the control group (column 1), diagnostic feedback or T1 group (column 2), and capacity building or T2 group (column 3). It also estimates the ITT effect of T1 and T2 with respect to control schools, using randomization fixed effects (columns 4-5). (2) Students were asked to indicate how frequently they or their teachers engaged in the activities above. The figures indicate the share of students who reported that these activities occurred two or more times a week, based on the two weeks prior to the round of data collection.
(3) False discovery rate (FDR) q-values (shown between brackets) were computed using the Benjamini-Hochberg's step-up procedure. (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.19: ITT effect of the interventions on student-reported teacher activity with FDR q-values (2015)

	(1)	(2)	(3)	(4)	(5)	(6)
	Control	T1	T2	Col. (2)-	Col. (3)-	F-test
	schools	schools	schools	Col. (1)	Col. (1)	$\beta_1 = \beta_2$
I used a textbook	.814	.872	.85	.058***	.036*	1.869
	(.389)	(.334)	(.357)	(.016)	(.016)	(.175)
				[.0044]	[.0612]	
My teacher assigned me homework	.936	.958	.949	.022**	.014	.888
	(.245)	(.2)	(.22)	(.008)	(.009)	(.348)
				[.0308]	[.2196]	
I copied from the blackboard	.912	.938	.913	.026*	.001	4.186
	(.283)	(.241)	(.281)	(.012)	(.01)	(.043)
				[.0612]	[.9491]	
I worked with a group	.916	.936	.921	.021	.004	1.506
	(.278)	(.245)	(.27)	(.014)	(.013)	(.223)
				[.2196]	[.8498]	
My teacher explained a topic	.96	.977	.969	.018**	.009	1.992
	(.196)	(.149)	(.174)	(.006)	(.006)	(.161)
				[.0185]	[.2334]	
My teacher asked me to take a practice test	.892	.911	.9	.02	.008	.758
	(.311)	(.285)	(.299)	(.014)	(.011)	(.386)
				[.2334]	[.6202]	
My teacher graded my homework	.958	.975	.956	.018**	-.004	9.084
	(.2)	(.156)	(.206)	(.006)	(.007)	(.003)
				[.0185]	[.6496]	
N (students)	4034	3014	2854	9902	9902	9902

Notes: (1) The table shows, for the 2015 school year, the means and standard deviations for the control group (column 1), diagnostic feedback or T1 group (column 2), and capacity building or T2 group (column 3). It also estimates the ITT effect of T1 and T2 with respect to control schools, using randomization fixed effects (columns 4-5). (2) Students were asked to indicate how frequently they or their teachers engaged in the activities above. The figures indicate the share of students who reported that these activities occurred two or more times a week, based on the two weeks prior to the round of data collection.
(3) False discovery rate (FDR) q-values (shown between brackets) were computed using the Benjamini-Hochberg's step-up procedure. (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.20: ITT effect of the interventions on student-reported teacher effectiveness with FDR q-values (2015)

	(1)	(2)	(3)	(4)	(5)	(6)
	Control	T1	T2	Col. (2)-	Col. (3)-	F-test
	schools	schools	schools	Col. (1)	Col. (1)	$\beta_1 = \beta_2$
Care (standardized score)	0	.186	.003	.191***	-.006	17.419
	(1)	(.854)	(.967)	(.05)	(.052)	(0)
				[.0019]	[.9022]	
Control (standardized score)	0	.128	.029	.133**	.02	4.108
	(1)	(.9)	(.949)	(.05)	(.053)	(.045)
				[.0213]	[.9022]	
Clarify (standardized score)	0	.163	.053	.168**	.048	4.386
	(1)	(.865)	(.951)	(.056)	(.053)	(.039)
				[.0111]	[.645]	
Challenge (standardized score)	0	.171	.017	.176**	.01	7.21
	(1)	(.887)	(1.01)	(.06)	(.057)	(.008)
				[.0111]	[.9022]	
Captivate (standardized score)	0	.152	.022	.157***	.017	8.523
	(1)	(.822)	(.943)	(.042)	(.049)	(.004)
				[.0019]	[.9022]	
Consolidate (standardized score)	0	.168	-.004	.177***	-.017	11.054
	(1)	(.881)	(.983)	(.054)	(.054)	(.001)
				[.0069]	[.9022]	
N (students)	4034	3014	2854	9902	9902	9902

Notes: (1) The table shows, for the 2015 school year, the means and standard deviations for the control group (column 1), diagnostic feedback or T1 group (column 2), and capacity building or T2 group (column 3). It also estimates the ITT effect of T1 and T2 with respect to control schools, using randomization fixed effects (columns 4-5). (2) Students were asked to indicate how frequently their teacher engaged in certain behaviors (e.g., treating them nicely when they ask questions) using a Likert-type scale, from 1 (never) to 5 (always). Their responses were then used to calculate a score for each teacher on seven domains, including: (a) demonstrating interest in their students; (b) managing the classroom; (c) clarifying difficult concepts/tasks; (d) challenging students to perform at their best; (e) capturing students' attention with their lessons; (f) engaging students in discussions; and (g) summarizing the material learned at the end of every lesson. Students' scores were standardized with respect to the control group in 2015. The scores for each domain are expressed in student-level standard deviations. (3) False discovery rate (FDR) q-values (shown between brackets) were computed using the Benjamini-Hochberg's step-up procedure. (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix B Additional information on instruments

B.1 Student assessments

B.1.1 Design

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

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

B.1.2 Scaling and linking

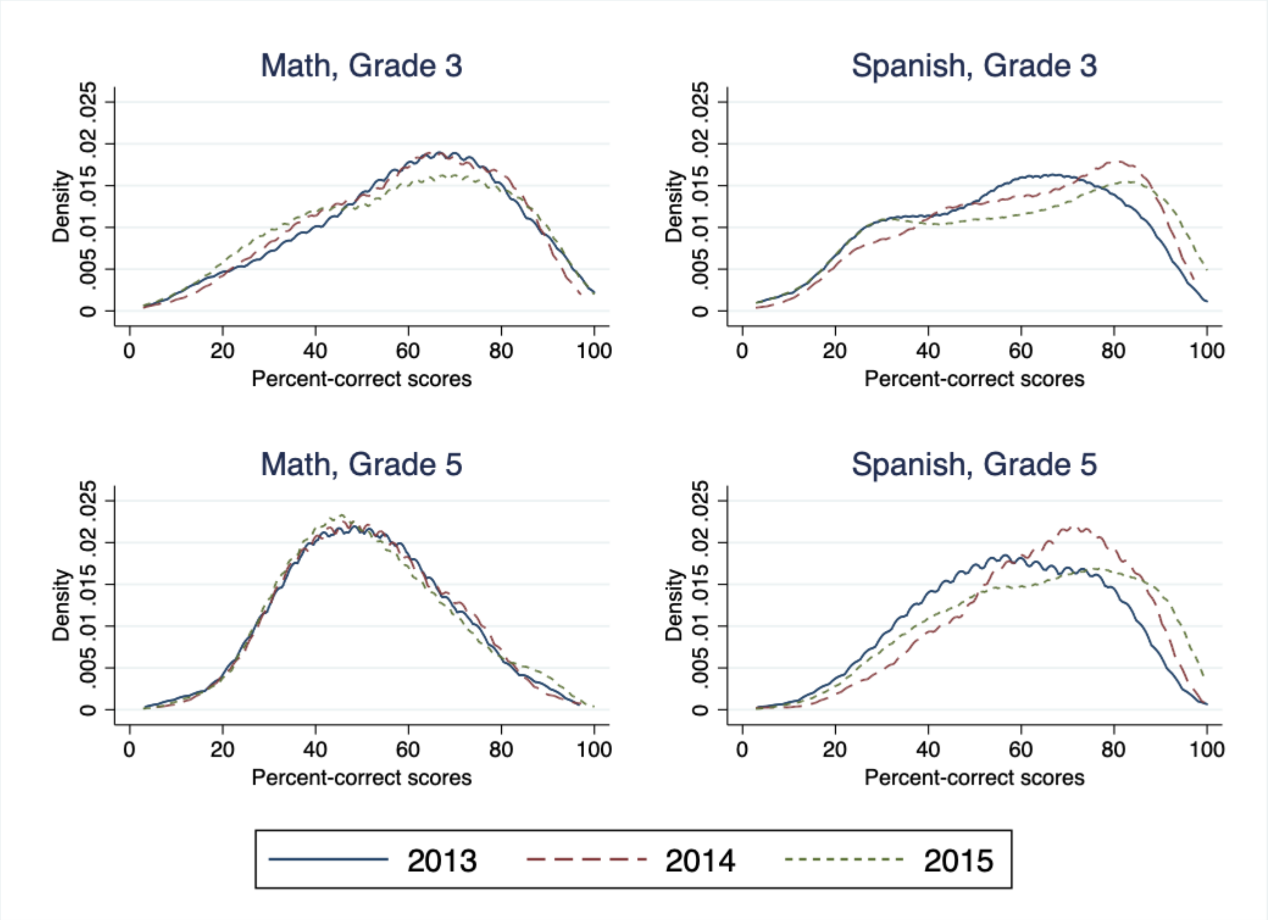
The think tank that designed the assessments scored all items dichotomously. We used a two-parameter logistic (2PL) Item Response Theory (IRT) model to scale and link the results (Harris 2005).[48] This model allows us to account for differences between items (specifically, differences in their difficulty and capacity to distinguish between students of similar ability). It also allows us to capitalize on common items across data collection rounds (within subjects and grades) to map assessment results for all three years of the study (2013 to 2015) onto the same scale. We standardized all IRT-scaled scores to have a mean of 0 and a standard deviation of 1 with respect to the control group in 2015.

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

The design, scaling, and linking processes described above were successful in producing well-behaved distributions in all grades, subjects, and years of the study, with little evidence of "floor" effects (i.e., a high concentration of students with no correct answers) or "ceiling" effects (i.e., a high concentration of students with perfect scores) (Figures B.1 and B.2).

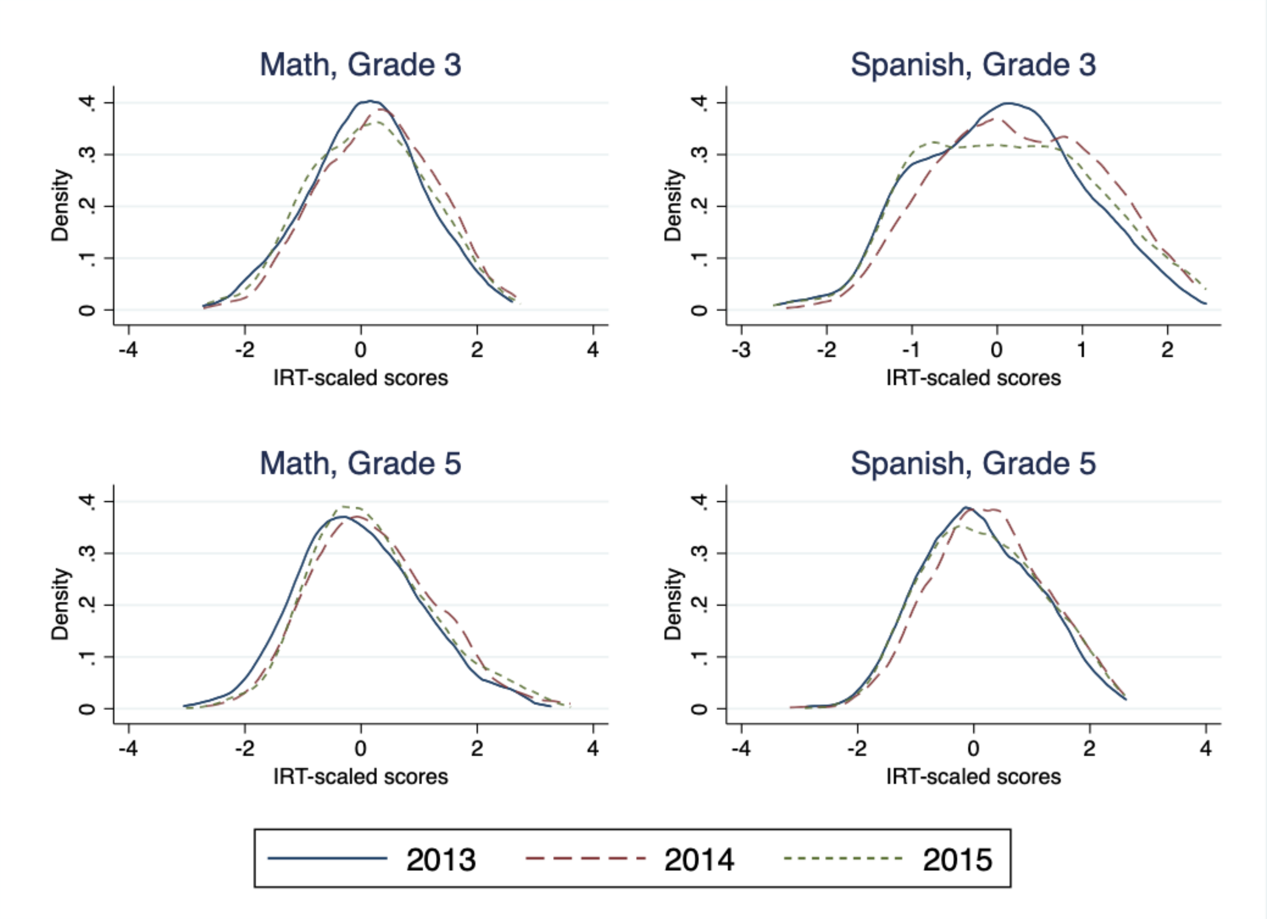
[48] We used the IRT program Stata 15 module (see Stata 2017). Our choice of a 2PL instead of a 3PL model was based partly on the sampling requirements for 3PL models discussed in Yen and Fitzpatrick (2006).

Figure B.1: Distribution of percent-correct scores on student assessments (2013-2015)



Notes: (1) This figure shows the distribution of percent-correct scores on the student assessments for this evaluation by subject and grade. (2) Each graph includes all students assessed on a given year (i.e., students in the diagnostic feedback or T1 and capacity building or T2 groups in 2013 and 2014, and students in those two groups and the control group in 2015) (see Table 1).

Figure B.2: Distribution of IRT-scaled scores on student assessments (2013-2015)



Notes: (1) This figure shows the distribution of IRT-scaled scores on the student assessments for this evaluation by subject and grade. (2) Each graph includes all students assessed on a given year (i.e., students in the diagnostic feedback or T1 and capacity building or T2 groups in 2013 and 2014, and students in those two groups and the control group in 2015) (see Table 1).

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

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

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