The World Bank Economic Review, 00(0), 2019, 1–22
doi: 10.1093/wber/lhz026
Article

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

Rafael de Hoyos, Alejandro J. Ganimian, and Peter A. Holland

Abstract

This article examines the impact of two strategies for using large-scale assessment results to improve school management and classroom instruction in the province of La Rioja, Argentina. In the study, 104 public primary schools were randomly assigned to three groups: a diagnostic-feedback group, in which standardized tests were administered at baseline and two follow-ups and results were made available to schools; a capacity-building group, in which workshops and school visits were conducted; and a control group, in which tests were administered at the second follow-up. After two years, diagnostic-feedback schools outperformed control schools by 0.33 standard deviations (σ) in mathematics and 0.36σ in reading. In fact, feedback schools still performed 0.26σ better in math and 0.22σ better in reading in the national assessment a year after the end of the intervention. Additionally, principals at feedback schools were more likely to use assessment results in making management decisions, and students were more likely to report that their teachers used more instructional strategies and to rate their teachers more favorably. Combining feedback with capacity building does not seem to yield additional improvements, but this could be due to schools assigned to receive both components starting from lower learning levels and participating in fewer workshops and visits than expected.

JEL classification: C93, I21, I22, I25

Keywords: diagnostic feedback, capacity building, large-scale assessments, Argentina

Rafael de Hoyos is a Lead Economist for Education at the World Bank; his e-mail is rdehoyos@worldbank.org. Alejandro J. Ganimian (corresponding author) is an Assistant Professor of Applied Psychology and Economics at New York University's Steinhardt School of Culture, Education, and Human Development; his e-mail is alejandro.ganimian@nyu.edu. Peter A. Holland is a Senior Specialist for Education at the World Bank; his e-mail is pholland@worldbank.org. The research for this article was financed by the World Bank. The authors thank Rita Abdala, Walter Flores, Silvia Romero, and Juan Vegas at the Ministry of Education of La Rioja, as well as 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, the authors thank Larry Aber, Felipe Barrera-Osorio, Elise Cappella, Joe Cimpian, Sean Corcoran, Andy de Barros, Emmerich Davies, Eric Edmonds, Andrew Ho, Jimmy Kim, Dan Koretz, Isaac Mbiti, Dick Murnane, Jeff Puryear, Jonah Rockoff, Jennifer Steele, Hiro Yoshikawa, seminar participants at APPAM, FRA, HGSE, NEUDC, NYU, and the World Bank, and two anonymous reviewers for useful comments. All views expressed are those of the authors and not of any of the institutions with which they are affiliated. A supplementary online appendix is available with this article at *The World Bank Economic Review* website.

1. Introduction

Over the past decade, developing countries have shifted their attention from expanding access to schooling to ensuring that 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 mathematics, reading, and science since 2004 (Cheng and Gale 2014). A similar effort found that 328 national and subnational school systems have participated in 37 international assessments of the same subjects from 1963 to 2015, with nearly half of them beginning their participation since 1995 (Ganimian and Koretz 2017).[2]

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

This article presents experimental evidence on two strategies of using large-scale assessment results for improvement in a developing country: diagnostic feedback and capacity building. In the study, 104 public primary schools in the province of La Rioja, Argentina, were randomly assigned to three groups: (a) a diagnostic-feedback (T1) group, in which standardized tests in math and reading comprehension were administered at baseline and two follow-ups and results were made available to schools through user-friendly reports; (b) a capacity-building (T2) group, in which professional development workshops and school visits for supervisors, principals, and teachers were also conducted; and (c) a control group, in which the tests were administered only at the second follow-up. This setup makes it possible to understand whether disseminating assessment results to schools is enough to prompt improvements in management and instruction or if principals and teachers need additional support to understand and act on this information.

Three sets of results are reported based on this experiment. First, 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 0.33 standard deviations (σ) in math and 0.36σ 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

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" (United Nations General Assembly 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*" (United Nations General Assembly 2015, emphasis added).

2 These figures are likely to increase as testing agencies develop assessments for low-income countries, which have so far 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, e.g., Betts, Hahn, and Zau 2017). However, it is not clear that the lessons from these studies would apply to developing countries, which have both less technically solid assessments and lower capacity to analyze and disseminate their results.

easier to improve in the short term, suggesting that the intervention did not result in a narrowing of the curriculum—or what is known as "teaching to the test." Further, the intervention led to improvements both on 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, students who were exposed to the intervention for two years still outperformed control peers by 0.26σ in math and 0.22σ in reading in the national student assessment a year after the end of the experiment, indicating that improvements are not test-dependent or short-lived.

Second, consistent with these effects, diagnostic feedback also led to changes in school management and classroom instruction. Principals in T1 schools were more likely than their control group counterparts to report using assessment results to inform management decisions (e.g., changing 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 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, combining information provision with workshops and school visits did not lead to additional improvements in school management, instruction, or learning. In fact, for most outcomes and potential mechanisms, there was no statistically significant impact of T2, and the possibility that T1 and T2 had the same effect cannot be discarded. However, these results should be interpreted with caution. First, it is possible that schools randomly assigned to T2 were already at a disadvantage relative to those assigned to T1 at baseline. The T1 and T2 groups had similar internal efficiency and student characteristics, but T2 schools performed 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, while a few T1 schools participated in these activities even though 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]

These 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 attaching stakes. 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, the findings of this study 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; see Muralidharan and Sundararaman [2010]) but holds promise in settings where the constraint is the intensive effort margin (raising productivity, conditional on attendance; see de Hoyos, García-Moreno, and Patrinos [2017]).

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."

6 Two results are consistent with this possibility. First, T2 had a statistically significant effect only on grade 5, where the initial difference between T1 and T2 schools was smaller. Second, the coefficients on T2 become larger if baseline covariates are introduced in grade 3, where the initial difference was larger.

7 Data from principal surveys and intervention monitoring verify that T1 and T2 schools received similar levels of support.

The rest of the article 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

Context

Schooling in Argentina is compulsory from the age of four until the end of secondary school. In 12 of the country's 24 provinces, including 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 preschool, 4.5 million in primary school, and 3.9 million in secondary school (DiNIECE 2015).

Argentina achieved near-universal access to primary education before most of Latin America (see Bassi, Busso, and Muñoz 2013). Yet, the relative performance of Argentina's primary school students in the region has deteriorated. In 1997, on the first regional student assessment in primary school, Argentine third graders 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 subnational (provincial) 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 are responsible 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 called the *Operativo Nacional de Evaluación* and currently known as *Aprender*), which has been in place since 1993, in conjunction with its counterparts in each province. To the authors' knowledge, only the Autonomous City of Buenos Aires conducts its own subnational student assessment regularly.

Argentina is an interesting setting in which to evaluate the impact of leveraging student assessments for diagnostic feedback and capacity building in schools. Over the past two decades, the country has taken multiple steps to limit 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) and then 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 contrast to 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 in Argentina has reversed some of these policies: (a) it adopted a new national assessment, to be administered annually, that covers all students at the end of primary and secondary school (grades 7 and 12) and a sample halfway through each level (grades 3 and 8), assessing the students in math, reading, and natural and social sciences (SEE-MEDN 2016);[10] and (b) it started

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

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 (Ganimian 2013; de Hoyos, Holland, and Troiano 2015).

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

sharing school-level assessment results with all principals using reports resembling the ones evaluated in this article.[11] Therefore, the questions examined in this article are not only of general interest to developing countries, but also of specific interest to Argentina.

This study was conducted in La Rioja for three reasons. First, it is one of the lowest-performing provinces in Argentina. In the latest national assessment, 41 percent of its sixth graders performed at the two lowest levels in math and 53 percent performed at the two lowest levels in reading (SEE-MEDN 2018). Second, La Rioja has one of the smallest subnational school systems in the country, so it is better positioned to implement a quality assurance mechanism; it is the seventh-smallest system in terms of schools (377 primary schools) and the fourth-smallest in terms of students (41,571 primary school students) (DiNIECE 2015). Third, La Rioja was one of the few provinces with the political will to experiment with a subnational assessment. The assessment was endorsed by the governor and the minister of education.

Sample

The sampling frame for the study included all 126 public primary schools in urban and semi-urban areas of La Rioja.[12] The frame was selected as follows. First, of the 394 primary schools in the province, all 29 private schools were excluded in order to focus on the effect of the interventions on public schools. Then, all 239 schools in rural areas were dropped because they are spread across the province, which would have limited the research team's capacity to implement the interventions.[13] A random sample of 104 urban and semi-urban public primary schools was drawn from this frame, stratified by enrollment in the 2013 school year.

The schools in the sample are comparable to all public primary schools in the province, and even more so to other urban and semi-urban public primary schools (see table S1.1 in the supplementary online appendix, available with this article at *The World Bank Economic Review* website). The average school in the sample enrolls more students, but this is mostly because rural schools, which are typically smaller, were excluded from the study. There are no statistically significant differences between in-sample and out-of-sample schools in the internal efficiency indicators collected by the school system, including pass, failure, dropout, and repetition rates.

The sample includes a *cross-section* of students and teachers 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, 4, and 5 participated; and in 2015, all students and teachers from grades 3 and 5 participated. All principals in selected schools participated in the study.

Randomization

The 104 public primary schools in the sample were randomly assigned to three groups: (a) a "diagnostic feedback" (T1) group, in which student assessments were administered at baseline and two follow-ups and results were made available to schools through user-friendly reports; (b) a "capacity building" (T2) group, in which professional development workshops and school visits were also provided for supervisors, principals, and teachers of the schools; and (c) a control group, in which the assessments were administered

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

12 Throughout this article, the term "semi-urban" refers to areas locally known as *rurales aglomeradas* and the term "rural" refers to 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 percent of the total), they serve only a small share of the students (less than 10 percent).

only at endline.[14] The randomization was stratified by enrollment in primary school to increase statistical power.[15]

This setup allows estimation of the effects 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).

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, the student assessments were administered at the end of each year (in only T1 and T2 schools in 2013 and 2014 and in all schools in 2015). School reports based on the prior-year assessments were delivered in T1 and T2 schools at the start of each year (in 2014 and 2015). Workshops and school visits for T2 schools were conducted during each school year (in 2014 and 2015).

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

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

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 discussed how to develop school improvement plans, one showed how to institute quality assurance

14 The rationale for this decision is discussed in section 2.4.3.

15 The randomization was conducted in June of 2013. Schools assigned to the T1 and T2 groups 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 The grades assessed in each year are given in table 1.

17 All scores were scaled and linked using a two-parameter item response theory (IRT) model.

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

Table 1. Timeline of the Study

Month (1)	Event (2)	School participation rates: Control schools (3)	T1 schools (4)	T2 schools (5)
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			

Source: Authors' own study design.
Note: The table shows the timeline for the interventions and rounds of data collection in 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).

mechanisms at the school level, and one focused on geometry instruction. The first four workshops were offered to supervisors and principals and the last one was offered 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 a report, including a diagnosis and some recommendations for improvement, which was shared with the school.[19]

As table 1 shows, participation in the workshops and school visits was lower than expected. Section 5.2 discusses the variation in dosage across T2 schools based on the endline data.

19 The workshops and school visits were based on the recommendations in Boudett, City, and Murnane (2005).

Control Group

Control schools were assessed only in 2015, at the end of the final 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), only the endline assessments were administered at these schools to estimate the impact of the interventions after two years.

3. Data

As table 1 shows, student assessments of math and reading and student and teacher surveys were administered in T1 and T2 schools for each year of the study (from 2013 to 2015). Principal surveys were conducted at the end of each intervention year (2014 and 2015). The assessments were administered in control schools only at the end of the study (in 2015).[21] Additionally, internal efficiency data from the census of schools (for 2013–2017) and learning outcomes data from the national student assessment (for 2016) were obtained.

Student Assessments

Student assessments of math and reading were administered before the interventions (in 2013) and after one and two years of the interventions (in 2014 and 2015) in T1 and T2 schools. The assessments were administered 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 had 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 four content domains (number, 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.

A two-parameter item response theory (IRT) model was used to scale the assessment results in a way that accounts for differences between items (their difficulty, capacity to distinguish between students of

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 subnational assessments (Ganimian 2015).

21 The grades assessed in 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, because they were present but excused from the tests, because they had dropped out of school, or because they had transferred to another school. However, there is no evidence of treatment schools having more students than control schools on such days; in fact, for 2015, the opposite is true and all differences are below 5 percentage points (see table S1.2 in the supplementary online appendix). The dropout rate in the sample is extremely low and varies little across experimental groups (tables S1.3 and S1.15 in the supplementary online appendix). School-level data on transfers or students being present and excluded on test day are not available, but there is 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 subnational curricular requirements.

similar ability, and propensity to be answered correctly by guessing) and that leverages common items across data collection rounds to link the results over time.

Supplementary online appendix S2 provides further details on the design, scaling, and linking of the assessments, as well as on the distribution of scores for all subjects, grades, and years of the study.[24]

Student Surveys

Surveys of students were also administered 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 collect information for describing the study sample. In 2015 (i.e., the second year of the interventions), the surveys 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, e.g., Ferguson 2010, 2012; Ferguson and Danielson 2014).[25]

Teacher Surveys

Surveys of teachers were conducted in the same years and groups as the student assessments. In 2013, the surveys asked about teachers' demographic characteristics, education and experience, professional development, and teaching practices to describe the study sample. In 2014 and 2015, the surveys 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]

Principal Surveys

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

National Assessments

The results from the national student assessment (called *Aprender*) for grade 6 students one year after the interventions (2016) were obtained from the ministry of education of the province. This is the only primary school grade for which the assessments cover all schools and students. It also corresponds to the cohort of students who were in grade 3 (one of the grades targeted by the interventions) in 2013, the first year of the study, and received two years of the interventions. These data are used to check whether there is evidence of fade-out.

Internal Efficiency

Finally, data on schools' internal efficiency (i.e., enrollment, pass, failure, repetition, and dropout rates) were obtained from the 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. The 2013 data are used to check balance across experimental groups, the 2014 and 2015 data to estimate the impact of the interventions, and the 2016 and 2017 data to check whether there is evidence of fade-out and/or dormant effects.

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 and 2015).

26 The surveys are available at http://bit.ly/20R1ni3 (2013) and http://bit.ly/1THNgr0 (2014 and 2015).

27 The survey is available at http://bit.ly/1TUkwyO (2014 and 2015).

4. Empirical Strategy

The effect of the offer (i.e., the intent-to-treat or ITT effect) of diagnostic feedback and capacity building after two years was estimated by fitting the model

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

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, $\alpha_{r(s)}$ 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 effect of each intervention relative to the control group. The null hypotheses are that the elements of β equal zero. Cluster-robust standard errors were used to account for within-school correlations across students in outcomes. The sensitivity of the estimates to the inclusion of X_{gs} was also tested.

Several variations of this model were also fitted, 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 sex (to estimate the heterogeneous effects of the interventions on subgroups of students); and (c) models in which the outcomes are measured at the student and school level in T1 and T2 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

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 (see table S1.3 in the supplementary online appendix). This is true regardless of whether they are compared based on all their primary school students or on students in the grades targeted by the interventions (grades 3 and 5 in 2013). The signs of the differences do not systematically favor any group,[31] and the magnitudes are small (less than 3.2 percentage points in all indicators). Only one of these differences is (marginally) statistically significant, which is less than would be expected given the number of hypotheses tested.

If groups are compared 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 (0.15σ in math and 0.16σ in reading) than in grade 5 (0.1σ and 0.12σ, respectively); see table S1.4 in the supplementary online appendix. Students at T1 schools also have marginally more educated parents, greater household assets, and more study supports than their T2 counterparts (table S1.5 in the supplementary online appendix). None of these differences is statistically significant, but, as discussed below, they may explain why the capacity-building intervention does not have a statistically significant effect on most outcomes.

28 These indicators include enrollment as well as pass, failure, dropout, and repetition rates.

29 It is not possible to 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.

30 In these models, it is not possible to account for students' outcomes before the intervention because neither students nor teachers were assigned unique identifiers to allow them to be tracked over time.

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

Implementation Fidelity

The interventions were implemented mostly as intended; see table 2. In surveys administered right before the endline (in 2015), the principals of nearly all diagnostic-feedback (T1) and capacity-building (T2) schools reported having administered student assessments at their schools that year (table 2 panel A, columns 2 and 3).[32] Some control group principals also reported administering assessments, but given that there were no national or subnational assessments in 2015, it is possible that these principals either thought that they were supposed to administer assessments and claimed to have done so even when they did not, or had in mind other types of assessments such as classroom tests (panel A, column 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 such 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]

Table 2. Administration and Use of Student Assessments, 2015

	Control schools (1)	T1 schools (2)	T2 schools (3)	Col. (2) – Col. (1) (4)	Col. (3) – Col. (1) (5)	*F*-test $\beta_1 = \beta_2$ (6)
Panel A: Principal survey						
Students at my school took national or subnational assessments	0.4	0.962	0.931	0.562***	0.529***	0.291
	(0.497)	(0.196)	(0.258)	(0.093)	(0.098)	(0.591)
I compared my school's results with those of the province	0.25	0.679	0.75	0.426***	0.501***	0.391
	(0.441)	(0.476)	(0.441)	(0.123)	(0.119)	(0.533)
I compared my school's results with those of other schools	0.25	0.741	0.519	0.47***	0.276**	2.215
	(0.441)	(0.447)	(0.509)	(0.122)	(0.129)	(0.141)
I compared my school's results over time	0.677	1	0.966	0.323***	0.29***	0.706
	(0.475)	(0)	(0.186)	(0.086)	(0.095)	(0.403)
Panel B: Implementation monitoring						
Number of reports received	1	3	3	2	2	0
	(0)	(0)	(0)	(0)	(0)	(0)
Number of workshops attended	0	0.233	3.4	0.22***	3.401***	334.158
	(0)	(0.43)	(0.855)	(0.075)	(0.156)	(0)
Number of visits received	0	0.733	1.467	0.758***	1.462***	24.782
	(0)	(0.583)	(0.571)	(0.1)	(0.105)	(0)

Source: Authors' analysis based on principal surveys administered by the research team and data from the Ministry of Education of La Rioja.
Note: The table shows, for the 2015 school year, the means and standard deviations of the control group (column 1), diagnostic-feedback or T1 group (column 2), and capacity-building or T2 group (column 3). It also estimates the intent-to-treat (ITT) effect of T1 and T2 with respect to control schools, using randomization fixed effects (columns 4 and 5). Finally, it reports 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 conducted in 2015, and Panel B uses implementation monitoring data from 2013–2015. In the surveys, principals were asked to indicate whether their schools used student assessment results for the purposes listed; each value represents the share of principals who reported that their school used assessments for the purpose shown in that row, based on the school year of data collection. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

32 The survey asked principals about whether their students had taken national or subnational assessments. It did not specifically refer to the intervention in order 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).

33 A key drawback of relying solely on self-reported data, and the reason that intervention monitoring data were also collected, 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 It is possible that the responses of principals of control schools to these three questions are subject to the same potential problems as responses to the previous question.

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 (table 2 panel B, columns 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 (panel B, column 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 though they were not supposed to do so (panel B, column 2). Therefore, the effect of T1 should be interpreted as that of diagnostic feedback with minimal capacity building.

Average ITT Effects

Student Achievement

Diagnostic feedback had a positive, large, and statistically significant effect on student achievement, but there is 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 0.07σ or 2 percentage points) and statistically insignificant (table 3, columns 1–3). After two years (in 2015), T1 and T2 schools outperformed control schools (columns 4–8). Only the differences between control and T1 schools are statistically significant, and they are large in both subjects (0.33σ or 6.3 percentage points in math and 0.36σ or 7.7 percentage points in reading), but the possibility that T1 and T2 had the same impact on student achievement (column 9) cannot be discarded.[35]

Table 3. ITT Effect of the Interventions on Student Achievement, 2014 and 2015

	2014			2015					
	T1 schools (1)	T2 schools (2)	Col. (2) – Col. (1) (3)	Control schools (4)	T1 schools (5)	T2 schools (6)	Col. (5) – Col. (4) (7)	Col. (6) – Col. (4) (8)	*F*-test $\beta_1 = \beta_2$ (9)
Math (percent-correct score)	55.779	54.941	−0.635	51.552	57.69	55.46	6.275**	3.939	0.538
	(18.059)	(19.614)	(2.432)	(19.417)	(19.599)	(20.761)	(2.422)	(2.757)	(0.465)
Math (IRT-scaled score)	0.278	0.243	−0.023	0	0.324	0.216	0.333**	0.216	0.481
	(0.992)	(1.075)	(0.139)	(1)	(1.036)	(1.104)	(0.131)	(0.145)	(0.489)
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)	(0.162)
Reading (IRT-scaled score)	0.312	0.212	−0.066	0	0.357	0.153	0.356***	0.166	1.987
	(0.969)	(0.966)	(0.105)	(1)	(1.018)	(1.035)	(0.102)	(0.111)	(0.162)
N (number of students)	3950	3588	7538	3446	3950	3588	10984	10984	10984

Source: Authors' analysis based on data from student assessments administered by the research team.

Note: 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 intent-to-treat (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 and 8). Finally, it reports the *F*-statistic and associated *p*-value for the null hypothesis that the coefficients of T1 and T2 are equal in 2015 (column 9). All test scores are shown as percent-correct scores and as scores scaled using item response theory (IRT), standardized with respect to the control group in 2015. Control schools were assessed only in 2015 (see sections 2.3 and 2.4 of the article). *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

35 There are no statistically significant heterogeneous ITT effects on learning by school size or location. These estimates are omitted from the article but are available from the authors upon request.

The results above are robust to checks for student attrition and multiple hypothesis testing. A similar pattern is observed when Lee (2009) bounds are computed to account for the potential influence of student attrition in 2015.[36] The bounds are very wide: for T1 schools, they range from -0.18σ to 0.89σ in math and from -0.17σ to 0.92σ in reading; for T2 schools, they range from 0.31σ to 0.76σ in math and from -0.36σ to 0.7σ in reading. Yet, they are mostly positive (table S1.13 in the supplementary online appendix).[37] Further, the statistical significance of the coefficients on T1 and T2 remains unchanged when false discovery rate (FDR) q-values are computed for 2014 and 2015 (table S1.14 in the supplementary online appendix).

The positive effect of diagnostic feedback is not limited to only certain topics or skills. In fact, T1 schools outperformed control schools in nearly all content and cognitive domains in both grades and both subjects (tables S1.7 and S1.8 in the supplementary online appendix). This finding is important because it suggests that the intervention did not lead schools to focus on more malleable skills in order 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 both on items that were repeated from prior assessment rounds (referred to as "familiar" items) and on items that were introduced for the first time on each assessment round ("unfamiliar" items); see table S1.9 in the supplementary online appendix. 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 as large as that of T1.[39] However, some evidence suggests that this may be due to the initial imbalance between T1 and T2.[40] First, the effect of T2 is commensurate with that imbalance: it is small (0.14σ in math and 0.11σ in reading) and statistically insignificant in grade 3, where the imbalance was greater, and it is large (0.29σ in math and 0.22σ in reading) and statistically significant in grade 5, where the imbalance was smaller (table S1.6 in the supplementary online appendix). Second, the effect of T2 becomes slightly larger in grade 3 if covariates are used (increasing from 0.14σ to 0.15σ in math and from 0.11σ to 0.12σ in reading), but it does not change in grade 5 (table S1.10 in the supplementary online appendix, column 8). Admittedly, however, these patterns are suggestive rather than conclusive.

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 in the first year of the study (2013), the low take-up of workshops and school visits among T2 principals and the participation of some T1 principals

36 All Lee (2009) bounds estimated in this article involve analytic standard errors. Therefore, they should be interpreted with caution, as they ignore the school-level error component.

37 The ITT estimates change little when absences on test day are accounted for (table S1.11 in the supplementary online appendix). Further, there is 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 0.1 for all grades and subjects. The coefficients for T2 schools are slightly larger, but this may be due to the baseline imbalance discussed in section 5.1. The differences in the magnitudes of the coefficients across grades are consistent with these hypotheses: coefficients are larger in grade 3 (where there is greater imbalance between T1 and T2 schools at baseline) and smaller in grade 5 (where there is less imbalance).

38 Capacity building did not have a larger effect on geometry. This could be because only 20 percent 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 because of 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 discussion of this imbalance.

41 See section 5.2 for discussion of these problems.

in those activities may have preserved that disadvantage.[42] The role of noncompliance on the effect of capacity building cannot, however, be ascertained.

National Assessments

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 though control schools had received one report at the end of the experiment, as described in section 5.2 (table 5, columns 4 and 5).[43] The effects were small to moderate (0.26σ in math and 0.22σ in reading among T1 schools and 0.17σ in math and 0.18σ in reading among T2 schools). The coefficients on T1 are larger than those on T2, but consistent with the results from the assessments, it cannot be ruled out that both interventions had the same effect (column 6).

Internal Efficiency

There is no clear evidence that the interventions improved schools' internal efficiency (table 4), either while they were being implemented (in 2014 and 2015, columns 1–6), or for up to two years after they concluded (in 2016 and 2017, columns 7–12). For 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). For other indicators and years, however, 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. A similar pattern emerges when effects are estimated separately by grade (table S1.15 in the supplementary online appendix).

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

Potential Mechanisms

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 were, respectively, 48 and 28 percentage points more likely than their control peers to report that they set goals for their schools based on assessment results, and were 30 and 23 percentage

42 This possibility is consistent with analyses in which random assignment to T2 was used to estimate the effect of take-up of capacity building, finding it to have 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.

44 As discussed in a prior footnote, principals were asked whether they used results from national or subnational 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 subnational 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 in a prior footnote, the results in this section should be interpreted with caution, as principals' reports do not always correspond with school practices.

Table 4. ITT Effect of the Interventions on Internal Efficiency, 2014–2017

	2014			2015			2016			2017		
		Difference			Difference			Difference			Difference	
	Control (1)	T1 (2)	T2 (3)	Control (4)	T1 (5)	T2 (6)	Control (7)	T1 (8)	T2 (9)	Control (10)	T1 (11)	T2 (12)
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	−0.626	−1.69*	97.155	−0.167	0.556	98.416	0.147	0.009	98.774	−0.358	−0.116
	(4.31)	(1.018)	(1.014)	(3.497)	(0.842)	(0.838)	(2.662)	(0.644)	(0.641)	(1.904)	(0.452)	(0.45)
Percentage of students who failed the grade	3.682	0.694	1.718*	2.571	−0.067	−0.67	1.432	−0.188	−0.202	1.175	0.313	0.017
	(4.32)	(1.02)	(1.016)	(3.412)	(0.824)	(0.82)	(2.516)	(0.606)	(0.603)	(1.869)	(0.444)	(0.441)
Percentage of students who dropped out of school	0.038	−0.067*	−0.028	0.274	0.234	0.115	0.152	0.041	0.193	0.051	0.045	0.099*
	(0.153)	(0.037)	(0.036)	(0.929)	(0.22)	(0.219)	(0.75)	(0.18)	(0.18)	(0.214)	(0.051)	(0.05)
Percentage of students who repeated the grade	2.649	0.581	−0.076	1.922	0.436	−0.107	1.453	−0.711	−1.09**	2.535	−0.025	−1.552*
	(3.191)	(0.771)	(0.767)	(2.567)	(0.615)	(0.612)	(2.279)	(0.536)	(0.533)	(3.593)	(0.851)	(0.847)
N (number of schools)	44	104	104	44	104	104	44	104	104	44	104	104

Source: Authors' analysis based on internal efficiency data collected through the annual census of schools and provided by the Ministry of Education of La Rioja.

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

Table 5. ITT Effect of the Interventions on National Student Assessment, 2016

	Control schools (1)	T1 schools (2)	T2 schools (3)	Col. (2) – Col. (1) (4)	Col. (3) – Col. (1) (5)	*F*-test $\beta_1 = \beta_2$ (6)
Math (IRT-scaled score)	0	0.259	0.164	0.257**	0.173**	0.503
	(1)	(1.082)	(1.038)	(0.107)	(0.079)	(0.48)
Reading (IRT-scaled score)	0	0.226	0.164	0.221**	0.177*	0.139
	(1)	(1.034)	(1.04)	(0.099)	(0.102)	(0.71)
N (number of students)	1661	1291	1217	3538	3538	3538

Source: Authors' analysis based on results of the national student assessments provided by the Ministry of Education of La Rioja.
Note: 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 intent-to-treat (ITT) effect of T1 and T2 with respect to control schools in 2015, using randomization fixed effects (columns 4 and 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). All test scores are shown as scores scaled using item response theory (IRT), standardized with respect to the control group in 2016. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

points more likely to report that they changed the curriculum based on the results (table 6, columns 4 and 5). All of these differences are statistically significant, and the possibility that T1 and T2 had the same effect on either indicator cannot be discarded (column 6).

Second, principals at these schools also used assessment results to make staffing decisions. Principals at T1 and T2 schools were, respectively, 28 and 37 percentage points more likely than their control counterparts to report being evaluated based on assessment results. Principals at T1 schools were also 23 percentage points more likely to report using assessment results to evaluate teachers, but this difference

Table 6. ITT Effect of the Interventions on Principal-Reported School Management, 2015

	Control schools (1)	T1 schools (2)	T2 schools (3)	Col. (2) – Col. (1) (4)	Col. (3) – Col. (1) (5)	*F*-test $\beta_1 = \beta_2$ (6)
My school set goals based on assessment results	0.483	0.964	0.963	0.475***	0.484***	0.021
	(0.509)	(0.189)	(0.192)	(0.103)	(0.103)	(0.885)
I made changes to the curriculum based on assessment results	0.7	1	0.931	0.295***	0.233**	1.665
	(0.466)	(0)	(0.258)	(0.086)	(0.099)	(0.2)
I am evaluated partly based on assessment results	0.333	0.615	0.704	0.284**	0.37***	0.405
	(0.479)	(0.496)	(0.465)	(0.135)	(0.127)	(0.526)
My teachers are evaluated partly based on assessment results	0.452	0.69	0.654	0.233*	0.208	0.035
	(0.506)	(0.471)	(0.485)	(0.128)	(0.133)	(0.853)
I assign students to sections based on assessment results	0.107	0.231	0.107	0.11	0	1.226
	(0.315)	(0.43)	(0.315)	(0.099)	(0.086)	(0.271)
I informed parents about the results of their children	0.452	0.857	0.931	0.423***	0.485***	0.542
	(0.506)	(0.356)	(0.258)	(0.11)	(0.104)	(0.464)
I made my school's assessment results public	0.207	0.519	0.519	0.318**	0.309**	0.004
	(0.412)	(0.509)	(0.509)	(0.126)	(0.126)	(0.952)
N (number of schools)	42	29	30	101	101	101

Source: Authors' analysis based on data from principal surveys administered by the research team.
Note: The table shows, for the 2015 school year, the means and standard deviations of the control group (column 1), diagnostic-feedback or T1 group (column 2), and capacity-building or T2 group (column 3). It also estimates the intent-to-treat (ITT) effect of T1 and T2 with respect to control schools, using randomization fixed effects (columns 4 and 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). Principals were asked whether their schools used student assessment results for the purposes listed; each value represents the share of principals who reported that their school used assessments for the purpose shown in that row, based on the school year of data collection. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

is only marginally statistically significant, insignificant in the case of T2 schools, and the null that T1 and T2 had the same effect on this indicator cannot be rejected.[46]

Finally, principals made assessment results available to others. Principals at T1 and T2 schools were, respectively, 42 and 49 percentage points more likely than principals at control schools to report that they informed parents of assessment results, and were 32 and 31 percentage points more likely to report making these results public. All of these differences are statistically significant, and the possibility that T1 and T2 had the same effect cannot be discarded. These findings suggest 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 S1.16 in the supplementary online appendix).

Classroom Instruction

Another way in which feedback may have improved learning is through classroom instruction. The student 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, arrive on time, or 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, column 4); however, none of these results is statistically significant once multiple hypothesis testing is accounted for (table S1.18 in the supplementary online appendix).

Table 7. ITT Effect of the Interventions on Student-Reported Teacher Time Use, 2015

	Control schools (1)	T1 schools (2)	T2 schools (3)	Col. (2) − Col. (1) (4)	Col. (3) − Col. (1) (5)	*F*-test $\beta_1 = \beta_2$ (6)
My teacher was absent from school	0.578	0.554	0.6	−0.025	0.026	2.206
	(0.494)	(0.497)	(0.49)	(0.031)	(0.03)	(0.141)
My teacher arrived late to school	0.394	0.364	0.413	−0.032	0.021	1.411
	(0.489)	(0.481)	(0.492)	(0.033)	(0.04)	(0.238)
My teacher started class late	0.443	0.394	0.446	−0.05*	0.005	1.763
	(0.497)	(0.489)	(0.497)	(0.029)	(0.036)	(0.187)
My teacher ended class early	0.5	0.441	0.492	−0.06**	−0.006	1.699
	(0.5)	(0.497)	(0.5)	(0.03)	(0.035)	(0.195)
My teacher left school early	0.449	0.389	0.433	−0.06*	−0.014	0.982
	(0.497)	(0.488)	(0.496)	(0.035)	(0.039)	(0.324)
N (number of students)	4034	3014	2854	9902	9902	9902

Source: Authors' analysis based on data from student surveys administered by the research team.

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

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. Hence, it is unlikely that they attached stakes to these results. Instead, it is more likely that schools 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.

Table 8. ITT Effect of the Interventions on Student-Reported Teacher Activity, 2015

	Control schools (1)	T1 schools (2)	T2 schools (3)	Col. (2) – Col. (1) (4)	Col. (3) – Col. (1) (5)	*F*-test $\beta_1 = \beta_2$ (6)
I used a textbook	0.814	0.872	0.85	0.058***	0.036**	1.869
	(0.389)	(0.334)	(0.357)	(0.016)	(0.016)	(0.175)
My teacher assigned me homework	0.936	0.958	0.949	0.022***	0.014	0.888
	(0.245)	(0.2)	(0.22)	(0.008)	(0.009)	(0.348)
I copied from the blackboard	0.912	0.938	0.913	0.026**	0.001	4.186
	(0.283)	(0.241)	(0.281)	(0.012)	(0.01)	(0.043)
I worked with a group	0.916	0.936	0.921	0.021	0.004	1.506
	(0.278)	(0.245)	(0.27)	(0.014)	(0.013)	(0.223)
My teacher explained a topic	0.96	0.977	0.969	0.018***	0.009	1.992
	(0.196)	(0.149)	(0.174)	(0.006)	(0.006)	(0.161)
My teacher asked me to take a practice test	0.892	0.911	0.9	0.02	0.008	0.758
	(0.311)	(0.285)	(0.299)	(0.014)	(0.011)	(0.386)
My teacher graded my homework	0.958	0.975	0.956	0.018***	−0.004	9.084
	(0.2)	(0.156)	(0.206)	(0.006)	(0.007)	(0.003)
N (number of students)	4034	3014	2854	9902	9902	9902

Source: Authors' analysis based on data from student surveys administered by the research team.

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

Second, teachers at T1 schools used more activities during lessons. According to the reports of students, teachers at T1 schools were 6 percentage points more likely to use a textbook, 2 percentage points more likely to assign homework, 3 percentage points more likely to write on the board, 2 percentage points more likely to explain topics, and 2 percentage points more likely to grade homework than teachers at control schools (table 8, column 4). The statistical significance of these coefficients remains unchanged once multiple hypothesis testing is accounted for (table S1.19 in the supplementary online appendix). However, teachers were no more likely to assign practice tests, which suggests that the intervention did not lead to "teaching to the test." For most of these outcomes, the possibility that T1 and T2 had a similar effect cannot be ruled out (column 6).

Finally, teachers at T1 schools were rated more favorably. Students at T1 schools were more likely to report that their teachers demonstrate interest in students (0.19σ), manage their classrooms (0.13σ), clarify difficult concepts or tasks (0.17σ), motivate students to perform at their best (0.18σ), deliver captivating lessons (0.16σ), engage students in discussions (0.14σ), and summarize the material at the end of a lesson (0.18σ) than in control schools (table 9, column 4). The statistical significance of all coefficients remains unchanged once multiple hypothesis testing is accounted for (table S1.20 in the supplementary online appendix). In fact, T1 schools outperformed T2 schools (column 6).

6. Conclusions

This article presents experimental evidence on the effects of diagnostic feedback and capacity building on public primary schools in La Rioja, Argentina, and finds that providing schools with information on their relative performance led to large test-score gains in math and reading. The results in tables 6–9 indicate that this improvement was driven by principals using assessment results to inform school management

Table 9. ITT Effect of the Interventions on Student-Reported Teacher Effectiveness, 2015

	Control schools (1)	T1 schools (2)	T2 schools (3)	Col. (2) − Col. (1) (4)	Col. (3) − Col. (1) (5)	*F*-test $\beta_1 = \beta_2$ (6)
Care (standardized score)	0	0.186	0.003	0.191***	−0.006	17.419
	(1)	(0.854)	(0.967)	(0.05)	(0.052)	(0)
Control (standardized score)	0	0.128	0.029	0.133***	0.02	4.108
	(1)	(0.9)	(0.949)	(0.05)	(0.053)	(0.045)
Clarify (standardized score)	0	0.163	0.053	0.168***	0.048	4.386
	(1)	(0.865)	(0.951)	(0.056)	(0.053)	(0.039)
Challenge (standardized score)	0	0.171	0.017	0.176***	0.01	7.21
	(1)	(0.887)	(1.01)	(0.06)	(0.057)	(0.008)
Captivate (standardized score)	0	0.152	0.022	0.157***	0.017	8.523
	(1)	(0.822)	(0.943)	(0.042)	(0.049)	(0.004)
Confer (standardized score)	0	0.129	−0.001	0.137**	−0.009	6.244
	(1)	(0.874)	(0.995)	(0.055)	(0.056)	(0.014)
Consolidate (standardized score)	0	0.168	−0.004	0.177***	−0.017	11.054
	(1)	(0.881)	(0.983)	(0.054)	(0.054)	(0.001)
N (number of students)	4034	3014	2854	9902	9902	9902

Source: Authors' analysis based on data from student surveys administered by the research team.

Note: The table shows, for the 2015 school year, the means and standard deviations of the control group (column 1), diagnostic-feedback or T1 group (column 2), and capacity-building or T2 group (column 3). It also estimates the intent-to-treat (ITT) effect of T1 and T2 with respect to control schools, using randomization fixed effects (columns 4 and 5). Students were asked to indicate how frequently their teacher engaged in certain behaviors (e.g., treating students 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: (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. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

decisions, teachers employing more instructional strategies, and improved 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 have recently begun administering large-scale assessments (see Cheng and Gale 2014; Ganimian and Koretz 2017), that these assessments account for only a small share of the countries' education budgets (see Wolff 2007), and that school reports can be automated and distributed at little to no cost (e.g., online), provision of diagnostic feedback promises to be a cost-effective way of improving student learning in these settings (World Bank 2018).

It is not possible to distinguish between the relative contributions of the various components of diagnostic feedback, but it is important to note that the intervention was based on assessments that were informative over a wide range of achievements and comparable over time. There are good reasons to believe that these aspects are fundamental for feedback purposes. Assessments that cover only the material that students are supposed to know in a given grade (i.e., without testing material from lower grades), which are 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, Singh, and Ganimian 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 IRT for linking purposes) may convey

inconsistent information about the relative and absolute performance of schools, leading users of such information to make incorrect decisions (see Barrera-Osorio and Ganimian 2016).

In the same vein, it seems equally important to highlight that the school reports featured different types of information that may prove useful for principals and teachers, including comparisons not only between the average performance of a school and that of the average school in the school system, but also between sections within a school, as well as of the school's overall performance in each content and cognitive domain of each subject assessed. It is not possible to determine which of these different types of information was most useful for schools, but the authors caution against expecting that reports that simply compare a school with the rest of the system would result in the large test-score gains observed here. Experimentation with the types of information presented in school reports, and with the format in which that information is presented, is an interesting area for further research.

Finally, this study's findings on the lower-than-expected take-up of capacity building illustrate the challenges of implementing meaningful professional development in the developing world. These results are consistent with those from evaluations of traditional teacher training programs in other developing countries, which have also found low take-up and limited effects on learning (see Yoshikawa et al. 2015; Zhang et al. 2013; Angrist and Lavy 2001; Yue et al. 2014). Experimentation with more innovative models of professional development (e.g., coaching) seems a promising direction for future research (see, e.g., 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): 1535–63.

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): 343–69.

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, Abdul Latif Jameel Poverty Action Lab (J-PAL), New Delhi, India.

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 (July): 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, Inter-American Development Bank, Washington, DC.

Bassi, M., C. Meghir, and A. Reynoso. 2016. "Education Quality and Teaching Practices." NBER Working Paper No. 22719, National Bureau of Economic Research, Cambridge, MA.

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 (July): 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." *Journal of Human Resources* 53 (2): 414–44.

Cheng, X., and C. Gale. 2014. "National Assessments Mapping Metadata." Washington, DC: fhi360, Education Policy and Data Center. http://bit.ly/2yxBeBd. Last accessed October 17, 2019.

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*. doi: 10.3368/jhr.55.3.0618-9538R1.

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 (June): 123–40.

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, World Bank, Washington, DC.

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

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

———. 2013. *Anuario Estadístico 2013*. Buenos Aires, Argentina: Dirección Nacional de Información de la Calidad Educativa.

———. 2015. *Anuario Estadístico 2015*. Buenos Aires, Argentina: Dirección Nacional de Información de la Calidad Educativa.

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. Unpublished manuscript.

———. 2012. "Can Student Surveys Measure Teaching Quality?" *Phi Delta Kappan* 94 (3): 24–8.

Ferguson, R. F., and C. Danielson. 2014. "How Framework for Teaching and Tripod 7Cs Evidence Distinguish Key Components of Effective Teaching." In *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*, edited by T. J. Kane, K. A. Kerr, and R. C. Pianta. San Francisco, CA: Jossey-Bass.

Ferrer, G. 2006. *Educational Assessment Systems in Latin America: Current Practice and Future Challenges*. Washington, DC: 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." Working Paper No. 100245, World Bank, Washington, DC.

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.

———. 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.

———. 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." Cambridge, MA: Harvard Graduate School of Education. https://bit.ly/33FrvUI. Last accessed October 17, 2019.

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*. Boston, MA: TIMSS & PIRLS International Study Center, Boston College Lynch School of Education, and the International Association for the Evaluation of Educational Achievement (IEA).

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

Lee, D. S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76 (3): 1071–102.

Mizala, A., and M. Urquiola. 2013. "School Markets: The Impact of Information Approximating Schools' Effectiveness." *Journal of Development Economics* 103: 313–35.

Muralidharan, K. 2012. "Long-Term Effects of Teacher Performance Pay: Experimental Evidence from India." Unpublished manuscript, University of California, San Diego, CA.

Muralidharan, K., A. Singh, and A. J. Ganimian. 2019. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India." *American Economic Review* 109 (4): 1–35.

Muralidharan, K., and V. Sundararaman. 2010. "The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India." *Economic Journal* 120 (546): F187–203.

———. 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.

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*. Buenos Aires, Argentina: Secretaría de Evaluación Educativa. Ministerio de Educación y Deportes de la Nación.

———. 2018. *Aprender 2017: Informe de resultados, secundaria*. Buenos Aires, Argentina: Secretaría de Evaluación Educativa. Ministerio de Educación y Deportes de la Nación.

Stata. 2017. *Stata 15 Item Response Theory Reference Manual*. College Station, TX: StataCorp LLC.

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

———. 2015. A/RES/70/1. Resolution adopted by the General Assembly on September 25, 2015, New York, NY.

Wolff, L. 2007. "The Costs of Student Assessments in Latin America." PREAL Working Paper No. 38, Partnership for Educational Revitalization in the Americas (PREAL), Washington, DC.

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 *Educational Measurement*, 4th ed., edited by R. Brennan. Westport, CT: American Council on Education and Praeger Publishers.

Yoshikawa, H., D. Leyva, C. E. Snow, and E. Treviño et al. 2015. "Experimental Impacts of a Teacher Professional Development Program in Chile on Preschool Classroom Quality and Child Outcomes." *Journal of Developmental Psychology* 51 (3): 309–22.

Yue, A., Y. Shi, F. Chang, and C. Yang et al. 2014. "Dormitory Management and Boarding Students in China's Rural Primary Schools." *China Agricultural Economic Review* 6 (3): 523–50.

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): 339–58.