Policy Research Working Paper 8156

Through the Looking Glass

Can Classroom Observation and Coaching Improve Teacher Performance in Brazil?

Barbara Bruns
Leandro Costa
Nina Cunha

WORLD BANK GROUP
Education Global Practice Group
July 2017

Policy Research Working Paper 8156

Abstract

This study conducted a randomized evaluation of a program in the Brazilian state of Ceará. The program was designed to improve teachers' effectiveness by increasing their professional interaction and sharing of classroom practice. In 175 of 350 secondary schools, teachers were provided with benchmarked feedback from classroom observations and access to expert coaching. Schools' uptake of the coaching program was high (85 percent). Over a single school year, the program increased teachers' time on instruction and student engagement and produced statistically significant gains in student learning on the Ceará state assessment and the national secondary school exit exam. Controlling for individual students' prior-year learning outcomes, schools exposed to the program had 0.05–0.09 standard deviation higher performance on the state test and 0.04–0.06 standard deviation higher scores on the national test. Implementation fidelity strongly boosted program impacts. In the 49 schools where the pedagogical coordinators achieved the highest certification at the end of the program, student scores were 0.13–0.23 standard deviation higher on the state test and 0.13–0.17 standard deviation higher on the national test. Coaching delivered by Skype kept the costs of the program low, $2.40 per student, and produced cost-effective impacts on learning in comparison with other rigorously evaluated teacher training interventions. The combination of classroom observation feedback and expert coaching appears to be a promising strategy for whole-school efforts to raise teacher effectiveness.

This paper is a product of the Education Global Practice Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at lcosta@worldbank.org.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Produced by the Research Support Team

Through the Looking Glass: Can Classroom Observation and Coaching Improve Teacher Performance in Brazil?

Barbara Bruns*, Leandro Costa**, and Nina Cunha***[1]

JEL: I21, I28, H43, C52, C93.

Keywords: classroom observation, teacher feedback, coaching, instructional time, student engagement, student learning, and randomized control trial.

* Center for Global Development
** World Bank
*** Stanford University

1. Introduction

A central education policy question is how to improve teachers' classroom effectiveness. Research in the United States (Jackson et al, 2014; Chetty et al, 2014; Hanushek and Rivkin, 2010; Rockoff, 2004) on teacher value-added and in Latin America (Araujo et al, 2016; Bruns and Luque 2014) on observed classroom practice has consistently documented large variations in teachers' practice and classroom-level results, even among teachers in the same school teaching the same grade and subject.

There is new research interest in observing teachers' classroom practice and unpacking what affects it. First, there is growing evidence that the quality of teachers' classroom practice, as measured through classroom observations, is important for student learning and other key outcomes, such as students' socio-emotional skills. The influential large-scale Measures of Effective Teaching study in the US found that classroom observations using three different instruments could predict differences in individual teachers' ability to produce classroom-level learning gains (MET, Kane and Staiger 2008). Other US researchers have also found that children exposed to teachers with better scores on the CLASS classroom observation instrument have higher learning gains, better self-regulation and fewer behavioral problems (Howes et al, 2008; Grossman et al, 2010). The only research to date in a developing country, by Araujo et al (2016) in Ecuador, has produced similar findings. By randomly assigning pre-school students to different teachers, Araujo and colleagues found that a one standard deviation increase in teachers' classroom quality, measured using the CLASS observation instrument, resulted in 0.11, 0.11 and 0.07 standard deviation higher student test scores in language, math and executive function.

Beyond these studies, which have directly linked teachers' classroom practice to classroom level outcomes, there is a larger body of research that has not measured teachers' classroom practice, but has linked classroom-level outcomes to individual teachers. This literature has established convincingly that individual teachers have large impacts on their students and that impacts on students' socio-emotional development and life outcomes may be even longer-lasting than impacts on learning (Chetty et al, 2014; Jackson et al, 2014; Jennings and DiPrete, 2010).

What factors cause some teachers to be so much more effective than others? There is substantial US research that "observable" teacher characteristics, such as age, education, qualifications, and contract status do not explain differences in individual teachers' ability to produce classroom level learning gains – except for a consistent finding that all teachers tend to be less effective during their first three-to-five years of teaching (Kane and Staiger, 2008). Araujo et al (2016) similarly found that differences in teachers' classroom practice are not explained by teacher background and status. Except for "rookie" teachers with less than three years of service, the quality of teachers' classroom practice in Ecuador was not correlated with teachers' tenure status, salary, and age, or even with an unusually rich set of data the researchers collected on teachers' IQ, Big Five personality traits, and executive function.

The accumulating evidence that teachers' classroom practice varies widely, has important impacts on student learning and socioemotional skills development, and cannot be predicted by the observable characteristics commonly used to hire and promote teachers implies at least two major policy challenges. First, school systems need better ways of identifying candidates with the potential for excellence and/or weeding out lower-potential teachers early in their careers. Second, school systems need effective strategies for improving the classroom practice of the existing stock of teachers.

This paper focuses on the second challenge: improving the effectiveness of teachers in service. We evaluate a program in the northeast Brazilian state of Ceará designed to improve teachers' effectiveness by using an information "shock" (benchmarked feedback on their classroom practice) and expert coaching to promote increased professional interaction among teachers in the same school. This is the first study we know of in a developing country context that rigorously measures the impact of a training program both on teachers' classroom practice and their students' learning outcomes. It contributes to the very scant evidence base on the impact and cost-effectiveness of teacher training programs in developing countries as well as the growing global research base on how teachers' classroom practice affects student learning.

The design of the program was inspired by the research evidence that there exists large variation in teacher quality *within* schools. In the US, Hanushek and Rivkin (2010) have documented that value-added learning gains of different classroom teachers in the same school can range from 0.5 to 1.5 years of curriculum mastery. In studies of teachers' use of class time across six different countries in Latin America and the Caribbean, Bruns and Luque (2014) found that the average variation *within* schools in the share of total class time different teachers spend on instruction is consistently very large, irrespective of the average level of teacher performance in the school or even in the school system. Around a mean of roughly 65 percent of class time spent teaching in school at the median of the distribution in a Latin American country, the lowest-performing teachers in that school spend on average less than 45% of class time on instruction and the best-performing over 85%. This is a striking degree of classroom level heterogeneity given the fact that within a given school all teachers serve a roughly homogenous student population, deliver the same curriculum, and work under the same set of management and institutional conditions. Gaps of this magnitude in the instructional time different students experience could be expected to affect learning outcomes at the classroom level.

One positive implication of the Latin America research is the scope for school-level performance gains through greater diffusion of the best teaching practices within schools. Indeed, the exchange of practice among teachers in a school is a core strategy in high-performing East Asian systems, such as Japan's lesson study (Easton, 2008; Yoshida, 1999; Doig and Groves, 2011), Singapore (OECD, 2013) and Shanghai (Liang, 2015). Sustained school-level learning improvements are also reported in Ontario, Canada through a program which provided schools with feedback on their teachers' classroom-level learning outcomes and external coaches who encouraged school personnel to work together to share practice and improve instruction (OECD, 2011; Mourshed et al. 2011; Fullan, 2013). Fullan calls this the creation of a "professional learning community" within the school.

A hypothesized theory of action is that promoting and supporting school-level professional interaction among teachers may improve results through four channels. First, by increasing the amount of transparency about differential teacher performance within a school it can create "lateral accountability" or peer pressures on teachers to exert more effort towards improving their performance. Second, it can provide teachers with "curated" pedagogical or classroom management techniques (used effectively by their peers) that are clearly relevant to their school context. Third, it can transfer knowledge through modeled practice, which may be inherently more effective in supporting the adoption of new practices and behaviors than off-site, lecture-based training. Fourth, it can guarantee continuous support and reinforcement for the new behaviors from the school director and peers if the "whole school" is engaged in and committed to achieving improved classroom practice. However, countervailing factors include possible unwillingness among teachers to acknowledge differences in classroom effectiveness and weak extrinsic (salary, promotion, managerial oversight) incentives to reward improvements. A final issue may be that even if teachers were able to improve their classroom practice by, for example, devoting more time to instruction, weaknesses in teachers' content mastery could limit the impacts on student learning.

This paper presents the results of a randomized evaluation of the Ceará program. We show that the program increased teachers' use of class time for instruction, by reducing the time spent on classroom management and time off-task. The program also increased teachers' use of questions during their lessons, consistent with the coaching program's goal of encouraging more interactive teaching practice. The treatment schools also registered an increase in student engagement. Finally, consistent with the program's strategy of promoting greater interaction among teachers, the improvements in schools' average results were achieved by reducing the within-school variation in teacher practice. There is encouraging evidence that in most dimensions, schools and classrooms with the weakest performance improved the most.

Over the 2015 school year, these changes in teacher practice raised student learning in math and Portuguese on both the Ceará state assessment, SPAECE, and the national secondary school exit exam, ENEM. Learning gains were statistically significant, as we can control for individual students' prior-year learning outcomes. Performance in the treatment schools was 0.05 SD higher in Portuguese and 0.08 SD higher in math on the SPAECE test and 0.055 SD higher in Portuguese and 0.04 higher in math on the ENEM test. For classrooms with the highest average time on classroom management, rather than instruction, at baseline, the learning gains were significantly higher: 0.12-0.17 higher on SPAECE and 0.14-0.15 SD higher on the ENEM test. Finally, implementation fidelity had a large impact on learning results. In schools whose pedagogical coordinator scored "excellent" on a final certification test, SPAECE scores were 0.13-0.23 SD higher and ENEM scores were 0.13-0.17 SD higher. These results suggest that the combination of teacher feedback and expert coaching is a promising strategy for raising school quality.

Section 2 describes the context, the intervention, and the research questions. Section 3 describes the instruments used and the sample. Section 4 presents the impacts on teacher practice and classroom dynamics and analyzes threats to the experiment. Section 5 analyzes the program's impacts on student learning. Section 6 presents a cost-effectiveness analysis. Section 7 summarizes our conclusions and their implications for education policy in Brazil and other settings.

2. Intervention and Experiment Design

The Northeast state of Ceará, with 8.9 million people, is the 8th most populous in Brazil. With a GDP per capita estimated at $2,500, it is also one of Brazil's poorest states. While municipalities manage the provision of pre-school and primary education (grades 1-9) in Brazil, states are responsible for the three-year cycle of secondary education. Ceará state's education secretariat manages 621 schools with a total of

340,766 students[2]. Despite its poverty, Ceará has enjoyed a reputation within Brazil for progressive and effective government and in 2015, Ceará's secondary schools ranked 10th of 27 Brazilian states[3] on the Ministry of Education's IDEB index of basic education quality (a combined index of national assessment test scores and promotion rates).

Over the 2015 school year, the state implemented an experimental program designed to test whether improvements in teacher practice can be stimulated by providing schools with performance feedback based on classroom observations and practical suggestions and coaching support for more effective pedagogy. Classroom observation research supported by the World Bank in Brazil and elsewhere (Bruns and Luque, 2014) suggests that teachers' failure to use class time intensively, heavy reliance on traditional "chalk and talk" teaching methods, and inability to keep students engaged may be important factors in repetition, dropout and low learning outcomes. A 2014 federal government policy mandating that schools free up significant teacher time (1/3 of total working hours) in the school week to enable them to engage in professional interaction has created an opportunity for technical assistance or coaching programs to help schools maximize the utility of this extra time.

Supported by the Lemann Foundation, a respected Brazilian NGO dedicated to education, the ELOS consulting group developed a 9-month (one school year) training course and coaching program designed to promote professional interaction among teachers and to promote good practice techniques for lesson planning, classroom management, and keeping students engaged. The Lemann Foundation also published a Portuguese translation (*Aula Nota 10*) of US educator Douglas Lemov's book *Teach Like a Champion*.

a. The intervention

The intervention had four components:

- *Performance feedback on teacher practice*. At the beginning of the 2015 school year, treatment schools each received a two-page info-graphic "Bulletin" (Annex figures A1 and A2), providing key results from classroom observations undertaken at the end of the prior school year, in November 2014. For each variable, the Bulletin compared the school average to the best school in its district, the state average, the average for Brazil, and to US benchmarks for good practice. The bulletins also included a table with results for individual classrooms, to help schools understand the range in practice that exists in their school, and to identify best practices. (Teachers were not identified by name, only by the class hour and subject.)
- *Self-help materials*. Each school's principal, pedagogical coordinator, and teachers received a copy of *Aula Nota 10*, which describes "high-impact" teaching practices that stimulate student learning. The intervention distributed 4,680 books in 175 schools. The book includes practical descriptions of useful techniques, and access to online video examples. The Lemann Foundation website includes examples of the same techniques filmed in Brazilian classrooms.
- *Face-to-face interaction with high-skill coaches*. Three different one-day workshops were delivered by eight members of the ELOS coaching team. The workshops exposed school directors and pedagogical coordinators to the goals of the program and how to understand the feedback bulletins and use the results. The pedagogical coordinators were trained on how to observe teachers in the classroom and how to hold individual coaching sessions with teachers to provide specific feedback on their teaching practice. They were also trained to film themselves providing feedback to teachers and to upload and share these videos with their coaches, for additional feedback. The workshops stressed that coordinators were responsible for using an online log book to report weekly on their activities and the implementation of the program in their school.
- *Expert coaching support via Skype* - One expert trainer from the São Paulo team interacted regularly with each school's pedagogical coordinator via Skype. Each coach supported 31-36 schools and was responsible for delivering four coaching sessions over the period to each school. Treatment schools accessed a private website with good practice videos, their own uploads and other materials. The website required weekly online feedback from every pedagogical coordinator about the number of classroom observation and feedback activities implemented in the school, specific issues identified and addressed, and an assessment of progress. The site encouraged teachers and pedagogical coordinators to post video examples of good teacher practices in their school – both classroom teaching examples and pedagogical coordinators giving teachers specific feedback after observing their classes. The online reports indicate that the average time spent on teacher observation, coaching and feedback over the 2015 school year was about 111 hours per school.

b. Research questions

[2] Instituto Nacional de Estudos e Pesquisas Educacionais Anisio Teixeira – INEP/MEC

[3] Secretaria da Educação do Ceará – SEDUC-CE

The Ceará education secretariat agreed to randomize the implementation of the program across approximately half of its schools during the 2015 school year, to evaluate rigorously the following research questions:

1. Can providing schools with individualized teacher feedback based on classroom observations plus support materials and coaching stimulate measurable changes in teacher practice in a relatively short period (a single school year)?

2. Can providing classroom observation feedback and coaching reduce variation in teacher practices within a school?

3. Are positive changes in teachers' classroom practice, such as higher time on instruction and higher levels of student engagement, positively correlated with student learning results?

4. Is the combined program developed in Ceará (classroom observation feedback and school-level coaching) cost-effective in producing learning results when compared with alternative teacher training programs?

3. Instruments and Data

a. The Stallings Instrument

Teachers' classroom practice was measured using the Stallings "classroom snapshot" method, technically called the Stanford Research Institute Classroom Observation System, developed by Professor Jane Stallings for research on the efficiency and quality of basic education teachers in the United States in the 1970s. (Stallings, 1977; Stallings and Mohlman, 1990). The Stallings instrument generates quantitative data on the interaction of teachers and students in the classroom with a high degree of inter-rater reliability (0.8 or higher) among observers with relatively limited training.

This is a principal advantage of the Stallings instrument, in contrast to observation instruments such as CLASS, which capture more dimensions of teacher quality but which require a high degree of observer training and skill to apply reliably. The Stallings instrument's relative simplicity makes it suitable for large scale samples in developing country settings (Jukes, 2006; Abadzi, 2009; DeStefano et al, 2010; Schuh-Moore et al, 2010; World Bank 2014). The instrument is language and curriculum-neutral, so results are directly comparable across different types of schools and country contexts. A growing body of comparative country data –from more than 20,000 teachers in eight developing countries as of end-2016 –has been collected in collaboration with the World Bank. Once research studies are completed and data are suitably anonymized, data are hosted on the World Bank's open data website for further research use and country benchmarking.

The strength of the Stallings method is that it is a way of converting the qualitative activities and interactions between a teacher and students that occur during a class into robust quantitative data on teachers' instructional practice and students' engagement. Observations are coded at ten different moments in every class, at exact intervals whose spacing depends on the length of the class; every 3 minutes in a 30-minute class, every 5 minutes in a 50-minute class, etc. It is essential that the observer be present in the classroom before the first official moment of class and stay through the official end time of the class, whether the teacher is present or not. Each observation consists of a 15 second scan of the classroom, starting with the teacher and proceeding clockwise around the room. Observers code what the teacher is doing, what materials s/he is using, and what the students are doing.

For the purposes of generating quantitative estimates of time on task, student engagement, teachers' use of available materials and their core pedagogical practices, the coded activities are grouped into four categories:

1. Instruction: Reading Aloud; Demonstration/Lecture; Discussion/Debate/Question and Answer; Practice & Drill; Assignment/Class Work; Copying

2. Classroom Management: Verbal Instruction; Disciplining student(s); Classroom Management with Students; Classroom Management Alone

3. Teacher Off-Task: Teacher in Social Interaction with Students; Teacher in Social Interaction with Outsiders or Teacher Uninvolved; Teacher out of the classroom

4. Students Off-Task: Students in Social Interaction; Student(s) Uninvolved

For the purposes of generating quantitative estimates of the intensity of teachers' use of available learning materials, the coding options are: No Materials; Textbooks; Workbooks; Blackboard or whiteboard; Learning aids (maps, blocks, science equipment, calculators); ITC (LCD projectors, computers, TV/radio).

The original Stallings instrument is a one-page coding grid with classroom materials listed across the top and activities down the left side. Within each resulting cell, there is one row labeled "T", for coding what the teacher is doing and what materials s/he is using at the moment of observation and one row labeled "P" for marking what the pupils are doing and what materials they are using. Each 15 second observation is coded on a single sheet, thus each class observed generates 10 coding sheets. As the paper-based version has no in-built consistency checks to guard against mistaken double-coding or inconsistent coding (for example, if a student is being disciplined, both the teacher row and the student row must be coded with this activity), a full week (40 hour) training course with substantial time practicing in schools has typically been required to achieve .80 inter-rater reliability among observers.

The November 2014 round of observations in Ceará was conducted using the paper coding sheets, with subsequent data entry by a survey research firm. In August 2015, the research team conducted a pilot study in ten schools of observers sitting side by side (but not able to see each other's coding instruments) to compare the paper based method with a newly-available version of the Stallings instrument on electronic tablet, using ODK software. The team found high consistency in coding across the two instruments and lower error rates with the tablet, which is much more intuitive and where the sequence of questions permits in-built consistency checks. The November 2015 observations were conducted on tablets.

b. Sample

Ceará has 573 secondary schools that offer the complete three-year cycle. Our budget was sufficient to finance the observation of roughly 400 schools, so we randomly selected a sample of 400 schools and stratified these by geographic location, size and quartile of the learning distribution. After demonstrating that this sample was representative of the state system, we were forced to trim it further because of budgetary constraints. From the 400- school sample, we randomly selected 175 schools for a treatment group, and another 175 schools for a control group.[4]

The baseline round of classroom observations was conducted over a period of five weeks in November and early December 2014. Schools were visited without any advance notice, although all schools received a letter from the Secretariat in October informing them that a research study involving school visits would be implemented in November and December, and their cooperation was requested. When observation teams arrived at the schools, they informed school directors and teachers that the classroom observations were for research purposes only and that teachers would remain anonymous. School directors were advised that they could decline to participate in the study, and individual teachers could decline as well. In the end, no schools declined to participate and the full number of teachers planned to be observed in each school was in fact observed. The only change to the final sample was the closure of two schools – one in the treatment group and one in the control group – after the 2014 school year. Thus, the program was implemented in 174 treatment schools, with a 348 school total sample.

For each school, a schedule of classrooms for observation was pre-identified in order to give priority to observing math and language (Portuguese) classes, since standardized tests are applied in these subjects. Other subjects observed were biology, chemistry, physics, history and geography. Among classrooms offering these subjects, the selection of teachers to be observed was random. In case the teacher for a class and period originally programmed was absent, observers had a list of two acceptable alternatives.

Table 1 – Classroom assignment protocol for Stallings observations, November 2014

School Type	Twin class (100 min.)	Regular class (50 min.)
EP or C	1 Math 1 Portuguese	1 Math 1 Portuguese 2 other subjects of the core curriculum
B	1 Math 1 Portuguese	3 Math 3 Portuguese 4 other subjects of the core curriculum
A	1 Math 1 Portuguese	6 Math 6 Portuguese 4 other subjects of the core curriculum

Depending on school size (Type A, B, or C) and whether it had any vocational classes (EP, *Educacão Profissional*), teams of 1-4 observers visited each school and fanned out to observe between 6 and 24 classrooms.[5] The goal was to observe at least one-third of the teachers in the school. In type A schools, 18 classrooms were observed, 12 classrooms in type B schools, 6 classrooms in type C and vocational school, as shown in Table 1.

Since a significant share of Ceará's secondary school classes are 100-minute long double classes (called "twin classes" in Portuguese), both these and regular classes of 50 minutes were observed.

[4] The selection of 175 schools to treatment and 175 schools to control was performed through a simple randomization and was not stratified by geographic location, size and quartile of the learning distribution.

[5] Type A schools have more than 1,000 students; type B schools have 600-1,000 students; type C schools have less than 600 students. Vocational (EP) schools typically have less than 600 students.

At endline, the objective was insofar as possible to conduct observations in the same classrooms observed at baseline. Since individual teachers were guaranteed anonymity, the protocol was to observe classrooms with the same three characteristics: grade, subject and shift. As some schools changed their schedules between November 2014 and November 2015, only 75% of the classrooms were able to be "matched" on these criteria at endline.

The observers were state pedagogical coordinators who received a 40-hour training course in the Stallings method and scored 80% or higher on a final test, in which they coded videos and answered other questions. The observer teams were drawn only from schools identified for the treatment sub-sample, to avoid any contamination of control schools from having someone at the school familiar with the Stallings observation method and/or the training program. Observers were organized by district and assigned to districts other than their own, to avoid any familiarity with the schools they observed. Each team was coordinated by a supervisor with advanced expertise in the Stallings method. Supervisors conducted at least two observations side by side with each observer to check consistency, and reviewed the coding sheets submitted by observers for inconsistencies. In the cases of major inconsistencies, supervisors were responsible for making a repeat visit to the school to conduct new observations.

Out of the 348 schools of the randomization, with 174 each planned for treatment and controls, only 292 schools were able to be observed in November 2014 and in November 2015. The full initial sample could not be observed due to disruptions in the school calendar in November 2014 (standardized tests and holidays) and a shortage of observers in the Fortaleza district. The final sample of 292 schools included 156 schools in the treatment group and 136 in the control group. Because the loss of schools from the treatment and control groups was uneven, we conducted a series of balance checks to test the randomization. The 18 schools in the original treatment sample that were not observed could not receive the information treatment (benchmarked classroom observation feedback for the teachers in their school). But these schools were given access to the other three components of the program -- self-help materials, face to face training and coaching, and we could include these 18 schools (and the 38 schools in the control group that also were not observed) in the analysis of learning outcomes.

c. Balance checks

To ensure that our final 292-sample was balanced, we perform three sets of tests. First, we compare summary statistics for available outcome variables at baseline for the initially defined treatment and control groups in the 348-school sample. Second, we compare the same statistics for the final sample of 292 schools. Third, we check for balance in data from the baseline round of classroom observations collected in November 2014 for the 292 schools.

The randomization was based on 2013 data on school demographics and outcomes. When 2014 data became available, we performed a new balance check. All variables represent school averages.

Table 2 presents results for the first two sets of tests, along with the results of t tests of mean differences across the treatment and control groups for each variable, as well as joint significance tests. The first set of balancing tests (random sample) shows that the treatment and the control groups are well balanced, although the treatment schools present a higher average math proficiency. A joint test for the joint significance of the variables in predicting treatment fails to reject that they are jointly equal to zero, supporting the notion of baseline balance in these outcome variables.

Table 2 - Pre-treatment covariate balance

	Random Sample (350 Schools)			Baseline Data (292 Schools)		
	Control Means	Treatment Means	Difference	Control Means	Treatment Means	Difference
2013 Covariates						
Portuguese proficiency	257.4	260.8	-3.245	256.9	261.4	-4.454
	[19.73]	[22.39]	[2.259]	[18.69]	[23.08]	[2.481]
Mathematical proficiency	267.4	272.2	-4.679	267.7	273.3	-5.562
	[23.81]	[29.77]	[2.882]	[22.67]	[30.72]	[3.199]
High School enrollment	641.4	588.9	55.15	676.3	575.3	101.0*
	[368.2]	[330.3]	[37.44]	[349.3]	[321.5]	[39.27]
High school enrollment - vocational	46.63	68.21	-21.18	47.11	76.08	-28.97
	[132.6]	[154.1]	[15.35]	[136.0]	[160.9]	[17.58]
Rural Area	0.0286	0.0517	-0.0229	0.0368	0.0577	-0.0209
	[0.167]	[0.222]	[0.0210]	[0.189]	[0.234]	[0.0251]
Pass rate	83.33	84.56	-1.248	84.46	85.57	-1.115
	[10.33]	[10.74]	[1.125]	[10.07]	[10.50]	[1.208]
Failure rate	6.938	6.311	0.649	6.398	6.051	0.347
	[5.614]	[5.283]	[0.582]	[5.620]	[5.227]	[0.635]
Dropout rate	9.731	9.129	0.600	9.144	8.375	0.769
	[7.179]	[7.002]	[0.757]	[6.896]	[6.637]	[0.793]
Students per class	34.06	34.00	0.0734	34.38	34.03	0.349
	[4.939]	[5.198]	[0.541]	[4.941]	[5.317]	[0.604]
Female principals	0.520	0.511	0.00571	0.485	0.519	-0.0339
	[0.501]	[0.501]	[0.0536]	[0.502]	[0.501]	[0.0588]
Experience as a principal (> 10 years)	0.543	0.517	0.0229	0.507	0.500	0.00735
	[0.500]	[0.501]	[0.0535]	[0.502]	[0.502]	[0.0589]
Principal with graduate degree	0.994	0.994	0	0.993	0.994	-0.000943
	[0.0756]	[0.0758]	[0.00808]	[0.0857]	[0.0801]	[0.00971]
Female teachers	0.551	0.515	0.0341	0.562	0.515	0.0476*
	[0.180]	[0.181]	[0.0193]	[0.184]	[0.183]	[0.0216]
Temporary teachers	0.995	0.994	0.00114	0.995	0.994	0.000713
	[0.0148]	[0.0188]	[0.00181]	[0.0155]	[0.0193]	[0.00207]
Teacher's age	35.00	30.34	4.609	35.34	30.15	5.197
	[27.09]	[63.98]	[5.239]	[25.52]	[67.22]	[6.117]
Experience as a teacher (>10 years)	0.816	0.814	0.00194	0.819	0.812	0.00749
	[0.0871]	[0.0850]	[0.00919]	[0.0858]	[0.0873]	[0.0102]
Low salary (< 2m.w.)	0.185	0.184	0.000229	0.194	0.183	0.0109
	[0.141]	[0.152]	[0.0157]	[0.146]	[0.155]	[0.0177]
High Salary (> 5 m.w.)	0.225	0.200	0.0253	0.219	0.187	0.0327
	[0.179]	[0.183]	[0.0194]	[0.183]	[0.179]	[0.0212]
Mother's education < middle school	0.472	0.485	-0.0115	0.490	0.488	0.00159
	[0.104]	[0.108]	[0.0114]	[0.0966]	[0.109]	[0.0122]
Mothers with graduate degree	0.0507	0.0523	-0.00143	0.0548	0.0546	0.000228
	[0.0301]	[0.0302]	[0.00322]	[0.0282]	[0.0305]	[0.00345]
2014 Covariates						
Portuguese proficiency	252.8	256.5	-3.675	252.3	257.1	-4.764*
	[17.72]	[20.53]	[2.053]	[17.76]	[21.24]	[2.311]
Mathematical proficiency	252.8	258.8	-5.972*	253.1	260.2	-7.082*
	[21.58]	[27.66]	[2.655]	[21.79]	[28.59]	[3.009]
Age-Grade distortion	33.72	32.06	1.662	31.63	30.66	0.964
	[15.21]	[15.47]	[1.642]	[14.04]	[15.18]	[1.720]
Proportion of students per teacher	0.0588	0.0593	-0.000576	0.0534	0.0586	-0.00526*
	[0.0214]	[0.0215]	[0.00230]	[0.0142]	[0.0208]	[0.00212]
Proportion of black and brown teachers	0.298	0.302	-0.00400	0.281	0.302	-0.0209
	[0.232]	[0.228]	[0.0246]	[0.238]	[0.231]	[0.0275]
Proportion of black and brown students	0.606	0.606	0.000215	0.595	0.607	-0.0115
	[0.216]	[0.230]	[0.0239]	[0.220]	[0.229]	[0.0264]
Joint test (p-value) - All Variables			0.620			0.18
Joint test (p-value) - Only proficiency variables			0.120			0.13
Joint test (p-value) - Other variables excluding proficiency			0.850			0.31
Number of schools	175	175		136	156	
Response Rate				78%	89%	0.11
p-value of the response rate difference						0.00

Note: Numbers in parentheses are standard deviations in the column of means and standard errors in columns of differences. * p<0.05 ** p<0.01 *** p<0.001

Table 3- Pre-treatment classroom dynamics balance

	No Weights		
	Control Means	Treatment Means	Difference
Instructional activities	0.656	0.674	-0.0184
	[0.101]	[0.102]	[0.0119]
Classroom management activities	0.250	0.228	0.0220*
	[0.0724]	[0.0812]	[0.00906]
Off-task activities	0.0940	0.0976	-0.00361
	[0.0618]	[0.0654]	[0.00748]
Student off-task	0.227	0.189	0.0383*
	[0.146]	[0.136]	[0.0165]
Instructional activities with all students engaged	0.194	0.236	-0.0424*
	[0.144]	[0.153]	[0.0174]
Reading aloud	0.0430	0.0432	-0.000226
	[0.0363]	[0.0351]	[0.00418]
Demonstration/Lecture	0.326	0.334	-0.00807
	[0.112]	[0.110]	[0.0130]
Discussion/Debate/Q&A	0.0972	0.0990	-0.00182
	[0.0590]	[0.0726]	[0.00781]
Practice & Drill	0.00431	0.00442	-0.000119
	[0.00874]	[0.0128]	[0.00131]
Assignment/Class work	0.122	0.132	-0.00984
	[0.0801]	[0.0994]	[0.0107]
Copying	0.0629	0.0613	0.00167
	[0.0431]	[0.0484]	[0.00540]
Verbal Instruction	0.0604	0.0569	0.00352
	[0.0351]	[0.0347]	[0.00409]
Discipline	0.0205	0.0167	0.00387
	[0.0190]	[0.0166]	[0.00209]
Classroom management	0.0807	0.0767	0.00395
	[0.0421]	[0.0450]	[0.00512]
Classroom management alone	0.0886	0.0779	0.0107
	[0.0573]	[0.0525]	[0.00643]
Social interaction	0.0156	0.0175	-0.00185
	[0.0229]	[0.0283]	[0.00305]
Teacher out of the room	0.0572	0.0581	-0.000815
	[0.0397]	[0.0478]	[0.00518]
Teacher uninvolved	0.0211	0.0221	-0.000941
	[0.0307]	[0.0274]	[0.00340]
No material	0.128	0.131	-0.00240
	[0.0777]	[0.0667]	[0.00845]
Textbook	0.101	0.0938	0.00731
	[0.0820]	[0.0811]	[0.00956]
Notebook	0.119	0.137	-0.0186
	[0.0738]	[0.117]	[0.0116]
Blackboard	0.271	0.270	0.000989
	[0.108]	[0.112]	[0.0130]
Learning aides	0.0255	0.0216	0.00386
	[0.0476]	[0.0354]	[0.00487]
TIC	0.0632	0.0686	-0.00543
	[0.0813]	[0.0813]	[0.00954]
Cooperative	0.00795	0.00859	-0.000640
	[0.0188]	[0.0234]	[0.00251]
Joint test (p-value)			0.81
Number of schools	136	156	

Note: Numbers in parentheses are standard deviations in the column of means and standard errors in columns of differences. ** p<0.05 ** p<0.01 *** p<0.001*

The second set of columns (baseline sample) shows that despite the reduction in the number of schools, characteristics of the treatment and control groups remained balanced. Although some of the differences between treatment and control schools are significant at a 5% level (enrollments, proportion of female teachers, Portuguese and Math proficiency, and student-teacher ratio), a joint test for significance of the full set of variables in predicting treatment fails to reject that they jointly equal zero, suggesting that our treatment and control groups are equal in expectations on both observed and unobserved characteristics. We nevertheless control for demographic characteristics in the analysis to account for any potential differences between the different groups.

Results for the third set of balance checks, on the classroom dynamics variables observed at baseline, are presented in table 3. Although the control schools spend more time on classroom management, less time on instructional activities with all students engaged, and have a higher share of time with students off-task, a joint significance test yields a p-value of 0.81, suggesting that the randomization is collectively balanced along the full set of classroom dynamics indicators we consider.

4. Impacts on Teacher Practice

a. Descriptive statistics

Table 4 presents key indicators of classroom practice that are captured by the Stallings instrument: i) teacher time on instruction; ii) teacher time on classroom management; iii) teacher time off-task; iv) teacher time on instruction with all students engaged; and v) time with a large group of students (six or more) off-task, meaning visibly not engaged in the activity being led by the teacher. The first four variables are expressed as a percentage of total official class time, while the last two variables are expressed as a percentage of total time the teacher was engaged in instructional activities.

Teachers' time on instruction increased significantly in the treatment schools, to 76% of class time, compared with 70% in the control schools, implying almost 10% more time on instruction in every class hour. Teachers in the program schools gained more time for instruction by significantly reducing time spent on classroom management, which fell to 18% of class time vis a vis 21% in control schools, and time off task, which fell to 5.8% in schools exposed to the program, compared with 8.4% in the control schools. The biggest driver of this change was a decline in the share of class time that teachers were out of the room. In treatment schools, this fell to 3%, compared to 5% in the control schools.

Table 4: Change in classroom dynamics from Nov. 2014 to Nov. 2015

	Baseline Means and Std			Endline Means and Std		
	All Sample	Control	Treatment	All Sample	Control	Treatment
Instructional activities	0.655	0.646	0.665	0.735	0.704	0.766
	[0.212]	[0.211]	[0.212]	[0.199]	[0.209]	[0.183]
Classroom management activities	0.244	0.255	0.233	0.194	0.211	0.176
	[0.176]	[0.176]	[0.176]	[0.157]	[0.166]	[0.145]
Off-task activities	0.101	0.0992	0.102	0.0718	0.0848	0.0587
	[0.132]	[0.132]	[0.133]	[0.1177]	[0.127]	[0.105]
o/w Teacher out of the room	0.0608	0.0611	0.0605	0.0402	0.0498	0.0306
	[0.0996]	[0.0998]	[0.0995]	[0.0766]	[0.0872]	[0.0629]
Instructional activities with all students engaged	0.200	0.183	0.217	0.267	0.265	0.269
	[0.263]	[0.251]	[0.273]	[0.302]	[0.302]	[0.303]
Student off-task	0.223	0.242	0.203	0.166	0.187	0.144
	[0.284]	[0.296]	[0.271]	[0.265]	[0.280]	[0.246]
Sample Size	3121	1560	1561	3121	1560	1561

Figures 1-5 illustrate the distribution across schools of these changes in classroom dynamics. The box plots show schools' average values with the median value (the horizontal line within the box), the lower and upper quartiles (the two edges of the box) and the extreme values (the two whiskers extending from the box).[6]

The figures confirm some expected results. First, benchmarked, individualized feedback should help focus teachers on the importance of maximizing instructional time and coaching support should improve teachers' capacity for planning lessons and conducting routine administrative processes more efficiently, as well as minimizing time off task. Second, the coaching program's emphasis on keeping students engaged with well-paced and more interactive (question and answer) lesson plans should be reflected in a lower share of class time with a large group (six or more) of students visibly tuned out or in social interaction (off-task). Third, promoting greater interaction among teachers in a school should reduce the variation *within schools* in teaching practices and Stallings measures, by bringing the performance of less proficient teachers into line with their more effective peers.

It is encouraging that there was a clear improvement in the bottom-performing treatment schools. As can be seen from Figures 1 and 2, *all* treatment schools raised the average time on instruction to 55% or more of class time, compared to the control group, where some schools continued to average only 40% of class time on instruction. *All* treatment schools reduced the share of class time spent on administrative activities to below 33% and time with teachers completely off task to below 15%. In contrast, the lower tail of control schools showed no improvement from the baseline; some schools continued average up to 40% of teacher time on administrative activities and up to 25% of total class time completely off-task (with teachers either

[6] Kernel and cumulative distributions are presented in Annex, figures A3 and A4, as well as statistics for classroom dynamic characteristics at baseline and endline, at the class observation level, table A1.

out of the classroom or in social interaction with students or visitors). The progress registered in the lowest performing treatment schools, shifting teacher time from classroom management and off-task activities towards increased instruction, is an important gain.

Figures 1 and 2– Box plot distribution for teacher time on instruction and classroom management

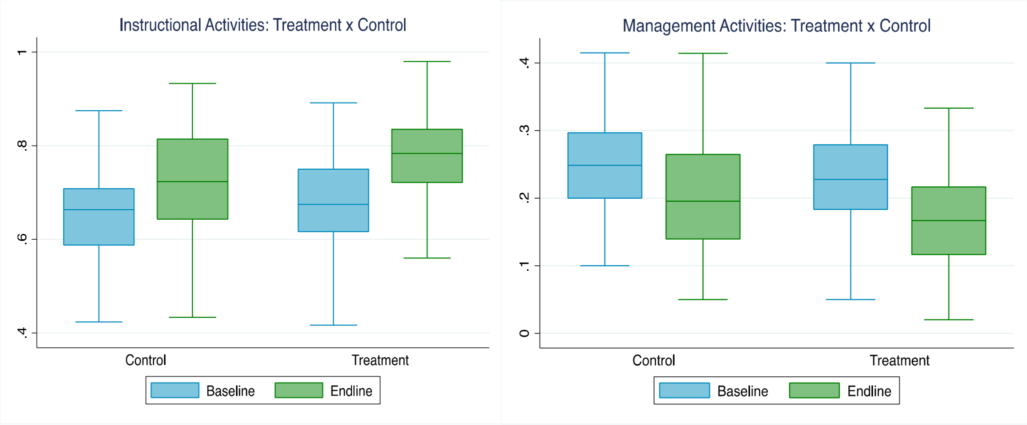
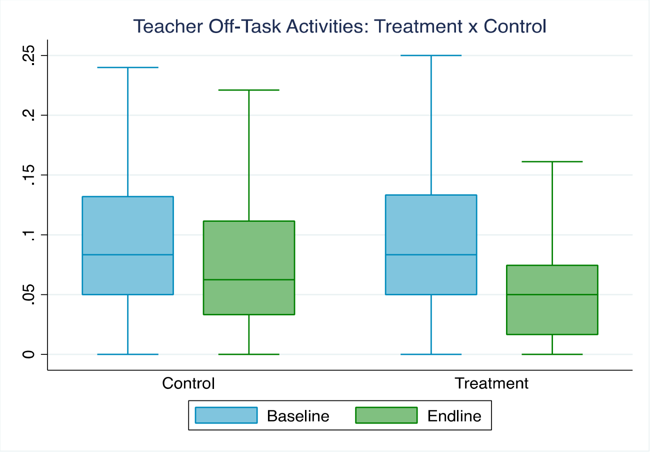
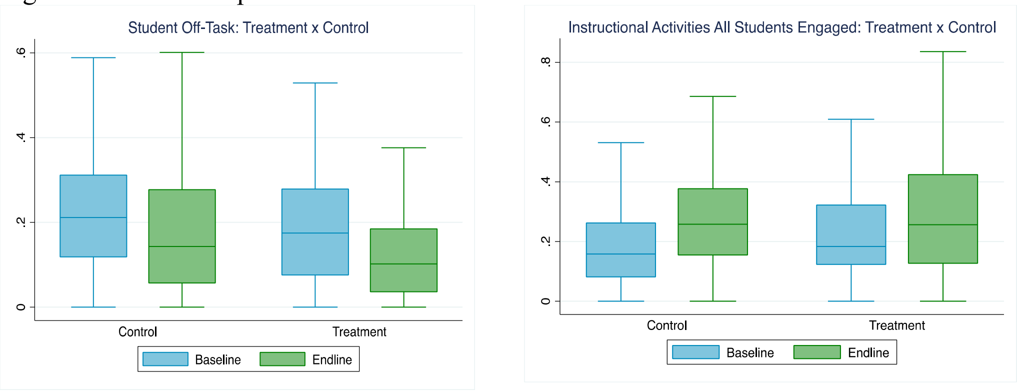


Figure 3– Box plot distribution for teacher time off task



Treatment schools also showed some improvement in student engagement. The Stallings instrument has two measures for this: i) the share of class time that a large group of students (defined as six or more) is not engaged with the teacher, either chatting with other students (in social interaction) or visibly tuned out (texting, sleeping, gazing out the window, etc.) and ii) the share of time that a teacher keeps the entire class engaged in the activity she is leading. The former captures the degree to which the teacher can minimize the number of students drifting off; in classes of typically 25 students, letting one-quarter of them tune out can compromise the lesson, especially if groups of students are chatting and raise the noise level in the classroom. But the latter indicator is quite challenging; teachers must either organize the class into groups working in parallel on assignments that keep them all engaged, or manage to keep the entire class focused on material she is presenting, with questions or discussion that draws in all students.

Figure 4 and 5 – Box plot distribution for students off-task and time on instruction with all students engaged



At baseline, a large group of students was off-task, on average, 20% of class time in treatment schools and 24% in control schools. Treatment schools brought this down to 14%, while in control schools it fell to

19%. As Figure 4 shows, at the end of the program, there were no treatment schools averaging more than 40% of class time with six or more students tuned out or in social interaction, while some control schools continued to average more than 50% of time with a large group off task. The feedback and coaching appears to have helped teachers in treatment schools adopt instructional practices that engage more students and achieve less disruptive classroom environments.

Treatment schools made less progress in raising the share of class time with all students engaged. Figure 5 shows that, while at baseline, no school in the entire sample averaged more than 65% of time with all students engaged, at endline the positive tail of treatment schools averaged over 80% of time on instruction with all students engaged; the best performing control schools averaged only 60%. However, the low tail of the distribution in both treatment and control schools at endline continued to include schools averaging less than 10% of time on instruction with all students engaged and the sample mean for treatment schools at endline was no better than for control schools. Finding instructional strategies that manage to engage all students in relatively large and diverse classrooms is clearly a challenge in Ceará's schools.

b. Intent to treat effects

To confirm that the feedback plus coaching intervention *caused* the observed impacts on teachers' classroom practices we first estimate intent-to-treat effects (ITT), an estimate of the impact of being offered a chance to participate in the experiment. We use a parsimonious set of controls to aid in precision and correct for any potential imbalance between treatment and control. The ITT effect is estimated from the equation below:

$$y_i = \beta_0 + \beta_1 y_{i,t-1} + \boldsymbol{X'}_{\boldsymbol{i}}\beta_2 + \alpha_0 Z_i + \varepsilon_i \quad (1)$$

where y_i is the dependent variable for classroom observation i; $y_{i,t-1}$ is the baseline classroom dynamic variable collected in November of 2014; $\boldsymbol{X_i}$ represents a vector of pre-intervention characteristics at the school level; Z_i is an indicator for whether the classroom observation was in school that was offered participation in the intervention; and ε_i if the error term, clustered at the school level. The coefficient of interest is α_0.

We estimate (1) using four sets of control variables: "no controls," i.e., excluding the baseline control and the $\boldsymbol{X_i}$ variables; "baseline controls", including only control for the baseline observation; "student, teacher and classroom controls" including the Xi variables at the school level for students and teachers and Xi variables at the classroom[7]; and "all controls" which includes all Xi controls.[8]

Results are presented in Table 5. Treatment effects are reported in percentage of class time and standard errors clustered at the school level are presented in brackets below each estimate.

The intervention increased the amount of time teachers spend on instructional activities, decreased the amount of time spent on classroom administration and off-task activities, and decreased the amount of time a large group of students is off-task while the teacher is teaching. Except for instructional activities with all students engaged, results are strong and significant in all four specifications. The share of class time on instruction increased 5.2- 6.2% (0.248 to 0.295 standard deviations); the share of time teachers' spent on classroom management fell -2.8-3.6% (-0.168 to -0.215 SD); time off-task was reduced by -2.5-2.6% (-0.192 to -0.204 SD); and the share of time that a large group of students is off-task declined between -3-4.3 (-0.108 to -0.152 SD). The estimates of α_0 change relatively little as the list of control variables changes, which is to be expected since treatment and control were randomly assigned.

[7] Controls for students include: Math and Portuguese proficiency in 2013 and 2014, pass rate, failure rate, dropout rate, mother's education below middle school, mothers with a graduate degree, age-grade distortion. Controls for teachers include: proportion of female teachers, proportion of temporary teachers, teacher's age, teacher's experience, teacher salary low, teacher salary high, proportion of black or brown teachers.

Controls for classroom include: discipline (Portuguese, Math, Social Sciences and Sciences), grade and tween classroom.

[8] All controls includes all of the student, teacher and classroom controls listed above plus: high school enrollment, high school vocational enrollment, rural area, average number of student per class, proportion of female principals, principal experience, principal with graduate degree, student-teacher ratio.

Table 5: Mean effect sizes on summary measures of classroom observation

Dependent variable	(1)	(2)	(3)	(4)	Control Average
A. Instructional activities	0.062*** [0.013]	0.061*** [0.013]	0.057*** [0.013]	0.052*** [0.013]	0.704
B. Classroom management activities	-0.036*** [0.009]	-0.035*** [0.009]	-0.033*** [0.009]	-0.028*** [0.009]	0.211
C. Off-task activities	-0.026*** [0.007]	-0.026*** [0.007]	-0.025*** [0.007]	-0.025*** [0.006]	0.085
D. Instructional activities all students engaged	0.003 [0.020]	-0.003 [0.019]	-0.005 [0.020]	-0.011 [0.020]	0.265
E. Big group (>6) of student off-task	-0.043** [0.017]	-0.036** [0.016]	-0.036** [0.016]	-0.030* [0.016]	0.187
Control for baseline		x	x	x	
Student, teacher and classroom covariates			x	x	
School covariates				x	

Note: Sample size 3121. Robust standard errors in brakets, clustered at the school level. Variables D and E only consider the time teacher was instructing. These variables assumes missing values if the teacher did not spend any time instructing. * p<0.10 ** p<0.05 *** p<0.01

Figures 6 – 10 unpack results for each of the five summary measures using specification (4) - OLS results with baseline and all covariates as control.[9] The figure displays coefficients and 90% confidence intervals for summary measures and for all individual outcomes under each category. The black line crosses at zero; results to the right of the zero line represent positive effects of the treatment and results to the left represent negative effects of the treatment.

The impact of the program on teacher time on instruction, as shown in Figure 6, was driven by statistically significant increases in time spent on "discussion/debate/Q&A" (2% of class time), "copying" (1,8%) and "demonstration/lecture" (2%). "Reading aloud" showed a statistically significant but smaller increase of 0.7%.

The use of more interactive teaching techniques, and especially the importance of using questions to probe students' understanding of the material being taught and to stimulate discussion are key elements of the coaching program and the *Teach Like a Champion* book. Notwithstanding the increase, teachers in treatment schools still used discussion/question and answer only 10.5% of the time at endline, and only 8.4% of time in control schools. Lecturing from the blackboard remained the dominant teaching mode – used on average 38% of the time in treatment schools and 34% of time in control schools. There was a statistically significant increase in copying in the treatment schools relative to control schools, but it still absorbed less than 10% of the time.

Figure 6 – Decomposition of effects on instructional time - all controls (90% confidence interval)

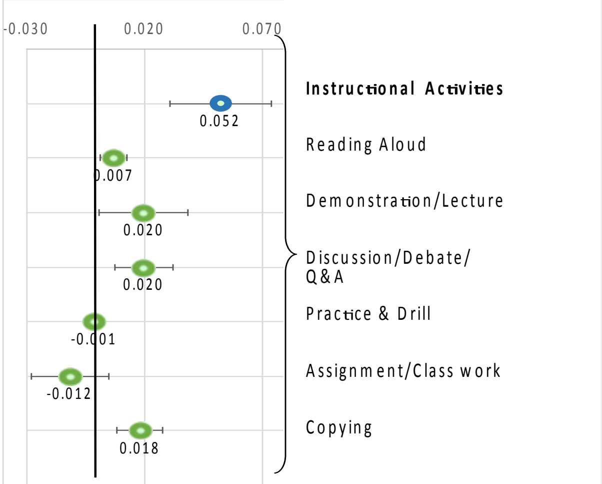


Figure 7 presents the average treatment effects for the four underlying activities that constitute classroom management. The improvement was driven by a sharp, 2% reduction in teacher time spent on classroom management alone (e.g., teacher at his/her desk grading papers). Declines in time spent on verbal instruction (teacher discussing non-academic matters, such as plans for school activities or dates for upcoming tests,

[9] Regression results for each outcome measure using the four specifications are shown in Appendix tables A2 to A6.

etc.), discipline, and classroom management with students (typically taking attendance, passing out papers, collecting homework, etc.) were not statistically significant.

Figure 7 – Decomposition of effects on classroom management - all controls (90% confidence interval)

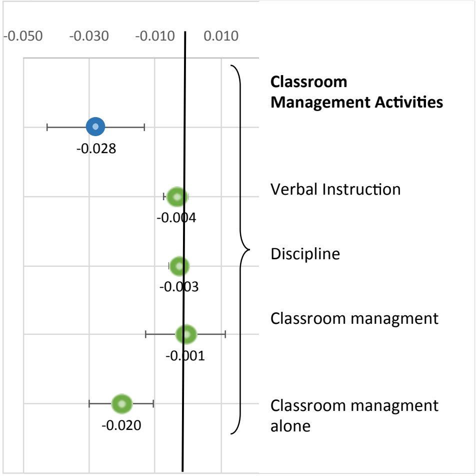
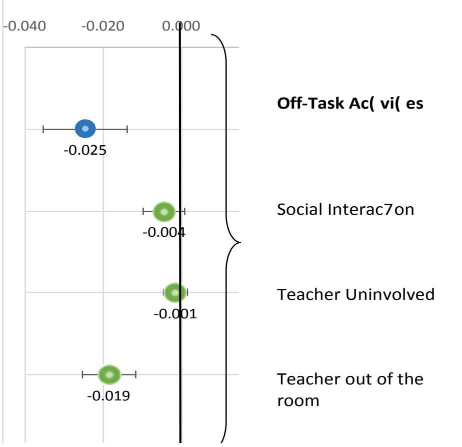


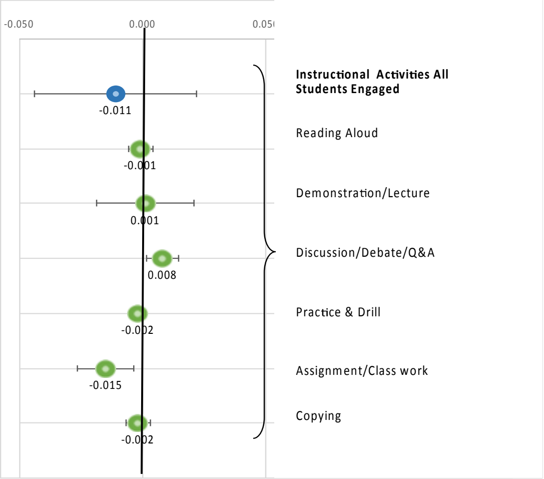
Figure 8 shows that the reduction in the share of time teachers in the treatment schools are off-task was entirely due to the large decline in time spent out of the classroom: the coefficient of -1.9% of class time is strong and significant. There were no significant impacts on teacher in social interaction with students and teacher uninvolved.

Figure 8 – Decomposition of effects on teacher off-task activities - all controls (90% confidence interval)



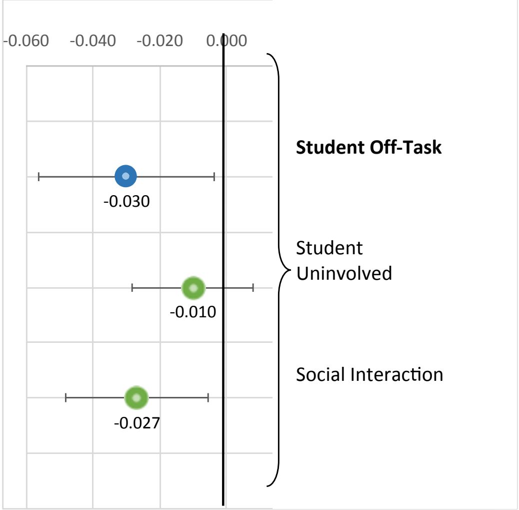
The treatment schools' improvement in the share of time teachers keep the entire class engaged was entirely associated with the increased time teachers spent on "discussion/debate/Q&A". The positive and significant result of 0.8% is consistent with goals of the coaching program and, in a sense, validates the program's emphasis on more interactive teaching practices to keep students engaged. Time spent on doing assignments in class declined by 1.5%, also consistent with the content of the coaching program.

Figure 9 – Decomposition of effects on instructional activities with all students engaged - all controls (90% confidence interval)



The effect on students off-task, Figure 10, is driven by a decrease of 2.7% in the share of class time that a large group of students is in social interaction. Having numerous students chatting in a classroom creates noise and distraction for other students and can undermine learning. An improvement in teachers' ability to maintain classroom discipline and reduce or eliminate student socializing is a potentially important change.

Figure 10 – Decomposition of effects on student time off-task - all controls (90% confidence interval)



Appendix Table A7 shows the program impacts on materials used by teachers. There was an increase in the amount of time teachers in the treatment schools led the class using no materials—1.3%—and using the blackboard—5%. The use of other materials—textbooks, notebooks, learning aides, ITC, and cooperative activities among students--was not significantly affected by the intervention.

c. Intent to treat effects– matched sample

In the endline data collection in November 2015, efforts were made to return to the same classrooms observed at baseline, with the understanding that the teacher might have changed, since our protocol did not allow for collecting teachers' names, codes or other identifying information. If in 2014 a 3rd year math class was observed during the sixth period of the day, the most precise measure of program impact on teaching practice would come from observing the same classroom, subject and time of day exactly one year later, in the expectation that in most cases we would be observing the same teacher. In practice, observers were only able to make matched repeat observations in 2,399 classrooms, 75% of those observed at baseline. Variations in the school calendar and logistical issues resulted in 25% of the 2015 observations being conducted in grades and subjects in the school that had not been observed at baseline.

Arguably, results for the whole sample of 3,121 classes may underestimate the real effects, since 25% of the observations were in classrooms not observed at baseline. By analyzing the 75% of classrooms where the full protocol was followed, we may expect measured impacts to be closest to the real impacts.

To test this, we first check the extent to which the matched 75% sample is different from the main sample. Table 6 shows balance tests for the matched sample, for pre-treatment covariates and for the classroom observation variables collected at baseline. The balance is quite similar to the baseline sample. Treatment and control schools present some differences in enrollments, proportion of female teachers, student-teacher ratio, classroom management activities, instructional activities with all students engaged, and students off-task, but a joint test for the joint significance of the variables in predicting treatment fails to reject that they are jointly equal to zero, supporting that the randomization is balanced for this subsample.

Table 6: Pre-treatment covariates and classroom dynamics balance - Matched sample

Covariates				Classroom Dynamics			
	Control Means	Treatment Means	Difference		Control Means	Treatment Means	Difference
2013 Covariates							
Portuguese proficiency	257.4	260.6	-3.209	Instructional activities	0.655	0.674	-0.0192
	[18.78]	[22.45]	[2.492]		[0.117]	[0.108]	[0.0134]
Mathematical proficiency	268.4	272.2	-3.799	Classroom management activities	0.252	0.226	0.0257*
	[22.65]	[29.51]	[3.178]		[0.0838]	[0.0859]	[0.0102]
High School enrollment	680.8	581.0	99.80*	Off-task activities	0.0930	0.0995	-0.00651
	[350.8]	[324.9]	[40.34]		[0.0725]	[0.0702]	[0.00853]
High school enrollment - vocational	49.28	69.40	-20.11	Student off-task	0.0387	-0.247	0.286*
	[138.8]	[153.7]	[17.59]		[1.089]	[1.012]	[0.125]
Rural Area	0.0385	0.0596	-0.0211	Instructional activities with all	-0.0323	0.236	-0.268*
	[0.193]	[0.238]	[0.0261]	students engaged	[1.071]	[1.105]	[0.130]
Pass rate	85.00	85.42	-0.412	Reading aloud	0.0477	0.0423	0.00543
	[9.568]	[10.45]	[1.202]		[0.0525]	[0.0424]	[0.00567]
Failure rate	6.172	6.117	0.0551	Demonstration/Lecture	0.325	0.337	-0.0119
	[5.058]	[5.273]	[0.619]		[0.127]	[0.129]	[0.0153]
Dropout rate	8.825	8.466	0.358	Discussion/Debate/Q&A	0.0978	0.0976	0.000239
	[6.789]	[6.570]	[0.798]		[0.0656]	[0.0773]	[0.00863]
Students per class	34.30	33.95	0.343	Practice & Drill	0.00375	0.00543	-0.00168
	[4.978]	[5.308]	[0.617]		[0.00933]	[0.0147]	[0.00150]
Female principals	0.477	0.510	-0.0330	Assignment/Class work	0.120	0.133	-0.0138
	[0.501]	[0.502]	[0.0600]		[0.0919]	[0.100]	[0.0115]
Experience as a principal (> 10 years)	0.515	0.510	0.00545	Copying	0.0612	0.0587	0.00250
	[0.502]	[0.502]	[0.0600]		[0.0481]	[0.0453]	[0.00558]
Principal with graduate degree	0.992	0.993	-0.00107	Verbal Instruction	0.0635	0.0571	0.00646
	[0.0877]	[0.0814]	[0.0101]		[0.0427]	[0.0384]	[0.00484]
Female teachers	0.568	0.515	0.0526*	Discipline	0.0200	0.0163	0.00371
	[0.181]	[0.185]	[0.0219]		[0.0194]	[0.0193]	[0.00231]
Temporary teachers	0.994	0.994	0.000665	Classroom management	0.0823	0.0789	0.00335
	[0.0158]	[0.0196]	[0.00215]		[0.0580]	[0.0478]	[0.00631]
Teacher's age	35.29	30.16	5.131	Classroom management alone	0.0862	0.0740	0.0122
	[26.10]	[68.33]	[6.360]		[0.0653]	[0.0555]	[0.00721]
Experience as a teacher (>10 years)	0.821	0.813	0.00724	Social interaction	0.0152	0.0180	-0.00275
	[0.0859]	[0.0878]	[0.0104]		[0.0260]	[0.0315]	[0.00348]
Low salary (< 2m.w.)	0.193	0.184	0.00940	Teacher out of the room	0.0558	0.0576	-0.00178
	[0.145]	[0.156]	[0.0180]		[0.0478]	[0.0498]	[0.00585]
High Salary (> 5 m.w.)	0.219	0.191	0.0287	Teacher uninvolved	0.0220	0.0239	-0.00198
	[0.186]	[0.180]	[0.0219]		[0.0371]	[0.0353]	[0.00432]
Mother's education < middle school	0.490	0.489	0.000752	No material	0.129	0.131	-0.00210
	[0.0978]	[0.111]	[0.0125]		[0.0894]	[0.0761]	[0.00988]
Mothers with graduate degree	0.0558	0.0544	0.00141	Textbook	0.105	0.0924	0.0123
	[0.0283]	[0.0305]	[0.00353]		[0.102]	[0.0922]	[0.0116]
2014 Covariates				Notebook	0.121	0.130	-0.00891
Portuguese proficiency	253.0	256.3	-3.289		[0.0799]	[0.117]	[0.0121]
	[17.84]	[20.45]	[2.308]	Blackboard	0.271	0.276	-0.00472
Mathematical proficiency	254.0	258.9	-4.931		[0.132]	[0.124]	[0.0152]
	[21.88]	[27.00]	[2.963]	Learning aides	0.0233	0.0231	0.000184
Age-Grade distortion	31.02	30.89	0.139		[0.0503]	[0.0400]	[0.00539]
	[13.75]	[14.92]	[1.722]	TIC	0.0609	0.0702	-0.00925
Proportion of students per teacher	0.0532	0.0582	-0.00497*		[0.0901]	[0.0872]	[0.0106]
	[0.0143]	[0.0208]	[0.00217]	Cooperative	0.00878	0.00904	-0.000253
Proportion of black and brown teachers	0.282	0.301	-0.0197		[0.0263]	[0.0308]	[0.00345]
	[0.242]	[0.232]	[0.0283]				
Proportion of black and brown students	0.607	0.611	-0.00392				
	[0.214]	[0.228]	[0.0265]				
Joint test (p-value) - All Variables			0.34				0.63
Joint test (p-value) - Only proficiency variables			0.41				
Joint test (p-value) - Other variables excluding proficiency			0.44				
Number of schools	130	151			130	151	

Note: Numbers in parentheses are standard deviations in the column of means and standard errors in columns of differences. * p<0.05 ** p<0.01 *** p<0.001

Table 7: Mean effect sizes on summary measures of classroom observation – Matched sample

Dependent variable	(1)	(2)	(3)	(4)	Control Average
A. Instructional activities	0.066*** [0.014]	0.066*** [0.014]	0.063*** [0.014]	0.056*** [0.013]	0.698
B. Classroom management activities	-0.038*** [0.010]	-0.038*** [0.010]	-0.038*** [0.010]	-0.032*** [0.010]	0.216
C. Off-task activities	-0.028*** [0.008]	-0.028*** [0.007]	-0.025*** [0.007]	-0.024*** [0.007]	0.086
D. Instructional activities all students engaged	0.003 [0.021]	-0.002 [0.021]	-0.001 [0.022]	-0.004 [0.022]	0.262
E. Big group (>6) of student off-task	-0.050*** [0.019]	-0.043** [0.017]	-0.042** [0.018]	-0.037** [0.018]	0.196
Control for baseline		x	x	x	
Student, teacher and classroom covariates			x	x	
School covariates				x	

Note: Sample size 3121. Robust standard errors in brakets, clustered at the school level. Variables D and E only consider the time teacher was instructing. These variables assumes missing values if the teacher did not spend any time instructing. * p<0.10 ** p<0.05 *** p<0.01

Table 7 shows ITT estimates for the matched sample, using the same model presented in equation (1). Results are similar to results for the whole sample, but slightly stronger. The treatment increased teachers' time on instruction 5.6-6.6% (0.262-0.313 SD), and reduced time on classroom management and time off-task by 3.2-3.8% (0.189-0.225 SD), and 2.4- 2.8% (0.185-0.221 SD), respectively. The share of time that a large group of students was off task went down in the range of 3.7% to 5% (0.129-0.173 SD). Except for instructional time with all students engaged, coefficients are strong and significant at a 5% level. As we would expect, estimates of α_0 change little as the list of control variables changes.

d. Intent to treat effect – intra-school variation

Given the program's emphasis on promoting diffusion of good practices within schools, an expected result is decreased variation in teacher practice within treated schools. To test this impact, we calculate the standard deviation of each of the main summary variables at the school level and use it as a dependent variable.

The ITT effect is then estimated from the equation below:

$$\mu_i = \beta_0 + \beta_1 \mu_{i,t-1} + \boldsymbol{X'}_{\boldsymbol{i}} \beta_2 + \alpha_0 Z_i + \varepsilon_i \qquad (2)$$

where μ_i is the standard deviation of the classroom observation variable for school i; $\mu_{i,t-1}$ is the baseline standard deviation of the classroom observation variable collected in November of 2014; $\boldsymbol{X_i}$ represents a vector of pre-intervention characteristics at the school level; Z_i is an indicator for whether the classroom was in a treatment school; and ε_i is the error term, clustered at the school level. The coefficient of interest is α_0. We estimate (2) using the same four sets of control variables described above. Results are reported in Table 8.[10]

Table 8: Intra-school variation in summary measures of classroom observation

Dependent variable	(1)	(2)	(3)	(4)	Control Average
A. Instructional activities	-0.022*** [0.008]	-0.022*** [0.007]	-0.018** [0.007]	-0.016** [0.007]	0.166
B. Classroom management activities	-0.017** [0.007]	-0.017** [0.007]	-0.014* [0.007]	-0.012* [0.007]	0.138
C. Off-task activities	-0.020*** [0.007]	-0.020*** [0.006]	-0.020*** [0.007]	-0.018*** [0.006]	0.093
D. Instructional activities all students engaged	-0.009 [0.012]	-0.017 [0.011]	-0.005 [0.012]	-0.003 [0.012]	0.243
E. Big group (>6) of student off-task	-0.037** [0.014]	-0.024* [0.014]	-0.024* [0.013]	-0.021 [0.013]	0.209
Control for baseline		x	x	x	
Student, teacher and classroom covariates			x	x	
School covariates				x	

Note: Sample size 292. Note: Robust standard errors in brackets, clustered at the school level. * p<0.10 ** p<0.05 *** p<0.01

[10] Regression tables unpacking intra-school variation for each of the summary measures using the four specifications are shown in the appendix, tables A8 to A12.

The program reduced the variation in teacher practice within schools for all three core measures of teacher time allocation. Results are strong and significant in most specifications. The standard deviation within schools in time on instruction fell by -0.016 in the strongest specification. For classroom management activities, the standard deviation fell by -0.012 and for teacher off-task it fell by -0.018. For time on instruction with all students engaged, which improved very little for the whole sample, the change within schools is not significant. For the share of teaching time with a big group of students off-task, results are significant in some specifications, but not with full controls.

e. Intent to treat - heterogeneous effects

To assess heterogeneity in treatment effects across the distribution of teachers observed, we use the baseline data to create quartiles for each of the five key measures. It is plausible that the intervention will affect teachers differently according to where they stand in the distribution of our main variables. For example, if a teacher already achieves high time on instruction, it may be hard to improve further. Positive changes may be easier at the bottom of the distribution, where there is large room for improvement. Conversely, if being at the bottom of the distribution is a proxy for low capacity and/or motivation, achieving measurable change in teacher practice – particularly in the space of a single school year – may be more difficult.

We use the following equations to estimate heterogeneous effects:

$$y_i = \beta_0 + \beta_1 y_{i,t-1} + \boldsymbol{X'}_{\boldsymbol{i}}\beta_2 + \alpha_1 Z_i Q1_{i,t-1} + \alpha_2 Z_i Q2_{i,t-1} + \alpha_3 Z_i Q3_{i,t-1} + \alpha_4 Z_i Q4_{i,t-1} + \beta_3 Q1_{i,t-1} + \beta_4 Q2_{i,t-1} + \beta_5 Q3_{i,t-1} + \varepsilon_i \quad (3)$$

where y_i is the dependent variable for classroom observation i; $y_{i,t-1}$ is the baseline classroom dynamic variable collected in November of 2014; $\boldsymbol{X_i}$ represents a vector of pre-intervention characteristics at the school level; Z_i is an indicator for whether the classroom observed was in a school offered treatment; $Q1_{i,t-1}$, $Q2_{i,t-1}$, $Q3_{i,t-1}$, and $Q4_{i,t-1}$ are the quartiles (0-25%; 25-50%; 50-75%; and 75-100%) of the baseline classroom dynamic variables $y_{i,t-1}$; and ε_i is the error term, clustered at the school level. The coefficients of interest are α_1, α_2, α_3 and α_4. Results using specification (4)—OLS results with baseline and all covariates as controls—are presented in the table 9.

Overall, the results in Table 9 show no consistent pattern of heterogeneity connected with teachers' starting performance. For time on instruction, the strongest effect was concentrated in the second quartile (increase of 8.3% of class time), while results for the other 3 quartiles are relatively homogeneous. For other dimensions, patterns varied. Time spent on classroom management and time off-task is relatively homogeneous, with a slightly stronger effect on quartile 3 for the first (3.1%) and on quartile 4 for the latter (2.6%). A large effect is observed on the share of time a large group of students is off-task in treatment classrooms at the 4th quartile (-7.4%), indicating that the intervention had strongest impacts in highly disorganized classrooms -- where students were disengaged a very high 75-100% of total class time. This is an interesting and potentially significant result, indicating how useful coaching support can be for teachers having particular difficulty in keeping students engaged.

The intervention was organized at the level of the school, so our hypothesis was that the program might impact schools differently, depending on their starting level of average classroom practice (or variance in classroom practice). Controlling for other factors, however, we found only weak evidence that the program affected schools with different average values at baseline differently. As seen in Annex Table 13, we found statistically significant correlations in expected directions for schools at the bottom of the distribution on several key variables: time on instruction rose 6.6 percentage points in schools in the bottom quartile of the distribution (which averaged between 35-59% of class time on instruction at baseline); time on classroom management fell the most for schools in the highest quartile of this distribution at baseline (averaging 29-49% of time on classroom management); teacher time off-task fell the most for the top two quartiles of this distribution (10-43% of time off-task); and instructional activities with all students engaged rose the most in schools in the bottom quartile of this distribution (only 0-10% of instructional time with all students engaged). However, p-tests fail to confirm consistent patterns of heterogeneity in the overall distribution of the school-average results.

Table 9: Effect sizes across the main summary variables distribution

Dependent variable		(Quartile 1)	(Quartile 2)	(Quartile 3)	(Quartile 4)	p-value
	Quartile	0.00- 0.50	0.56- 0.70	0.70- 0.80	0.80- 1.00	
A. Instructional activities		0.046** [0.018]	0.084*** [0.017]	0.043*** [0.016]	0.036* [0.019]	0.059
	N	850	737	879	655	
	Quartile	0.00- 0.10	0.11- 0.20	0.22- 0.30	0.33- 1.00	
B. Classroom management activities		-0.023* [0.012]	-0.028** [0.013]	-0.032** [0.014]	-0.030** [0.014]	0.946
	N	1045	754	610	712	
	Quartile	0.00- 0.00	0.10- 0.10	0.11- 0.11	0.20- 1.00	
C. Off-task activities		-0.026*** [0.008]	-0.023*** [0.008]	-0.006 [0.042]	-0.025** [0.011]	0.963
	N	1422	912	14	773	
	Quartile	0.00- 0.00	0.10- 0.11	0.12- 0.33	0.38- 1.00	
D. Instructional activities all students engaged		-0.017 [0.024]	-0.040 [0.059]	-0.009 [0.027]	0.007 [0.031]	0.825
	N	1490	129	746	720	
	Quartile	0.00- 0.00	0.10- 0.12	0.14- 0.38	0.40- 1.00	
E. Big group (>6) of student off-task		-0.020 [0.016]	0.013 [0.032]	-0.025 [0.024]	-0.072** [0.031]	0.138
	N	1393	250	714	728	
Control for baseline		x	x	x	x	
Student, teacher and classroom covariates		x	x	x	x	
School covariates		x	x	x	x	

Note: total sample size 3121. Robust standard errors in brakets, clustered at the school level. Variables D and E only consider the time teacher was instructing. These variables assumes missing values if the teacher did not spend any time instructing. P-value tests if Quartile 1=Quartile 2=Quartile 3=Quartile 4. * p<0.10 ** p<0.05 *** p<0.01

f. Compliance effects

The treatment relied on teachers' ability to adopt changes in their practice in response to the feedback and supports provided. Most crucially, it relied on the pedagogical coordinator in each school, who was the interface between his or her school's teachers and the external coaches. The pedagogical coordinators were responsible for observing the teachers in their school, sharing their assessments with their assigned coach via Skype calls, and conveying recommended strategies and techniques back to the teachers in their school. The pedagogical coordinators were required to upload videos of themselves working with individual teachers in order to get feedback on these from their coach.

The evaluation team placed substantial emphasis on gathering monitoring data on the coordinators' and teachers' participation in the scheduled activities as well as direct measures of the skills they acquired, since both are critical issues for the effectiveness of the intervention. The coaching team kept records of all school-level activities that were reported as well as their own log of Skype conferences conducted, videos uploaded and reviewed, and feedback shared. They also asked coordinators to take an exam at the end of the program, offering certification to coordinators who had participated in at least 80% of the face to face and online activities and who achieved a score of 80% or higher on the exam. As Table 10 shows, of the 156 pedagogical coordinators in treatment schools, 138 achieved certification. The average attendance rate at the four face-to-face workshops was 86% and 68% of the coordinators achieved scores of "good" or "excellent" on the final test. Although these participation rates are reasonably high, there was clearly scope for partial compliance to hamper the impact of the program in some treatment schools (Glennerster, Takavarasha, 2013).

Table 10: Participation and certification rates for Pedagogical Coordinators (treatment schools only), 2015

Certification by ELOS

Certifield	Total	156 Sample
No	21	18
(%)	12.07%	11.54%
Yes	153	138
(%)	87.93%	88.46%
Total	174	156
(%)	100%	100%

Grade for Certification by ELOS

Grade	Total	156 Sample
Bad	21	18
(%)	12.07%	11.54%
Regular	39	30
(%)	22.41%	19.23%
Good	63	59
(%)	36.21%	37.82%
Excelent	51	49
(%)	29.31%	31.41%
Total	174	156
	100%	100%

To assess the degree to which partial compliance affected program results, we estimated the effects of the program on the schools of different pedagogical coordinators using an Instrumental Variables model. This estimate tells us the impact of the program on those schools that received the complete intervention (e.g., their pedagogical coordinator acquired the key skills imparted by the training) as compared with our Intent-to-Treat estimates (Section 4), which show the average impacts of the program on all 174 schools that were offered participation. The IV estimation uses the randomized assignment into the program (e.g., offered participation) as an instrument to predict the expected degree of full engagement in the program.

We use three measures of full engagement: (i) the pedagogical coordinator achieved certification (he or she scored adequate (regular), good or excellent on the final test); (ii) the pedagogical coordinator achieved a score of good or excellent on the final test; and (iii) the pedagogical coordinator scored excellent on the final test. We expect the magnitude of the results to increase from specification (i) to (ii) and (iii).

The IV estimate is conducted in a two-stage least-squares (2SLS) setup as initially used to adjust partial compliance in experiments by Angrist et al. (1996)[11]. In the first stage regression, we predict the degree of full engagement in the program from the random assignment. In the second stage, we regress our outcome variables on the predicted full engagement that we found in the first stage. The assumption is that a pedagogical coordinator receiving certification satisfies the exclusion restriction in an instrumental variables (IV) setup. This leads to the 2SLS estimation of the equation:

$$y_i = \beta_0 + \beta_1 y_{i,t-1} + \boldsymbol{X'}_{\boldsymbol{i}}\beta_2 + \alpha_0 c_i + \tau_i \qquad (4)$$

where ci is a dummy for being certified, and Xi is the vector of covariates. The associated first-stage relationship using Zi as an instrument is

$$c_i = \boldsymbol{X'}_{\boldsymbol{i}}\gamma_1 + \pi Z_i + \mu_i \qquad (5)$$

The estimates use the 4th specification of the ITT analysis (including all controls). The estimate of π is statistically significant for all three measures of full engagement: 0.89, 0.67 and 0.4 for the measures (i), (ii) and (iii), respectively (Annex Table A.14).

Table 11 confirms that the effects of the intervention are consistent with the regression estimates presented in section 3. The program had a significant and positive impact on the share of class time teachers devoted to instruction, increasing in the treatment schools by 5.8% to 15.8% (0.279 to 0.754 SD) as we go from specification (i) to (iii). The program helped teachers reduce the time spent on classroom management by 3.1% to 8.5% (0.188 to 0.510 SD), and time off-task from 2.8% to 7.5% (0.216 to 0.584 SD). The reduction on the percentage of time a big group of students is off task ranged from 3.4% to 9.2% (0.121 to 0.327 SD). The only variable that was not significantly affected was "time on instruction with all students engaged".

These results are quite dramatic. They suggest that pedagogical coordinators played a key role in the implementation of this program, and hint at the importance they may have more broadly for school quality. Improvements in *all* dimensions of teachers' classroom practice were consistently larger in schools where the pedagogical coordinator had stronger mastery of the concepts and techniques presented in the *Teach Like a Champion/Aula Nota 10* book and by the ELOS coaches. Program impacts in the 49 schools where pedagogical coordinators scored at the highest level (excellent) are very large. Instruction averaged 86% of class time – higher than the 85% Stallings good practice benchmark. Time on classroom management was

[11] Angrist et al (2002) is an example of using IV models to adjust for partial compliance in an RCT of a voucher program in Colombia.

13% of class time – lower than the 15% Stallings benchmark. Time off task was only 1% of class time, and time with a large group of students off-task was cut in half, to 9% of class time. Overall, the strongly progressive connection between coordinators' engagement and mastery and the magnitude of changes in teachers' practice suggests that the feedback and coaching program, if well implemented, holds considerable promise.

Table 11: 2SLS estimates of the effect on summary measures of classroom observation

Dependent variable	Certificates	Score: excellent or good	Score: excellent	Control Average
A. Instructional activities	0.058*** [0.014]	0.077*** [0.019]	0.158*** [0.039]	0.704
B. Classroom management activities	-0.031*** [0.010]	-0.041*** [0.014]	-0.085*** [0.028]	0.211
C. Off-task activities	-0.028*** [0.007]	-0.036*** [0.010]	-0.075*** [0.020]	0.085
D. Instructional activities all students engaged	-0.012 [0.022]	-0.016 [0.029]	-0.033 [0.061]	0.265
E. Big group (>6) of student off-task	-0.034* [0.018]	-0.045* [0.024]	-0.092* [0.047]	0.187
Control for baseline	x	x	x	
Student, teacher and classroom covariates	x	x	x	
School covariates	x	x	x	

Note: Sample size 3121. Robust standard errors in brackets, clustered at the school level. Variables D and E only consider the time teacher was instructing. These variables assumes missing values if the teacher did not spend any time instructing.
* p<0.10 ** p<0.05 *** p<0.01

f. Experiment Threats and Robustness checks

i.Attrition

The evaluation was designed to measure key elements of classroom dynamics in a representative sample of secondary schools spread across all 21 regional administration units of the Ceará state education system. Data collection posed significant technical and logistical challenges, from the need to train 60 of the state's pedagogical coordinators in the Stallings observation method to the logistics of reaching remote rural schools for one or more days of observations. Principals and teachers were allowed to decline participation; thus, the experiment relied on schools and teachers' willingness to be observed by outsiders, something the schools had never experienced before. Pedagogical coordinators were always assigned to school districts other than their own, so they were unfamiliar to the directors and teachers of the schools they observed.

Due to constraints during fieldwork, 56 schools from the original sample (18 treatment and 38 control) could not be observed in the baseline. The 18 schools that were originally assigned for treatment but were not observed still participated in the coaching program, but without the "information shock" of school-level feedback from classroom observations.

The loss of data for 56 schools at baseline meant an overall attrition rate of 16% that could be a source of bias, because the rates of attrition were different in the treatment and control groups, 11% and 22% respectively. Attrition was most concentrated in the state's capital city, Fortaleza, with 16 treatment and 38 control schools that could not be observed. The two major reasons were actions by the teachers' union to mobilize against the classroom observations, which caused several pedagogical coordinators from that district to decline participation in the program, and the refusal of some of the observers that remained to travel to schools in dangerous slum areas. Fortaleza's population is 4 million and many of its public high schools are located in high-risk neighborhoods. The correlation between low school socioeconomic status and probability of not being observed constitutes a clear potential source of selection bias for the key classroom observation indicators.

We carried out three strategies for dealing with attrition. First, we used Heckman's strategy for modeling the sample selection under very strong assumptions, in order to adjust for selection bias (Heckman, 1979). This approach estimates a two-stage model in which the first stage predicts selection into the program based on observed variables. The second stage regresses outcomes on the predicted selection into the program. As shown in Table 12, this approach produces results consistent with our intent-to-treat estimates, when we control for covariates of students, teachers and schools. This suggests that our sample attrition did not introduce any significant selection bias that could invalidate the experiment.

Table 12: Heckman model to adjust for Selection Bias due to Attrition

Dependent variable	Selection variables: all controls
A. Instructional activities	0.0615***
	(0.00706)
B. Classroom management activities	-0.0279***
	(0.00580)
C. Off-task activities	-0.0270***
	(0.00420)
D. Instructional activities all students engaged	-0.00314
	(0.0109)
E. Big group (>6) of student off-task	-0.0358***
	(0.00942)

Note: Sample size 3178. Variables D and E only consider the time teacher was instructing and thus assumes missing values if the teacher did not spend any time instructing. * p<0.10 ** p<0.05 *** p<0.01

A second robustness check is to estimate bounds for the average treatment effects based on weaker assumptions about the sample selection process. We estimated the bounds using the Lee trimming method that relies on the monotonicity of the outcomes if the individuals participate in the treatment (Lee, 2002 and 2009). These bounds involve excluding a fraction of the observations from the part of the sample that had less attrition (in this case, the treatment group) to equalize its size with that of the control group. In other words, the Lee bounds are generated by trimming the sample to equalize attrition rates between the treatment and control groups (Fryer, 2013). The excluded observations are the ones most likely to bias the results.

Table 13: Results for sample trimmed with Lee bounds

			Lee Bounds - No Tight		Lee Bounds - With Tight	
	ITT (no controls)	ITT (all controls)	Lower Bound	Upper Bound	Lower Bound	Upper Bound
Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)
A. Instructional activities	0.0589***	0.0470***	0.0375***	0.0816***	0.0555***	0.0555***
	(0.0124)	(0.0122)	(0.0139)	(0.0142)	(0.0127)	(0.0127)
B. Classroom management activities	-0.0334***	-0.0235***	-0.0515***	-0.0195*	-0.0310***	-0.0280***
	(0.00896)	(0.00897)	(0.0111)	(0.00996)	(0.00950)	(0.00935)
C. Off-task activities	-0.0254***	-0.0269***	-0.0378***	-0.0175**	-0.0258***	-0.0245***
	(0.00650)	(0.00647)	(0.00768)	(0.00773)	(0.00680)	(0.00672)
D. Instructional activities all students engaged	0.00347	-0.0147	-0.0454*	0.00125	0.0402	0.00255
	(0.0216)	(0.0213)	(0.0265)	(0.0217)	(0.0291)	(0.0220)
E. Big group (>6) of student off-task	-0.0492***	-0.0344**	-0.0850***	-0.0299*	-0.0473***	-0.0448***
	(0.0159)	(0.0149)	(0.0194)	(0.0170)	(0.0172)	(0.0172)
Sample Size	292	292	350	350	350	350

Note: Standardized dependent variables (z-scores). * p<0.10 ** p<0.05 *** p<0.01

Table 13 presents the lower and upper bounds results for two specifications of the Lee bounds, without any covariates and with dummy or categorical covariates that allow for tightening the bounds (Lee, 2002). We used the quintile distribution of schools based on student performance in the previous year. In the model with no tightening, columns (3) and (4), the lower bounds are significant for instructional activities (0.0375), for classroom management (-0,0515), off-task activities (-0.0454), and student off-task (-0,085). The upper bounds are significant for instructional activities (0.0816), for classroom management (-0,0195), off-task activities (-0.0175), and student off-task (-0.0299). All of our ITT estimates fall within the intervals of the Lee bounds, which suggests that selection bias due to attrition did not affect our impact estimates.

In the model with tight bounds, shown in columns (5) and (6), the estimates were positive for instructional activities and negative for classroom management, off-task activities, and big group of students off-task. In a few model specifications, our ITT estimates are not within the bounds intervals. However, all of the bounds estimates have the same sign and are close to the ITT estimates. In summary, this analysis provides further assurance that our sample attrition did not bias the comparability of our treatment and control groups, because we are able to control for a large range of observable covariates.

Finally, we conducted a more intuitive exercise to adjust for attrition. In the balance check analysis in Section 2, we showed that the treatment and control groups are balanced across covariates. However, 20 more schools were observed in the treatment group than in the control group. The fact that a higher share of our initially-defined treatment sample was observed (89% of the original treatment sample against 78% of the original control sample) could lead to a bias in unobservable characteristics. To test for this, we perform a simple exercise: instead of looking to the total of 58 schools not observed, we focus on the difference in the participation rate (the 20 school or 11% differential) and we make a series of assumptions about the possible distribution of our core variables if 20 additional control schools *had been* part of the sample. This enables us to set some bounds on how our results might have been impacted.

First, we suppose that the 20 missing schools had perfect teaching and the variable for teacher time on instruction was at the 90% point of the distribution in all cases. In this case, the mean difference between treatment and control would have been 2.8% and with the standard deviation of 0.013—from specification (4) in table 5—we would still have found a significant effect at the 5% level. On the other hand, if we assume the opposite scenario, with time on instruction at the 10% point of the distribution in these 20 schools, we would have found a large effect of 7.9% at a 1% level of significance.

We perform this exercise for the 5 summary variables, playing with the different assumptions about where the average values for these 20 schools could have fallen in the distribution – 90%, 75%, 50%, 25% and 10%. Table 14 shows the results under four of our model specifications (OLS results with baseline and all covariates as controls).

Overall, the exercise confirms the robustness of our results: we would have still found sizeable and significant effects in most of the scenarios. For teacher time on instruction, coefficients range from 2.8% to 7.9% and are always significant at the 5% level. Results for teacher off-task range from -1.4% to -3.9% and are always significant. Classroom management results range from -2.7% to -5.3% and would be significant in all cases except the assumption of baseline performance at the lowest points in the distribution (25% and 10%).

The impacts of the program on the two student engagement measures – share of time with all students engaged and share of time with a large group of students off-task – would be affected if all the missing schools were at extremes of the distribution. For instructional activities with all students engaged, if the missing schools had been above the 75% point in the initial distribution (very high in comparison with the overall sample mean of 21%), the model predicts the intervention actually would have had a negative effect. For the share of time with a large group of students off-task, results would have remained significant only if the missing schools were above the very high 75% point in the distribution (compared with the sample mean of 20%).

This exercise provides additional assurance that our results are not affected by the loss of 20 schools from the control group. Given characteristics of the overall sample, it is highly unlikely that all 20 schools would rest at either extreme of the distribution on any of these variables, and under all other scenarios our impact estimates are not affected.

Table 14: Robustness check exercise

Percentile assumption for missing values on control schools	A. Instructional activities	B. Classroom management activities	C. Off-task activities	D. Instructional activities all students engaged	E. Big group (>6) of student off-task
	(1)	(2)	(3)	(4)	(5)
90%	0.028**	-0.053***	-0.039***	-0.071***	-0.081***
75%	0.028**	-0.04***	-0.027***	-0.039**	-0.04**
50%	0.041***	-0.027***	-0.014**	0.007	-0.008
25%	0.066***	-0.015	-0.014**	0.025	-0.008
10%	0.079***	-0.002	-0.014**	0.025	-0.008

Note: 1 to 5 consider coefficients from OLS results with baseline and all covariates as control for the robustness check exercise. * p<0.10 ** p<0.05 *** p<0.01

ii.Spillover

Since the treatment was allocated at the school level, and the sample was drawn across different municipalities state-wide, teachers in the control schools were not likely to know about or participate in any part of the treatment. Only pedagogical coordinators from treatment schools were trained in the Stallings method and participated in the data collection. The online website for the coaching program could only be accessed with a school code.

Nevertheless, it is possible that some regional supervisors, who were aware of the intervention, conveyed information about the program to principals of control schools, even though they were asked to avoid this. If this happened, it could create spillover effects that reduced the quality of the counterfactual because outcomes in the control schools were also affected by the program.

Data from a questionnaire applied to principals at baseline and endline provide mixed evidence on the possibility that some control schools became aware of the program. Principals were asked to identify the single most important of six possible strategies (including the option of doing nothing) to improve teachers' effectiveness (Table 15). On the one hand, the share of principals in treatment schools that identified "feedback based on classroom observation" as most important rose from 11% to 22%, compared to an increase among control school principals of only 3.5 percentage points. However, the increase in the number of control school principals who named "coaching of teachers" as the most important strategy was as large as the increase among treatment school principals. As discussed in the next section on contamination, there was in fact another teacher coaching program implemented in Ceará in 2014 and 2015, which we believe is a more likely explanation for this result.

Table 15: Principals' Survey: Which of these instruments are most important for raising teacher quality?

	Spoken guidance	Written guidance	Feedback on lesson planning	Feedback based on classroom observation	Coaching of teachers	Nothing	Total
Baseline							
Control	84	11	23	17	0	2	136
Treatment	104	11	18	19	3	1	156
Total	188	22	41	36	3	2	292
Endline							
Control	69	10	20	22	15	0	136
Treatment	76	13	15	35	16	1	156
Total	145	23	35	57	31	1	292

A final source of potential spillover is the fact that 8% of teachers in the control group and 10% in the treatment group work in more than one school. As most secondary schools run morning and afternoon and sometimes evening shifts, teachers may work the different shifts in two different schools. If a teacher in a treatment school also works in a control school, it would be natural to share information about the program, including teaching practices recommended by the coaches, with the control school's pedagogical coordinator.[12] To investigate the scope for this, we verified the school assignments of teachers in our sample. Only 3% of our control schools and 1.7% of our treatment schools have teachers that work in both treatment and control schools.

Table 16: Possibility of spillover from teachers working in more than one school

	Number of schools where teachers work				
	1	2	3	4	Total
Control	107	21	7	1	136
(%)	30.66	6.02	2.01	0.29	38.97
Treatment	119	30	7	0	156
(%)	34.1	8.6	2.01	0	44.7
Not Observec	42	10	5	0	57
	12.03	2.87	1.43	0	16.33
Total	268	61	19	1	349
	76.79	17.48	5.44	0.29	100

We implemented two strategies to check the robustness of our impact estimates in the context of likely spillover. First, we assumed that the municipalities with the largest number of classrooms observed, whether from treatment or control schools, are more susceptible to spillover. The larger the absolute number of treatment teachers participating in the program, the higher the chance that some of these may know teachers in control schools in the same municipality. They might understandably think these colleagues could also benefit from the coaching advice and training materials imparted by the program, such as the *Aula Nota 10* (*Teach Like a Champion*) book. In fact, in focus group discussions at the end of the program, several control school principals confirmed that some of their teachers had heard about the *Aula Nota 10* book and purchased it on their own initiative. We were unable to get systematic data on these anecdotes, but we tested the potential impact of between-school spillovers by including in our regressions a variable on the number of classrooms in the municipality and a variable on the number of treated classrooms in the same municipality.[13] The results in table 17 suggest that the additional variables we used to test for spillovers were not significant. The intervention effect was greater than the mean effect size in all specifications of the model. Except for time on instruction with all students engaged, all the other four variables are significant at the 1% level. The effect of the intervention is 5,79% on time spent on instructional activities; -3,24% on classroom management activities; -2,55% on teacher time off task is -2,55; and -4,85% on big group of students off-

[12] Regarding the pedagogical coordinators, they are usually only assigned to work in one school.

[13] We adapted this strategy based on the work of Miguel and Kremer (2004).

task. We take this as evidence that while some spillovers appear to have occurred, their effects were not significant enough to threaten our conclusions about the impact of the intervention.

Table 17: Test for spillovers in municipalities with high concentration of teachers in the program

	A. Instructional activities	B. Classroom management activities	C. Off-task activities	D. Instructional activities all students engaged	E. Big group (>6) of student off-task
	(1)	(2)	(3)	(4)	(5)
Treat	0.0579***	-0.0324***	-0.0255***	-0.00575	-0.0485***
	(0.0147)	(0.0105)	(0.00747)	(0.0231)	(0.0183)
Classroons in the municipality*Treat	-0.0000544	0.0000465	0.00000791	-0.0000577	0.000187*
	(0.0000682)	(0.0000580)	(0.0000442)	(0.000129)	(0.0000960)
Classroons in the municipality	-0.0000410	-0.0000226	0.0000636	0.0000604	-0.0000489
	(0.0000683)	(0.0000563)	(0.0000419)	(0.000128)	(0.000105)
Sample Size	3121	3121	3121	3085	3085

Note: Robust standard errors in parentheses, clustered at the school level. Variables D and E only consider the time teacher was instructing and thus assumes missing values if the teacher did not spend any time instructing. * p<0.10 ** p<0.05 *** p<0.01

Our second strategy relied on information from the principals' questionnaire about their perceptions of the most important tool for raising teacher quality. Two of the options we offered related to our program – providing feedback from classroom observations and offering coaching to teachers. Our assumption is that if regional supervisors shared information about the main elements of the program with principals in control schools, these principals may have tried to implement similar activities for their schools. We tested this hypothesis by adding to our regressions a dummy variable related to principals' responses about these key instruments and testing their interaction with the treatment variables.

Table 18: Test for possible spillover of program elements to control school principals

	A. Instructional activities	B. Classroom management activities	C. Off-task activities	D. Instructional activities all students engaged	E. Big group (>6) of student off-task
	(1)	(2)	(3)	(4)	(5)
Treat	0.0613***	-0.0359***	-0.0254***	0.00185	-0.0392**
	(0.0141)	(0.0101)	(0.00742)	(0.0223)	(0.0178)
Feedback & Coaching *Treat	-0.0307	0.0262	0.00448	-0.0418	0.0295
	(0.0299)	(0.0201)	(0.0161)	(0.0419)	(0.0343)
Feedback & Coaching	0.0310	-0.0196	-0.0114	-0.00252	0.0138
	(0.0255)	(0.0165)	(0.0139)	(0.0336)	(0.0261)
Sample Size	3121	3121	3121	3085	3085

Note: Robust standard errors in parentheses, clustered at the school level. Variables D and E only consider the time teacher was instructing and thus assumes missing values if the teacher did not spend any time instructing. * p<0.10 ** p<0.05 *** p<0.01

The results, in table 18, are similar to those of Table 17 on the spillover threat linked to the number of classrooms in a municipality. First, the additional variables are not significant. Second, the intervention still shows strong and statistically significant effects in all of the model specifications. This provides additional evidence that, while information about the program may have spilled over to principals in control schools, its effects were not significant enough to change our conclusions about the impact of the intervention.

iii.Treatment Contamination

The Ceará government is known for its emphasis on education, and in addition to the Teacher Feedback and Coaching program, the Secretariat implemented three other important programs aimed at raising secondary school quality over the 2015 school year. A "socioemotional learning program" supported school administrators and teachers in delivering a special curriculum designed to strengthen the socioemotional skills of both teachers and students. It was offered to 80 secondary schools; 23 of these fell in our control group and 18 in our treatment group.[14]

A second program, called *Tutoria Pedagogica*, has very similar objectives to the Teacher Feedback and Coaching program. *Tutoria Pedagogica* aims at developing professional learning communities, based on models in New York City and Ontario Canada. However, this program was in a pilot phase during 2015 and

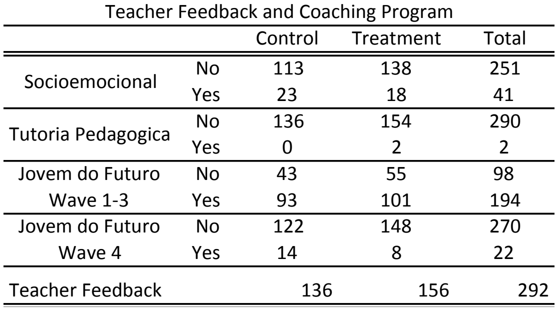
[14] This intervention was designed by University of São Paulo researchers and was financed by the Itau Social Foundation.

was only implemented in 10 schools; two of these fell in our treatment group, but none were among our control schools.

The third program, *Jóvem do Futuro*, aims at improving school management and accountability. JF has much wider coverage; it has been going on for 4 years and has reached 216 secondary schools. Waves 1 to 3 of the program cannot contaminate our treatment group because they were implemented before our randomization assigned treatment and control schools.[15] However, the 2015 wave of the program could contaminate our results, as it was rolled out at the same time as the Teacher Feedback program. The 4th wave of *Jovem do Futuro* covered 22 schools, 14 in our control group and 8 in our treatment group.

To adjust for possible treatment contamination, we ran the main regressions controlling for each program and the interaction with our treatment. We did not find differences, as shown in Annex Table A.15. However, statistical power was low, because the degree of crossover between the treatment and the other programs was very low. This implies large standard errors for estimation of the interaction effect. We are therefore unable to reject any hypotheses about the interaction.

Table 19: Overlap in education quality programs implemented in Ceará secondary schools, 2015



		Teacher Feedback and Coaching Program		
		Control	Treatment	Total
Socioemocional	No	113	138	251
	Yes	23	18	41
Tutoria Pedagogica	No	136	154	290
	Yes	0	2	2
Jovem do Futuro Wave 1-3	No	43	55	98
	Yes	93	101	194
Jovem do Futuro Wave 4	No	122	148	270
	Yes	14	8	22
Teacher Feedback		136	156	292

iv.Evaluation-Driven Effects

Social experiments are often exposed to the risk of evaluation-driven effects that hinder the identification of program impacts. The mere fact of being part of an evaluation can motivate individuals to change their behavior. In the case of the Teacher Feedback and Coaching program, there is clear scope for Hawthorne effects because data collection requires the presence of an outside observer in the classroom, which is out of the ordinary in Brazilian schools.

Teachers in both the treatment and control schools are likely to try to exhibit their best teaching practice, perhaps especially during the endline round of observations if they believe they are being compared to an earlier observation. Endline measures of classroom dynamics in the control schools also "improved" modestly over their baseline values, and Hawthorne and/or John Henry effects are clearly a plausible explanation.

A bigger issue for the evaluation is that teachers in the treatment schools were especially susceptible to evaluation-driven effects. Over the 2015 school year, they were observed several times and received feedback from their pedagogical coordinators. At endline, they had a much better idea than teachers in control schools of why someone was coming to observe them and what things the observer would measure. They were also more knowledgeable about what good classroom practice is and how important it is to use class time effectively and keep students engaged.

Observer teams in both the baseline and endline rounds reported some instances where students commented after class that the teacher had repeated the previous day's lesson, clearly a change in teacher behavior induced by the evaluation. On the other hand, observers trained in the Stallings method generally concur that it is difficult for any teacher to sustain unfamiliar teaching practices for a full class hour or 100 minutes. Indeed, our results show that even where teachers improved the efficiency of classroom administrative processes and freed up more time for instruction, they were not able to sustain interactive question and answer/discussion activities during all of the incremental time. Treatment school teachers also increased the share of class time that students spent "copying", either from the blackboard or textbooks.

Evaluation-driven effects can be expected to introduce some upward bias into the Stallings measures of classroom dynamics at any point. But while there is no reason to suppose a differential effect on control and treatment schools at baseline, we can clearly expect differential effects at the endline, and some part of the improvement in treatment schools' measured classroom dynamics likely was evaluation driven.

However, we can exclude the possibility that *all* of the changes in teacher behavior were evaluation driven, as the program also produced significant impacts on student learning, discussed in the next section. It is

[15] The Unibanco Institute developed this program. The 4 waves of the program have slightly different designs.

implausible that improvements in students' test scores could be the result of one day during the school year when teachers were observed and changed their practice.

5. Impacts on Student Learning

a. Sample and balance

Ceará administers an annual, state-wide, standardized learning assessment, the *Sistema de Avaliação Padronizada de Aprendizagem do estado do Ceará,* SPAECE. It covers all students in 2nd, 5th and 9th grade of elementary school and until recently covered all three grades of secondary school. It tests Portuguese language and mathematics. The test also includes a student survey, and surveys of teachers and the school director.

Ceará's secondary school students also take a national exam at the end of high school, called ENEM (*Exame Nacional do Ensino Médio*). The ENEM was designed in 1998 as a low-stakes high-school leaving exam that could serve as a tool for monitoring education system quality. In 2009, however, Ministry of Education established ENEM as the official university entrance exam, and most universities now either require the ENEM alone or as a supplement. It has thus become a high-stakes exam for graduating students and a growing number of first and second year students also take the exam, to gain practice.

The ENEM is administered over two days, with the first session lasting four hours and 30 minutes and the second lasting five hours and 30 minutes. It covers all subjects: Natural Sciences (Biology, Physics, Chemistry), Human Sciences (History, Geography, Philosophy, Sociology), Languages and Codes (Portuguese Language, Literature, Foreign Languages, Information technology, and Communication), Math (Algebra, Geometry), and Text writing. The final score is a weighted average of the different tests. Almost 6 million Brazilian teenagers sat the exam in 2015.

In 2013 the Ceará Secretariat began applying the state assessment, SPAECE, on a sample basis[16] to second and third year secondary students, given that most now take the ENEM exam. In 2015, students in the 2nd year of secondary school were not tested at all and testing of 3rd year students was limited to schools participating in the *Jovem do Futuro* program. Beginning in 2016, state policy is to apply SPAECE only to students in the first year of secondary school.

Table 20 - Test Participation Rates in Ceará secondary schools, 2013-2015

	Total	1st Grade	2nd Grade	3rd Grade
2013				
Enrollment	214,822	82,850	70,228	61,744
Spaece	89767	60113	15559	14095
%	42%	73%	22%	23%
ENEM	92126	7921	40024	44181
%	43%	10%	57%	72%
Spaece & ENEM	28422	7158	10321	10943
%	13%	9%	15%	18%
2014				
Enrollment	206,236	78,548	67,508	60,180
Spaece	92692	56433	18828	17431
%	45%	72%	28%	29%
ENEM	88601	1	41994	46606
%	43%	0%	62%	77%
Spaece & ENEM	28625	0	13562	15063
%	14%	0%	20%	25%
2015				
Enrollment	191,968	73,143	62,917	55,908
Spaece	71103	59805	543	10755
%	37%	82%	1%	19%
ENEM	87027	7146	34043	45838
%	45%	10%	54%	82%
Spaece & ENEM	15920	6807	8	9105
%	8%	9%	0%	16%

Table 20 shows test participation rates between 2013 and 2015 for our sample of 348 schools. In 2015, 82% of first year students and 19% of 3rd year students took the SPAECE exam. ENEM participation was almost the inverse, with 10% of first year students, 54% of 2nd year and 82% of 3rd year students taking the test in 2015. Because ENEM was designed to evaluate 3rd year students and since the participation of younger students is unrepresentative (biased towards stronger students) our analysis focuses on the ENEM results for 3rd year students. We also focus on the scores for Math and Portuguese, for two reasons. First, these are the

[16] The sample of students tested was random and representative at the school level.

subjects with the most prominence in the curriculum. Second, our classroom observation sample prioritized the observation of math and Portuguese teachers, to increase the comparability of the types of lessons observed. For SPAECE, we concentrate on the results for 1st year students, which have the broadest sample.

Table 21 – Sample Balance in Student Characteristics, pre-treatment, 2013-2015 - Ceará census and administrative data

		2013			2014			2015		
	Grade	N	Control Means (T=0)	Diff. (T[1]-T[0])	N	Control Means (T=0)	Difference (T[1]-T[0])	N	Control Means (T=0)	Diff. (T[1]-T[0])
Covariates										
Female	1	82850	0.50 [0.00]	0.01 [0.01]	78548	0.51 [0.00]	0.00 [0.01]	73143	0.49 [0.00]	0.00 [0.01]
Female	2	70228	0.54 [0.00]	0.00 [0.01]	67508	0.53 [0.00]	0.00 [0.01]	62917	0.53 [0.00]	-0.00 [0.01]
Female	3	61744	0.55 [0.00]	-0.01 [0.01]	60180	0.55 [0.00]	0.00 [0.01]	55908	0.54 [0.00]	0.00 [0.01]
White	1	82850	0.13 [0.01]	-0.01 [0.01]	78548	0.12 [0.01]	-0.00 [0.01]	73143	0.11 [0.01]	0.00 [0.01]
White	2	70228	0.13 [0.01]	-0.00 [0.01]	67508	0.13 [0.01]	-0.01 [0.01]	62917	0.12 [0.01]	-0.00 [0.01]
White	3	61744	0.14 [0.01]	0.01 [0.01]	60180	0.14 [0.01]	-0.00 [0.01]	55908	0.14 [0.01]	-0.00 [0.01]
Black	1	82850	0.01 [0.00]	0.00 [0.00]	78548	0.01 [0.00]	0.00 [0.00]	73143	0.01 [0.00]	0.00 [0.00]
Black	2	70228	0.01 [0.00]	0.00 [0.00]	67508	0.01 [0.00]	0.00 [0.00]	62917	0.01 [0.00]	0.00 [0.00]
Black	3	61744	0.02 [0.00]	0.00 [0.00]	60180	0.01 [0.00]	0.00 [0.00]	55908	0.01 [0.00]	0.00 [0.00]
Brown	1	82850	0.60 [0.02]	0.01 [0.03]	78548	0.62 [0.02]	-0.01 [0.03]	73143	0.59 [0.02]	-0.01 [0.03]
Brown	2	70228	0.59 [0.02]	0.01 [0.03]	67508	0.60 [0.02]	-0.00 [0.03]	62917	0.63 [0.02]	-0.02 [0.03]
Brown	3	61744	0.57 [0.02]	0.01 [0.03]	60180	0.60 [0.02]	0.01 [0.03]	55908	0.62 [0.02]	-0.01 [0.03]
Age in years	1	82850	17.07 [0.07]	-0.01 [0.10]	78548	17.08 [0.06]	0.01 [0.09]	73143	17.07 [0.06]	-0.03 [0.09]
Age in years	2	70228	17.95 [0.07]	0.02 [0.10]	67508	17.87 [0.06]	-0.01 [0.09]	62917	17.88 [0.06]	-0.01 [0.09]
Age in years	3	61744	18.92 [0.07]	-0.04 [0.10]	60180	18.89 [0.07]	-0.03 [0.10]	55908	18.81 [0.06]	-0.02 [0.09]
Students per class	1	82850	40.18 [0.88]	-1.43 [0.98]	78548	37.90 [0.37]	-0.06 [0.57]	73143	36.88 [0.41]	-0.36 [0.59]
Students per class	2	70228	38.18 [0.96]	-1.33 [1.09]	67508	36.19 [0.39]	-0.38 [0.58]	62917	35.23 [0.47]	-0.01 [0.65]
Students per class	3	61744	37.76 [1.02]	-1.10 [1.10]	60180	36.04 [0.41]	-0.71 [0.59]	55908	34.79 [0.46]	-0.34 [0.62]

Note: Standard errors in brakets, clustered at the school level . Statistical significance levels * p<0.1 ** p<0.05 *** p<0.01

Out of our original randomized sample of 350 schools, one treatment school and one control school closed between 2014 and 2015. Beyond this, 18 treatment schools were not able to be observed at baseline and thus could not receive the "information shock" of observation results at the beginning of 2015. However, all 174 schools received the other two elements of the intervention – coaching and self-help materials – and we have learning outcomes for the full sample of 348 schools.

Tables 21 and 22 examine the balance on pre-treatment covariates at the student level for the 2013, 2014 and 2015 cohorts of students. Table 21 uses administrative data from the National Educational Census, while Table 22 uses data from the SPAECE student survey.

Imbalance on these variables might indicate selection bias in the students taking the exam, and there are some small imbalances. For race, there were approximately 1% fewer black students in our treatment group in 2014 (among 2nd year students) and in 2015 (among 3rd year students) when compared to the control group, and 2% more brown students in several years (among 1st year students in 2013, 2nd year students in 2014 and 3rd year students in 2015). The proportion of parents who had not completed high school was 1% smaller in the treatment group for 1st year students in 2014. Somewhat countervailing this, the proportion of parents who had not completed primary school was 1% higher in our treatment group for 2nd year students in 2015. Overall, however, the students in our treatment and control groups are quite well balanced on demographic and background characteristics.

Table 22 – Sample Balance in Student Characteristics, Pre-treatment, 2013-2015 - SPAECE student survey data.

		2013			2014			2015		
	Grade	N	Control Means (T=0)	Diff. (T[1]-T[0])	Sample	Control Means (T=0)	Difference (T[1]-T[0])	N	Control Means (T=0)	Diff. (T[1]-T[0])
Covariates										
White	1	61909	0.19 [0.00]	-0.01 [0.01]	56668	0.18 [0.00]	-0.00 [0.01]	64032	0.20 [0.00]	0.00 [0.01]
White	2	25335	0.18 [0.00]	0.00 [0.01]	52383	0.19 [0.00]	-0.01 [0.01]	47462	0.18 [0.00]	-0.00 [0.01]
White	3	12643	0.19 [0.01]	-0.01 [0.01]	24184	0.17 [0.00]	0.01 [0.01]	44422	0.19 [0.00]	-0.01 [0.01]
Black	1	61909	0.13 [0.00]	-0.00 [0.00]	56668	0.10 [0.00]	0.00 [0.00]	64032	0.08 [0.00]	-0.01 [0.00]
Black	2	25335	0.11 [0.00]	-0.01 [0.01]	52383	0.12 [0.00]	-0.01** [0.00]	47462	0.10 [0.00]	0.00 [0.00]
Black	3	12643	0.11 [0.00]	0.01 [0.01]	24184	0.10 [0.00]	-0.00 [0.01]	44422	0.12 [0.00]	-0.01* [0.00]
Brown	1	61909	0.58 [0.01]	0.02* [0.01]	56668	0.62 [0.01]	0.01 [0.01]	64032	0.57 [0.00]	0.01 [0.01]
Brown	2	25335	0.64 [0.01]	0.01 [0.01]	52383	0.60 [0.01]	0.02*** [0.01]	47462	0.62 [0.01]	0.01 [0.01]
Brown	3	12643	0.62 [0.01]	0.00 [0.01]	24184	0.65 [0.01]	0.00 [0.01]	44422	0.61 [0.01]	0.02** [0.01]
Elementary School Incomplete (EF I)	1	61164	0.17 [0.01]	0.00 [0.01]	56341	0.15 [0.01]	0.01 [0.01]	63320	0.17 [0.01]	0.01 [0.01]
Elementary School Incomplete (EF I)	2	25101	0.19 [0.01]	-0.01 [0.01]	51814	0.16 [0.01]	0.00 [0.01]	47233	0.14 [0.01]	0.01* [0.01]
Elementary School Incomplete (EF I)	3	12442	0.20 [0.01]	-0.01 [0.01]	23977	0.19 [0.01]	-0.00 [0.01]	43943	0.16 [0.01]	0.01 [0.01]
Secondary School Incomplete (EF II)	1	61164	0.29 [0.00]	0.00 [0.01]	56341	0.28 [0.00]	0.00 [0.01]	63320	0.25 [0.00]	-0.00 [0.01]
Secondary School Incomplete (EF II)	2	25101	0.30 [0.01]	0.00 [0.01]	51814	0.29 [0.00]	0.00 [0.01]	47233	0.28 [0.01]	-0.00 [0.01]
Secondary School Incomplete (EF II)	3	12442	0.31 [0.01]	0.01 [0.01]	23977	0.30 [0.01]	-0.00 [0.01]	43943	0.28 [0.01]	0.00 [0.01]
High School Incomplete	1	61164	0.16 [0.00]	-0.00 [0.01]	56341	0.17 [0.00]	-0.01* [0.01]	63320	0.17 [0.00]	-0.01* [0.00]
High School Incomplete	2	25101	0.16 [0.00]	-0.00 [0.01]	51814	0.16 [0.00]	0.00 [0.01]	47233	0.17 [0.00]	-0.01 [0.01]
High School Incomplete	3	12442	0.15 [0.01]	-0.00 [0.01]	23977	0.16 [0.00]	-0.00 [0.01]	43943	0.16 [0.00]	0.00 [0.01]
High School Complete	1	61164	0.18 [0.01]	0.00 [0.01]	56341	0.18 [0.01]	-0.00 [0.01]	63320	0.18 [0.01]	0.00 [0.01]
High School Complete	2	25101	0.18 [0.01]	0.01 [0.01]	51814	0.19 [0.01]	-0.00 [0.01]	47233	0.19 [0.01]	-0.00 [0.01]
High School Complete	3	12442	0.18 [0.01]	-0.00 [0.01]	23977	0.18 [0.01]	0.01 [0.01]	43943	0.19 [0.01]	-0.00 [0.01]
Paved Street	1	61822	0.65 [0.01]	-0.01 [0.02]	56231	0.65 [0.01]	-0.02 [0.02]	63354	0.65 [0.01]	-0.02 [0.02]
Paved Street	2	25316	0.64 [0.01]	-0.00 [0.02]	52237	0.65 [0.01]	-0.00 [0.02]	47168	0.65 [0.01]	-0.01 [0.02]
Paved Street	3	12602	0.65 [0.01]	-0.00 [0.02]	24137	0.64 [0.01]	0.00 [0.02]	44239	0.65 [0.01]	-0.00 [0.02]
Cash Transfer	1	62021	0.70 [0.01]	0.00 [0.01]	56155	0.69 [0.01]	-0.00 [0.01]	63965	0.62 [0.01]	0.00 [0.01]
Cash Transfer	2	25389	0.66 [0.01]	-0.01 [0.01]	52457	0.69 [0.01]	-0.00 [0.01]	47063	0.69 [0.01]	-0.00 [0.02]
Cash Transfer	3	12678	0.64 [0.01]	-0.00 [0.01]	24193	0.64 [0.01]	-0.01 [0.01]	44490	0.66 [0.01]	-0.00 [0.01]

Note: Standard errors in brakets, clustered at the school level . Statistical significance levels * p<0.1 ** p<0.05 *** p<0.01

Notwithstanding the balance on demographic characteristics and background, there is a marked imbalance in prior test performance. As Table 23 shows, in both years prior to the intervention, 1st year students in the treatment schools had significantly higher SPAECE scores in both Math and in Portuguese, of around 0.08 to 0.1 SD. In 2013, third year students in treatment schools also performed better in both subjects on ENEM. In 2014, however, the ENEM performance of students in our treatment and control schools was balanced. Although all of these test scores are for different cohorts of students than the cohort being evaluated, this degree of imbalance raises questions of bias in unobservable characteristics of these schools. Our econometric analysis uses different tests to deal with this imbalance.[17]

[17] Annex table A16 shows descriptive statistics for SPAECE and ENEM 2015 scores, by treatment group.

Table 23. Pre-treatment balance in student test scores, 2013-2014 (prior cohorts)

		2013			2014		
	Grade	N	Control Means (T=0)	Diff. (T[1]-T[0])	N	Control Means (T=0)	Difference (T[1]-T[0])
Covariates							
Spaece - Mathematics	1	60113	-0.00 [0.03]	0.10** [0.05]	56433	-0.00 [0.03]	0.11** [0.05]
Spaece - Portuguese	1	60125	-0.00 [0.03]	0.09** [0.04]	56435	0.00 [0.03]	0.08* [0.04]
ENEM - Portuguese	3	44181	0.00 [0.03]	0.08* [0.05]	46606	0.00 [0.03]	0.03 [0.04]
ENEM - Mathematics	3	44181	0.00 [0.03]	0.08* [0.04]	46606	-0.00 [0.02]	0.04 [0.04]

Note: Standard errors in brakets, clustered at the school level . Statistical significance levels * p<0.1 ** p<0.05 *** p<0.01

Tables 24 and 25 analyze the school-level averages for key variables over the 2013-2015 period. Table 24 presents school and student characteristics, by grade, and the final column looks for possible trends in the school-level averages. Table 25 presents school-average test scores, for SPAECE and ENEM in 2013 and 2014.

Table 24 – Pre-treatment balance in school and student background characteristics, by grade, 2013-2014 (previous cohorts)

		2013			2014			Difference 2014-2013		
	Grade	N	Control Means (T=0)	Difference (T[1]-T[0])	N	Control Means (T=0)	Difference (T[1]-T[0])	N	Control Means (T=0)	Difference (T[1]-T[0])
Covariates										
Enrollment	1	350	252.22 [10.52]	-31.01** [14.88]	349	237.60 [9.49]	-25.14* [13.44]	699	-14.62 [14.17]	5.87 [20.05]
Enrollment	2	350	209.47 [8.85]	-17.65 [12.51]	349	204.62 [8.35]	-22.43* [11.83]	699	-4.86 [12.16]	-4.79 [17.22]
Enrollment	3	350	182.46 [7.90]	-12.09 [11.17]	349	179.57 [7.47]	-14.31 [10.59]	699	-2.89 [10.88]	-2.22 [15.39]
Female	1	350	0.51 [0.00]	0.01 [0.01]	349	0.51 [0.00]	0.00 [0.01]	699	-0.00 [0.01]	-0.01 [0.01]
Female	2	350	0.54 [0.00]	0.00 [0.01]	349	0.53 [0.00]	0.01 [0.01]	699	-0.01 [0.01]	0.00 [0.01]
Female	3	350	0.55 [0.00]	-0.01 [0.01]	349	0.55 [0.00]	0.00 [0.01]	699	-0.01 [0.01]	0.01 [0.01]
White	1	350	0.12 [0.01]	0.00 [0.01]	349	0.12 [0.01]	0.01 [0.01]	699	-0.01 [0.01]	0.00 [0.02]
White	2	350	0.12 [0.01]	0.01 [0.01]	349	0.13 [0.01]	0.01 [0.01]	699	0.01 [0.01]	-0.01 [0.02]
White	3	350	0.13 [0.01]	0.03* [0.01]	349	0.13 [0.01]	0.01 [0.01]	699	0.00 [0.01]	-0.02 [0.02]
Black	1	350	0.01 [0.00]	0.00 [0.00]	349	0.01 [0.00]	0.00 [0.00]	699	-0.00 [0.00]	-0.00 [0.00]
Black	2	350	0.01 [0.00]	0.00 [0.00]	349	0.01 [0.00]	0.00 [0.00]	699	-0.00 [0.00]	0.00 [0.00]
Black	3	350	0.02 [0.00]	0.00 [0.00]	349	0.01 [0.00]	0.00 [0.00]	699	-0.00 [0.00]	-0.00 [0.00]
Brown	1	350	0.58 [0.02]	0.01 [0.03]	349	0.60 [0.02]	-0.01 [0.03]	699	0.02 [0.03]	-0.02 [0.04]
Brown	2	350	0.59 [0.02]	0.00 [0.03]	349	0.58 [0.02]	0.00 [0.02]	699	-0.00 [0.03]	0.00 [0.04]
Brown	3	350	0.56 [0.02]	0.01 [0.03]	349	0.59 [0.02]	0.01 [0.02]	699	0.03 [0.03]	-0.01 [0.04]
Age in years	1	350	17.09 [0.08]	-0.05 [0.11]	349	17.11 [0.07]	-0.05 [0.10]	699	0.01 [0.11]	-0.01 [0.15]
Age in years	2	350	17.98 [0.08]	-0.01 [0.11]	349	17.92 [0.07]	-0.03 [0.10]	699	-0.06 [0.10]	-0.02 [0.15]
Age in years	3	350	18.93 [0.08]	-0.03 [0.11]	349	18.93 [0.07]	-0.05 [0.11]	699	0.00 [0.11]	-0.02 [0.15]
Students per class	1	350	38.83 [0.53]	-0.83 [0.75]	349	36.99 [0.42]	-0.06 [0.60]	699	-1.84 [0.68]	0.77 [0.96]
Students per class	2	350	36.49 [0.54]	-0.87 [0.76]	349	34.79 [0.42]	-0.20 [0.59]	699	-1.71 [0.68]	0.67 [0.96]
Students per class	3	350	35.77 [0.57]	-0.38 [0.80]	349	34.78 [0.43]	-0.49 [0.62]	699	-0.99 [0.71]	-0.11 [1.01]

Note: Standard errors in brakets. Statistical significance levels * p<0.1 ** p<0.05 *** p<0.01

Table 25 shows that a slightly higher proportion of treatment school students took the SPAECE in 2013 and 2014 (3%) and ENEM in 2014 (3%). What is more concerning is the consistently higher SPAECE performance of students in the treatment schools, in both Math (0.1-0.14 SD) and Portuguese (0.10) in 2013 and 2014. This is a very big gap in performance and hard to explain given how well matched our treatment and control schools are on student background and school characteristics. On the other hand, once again, there are no differences in 2014 ENEM scores. The final column confirms that there are no trends in the differences across our sample.

Table 25. Pre-treatment balance in student test scores, across schools, 2013-2014 (previous cohort)

		2013			2014			Difference 2014-2013		
	Grade	N	Control Means (T=0)	Difference (T[1]-T[0])	N	Control Means (T=0)	Difference (T[1]-T[0])	N	Control Means (T=0)	Difference (T[1]-T[0])
Covariates										
% Students who took Spaece	1	350	0.72 [0.01]	0.03** [0.02]	349	0.71 [0.01]	0.03* [0.01]	699	-0.01 [0.01]	-0.00 [0.02]
Spaece - Mathematics	1	350	0.01 [0.04]	0.11** [0.05]	349	-0.02 [0.04]	0.14** [0.06]	699	-0.03 [0.05]	0.03 [0.08]
Spaece - Portuguese	1	350	0.01 [0.03]	0.10** [0.05]	349	-0.02 [0.03]	0.10** [0.05]	699	-0.03 [0.05]	0.00 [0.07]
% Students who took ENEM	3	350	0.72 [0.01]	0.01 [0.02]	349	0.76 [0.01]	0.03** [0.01]	699	0.04 [0.02]	0.02 [0.02]
ENEM - Portuguese	3	350	0.01 [0.03]	0.06 [0.04]	349	-0.00 [0.02]	0.01 [0.03]	699	-0.02 [0.04]	-0.04 [0.05]
ENEM - Mathematics	3	350	0.00 [0.03]	0.06 [0.04]	349	-0.01 [0.03]	0.03 [0.04]	699	-0.01 [0.04]	-0.04 [0.06]

Note: Standard errors in brakets. Statistical significance levels * p<0.1 ** p<0.05 *** p<0.01

b. Intention to treat effects

Our estimates of the impact of the feedback and coaching intervention on student learning must account for the test score imbalance presented in the previous section. We perform several different analyses to address this.

First, we estimate intent-to-treat effects (ITT), using a parsimonious set of controls. The ITT effect is estimated from the equation below:

$$y_i = \beta_0 + \boldsymbol{X'_i}\beta_1 + \alpha_0 Z_i + \varepsilon_i \qquad (6)$$

where y_i is achievement for student i; $\boldsymbol{X_i}$ represents a vector of pre-intervention characteristics; Z_i is an indicator for whether the student was in a school that was offered participation in the intervention; and ε_i if the error term, clustered at the school level. The coefficient of interest is α_0.

Ideally, we would control for all students' previous achievement. However, prior year test scores are not available for all students, so we must use student achievement averaged at the school-grade level as one set of controls. We estimate (6) using five sets of control variables: the first specification does not include any controls, i.e., excludes $\boldsymbol{X_i}$ variables; the second includes controls for students' age, sex, race, the number of student per class and school average pass, failure and dropout rates for that grade in 2013 and 2014, which we denominate "controls 1"; the third specification includes "controls 1" plus the school-grade average for these outcome variables in 2013[18]; the fourth includes "control 1" plus school-grade averages for these outcome variables in 2014[19]; and the fifth includes all the controls.

Results are presented in Table 26. Outcome variables (y_i) are normalized relative to the distribution of the test score in the comparison group in each grade and year[20] and standard errors clustered at the school level are presented in brackets below each estimate. As we add controls for previous achievement averaged at the school-grade level, we see a drop in the magnitude of effect size but an increase in significance: results across specifications 3-5 are consistent and strongly statistically significant. Looking at specifications 3 to 5, we see positive and significant results in Math and Portuguese both for SPAECE 1st grade and ENEM 3rd grade. The magnitude of the effects is modest, however, ranging from 0.01 to 0.04 standard deviation (SD).

To take advantage of our three years of data at the school-grade level, our second strategy is to analyze a difference-in-differences model, comparing student achievement trends in our treatment and control groups in 2014-2015 minus the 2013-2014 trends. As Table 25 showed, the trends on all variables were balanced over 2013-2014.

[18] When analyzing results for SPAECE, we add the school-grade average of SPAECE on Math and Portuguese in 2013; and when analyzing results for ENEM, we add school-grade average for all areas of the test on 2013.

[19] When analyzing results for SPAECE, we add the school-grade average of SPAECE on Math and Portuguese in 2014; and when analyzing results for ENEM, we add school-grade average for all areas of the test on 2014.

[20] Scores are normalized for each grade and year such that the mean and standard deviation of the comparison group are zero and one, respectively. (We subtract the mean of the control group, and divide by the standard deviation.)

Table 26: Impact on student learning – Intent to Treat estimation - full sample of 348 schools

Dependent variable		(1)		(2)		(3)		(4)		(5)	
	Grade	N	Results	N	Results	N	Results	N	Results	N	Results
Spaece - Mathematics	1	59805	0.12** [0.05]	59805	0.09** [0.04]	59805	0.02** [0.01]	59805	0.03*** [0.01]	59805	0.02*** [0.01]
Spaece - Portuguese	1	59806	0.08* [0.05]	59806	0.06* [0.04]	59806	-0.00 [0.01]	59806	0.01* [0.01]	59806	0.00 [0.01]
ENEM - Portuguese	3	45838	0.07* [0.04]	45838	0.05 [0.03]	45838	0.02** [0.01]	45838	0.04*** [0.01]	45838	0.03*** [0.01]
ENEM - Mathematics	3	45838	0.06 [0.04]	45838	0.04 [0.03]	45838	0.01 [0.01]	45838	0.02** [0.01]	45838	0.02* [0.01]
Controls 1*					x		x		x		x
2013 school-grade average test score							x				x
2014 school-grade average test score									x		x

Note: Standard errors in brakets, clustered at the school level . Statistical significance levels * p<0.1 ** p<0.05 *** p<0.01

* Controls 1: age, female, black, brown, students per class, 2013 and 2014 school-grade average pass, failure and drop out rate.

We estimate the following model:

$$y_{i,t} = \beta_0 + \beta_1 T_t + \ \beta_2 Z_i + \alpha_0 T_t Z_i + \varepsilon_i \qquad (7)$$

where y_i is student achievement averaged at the school-grade i; $\boldsymbol{T_t}$ is a dummy variable for the period, which assumes a zero value for the 2013-2014 trend and one for the 2015-2014 trend; Z_i is an indicator for whether the school was offered participation in the intervention; and ε_i if the error term, clustered at the school level. In this model, the average pre-treatment trend across the units serves as the counterfactual for the average post-treatment trend. The effect is estimated by comparing the average pre-treatment trend to the average post-treatment trend, given by the coefficient of interest α_0.

Table 27: Impact on student learning, difference-in-differences estimation - full sample of 348 schools

		ITT - No controls		Dif-Dif	
Dependent variable		(1)		(2)	
	Grade	N	Results	N	Results
% Students who took Spaece	1	348	0.02 [0.01]	697	-0.01 [0.01]
Spaece - Mathematics	1	348	0.13** [0.06]	697	-0.04 [0.03]
Spaece - Portuguese	1	348	0.08 [0.05]	697	-0.03 [0.03]
% Students who took ENEM	3	348	0.00 [0.01]	697	-0.04*** [0.02]
ENEM - Portuguese	3	348	0.06* [0.03]	697	0.10*** [0.03]
ENEM - Mathematics	3	348	0.06* [0.04]	697	0.07*** [0.03]

Note: Standard errors in brakets. Statistical significance levels * p<0.1 ** p<0.05 *** p<0.01

* Controls: enrollment, school-grade average for age, female, black, brown and students per class

Table 27 shows results for the school-grade level analysis. The first column presents a simple ITT result, with no controls. The second column presents results for the difference-in-differences analysis. Results from the difference-in-differences analysis show no impact on SPAECE scores, but positive effects on ENEM scores, of 0.1 SD in Portuguese and 0.07 SD in Math.

Our last strategy to account for the imbalance in our sample is to use student-level data to control for previous achievement in t-1 for the students we can find in both 2015 and 2014. While we lack data for all students, we are fortunate to be able to match two large sub-samples of students, which give us very strong controls:

(a) Students who took SPAECE in 2015 in the 1st year of secondary school and in 2014 while in 9th grade;

(b) Students who took ENEM in 3rd grade in 2015 and in 2nd grade in 2014.

For each of these sub-samples, we run five specifications:

(i) To analyze how each sub-sample compares with the total sample, we repeat the ITT results from table 26 (for the full sample, with no controls);

(ii) We then show ITT results for the sub-sample, with no controls;

(iii) Next, we add a control for students' previous achievement in t-1;

(iv) In addition to prior test scores, we include student demographic and background controls from the SPAECE survey (sex, age, race, parents' education, domicile on paved street, and recipient of *Bolsa Familia* cash transfer);

(v) To check the robustness of the results from specification (iv), we run a propensity score analysis. We first calculate the propensity score including all covariates used as controls in specification (iv) (previous achievement, sex, age, race, parents' education, paved street and cash transfer). We then use the method of nearest neighbor matching to match each treatment student to one control student, based on his or her propensity score. Finally, we calculate ITT results.[21]

Table 28 shows the results of specifications (i) to (v) for each sub-sample (a) and (b). Annex tables A17 and A18 show the balance for each of the subsamples.

Table 28: Impact on student learning, controlling for prior achievement- Full sample of 348 schools

Dependent Variable	(1)	(2)	(3)	(4)	(5)
SPAECE 1st Grade					
Spaece - Mathematics	0.124**	0.130**	0.069**	0.071**	0.079***
	[0.054]	[0.054]	[0.029]	[0.029]	[0.010]
Spaece - Portuguese	0.083*	0.087*	0.036	0.037*	0.055***
	[0.048]	[0.046]	[0.023]	[0.022]	[0.010]
N	59806	41817	41817	41817	41817
ENEM 3rd Grade (controling for ENEM t-1)					
ENEM - Portuguese	0.065*	0.064	0.045*	0.045*	0.038***
	[0.039]	[0.041]	[0.027]	[0.024]	[0.013]
ENEM - Mathematics	0.056	0.059	0.039	0.040	0.046***
	[0.037]	[0.041]	[0.027]	[0.025]	[0.014]
N	45838	27958	27958	27958	27958
Restricted sample		x	x	x	x
Control Spaece			x	x	
Previous achievement				x	
Matched sample					x

Note: Standard errors in brakets, clustered at the school level . Statistical significance levels * p<0.1 ** p<0.05 *** p<0.01. Control Spaece includes control for students' sex, age, race, parental education, paved street and cash transfer.

Overall results in Table 28 look robust, with controls. Estimated impacts on student learning are positive and significant for SPAECE 1st year scores. The magnitudes are quite similar from specifications 3 through 5: for Math, they range from 0.069 to 0.079 SD; for Portuguese, they range from 0.037 to 0.055 SD. We also have positive and significant results on the ENEM test from specifications 3-5 for Portuguese, ranging from 0.038 to 0.045 SD; and specification (5) also shows a positive and significant result for Math (0.046 SD).

Overall, we find impacts on student learning that are positive and strongly significant. Results range from 0.01-0.04 SD in the ITT analysis with student level data, using the full sample and including all controls, to 0.07-0.10 SD with school-grade level data, and 0.037-0.079 SD on the ITT for the sub-sample of students with prior test score data.

Thus far, we have analyzed results for the full initial randomized sample of 348 schools. However, 18 treatment schools were not observed at baseline and therefore did not receive the "information shock" of the school feedback bulletin. In discussions at the end of the program, regional supervisors, school directors and pedagogical coordinators reported that the "information shock" had provoked considerable discussion in the treatment schools. Although individual teachers were not identified, school personnel said that teachers and administrators could "figure out" who's results were whose. Coordinators believed the bulletin generated pressures to improve among schools with poor results and among teachers with the weakest classroom performance within schools.

Table 29: Impact on student learning controlling for prior student achievement- (292 school sample)

Dependent Variable	(1)	(2)	(3)	(4)	(5)
SPAECE 1st Grade					
Spaece - Mathematics	0.148**	0.148**	0.078**	0.081**	0.080***
	[0.059]	[0.059]	[0.032]	[0.032]	[0.011]
Spaece - Portuguese	0.105**	0.104**	0.044*	0.047*	0.055***
	[0.051]	[0.051]	[0.025]	[0.024]	[0.011]
N	50463	36389	36389	36389	36389
ENEM 3rd Grade (controling for ENEM t-1)					
ENEM - Portuguese	0.087**	0.087**	0.055**	0.055**	0.056***
	[0.039]	[0.042]	[0.026]	[0.024]	[0.014]
ENEM - Mathematics	0.068*	0.071	0.039	0.041	0.038**
	[0.039]	[0.043]	[0.028]	[0.027]	[0.015]
N	38985	24346	24346	24346	24346
Restricted sample		x	x	x	x
Control Spaece			x	x	
Previous achievement				x	
Matched sample					x

Note: Standard errors in brakets, clustered at the school level . Statistical significance levels * p<0.1 ** p<0.05 *** p<0.01. Control Spaece includes control for students' sex, age, race, parental education, paved street and cash transfer.

[21] We use the program *teffects psmatch* in Stata to calculate ITT results. The command takes into account the fact that propensity scores are estimated rather than known when calculating standard errors.

We explored whether the information shock contributed incremental impact to the program. Table 29 presents results for the 292 sub-sample that includes only the 156 schools that received the complete treatment. We can see that results are quite similar, but slightly higher. On SPAECE, results for Math range from 0.078 to 0.081 across specifications 3 to 5; for Portuguese, they range from 0.044 to 0.055. For the sub-sample of students who took ENEM in 2015 and 2014, results for Portuguese range from 0.055 to 0.056, and for Math the effect is 0.038 in specification 5.

These higher results reinforce the importance of the feedback bulletin. In the subsequent analysis, we focus on the 292-school sample.

c. Intent to treat - heterogeneous effects

The design of this program was inspired by evidence from large-scale classroom observations in Brazil and six other Latin American countries showing that average classroom practice varies tremendously across schools and across teachers within a school. The program's goals were to help schools with the weakest average practice improve and to help the weaker teachers within schools to improve. Thus, interesting questions are whether the program differentially impacted teachers and schools at different levels of baseline performance – both in terms of classroom management and student learning.

We used quartile analysis to test two main hypotheses: that the program might have strongest effects on schools with: i) the lowest student learning performance at baseline; and ii) the worst classroom dynamics at baseline. Our complete results are included in Annex tables A19 to A23.

We did not find evidence that schools in the bottom quartiles of the learning distribution improved more (Annex Table A19). Although the logic of the teacher feedback and coaching program is that improvements in teacher practice can raise student learning, this program did not focus directly on the sources of low learning performance in schools or provide tailored strategies for improving learning in those contexts. It could be that schools with persistently low learning outcomes need more intensive interventions, and/or more sustained support.

However, we did find differential impacts on schools stratified by their classroom dynamics at baseline. The strongest finding is that learning gains were highest for schools with the worst performance on classroom management at baseline (i.e., top quartile of this distribution). Schools in the top quartile *averaged* a very high 34% percent of total class time on classroom management, compared with the good practice indicator of 15% and the sample average of 24%. As Table 30 shows, schools in this top quartile of this distribution at baseline saw very significant increases on SPAECE of 0.17 SD in Math and 0.12 SD in Portuguese on SPAECE and on ENEM of 0.14 SD in Math and 0.15 SD in Portuguese as a result of the program.

Table 30: Heterogeneous Impact of Program on Schools, by Share of Time on Classroom Management at Baseline (292 school sample)

Dependent Variable	(Quartile 1)	(Quartile 2)	(Quartile 3)	(Quartile 4)	p-value
	0.00- 0.19	0.19- 0.23	0.23- 0.28	0.28- 0.49	
SPAECE 1st Grade					
Spaece - Mathematics	0.007	0.035	0.057	0.167***	0.341
	[0.097]	[0.055]	[0.060]	[0.061]	
Spaece - Portuguese	0.060	-0.042	0.022	0.117***	0.064
	[0.066]	[0.040]	[0.051]	[0.045]	
N	7443	9777	9396	9773	
ENEM 3rd Grade (controling for ENEM t-1)					
ENEM - Portuguese	0.045	-0.016	0.024	0.146***	0.098
	[0.053]	[0.048]	[0.043]	[0.047]	
ENEM - Mathematics	0.014	-0.035	0.025	0.137***	0.103
	[0.056]	[0.053]	[0.057]	[0.049]	
N	5220	6829	6048	6249	

Note: All specifications include student controls for age, sex, race, parental education, paved street, cash transfer and previous achievment. Standard errors in brakets, clustered at the school level . P-value tests if Quartile 1=Quartile 2=Quartile 3=Quartile 4. Statistical significance levels * p<0.1 ** p<0.05 *** p<0.01.

Even though there is an arithmetic inverse relationship between the average time spent on classroom management and time off-task and the average time spent on instruction, we did not find equally strong impacts on schools stratified by these other, related, variables. Schools in the second-to-bottom quartile of the distribution on the share of time "off-task" at baseline seem to have improved the most, and schools in the third quartile on the share of time on instruction at baseline improved seems to have improved most, but these impacts are not significantly different from the other quartiles (Annex Tables A20 and A21).

Given the program's goal of stimulating more consistent teacher practice within schools, we also asked whether the program had stronger impacts in schools with larger variance on time on instruction at baseline, but we did not find evidence to support this (Annex table A22). Finally, we looked at schools with the highest indices of a big group of students off-task (Annex table A23). Although we found significant impacts on SPAECE math scores for schools in the middle of the distribution, they are not significantly different from the rest of the distribution.

We also explored whether teachers' age made a difference, thinking that new teachers might have the most to gain from this program. As seen in Table 31, the youngest quartile of teachers (age 22-31) did not register statistically significant learning gains in their classrooms, but teachers in the second quartile (age 31-35) did achieve big gains in ENEM scores and SPAECE math scores (this heterogeneous effect is significant for the ENEM scores). This is consistent with evidence from other studies that teacher effectiveness improves dramatically after the first five years of experience, suggesting that in this case that teachers who are still young but have accumulated a critical level of experience may be best positioned to benefit from this type of program.

Table 31: Heterogeneous Impact of Program on Teachers, by Age (math and Portuguese teachers only)

Dependent Variable	(Quartile 1)	(Quartile 2)	(Quartile 3)	(Quartile 4)	p-value
	22.00- 31.00	31.33- 35.50	35.67- 40.50	40.67- 64.00	
SPAECE 1st Grade					
Spaece - Mathematics	0.055	0.098*	0.087*	0.080	0.924
	[0.052]	[0.051]	[0.047]	[0.053]	
Spaece - Portuguese	0.034	0.043	0.055	0.056	0.968
	[0.034]	[0.038]	[0.040]	[0.040]	
N	10699	10204	8441	7045	
ENEM 3rd Grade (controling for ENEM t-1)					
ENEM - Portuguese	0.032	0.161***	0.011	0.020	0.023
	[0.040]	[0.040]	[0.043]	[0.042]	
ENEM - Mathematics	0.035	0.154***	-0.017	-0.009	0.02
	[0.048]	[0.044]	[0.043]	[0.045]	
N	7815	6049	5828	4654	

Note: All specifications include student controls for age, sex, race, parental education, paved street, cash transfer and previous achievment. Standard errors in brakets, clustered at the school level . P-value tests if Quartile 1=Quartile 2=Quartile 3=Quartile 4. Statistical significance levels * p<0.1 ** p<0.05 *** p<0.01.

d. *Compliance Effects*

As discussed in Section 3.f, implementation of the teacher feedback and coaching intervention depended importantly on the engagement of the pedagogical coordinator in each school, as these were interface between teachers and the ELOS expert coaches. To generate a quantitative measure of school engagement, the coordinators and school directors maintained an online registry on the number of Skype conference calls, teacher observations, and feedback meetings held each week. For a qualitative measure, the ELOS team administered a final certification test to all coordinators who had participated in at least 80% of all online and face-to-face activities.

We use these qualitative measures to estimate the impact of program compliance on student learning outcomes, parallel to our methodology for estimating its impact on classroom dynamics. Models (4) and (5) (section 3.f) use three benchmarks as a proxy measure of the intensity of pedagogical coordinators' engagement with the program and likely capacity to support teachers in mastering the new teaching techniques.

(i) *certification* - the pedagogical coordinator achieved the minimal score on the final test necessary for certification (e.g., adequate (regular), good or excellent) – 138 coordinators achieved this

(ii) *good or excellent* - 108 coordinators achieved this; and

(iii) *excellent* – 49 coordinators achieved this

Results are shown on Table 32[22]. As expected, we find higher magnitudes on the 2SLS when compared to ITT analysis. The results suggest strongly that coordinators' mastery of the content of the program enhances their ability to support teachers in adopting the new practices and that these practices, if adopted faithfully, can have significant impact on student learning. On both SPAECE and ENEM, the size of the positive impact on student scores rises in parallel with coordinator qualifications. For the 49 schools whose coordinators achieved excellence, the learning impacts are quite strong – 0.232 SD on SPAECE math, 0.134 on SPAECE Portuguese, and 0.175 on ENEM Portuguese. The only result that is not statistically significant is on ENEM math.

Table 32: 2SLS estimates of the effect on student learning (292 school sample)

[22] 1st stage estimates are shown in Annex table A24.

Dependent Variable	Certificates	Score: excellent or good	Score: excellent
SPAECE 1st Grade			
Spaece - Mathematics	0.091***	0.118***	0.232**
	[0.035]	[0.046]	[0.092]
Spaece - Portuguese	0.052*	0.068*	0.134*
	[0.027]	[0.035]	[0.070]
N	36389	36389	36389
ENEM 3rd Grade (controling for ENEM t-1)			
ENEM - Portuguese	0.060**	0.080**	0.175**
	[0.026]	[0.035]	[0.084]
ENEM - Mathematics	0.045	0.060	0.130
	[0.029]	[0.039]	[0.089]
N	24346	24346	24346
Control Spaece	x	x	x
Previous achievement	x	x	x

Note: Sample of 292 schools. Standard errors in brakets, clustered at the school level . Statistical significance levels * p<0.1 ** p<0.05 *** p<0.01. Control Spaece includes control for students' sex, age, race, parental education, paved street and cash transfer.

These results are encouraging. They suggest that the ELOS certification test is a meaningful measure of the capacity of pedagogical coordinators to implement the coaching program effectively. They also demonstrate that the program has the potential to raise student learning significantly if schools achieve high fidelity implementation. These factors provide useful guidance to school systems considering this type of intervention.

6. Cost-Effectiveness

The teacher feedback and coaching program was designed to be low-cost and scalable. Table 33 shows costs for all key elements of the program and the evaluation. The largest program cost elements were logistics costs, both for the school visits for the baseline classroom observations, which provided the basis for the information shock to schools, and for the four one-day face-to-face training sessions that the ELOS team carried out for pedagogical coordinators and school directors, each of which was delivered in three different locations across the state. Even though the same number of schools was observed at both baseline and endline, the first round costs were twice as high, because a consulting firm was recruited to plan the logistics of school visits and supply transport, and the observations used paper coding sheets, which added costs to data recovery and digitization. For the second round, the same field plan was re-used, observers used their own cars, and observations were made on tablets that were property of the Secretariat.

Table 33: Costs of Ceará Teacher Feedback and Coaching Program

Cost Element	R$	US$	R$/ Student	US$/ Student
Program Costs				
Planning activities	64000	16000	0.52	0.13
Information shock:				
Classroom observations in 156 treatment schools	363287.67	90821.92	2.95	0.74
Preparation of schools bulletins	16000	4000	0.13	0.03
Coaching				
Aula nota 10 book for 174 schools	117000	29250	0.95	0.24
Logistics of face-to-face training sessions (Transport, lodging, subsistence for 400 participants)	152000	38000	1.24	0.31
ELOS training team (110 hours of training and coaching support by Skype)	468000	117000	3.80	0.95
Subtotal	**1180287.67**	**295071.92**	**9.60**	**2.40**
Evaluation Costs				
Planning activities and analysis	140000	35000		
Baseline data collection, Classroom observations in 136 control schools in Nov 2014	316712.3	79178.1		
Endline data collection, Classroom observations in 292 schools in Nov 2015	320000	80000		
Subtotal	**776712.3**	**194178**		
GRAND TOTAL	**1957000**	**489250**		

*Note: The number of students in the treatment schools at the beginning of the intervention was 123000 and the exchange rate was R$ /US$ = 4.0.

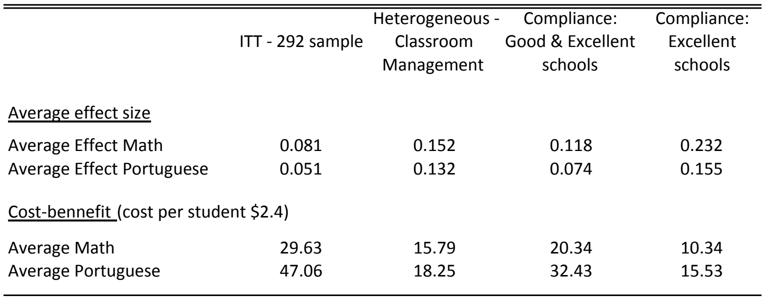
Other costs were the time of the coaches, and the books given to schools' pedagogical coordinators. Skype communications costs were minimal; all participating schools either had functioning internet and computers at school, or the pedagogical coordinator had a computer and internet access at home. Overall, the program cost R$ 1,180,287 (US$ 295,072), or R$ 9.60 (US$ 2.40) per student in the treatment schools.

We did not cost the time that teachers and pedagogical coordinators spent working together within schools, as a new federal government policy mandates that teachers spend part of their existing contract hours on collaborative planning.

The main evaluation cost was the cost of classroom observations at baseline and endline in the 138 control schools. Total evaluation costs were R$ 776,712 (US$ 194,178). The World Bank's SIEF program financed the evaluation plus part of the costs of the baseline round of classroom observations, for which a consulting firm supplied logistical support. All program costs were financed by the Secretariat and the Lemann Foundation.

Table 34 repeats our learning impacts for the 292 sample (specification 4, Table 29), which we believe to be our most robust analysis: .081 for Math and 0.051 for Portuguese, across both the SPAECE and ENEM tests[23]. For purposes of comparing the cost-effectiveness of this program to other evaluated programs, we follow JPAL's method of estimating the cost per student of increasing learning outcomes by 1 standard deviation. Although it is unlikely that any single intervention in education will increase test scores by a standard deviation, this comparison is a useful guide for policy makers considering different programs for the achievement of a policy goal.

Table 34: Effect Sizes and Cost-Effectiveness Estimates for Ceará Teacher Feedback and Coaching Program



	ITT - 292 sample	Heterogeneous - Classroom Management	Compliance: Good & Excellent schools	Compliance: Excellent schools
Average effect size				
Average Effect Math	0.081	0.152	0.118	0.232
Average Effect Portuguese	0.051	0.132	0.074	0.155
Cost-bennefit (cost per student $2.4)				
Average Math	29.63	15.79	20.34	10.34
Average Portuguese	47.06	18.25	32.43	15.53

Our average treatment effect and unit cost of $2.40 per student result in estimates that a 1 SD increase in test performance would cost $29.63 per student for math and $47.06 per student for Portuguese. As shown in Figure 11, these estimates place this program towards the middle of the distribution of programs by reported cost-effectiveness.

Given our finding that the learning impacts were strongest for classrooms in the worst-quartile of classroom management at baseline, we also present the effect sizes for this group. These program impacts are much more cost-effective, with 1 SD increase in math costing $15.79 and in Portuguese $18.25. An implication is that targeting the feedback and coaching program to classrooms with poor instructional time management could be a cost-effective strategy.

[23] When results were significant for both, math and Portuguese score, we calculated the average between the two. Otherwise, we used the significant result (as it is the case for Math, where results are only significant for the SPAECE test).

Finally, we present the substantially higher learning gains for schools whose pedagogical coordinators demonstrated strongest mastery of the program content: the roughly two-thirds of schools whose coordinators scored “good or excellent” and the one-third of schools whose coordinators scored “excellent”. For the latter group, the results in the range of $10-15 are among the most cost-effective results yet reported in the literature.

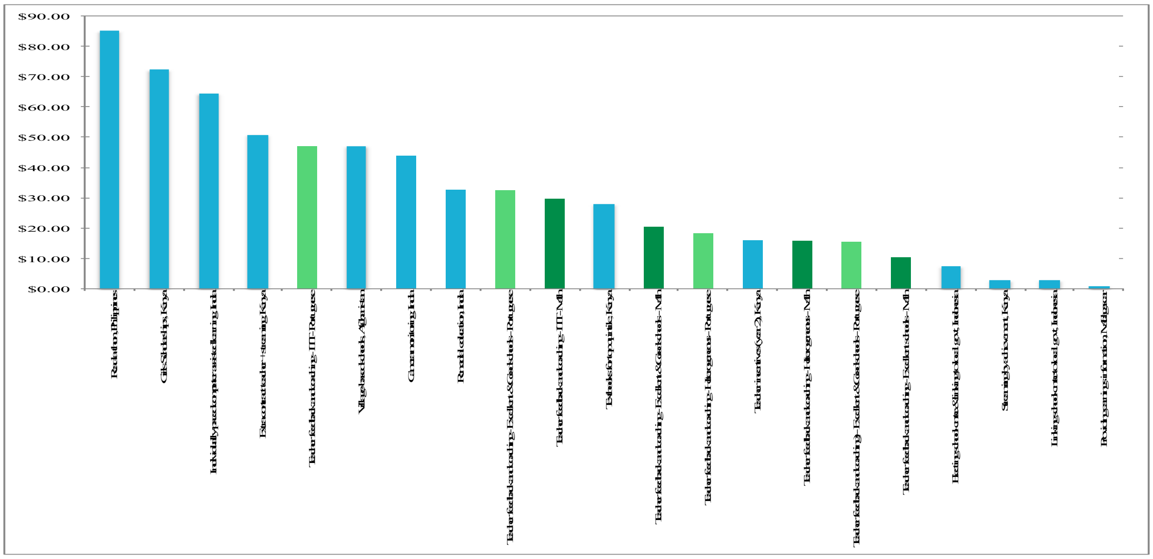
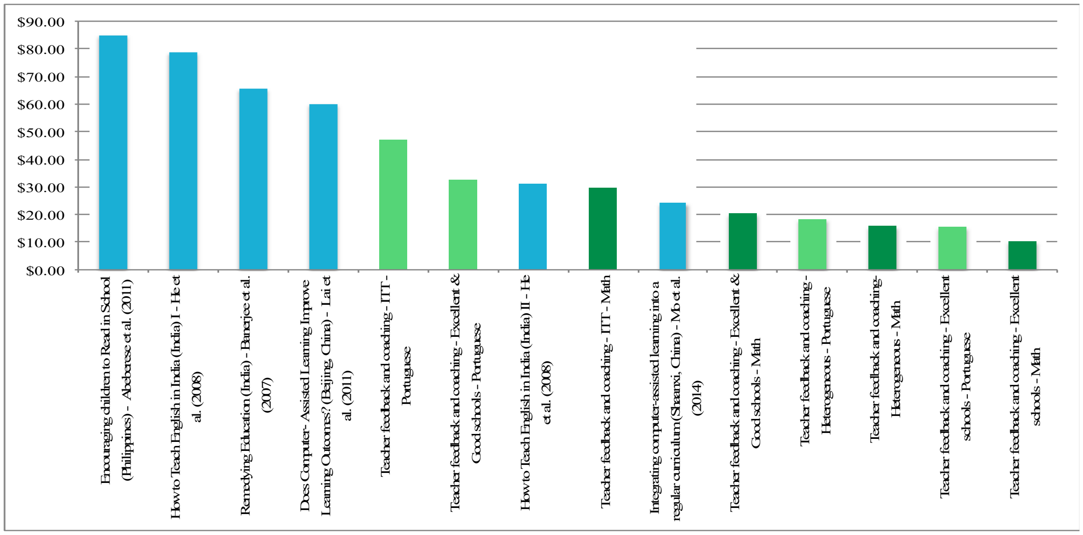


Figure 11: Comparative cost-effectiveness of Ceará Teacher Feedback and Coaching Program (Education programs evaluated by JPAL)

The JPAL evaluations cover a wide range of education interventions, so we looked for analyses that would situate our results within the narrower range of evaluated programs that focus on teacher training. Although literally thousands of different approaches exist, and hundreds of millions of dollars are spent by developing countries and donors on teacher development every year, the number of programs that have been rigorously evaluated is miniscule.[24] Among these, there is a huge publication bias; programs with negative or negligible impacts are almost never reported. And among programs with positive impacts, cost data not always reported.

Figure 12 compares our results to the set of evaluated interventions identified by Popova et al. in a comprehensive 2016 study. Out of 26 programs (23 papers) only 6 reported gains in test scores and the cost



[24] In a U.S.-focused review, Yoon et al. (2007) report that out of 1,300 studies identified as potentially addressing the effect of teacher professional development on student achievement, only 9 were identified as having pre- and post-test data and some sort of control group. Similarly, a 2014 review found 643 studies of math professional development interventions for K-12 teachers, but only 32 used any research design to measure learning impacts, and only 5 were high-quality randomized trials, and only 2 of these found positive results (Gersen et al. 2014).

of the program per student. (Appendix Table A26 provides details on the papers). In this context, the results for Ceará are extremely impressive. Only two evaluated programs show greater cost-effectiveness than our ITT results for Ceará. And *no* program evaluated to date has had more cost-effective impact on student learning than the Ceará program had in schools with high fidelity implementation.

The Ceará program has important advantages over many other models of in-service teacher development. First, it has lower unit costs, as it avoids the logistics expenses of off-site programs and makes efficient use of IT (video analysis and Skype interactions) to connect schools with expert coaches. Second, it leverages skills already present in schools. Third, by building the expertise of pedagogical coordinators and stimulating interaction among teachers, the program creates a platform for gradual and continuous reinforcement of improved teaching techniques. Given the reality that changing adult behavior is difficult, this may prove to be a necessary condition for improving teacher practice.

7. Conclusions

Middle-income developing countries in Latin America such as Brazil are investing heavily in education and adopting major education reforms, particularly aimed at raising teacher quality. Results of the 2015 PISA exam show that several countries in the region, such as Chile, Colombia and Peru, are making steady progress, but, overall, the eight Latin American countries (plus the City of Buenos Aires) that participated in the test remain in the bottom half of the 72-country sample. Among LAC countries, Brazil faces particularly big challenges. Brazilian 15-year-olds performed third-to-last (ahead of only Peru and Dominican Republic) in reading and science and second to last (ahead of only the Dominican Republic) in math. The Brazilian math score, 113 points below the OECD average, implies a lag almost three full years in math skills vis-a-vis the OECD and more than four years behind the top performing East Asian countries, Singapore, China and Japan.

A compelling body of global evidence shows that teachers' effectiveness is the key in-school determinant of student learning and Brazil, like other countries, is looking for strategies to raise teacher effectiveness. Documented issues in Brazil are the low academic caliber of entering teachers and the prevalence of ineffective classroom practice. On previous PISA tests, Brazilian 15-year-olds who described themselves as future teachers scored 50 points below the national average and 100 points below future engineers in math; on University of São Paulo entrance exams, the highest scoring teacher-education candidates perform below the lowest-scoring medical school entrants. (Bruns and Luque, 2014). Classroom observation research in several different states and municipalities in Brazil (Bruns and Luque, 2014) has evinced the same issues seen in Ceará: teachers' failure to use class time effectively, heavy reliance on traditional "chalk and talk" teaching methods, and inability to keep students engaged. While deep weaknesses in teachers' content mastery may be not be amenable to short-term improvement through in-service training, the Ceará experiment demonstrates that teachers' classroom practice is malleable, with potentially important impacts on student learning.

The design of the program was inspired by the research evidence, both from classroom observations of teacher practice and research on teacher value-added, of large variations in teacher quality *within* schools. Leveraging the teaching skills that exist within schools by promoting greater collaboration and exchange of practice among teachers offers a low-cost strategy for raising teachers' effectiveness. The Ceará program had four elements: an "information shock", giving teachers benchmarked feedback about their practice based on classroom observations using the Stallings method; three face-to-face orientation sessions for school directors and pedagogical coordinators with a high-skill team of trainers; ongoing Skype interactions throughout the 2015 school year with the training team; and self-help materials, notably the Portuguese language version (*Aula Nota 10*) of the book by US educator Douglas Lemov, *Teach Like a Champion*.

The information shock was intended to show schools they had room for improvement as well as to identify some of the individual teachers (identified by the subject and hour of the class they taught rather than by name) who managed class time most effectively, used interactive (question and answer) teaching practices, and kept students engaged. The coaching program aimed at turning the pedagogical coordinator in each school into a stronger resource for school improvement, by developing her ability to observe teachers' classroom practice and provide useful feedback, and to promote collaboration and exchange of practice among teachers.

To assess program impact rigorously, 175 treatment and 175 control schools were randomly drawn from a representative sample of 400 schools that was stratified by geographic region, school size, and quartile of learning results. A shortage of observers during the baseline data collection reduced the final number of schools observed to 156 treatment and 136 control schools, and one school from each group closed between 2014 and 2015. Despite the uneven attrition, school, student and teacher demographic and background characteristics, as well as baseline classroom observations showed that the final treatment and control groups were balanced on observables. A set of additional robustness tests described in Section 4 provided reassurance that bias in the sample was unlikely.

Monitoring data show that pedagogical coordinators in the program schools *did* increase the amount of time they spent observing teachers and giving them feedback. At baseline, coordinators reported that they did not do this routinely; reports compiled by the coaches shows that all of the pedagogical coordinators in the program conducted at least 3 observations and 3 feedback sessions with every teacher in the school. A test applied at endline showed that 88% of the pedagogical coordinators had a good understanding of the importance of maximizing instructional time, as well as specific techniques for planning effectively paced lessons and keeping students engaged, such as "cold calling".

The feedback and coaching program produced a statistically significant .25 SD increase in time on instruction. Program schools' teachers increased time on instruction to 76% of each class, compared with 70% in control schools. This implies 15 more minutes of instruction across five classes per day and 50 more hours – two additional weeks -- of learning time per year. Differences of this magnitude, all other things equal, can be expected to have consequences for student learning.

Teachers in the program schools freed up time for instruction by reducing the time they spent on routine classroom administrative processes (taking attendance, cleaning the blackboard, passing out papers) and especially by reducing their time off-task. Time spent on classroom management fell to 18% of class time in program schools, compared with 21% in control schools, a -.17 SD change. Time off-task fell from 8.6% to 6.2% of total time, a -.19 SD larger decline than in control schools. The biggest driver was less time absent from the classroom.

Teachers in the program schools also increased their use of questions during their lessons, consistent with the coaching program's goal of encouraging more interactive teaching practice, although lecture/demonstration continued to be the dominant teaching mode. They also kept students more engaged. Program schools reduced from 19% to 16% (a -.11 SD larger decline) the share of time that a large group of students (six or more) was visibly off-task while the teacher was teaching. The only dimension in which treatment schools' improvement was not statistically significant was the share of time on instruction with all students engaged. Although the data show that a few program schools achieved some impressive gains in this indicator (see box plot Figure 5), many schools, both treatment and control, continued to have the entire class engaged, on average, less than 20% of the time teachers are teaching.

Positive changes in teacher practice were slightly more pronounced for the 75% of classrooms where the November 2015 repeat observation "matched" – in terms of subject, grade and time of day -- the November 2014 one. Because teachers remained anonymous, we cannot guarantee that the same teacher was observed both times, but Secretariat officials expect this case for a high share of the classrooms. Teachers in the matched subsample showed a .26 SD increase in time on instruction vis a vis control schools, a -.19 SD reduction in classroom management, and a -.19 SD decline in time with a big group of students off task.

Finally, consistent with the core goal of getting teachers within the school to learn from each other, the program reduced the variation in teacher practices within schools. Compared with the control schools, the program schools saw a -0.016 decline in the variation in time on instruction, a -0.012 and -0.018 decline in the variation in classroom management and teacher off-task activities, respectively, meaning that teachers within schools began achieving more consistent practice.

The results suggest that providing schools with concrete, benchmarked feedback about their teaching practice plus access to high quality coaching support can produce significant improvements in teachers' time on instruction and ability to keep students engaged over the course of just one school year. The program helped schools achieve more consistent teacher practice and increased teachers' use of more interactive pedagogical techniques, such as question and answer. This evaluation is the first developing country study to generate rigorous evidence on the impact of a teacher development program aimed at improving teachers' classroom practice. Prior to this experiment, it was unknown how much variation in the measures of classroom dynamics captured by the Stallings instrument was even possible over the course of a single school year.

The impacts the program had on student learning are evidence that the observed changes in teacher practice were not all evaluation-driven. It is implausible that student test scores could increase as the result of one day during the school year when teachers were observed and altered their practice.

In the 292-school sample that received the full treatment, controlling for prior student achievement, the program produced increases of 0.08 SD in math and 0.55 SD in Portuguese on the state assessment and 0.04 SD in Math and 0.06 SD in Portuguese on the national test, ENEM. These gains are statistically significant, as we control for students' prior test scores. Program impacts were strongest in the quartile of classrooms with the weakest classroom dynamics (highest share of class time on classroom management, rather than instruction) at baseline (0.117-0.167 SD).

Implementation fidelity strongly affected program results. Within schools, the pedagogical coordinator played a pivotal role as the interface between teachers and the expert coaches. Across both tests and all subjects, student learning impacts were systematically, and substantially, higher in proportion to the performance of the pedagogical coordinator on the final certification exam, which we take as a proxy for the coordinator's engagement with the program and mastery of its content. In schools where coordinators achieved the basic level of certification, math and Portuguese scores on state and national tests were 0.05-0.09 SD higher; where coordinators achieved "good or excellent" certification, scores were 0.07-0.12 SD higher, and where coordinators scored in the top category of "excellent", scores were 0.13-0.23 SD higher. These are significant impacts over the course of a single school year. Given the low program cost of $2.40 per student, the learning results for the one-third of schools with "excellent" implementation are the most cost-effective yet reported in the literature for a rigorously evaluated teacher development program. These impacts raise the possibility of even larger cumulative effects if teachers' use of new and more effective classroom practices is reinforced and refined over time and professional interaction in these schools becomes a habit. Whether this outcome is more likely than fade-out of the new practices is something that requires further research. But the one-year results strongly suggest that well-implemented programs to raise teachers' classroom effectiveness through observation feedback and coaching are a promising strategy for "whole school" improvement.

References

Abadzi, H. (2009). Instructional Time Loss in Developing Countries: Concepts, Measurement, and Implications. World Bank Research Observer, 24(2).

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of Causal Effects Using Instrumental Variables," Journal of the American Statistical Association, 91, 444{455.

Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2002). Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment. American Economic Review, 92(5), 1535-1558.

Araujo, M. C., Carneiro, P. M., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher Quality and Learning Outcomes in Kindergarten. IDB working papers.

Bruns, B., & Luque, J. (2014). Great teachers: How to raise student learning in Latin America and the Caribbean. World Bank Publications.

Chetty, R, J. N. Friedman, and J. E. Rockoff. (2014). "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." American Economic Review, vol. 2014, nº 9, p. 2593-2632.

DeStefano, J., E. Adelman, and A.-M. Schuh Moore. (2010). Using Opportunity to Learn and Early Grade Reading Fluency to Measure School Effectiveness in Nepal. Washington, DC: EQUIP2, AED, and USAID.

Doig, B. and S. Groves. (2011). "Japanese Lesson Study: Teacher Professional Development through Communities of Inquiry." Mathematics Teacher Education and Development, vol. 13.1, p. 77–93.

Easton, L. B. (2008). From professional development to professional learning.Phi Delta Kappan, 89(10), 755.

Fryer Jr, R. G. (2013). Information and student achievement: evidence from a cellular phone experiment (No. w19113). National Bureau of Economic Research.

Fullan, M., N. Watson, and S. Anderson. (2013). Ceibal: Next Steps. Toronto: Michael Fullan Enterprises, http://www.ceibal.org.uy/docs/FULLAN-Ceibal-English.pdf.

Gersten, R., Taylor, M. J., Keys, T.D., Rolfhus, E., and Newman-Gonchar, R. (2014). Summary of research on the effectiveness of math professional development approaches. (REL 2014–010). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from http://ies.ed.gov/ncee/edlabs.

Glennerster, R., & Takavarasha, K. (2013). Running randomized evaluations: A practical guide. Princeton University Press.

Grossman, P., and S. Loeb, J. Cohen, K. Hammerness, J. Wyckoff, D. Boyd, and H. Lankford. (2010). "Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value Added Scores." NBER Working Paper 16015.

Hanushek, E., and S. Rivkin. (2010). "Using Value-Added Measures of Teacher Quality." Policy Brief 9, National Center for Analysis of Longitudinal Data in Education Research, Washington, DC.

Heckman, J. J. (1979), "Sample Selection Bias as a Specification Error," Econometrica, 47, 153-161.

Howes, C., and M. Burchinal, R. Pianta, R., D. Bryant, D. Early, R. M. Clifford, and O. Barbarin. (2008). "Ready to Learn? Children's pre-academic achievement in pre-kindergarten programs." Early Childhood Research Quarterly, 23, 17-50.

Jackson, C. Kirabo, J. Rockoff and D. O. Staiger. (2014). “Teacher Effects and Teacher-Related Policies.” Annual Review of Economics 6:34. 1-34.

Jennings, J. L., and T. A. DiPrete. (2010). “Teacher Effects on Social and Behavioral Skills in Early Elementary School.” Sociology of Education, April 2010 83: 135-159.

Jukes, M., S.B. Vagh, and Y.S. Kim. (2006). “Development of Assessments of Reading Ability and Classroom Practice”. Unpublished manuscript. World Bank, Washington, D.C.

Kane, T. J., J. E. Rockoff, and D. O. Staiger. (2008). “What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City.” Economics of Education Review 27 (6): 615–31.

Kane, T. J., and D. O. Staiger. (2008). “Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation.” Working Paper 14607, National Bureau of Economic Research, Cambridge, MA.

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

Lee, D. S. (2002). Trimming for Bounds on Treatment Effects with Missing Outcomes, Working Paper 51.

Lemov, D. (2010). Teach Like a Champion. San Francisco: Josey Bass.

Lemov, D. (2011). Aula Nota 10. Sao Paulo. Fundacao Lemann.

Mourshed, M., C. Chijioke, and M. Barber. (2011). How the World’s Most Improved School Systems Keep Getting Better. Londres: McKinsey.

Miguel, E., & Kremer, M. (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. Econometrica, 72(1), 159-217.

OECD (Organisation for Economic Co-operation and Development). (2005). Teachers Matter: Attracting, Developing and Retaining Effective Teachers. Paris: OECD Publishing.

———. (2009). “Teaching Practices, Teachers’ Beliefs and Attitudes.” In Creating Effective Teaching and Learning Environments: First Results from TALIS, 88–120. Paris: OECD Publishing.

———. (2010). Vol. 1 of PISA 2009 Results: What Students Know and Can Do—Student Performance in Reading, Mathematics and Science. Paris: OECD Publishing.

———. (2013a). Education at a Glance 2013: OECD Indicators. Paris: OECD Publishing. http://dx.doi.org/10.1787/eag-2013-en.

———. (2013b). Vol. I of PISA 2012 Results: What Students Know and Can Do—Student Performance in Mathematics, Reading and Science. Paris: OECD Publishing. http://www.oecd-ilibrary.org/education/pisa-2012-results-what-

Popova, Anna; Evans, David K. and Arancibia, Violeta. (2016) Training Teachers on the Job: What Works and How to Measure It. *Policy Research Working Papers*.

Rockoff, J. E. (2004). “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.” American Economic Review 94 (2): 247–52.

Schuh Moore, A.-M., J. DeStefano, and E. Adelman. (2010). Using Opportunity to Learn and Early Grade Reading Fluency to Measure School Effectiveness in Ethiopia, Guatemala, Honduras, and Nepal. Washington, DC: EQUIP2, AED, and USAID.

Stallings, J. A. (1977). Learning to look: A handbook on classroom observation and teaching models. Belmont, CA: Wadsworth Publishing.

Stallings, J. A., e Mohlman, G. G. (1990). Issues in qualitative evaluation research: Observation techniques. In H. J. Walberg & G. D. Haertel (Eds.), The international encyclopedia of educational evaluation (pp. 639-644). New York: Pergamon Press.

Liang, X. (2015). How Shanghai Does It: Scoring Highest in the Programme for International Student Assessment. World Bank.

World Bank. (2014). Conducting Classroom Observations Using the Stallings Classroom Snapshot Method: Manual and User Guide. Washington, DC: World Bank.

Yoon, K.S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). Reviewing the evidence on how teacher professional development affects student achievement (Issues & Answers Report, REL 2007–No. 033).

Yoshida, M. (1999). Lesson study: A case study of a Japanese approach to improving instruction through school-based teacher development. Doctoral Dissertation: University of Chicago.