CREDIT Research Paper

No. 17/08

# Training to teach science: experimental evidence from Argentina

by

Facundo Albornoz, Facundo Albornoz; María Victoria Anauati; Melina Furman; Mariana Luzuriaga; María Eugenia Podestá; Inés Taylor

## Abstract

This paper uses a RCT implemented in state schools in Argentina to estimate the learning impact and cost-effectiveness of different teacher training methods: structured curricula and coaching. Our findings suggest that there is a substantial gain in terms of learning for students with teachers being trained using structured curricula and coaching (between 55% and 64% of a standard deviation more than those students in the control group). Coaching teachers does not appear as a cost-effective intervention since the unit cost per 0.1 standard deviation is more than twice the cost of using a structured curriculum only. However, additional coaching is particularly relevant for relatively inexperienced teachers. A structured curriculum and coaching also affect perceptions: teachers enjoyed more teaching Science, they taught more hours of Science and students learned more and developed more skills.

**JEL Classification:** C93, I21, I28

**Keywords:** Science education; Teacher training; Experimental study

**Centre for Research in Economic Development and International Trade, University of Nottingham**

CREDIT Research Paper

No. 17/08

# Training to teach science: experimental evidence from Argentina

by

**Facundo Albornoz, Facundo Albornoz; María Victoria Anauati; Melina Furman; Mariana Luzuriaga; María Eugenia Podestá; Inés Taylor**

Outline

Facundo Albornoz (University of Nottingham and CONICET, facundo.albornoz@nottingham.ac.uk); María Victoria Anauati (University de San Andrés and CONICET); Melina Furman (Universidad de San Andrés and CONICET); Mariana Luzuriaga (University de San Andrés); María Eugenia Podestá (University de San Andrés) and Inés Taylor (University de San Andrés).

**Acknowledgements**

Funding for this study was provided by CIAESA (Centro de Investigación Aplicada en Educación San Andrés). We are very grateful to Samuel Berlinski, Guillermo Cruces, Alejandro Ganimian, Lucila Minvielle and Abhijeet Singh for comments and suggestions.

Research Papers at www.nottingham.ac.uk/economics/credit/

## 1. Introduction

Teacher training programs are ubiquitous across educational systems and constitute an essential tool to improve student learning and, thus, promote economic growth and development. Surprisingly, however, current approaches to teacher training are mainly uninformed by high quality evidence of their impact (Yoon, Duncan, Lee, Scarloss and Shapley, 2007). This is a serious issue especially because the different way a program can be designed and implemented involves substantial variation in costs. In Latin-American countries, for example, total investments in teacher training represent the major element of non-salary public spending in education, but there are no rigorous evaluations of their impact on learning (Bruns and Luque, 2015), let alone an evaluation of the cost-effectiveness of different designs to implement them. Thus, how to design cost-effective teacher training programs becomes one of the central questions of education policy. This paper provides experimental evidence on the impact and cost-effectiveness of different teacher professional development interventions on student Science learning from a specifically designed large scale study, implemented in state primary schools in the city of Buenos Aires, Argentina. While the experiment is specific to the instruction of Science in Argentina, our results may have broad relevance for other curricula and other contexts.

A typical training program consists of a short training session (Darling-Hammond, Wei, Andree, Richardson, and Orphanos, 2009). In our field experiment, we assess the marginal gain of complementing this basic training with two distinct teacher training models which provide different degrees of scaffolding. The first treatment is the provision of a structured curriculum unit (SC henceforth), which guides teachers in the organization, content and pedagogy of their lessons. The second treatment is supplementing the first two treatments (short training sessions plus SC unit) with weekly coaching. This allows us to study a comparison of a basic teacher training with no follow-ups with two distinct models with different degrees of support and associated costs.

More specifically, we report the main findings and associated policy lessons of a randomized controlled experiment designed to assess the effect of different working modalities with in-service teachers on student learning in Science. As primary education is considered of key importance to lay the foundations of scientific literacy (Näslund-Hadley and Bando, 2016; Novak, 2005), we focus on 7th grade –the last level of primary school in CABA. The study involves 70 schools that constitute a representative sample of CABA state primary schools. Although seeking to provide experimental evidence of teacher training in general, our focus in Science has interest in its own right. Over the last decades, many governments and international organizations have advocated Science, Technology, Engineering and Maths (STEM) subjects and degrees to promote economic growth in a context of highly technological and rapidly changing societies and jobs. The promotion of scientific literacy has also been emphasized by standardized international student assessment programs such as the Programme for International Student Assessment –PISA

(OECD, 2016).

The interest in our specific educational setting is easy to explain. Argentina, like the rest of Latin-American countries, is a perfect setting to study the effect of different strategies to train teachers in Science. Despite several government initiatives aimed at encouraging Science education (see e.g. Serra, 2001; Argentine Ministry of Education, 2007), the performance of Argentinian students in standardized assessments is still poor (UNESCO, 2016; Vegas, Ganimian and Bos, 2014). Even in CABA, the best performing Argentinian district, 41% percent of students only achieved the minimum level in Science, placing them as one of the lowest performing groups in the world (Martin et al., 2016; OECD, 2016).

The participating schools were randomly assigned to three groups. Teachers in the three groups received a short-term training. Besides being widely used in other countries, this approach to teacher professional development is also the most common one in Argentina (Argentine National Institute of Teacher Training, 2016). Teachers that only received this short-term training form the control group. However, the literature indicates that gains in student achievement are weak and they can only be observed in longer training interventions with ongoing support (Yoon, Duncan, Lee, Scarloss and Shapley, 2007).[1] In a recent review, Arancibia, Popova, and Evans (2016) conclude that only a few characteristics of teacher training programs, such as the inclusion of supplemental materials, follow-up visits, and focus on a specific subject, are positively associated with student test score gains. Our treatments include such characteristics.

In our second group (Sequence Group henceforth), teachers received the same short-term training but complemented with on-going support through the use of a structured curriculum unit, which guided teachers in the organization, content and pedagogy of a given topic. Research shows that well-designed structured curriculum units can enhance training sessions by providing concrete ways of taking the approaches learnt in training directly to the classroom and serving as catalysts for local customization (Brown, 2009). Developing curriculum units is a key strategy followed by the Argentinian education authorities as part of their efforts to improve teaching (Argentine Ministry of Education, 2017; Educ.ar, 2005). However, the literature also highlights challenges associated with the use of structured curriculum units. In some cases teachers adapt these units, making the lessons easier and more aligned to their regular practice, which lowers in turn their cognitive load (Davis, Janssen and Van Driel, 2016). Additionally, many factors may influence how and why teachers choose to adapt curriculum units, such as their previous teaching experience, knowledge, beliefs about science and education, among others (Arias, Davis, Marino, Kademian, and Palincsar, 2016; Forbes and Davis, 2010).

[1] Some studies even show that a minimum of 50 or even 80 hours of training and continuous post-training support are required to observe any result (Gulamhussein, 2013).

One way to bridge the gap between structured curriculum units and the classroom and to help teachers truly understand the rationale behind each activity proposed in structured curriculum units is by providing teachers with pedagogical support (Kraft and Blazar, 2016). Thus, in the third group of our study (Coaches Group henceforth), teachers received the same short-term training complemented with on-going support through the use of the same structured curriculum unit than the Sequence Group plus the tutoring of pedagogical coaches. Coaches worked with teachers on a weekly basis to promote that the full nature of the activities proposed in the curriculum unit was understood, and provided extra support, explanations and feedback depending on each particular teachers´ needs. The literature shows that coaches seem to increase the fidelity of implementation and improve teacher and student performance (Kretlow and Bartholomew, 2010) Kraft, Blazar and Hogan (2016) estimate that coaching raised student performance on standardized tests by 0.15 standard deviations and improved instructional practice by 0.58 standard deviations based on effect sizes reported in 44 studies that used experimental or quasi-experimental designs. This effect compares favorably when contrasted with the larger body of literature on teacher training (Yoon, Duncan, Lee, Scarloss and Shapley, 2007; Garet, Wayne, Stancavage, Taylor, Eaton, Walters, Song, Brown, Hurlburt, Zhu, Sepanik, and Doolittle, 2011).

Comparing these two treatment groups with the control group allows us to confidently establish the marginal effect of complementing training sessions with either just a structured curriculum unit or with additional coaching effort. Our first set of results clearly suggests that there is a gain in terms of learning. Specifically, students in the Sequence and Coaches Group learned between 55% and 64% of a standard deviation more than those students in the control group, respectively. This is equivalent to an average increase in student achievement from the 50th to the 66th (70th) percentile, approximately, for a student moved from the control condition to the structured curriculum condition (coaches condition). The marginal costs of doing so are also relative low compared to the benefits. Complementing training sessions with a structured curriculum unit costs (per student) 0.84 dollars per 0.1 standard deviations; in other words it costs 0.84 dollars to move a child from the 50th to the 53th percentile approximately, while complementing it with additional coaching effort costs (per student) 2.28 dollars per 0.1 standard deviations.

Establishing empirically the additional effect of coaching with respect to a structured curriculum unit is also relevant in terms of policy. Although, in general terms, coaches seem to increase the impact of teacher professional development public policies, hiring, training and providing coaches is an expensive and human-resource intensive approach.[2]

[2] In Argentina, exact figures and numbers are not publicly available, but many in-service teacher professional development programs – in particular those which provide support for rural or non-central provinces – include and finance the training and deployment of coaches. For instance, a recent national initiative involved the hiring of coaches to support the work of 800.000 teachers (Argentine Ministry of Education, 2015).

According to our results, there is no general additional benefit in terms of student learning between the structured curriculum unit with and without on-going coaching. However, qualifying this result is another contribution of our paper. We find that additional coaching does make a difference for relatively inexperienced teachers. Specifically, student in the Coaches Group learned 82% of a standard deviation more than students in the Sequence Group when we consider the least experienced teachers. Therefore, tutors add a value for those teachers who are relatively inexperienced in teaching science and this is particularly true when we consider higher-order skills, which require more intensive teaching. This suggests that improving teaching in Science is not a matter of choosing the best strategy but the one that suits best the different types of teachers and learning goals.

The effect of teaching training on the learning experience goes beyond scores. Their impact on other dimensions of the learning process is of independent interest, but our study allows us to shed light on how the effective adoption of evidence-based Science teaching techniques affects the perceptions of students and teachers. We find that the structured curriculum unit seems to be an effective instrument to enhance curiosity and interest among students. In particular, using an index that captures these aspects, we find that pupils in the Sequence Group present a scale 20% of a standard deviation higher than those in the control group. Results also show that both treatments favorably change teacher perceptions of their practices and their expectations of student learning. Compared to the Control Group, teachers in the Sequence Group and Coaches Group present a scale between 63% and 100% higher in their perception that their teaching practices meaningfully changed, that they enjoyed more teaching Science, that they taught more hours of Science and that students learned more and developed more skills.

There is a growing body of the literature in economics devoted to evaluate the impact of different policy interventions at the school level. Most of this effort has gone into identifying the causal effects of two broad categories of interventions: (a) improving school inputs, such as textbooks or classroom libraries (Abeberese, Kumler, and Linden, 2014[3]; Glewwe, Kremer, and Moulin, 2009; He, Linden and Margaret, 2009[4]), remedial education and/or assistant teachers (Banerjee, Cole, Duo, and Linden, 2007; Jacob and Lefgren 2004a), computers and computer-aided instruction (Linden 2008, Barrera and Linden, 2009; Cristia, Ibarraran, Cueto, and Severin, 2012; Mo, Zhang, Luo, Qu, Huang, Wang, Qiao, Boswell, and Rozelle 2014; Muralidharan, Singh and Ganimian, 2016; Berlinski and Busso, 2017), and other instructional technology, like flashcards (He, Linden, and

[3] The main component of the program evaluated by these authors was providing schools with a set of age-appropriate books. This component was completed with training teachers to incorporate reading in the curriculum, and with a 31 day “read-a-thon” to encourage children to read and supporting teachers as they incorporate reading into their classes.

[4] He, Linden and Margaret (2009) assessed a program that consisted of two main components: the child library and the activities carried out in class, which included using story books, flash cards for word and letter recognition, and charts to instruct children.

MacLeod, 2008) or flipcharts (Glewwe, Kremer, Moulin, and Zitzewitz, 2004); and (b) providing additional educational resources and their management, including the effect of voucher programs (Angrist, Bettinger, Bloom, King and Kremer, 2002) or lumps sum grants to schools (Das, Dercon, Habyarimana, Krishnan, Muralidharan, and Sundararaman, 2013), as well as organizational changes like, for example, curricular design (Harris, Penuel, DeBarger, D'Angelo and Gallagher, 2014), reducing class size (Angrist and Lavy, 1999; Urquiola, 2006; Krueger and Whitmore, 2002; Fredriksson, Ockert, and Oosterbeek, 2012), group tracking (Duflo, Dupas, and Kremer, 2011), and enhancing teacher incentives (Duflo, Hanna, and Ryan, 2012; Glewwe, Ilias and Kremer, 2010). We see our paper as a contribution to both literatures insofar as training teachers has a direct effect on school inputs and we are able to identify and evaluate alternative ways to organize and deliver this training.

The identification of the causal effect of on-the-job or in-service teacher training has received far less attention.[5] Most of this research in education economics uses regression discontinuity strategies to estimate the effect of different training programs. For example, Jacob and Lefgren (2004b) find no evidence on student achievement of an in-service training program targeting teachers of math and reading in elementary schools located in, relatively poor areas, in the United States. Angrist and Lavy (2001) estimate the effect of in-service teacher training on achievement in Jerusalem elementary schools. In this case, results are more encouraging. They find that the training program improved test scores by 0.2 to 0.4 standard deviations in secular schools but that seemed to have no effect in religious schools (which were poorly organized). Finally, Bassi, Meghir, and Reynoso (2016) use a randomized controlled trial to estimate the effectiveness of guided instruction methods in under-performing schools in Chile. Teachers in treated schools received detailed classroom guides and scripted material to follow in their lectures. They find that only the most advantaged students within treated schools (students from higher income families within our lower income population) benefit from the program, improving test scores by almost 0.2 of a standard deviation. Our contribution to this literature is twofold: we identify the learning effect and cost-effectiveness of different forms of delivering teacher training programs.

The remainder of the paper is organized as follows: Section 2 presents the research context and Section 3 explains the design of the experiment, describes the components of the

[5] There are a number of papers in the education literature studying the effect of on-the-job or in service teacher training programs. This literature has been recently reviewed by McEwen (2015), who concludes that most of these studies do not identify the pure effect of training as it usually overlaps with other type of treatments, such as class size reductions or other institutional treatments. Also these papers are based on small scale studies. An example of a RCT study on the effect of teacher training in Science is Sloan (1993), which involved a sample of 173 students and whose results on the positive effects of the intervention were later discarded by Yoon, Duncan, Lee, Scarloss and Shapley (2007) for not addressing clustering and multiple outcomes.

intervention and explains the data collection process. Section 4 presents the research sample. Section 5 presents descriptive statistics for the main variables. Section 6 discusses our identification strategy and Section 7 shows the main results of the paper. Finally, Section 8 concludes and reflects on the implications in educational policy.

## 2. Research Context

In Argentina, education from primary school through high school education is compulsory and free of charge. The country has one of the highest rates of literacy (98%) and school-life expectancy (16 years) in the world (World Bank, 2014). Although attendance and completion at the secondary school level remains an issue, primary education is considered to be universal.

According to official statistics, Argentina has 11 million students enrolled in four education levels: pre-school (ages 3-5, 15.6%); primary (ages 6-11, 41%); secondary (ages 12-17, 35.5%) and tertiary (ages 18-22, 7.9%). The majority of these students (71%) attend public schools (DINIECE, 2015)[6].

Between 2003 and 2013 student numbers increased by approximately 10%, while the number of teachers increased by more than 20% over roughly the same period (DINIECE, 2004; DINIECE, 2015). This allowed Argentina to reach a pupil-teacher ratio of 11, the lowest in Latin America after Cuba (OECD, 2016), although it is worth noting that this ratio varies considerably across provinces.

Although there have been large successes in terms of increasing coverage, the Argentinean education system fails to provide high quality education (at least as measured by standardized test scores). While other countries in the region improved learning outcomes since 2000 - measured by the OECD's Program for International Student Assessment (PISA) - Argentina's scores show no progress (at best) in Science, or even experienced a marginal decline between 2000 and 2012.[7] According to the 2012[8] study, Argentina ranked amongst the lowest of the participating countries (59 out of a possible 65) (OECD, 2014). [9]

CABA, being the wealthiest jurisdiction in Argentina, exhibits some specific features; namely a lower share of students attending public schools (49%) and higher levels of

[6] These figures do not include special and adult education.

[7] According to De Hoyos, Holland and Troiano (2015), there is a gradual increase in Argentina's Science scores between 2006 and 2012, but it is not statistically significant at conventional levels (95%).

[8] Argentina participated in PISA 2015, but its results were excluded from the main report due to problems with sample design. However, CABA participated as an adjudicated region and was included in the results.

[9] In a similar vein, results of the UNESCO Second Regional Comparative and Explanatory Study (SERCE), applied to students of 3rd and 6th graders, show that only 11.4% of students of 6th grader were able to explain everyday situations based on scientific evidence, use models to explain natural phenomena or draw conclusions based on data (UNESCO, 2009).

student achievement (DINIECE, 2015). Despite this, international assessments show that the level of achievement of CABA students in Science is still well below the OECD average (OECD, 2016). Problems with teaching and learning Science in CABA were also highlighted by the last wave of the Trends in International Mathematics and Science Study (TIMSS), according to which CABA students (4th- and 8th-graders) place at the bottom of the world ranking, just above Egypt and South Africa (Martin, Mullis, Foy and Hooper, 2016). These results are not surprising given the reality of science education in Argentina, where lessons are mostly teacher-centered and focused on the transmission of encyclopedic content, far from competency-based international learning standards, such as those assessed by TIMMS and PISA (Argentine Ministry of Education, 2007). Countries with high levels of scientific literacy tend to implement inquiry-based approaches which position students as active knowledge producers in a classroom community of practice, placing importance on the development of specific science skills and deep understandings (OECD, 2016).

## 3. Experimental design

We carried out a randomized controlled experiment to assess the effect of different teacher professional development approaches on student learning in Science. The intervention focused on a compulsory unit of the seventh grade Science national curriculum: the Human Body.

The intervention consisted of a random allocation of 70 CABA state primary schools to one of three experimental groups (Appendix 1 describes the intervention's timeline). Thus, the unit of randomization was the school.[10] All teachers involved received one in-service 4-hour training session for all 7th grade Science teachers, and were then asked to teach the Human Body Science unit, according to national curriculum guidelines, over the following 12 weeks. During the training session, teachers discussed and took part in inquiry-based activities related to the teaching of the Human Body topic.[11] This session was designed and run by specialists in Science education at the School of Education, University of San Andrés (Argentina). Those teachers receiving only this training form the control group.

In the first treatment group (Sequence Group), teachers received the same 4-hour training session and a structured curriculum unit that outlined how to teach the Human Body using an inquiry-based approach.[12] The structured curriculum unit focused not only on Human Body content, but also on the development of Science competencies, as defined by the

[10] The randomization was at the school level and not at the classroom level because 46% of the schools in our sample shared the same science teacher for at least one of their classrooms. Therefore, assigning classrooms to different treatments was operationally impossible.

[11] Details on the specific activities carried out during this session are available upon request.

[12] Given that inquiry-based pedagogies have been shown to promote Science competencies (Minner, Levy and Century, 2009), this approach was chosen for this unit with a particular emphasis on active learning.

ability to explain phenomena scientifically, evaluate and design scientific inquiry and interpret data and evidence scientifically (OECD, 2016). This document included experiential learning activities, which are a departure from more common and traditional teaching methods[13], along with questions, approaches and worksheets for students. The structured curriculum unit was designed by Science Education specialists with a group of seventh grade teachers who were not part of the schools selected for the study. Teachers were expected to adapt and implement these activities over the following 12 weeks.

In the second treatment group (Coaches Group), teachers also received the same 4-hour training session and structured curriculum unit, but their training was complemented with weekly sessions with a pedagogical coach. The coaches met with teachers at their schools during planning periods of 60 minute sessions over 12 weeks with the aim to guide and support teachers on how to implement the structured curriculum unit, as well as to enhance teacher reflection on their practice. The pedagogical coaches were recruited by the School of Education, University of San Andrés. They were selected based on their knowledge and prior experience in Science education, as well as their potential to create a positive working relationship with participating teachers. They all held at least a bachelor's degree in Science and/or pedagogical certification (Table A2.1 of Appendix 2 reports their main demographic characteristics). In addition, coaches received regular training sessions every fortnight throughout the intervention (a total of eight 3-hour meetings) and they were given access to an extensive library of guiding documents and resources to support their work

### 3.1. Design of the assessment instrument

A central part of the design of any experiment is to determine the outcome measure, which in this case is student achievement in Science. Together with university specialists in Science education, an assessment instrument (hereafter referred to as the “Science test”) was developed to measure learning and to potentially distinguish the gains from the different working modalities with teachers.

First, following an in-depth analysis of the unit of Human Body, the topics included in the Science test were determined. In addition to this, three levels of skills were outlined: (i) basic skills, which required students to recall scientific content (such as identifying organs) and read simple tables and graphs; (ii) medium-order skills, which required students to explain scientific phenomena and develop conclusions based on simple experimental data; and (iii) higher-order skills, which required students to describe how different body systems work together, identify researchable questions, design experiments to address hypothesis, explain scientific phenomena and draw conclusions based on more complex experimental

[13] Examples of these activities include investigating changes in heart rate, measuring lung capacity, dissecting organs and evaluating historical experiments. The structured curriculum unit designed for this study is available upon request.

data (Appendix 3 details the differences between these skills).

We developed the Science test using the following procedure: (1) we created a pool of items for basic, medium- and higher-order content areas following the structure of PISA and TIMMS Science questions; (2) experts reviewed the items; (3) we piloted the test in two 7th grade classrooms at schools not participating in the project, and performed think-aloud exercises with students to better understand their answers and make adjustments; and (4) a panel of experts reviewed the final assessment instrument.

The outcome of this process was an 11-items Science test of approximately one hour of duration. It consisted of both multiple- choice and open-ended questions. This combination allowed us to capture a wider range of student responses, including stronger evidence of critical thinking skills, than is typically associated with only multiple-choice tests (Stanger-Hall, 2012).The test was administered at the end of the intervention at each school by external observers to guarantee the fidelity of its implementation under strict exam conditions. The test had sound psychometric properties. The scale reliability coefficient (Cronbach's alpha) is 0.79 in the full sample data and 0.76 in the Control Group.

The Science test questions were weighted according to difficulty, with higher-order questions scoring 3 points, medium-order questions scoring 2 points and basic-skills questions scoring 1 point. Answers were classified as either "Correct", for which they achieved full marks; "Partially correct", for which they achieved half of the maximum marks for the given question; "Incorrect", for which no marks were given; and "Omitted", when no answer was given and for which no marks were given. Tests were corrected by specialists using a common rubric, which was shared and discussed during a half-day training session. Answers being challenging to classify were then discussed and determined by multiple assessors.

### 3.2. Experimental Data

We collected data on all students, teachers and schools. We conducted a student survey to collect socio-demographic data in order to check whether the randomly created groups of schools were comparable.[14] We also administered the Science test after schools had completed their 12 week intervention, followed by a student questionnaire designed to measure fidelity of implementation, students' perceptions on the teaching they had experienced and their attitudes towards Science. These questions were later used to build an index that measured if learning was interesting and relevant as well as if teaching practices

[14] The student survey was collected in the classroom and contained information on students' characteristics, their family and socioeconomic background.

inspired curiosity (See Appendix 4 for a detailed description of the index).[15] These questions were based on a validated instrument, the *Tripod Survey for Students* (Bill and Melinda Gates Foundation, 2012).

Before and after the intervention took place, we obtained additional information about the teachers. As a baseline, we gathered background characteristics of teachers and general information about their Science class. In the post-intervention survey, teachers responded a set of questions to assess the fidelity of implementation of the intervention, as well as perceived changes in class dynamics and teaching practices.

Finally, we collected administrative information at the school level. This information included data on school and seventh grade enrollment, number of classrooms and teachers, repetition rate, promotion rate, over-aged rate, location of the schools and the Language score in the local end-of-primary exam of 7th grade (FEPBA, for its Spanish acronym).

## 4. The sample

Our sample consisted of 70 state primary schools from 6 (out of 21) school districts within CABA, giving a representative sample of state primary schools in the jurisdiction.[16] These 70 schools involved about 3,000 students, grouped into 136 seventh grade classes, and 99 Science teachers (Table 1). Participating schools were individually randomized into a Control Group of 24 schools and two treatment groups, each of them composed of 23 schools.

**Table 1. Background characteristics of the sample**

| | All simple | Control Group | Sequence Group | Coaches Group |
|---|---|---|---|---|
| Number of schools | 70 | 24 | 23 | 23 |
| Average students per school | 301 | 316 | 289 | 297 |
| Number of class divisions in 7th grade | 136 | 50 | 44 | 42 |
| Number of students in 7th grade | 2965 | 1086 | 917 | 962 |
| Number of Science teachers in 7th grade | 99 | 36 | 32 | 31 |

On average, schools in the experiment were comparable to the rest of primary state schools throughout CABA. Table A5.1 of Appendix 5 compares the average characteristics of the 70 participating schools with the characteristics of the non-participating primary state

[15] See, for example, the Measures of Effective Teaching (MET) project where results show that student surveys produce more consistent results than classroom observations or achievement gain measures (Bill and Melinda Gates Foundation, 2012).

[16] Currently, there are 455 state primary schools in CABA. Thus, the share of schools included in our sample is 15%.

schools in CABA. As shown, there are no significant statistical differences between participating and non-participating schools in their size, seventh grade size, and seventh grade classrooms, as well as in student's promotion rate, over-aged rate and drop-out rate per school.[17] We find only one statistically significant difference, at 90% of confidence, in student's repetition rate per school, which is slightly higher for the schools included in the experiment. However, we do not find any statistically difference in the FEPBA results.[18] This is also reflected in Figure A4.1 of Appendix 4 that shows that the distributions of study participants and non-study participants' scores share a substantial common support. Finally, there are no statistically significant differences in the Social Vulnerability Index at the school district level, which ranks houses in each school district according to their degree of vulnerability in terms of material and non-material assets.[19] Based on these results, we can confidently state that the participating schools constitute a representative sample of the CABA primary state schools.

## 5. Randomization and Descriptive Statistics

Table 2 presents pre-treatment main sample means and standard deviations for the full sample and experimental groups. Half of the students in the research sample are female and on average they are approximately 12 years old. The majority were born in Argentina (86%). In relation to their socioeconomic background, approximately 70% of the students have parents with secondary education. In addition to this, 90% of the students have access to Internet in their homes; and 59% and 64% of them also owning at least one air conditioning and one car in their homes, respectively.. Finally, about 65% of students missed, at most, one class per month since the beginning of this investigation (see Appendix 6 for more details on other variables) .In terms of participating teachers, Table 2 shows that 88% of them are female with an average age of 42 years. Near 45% of teachers have a post-graduate certificate and, on average, they have gained about 12 years of teaching experience and 6.5 years of teacher experience in Science. Almost all of them participated in some form of teacher training in the last two years and half of them have ever used a structured curriculum unit.

Table 2 shows that schools enroll an average of 301 students and 42 students in 7th grade,

[17] Over-aged students are those who are older than the normal age for a grade level, as defined by law.

[18] The FEPBA test was prepared and administered to 7th grade students of both private and public schools in CABA by the Ministry of Education of CABA in 2014.

[19] The Social Vulnerability Index (SVI) is a weighted index, calculated by the Ministry of Education of CABA, which assigns a value to each household according to its characteristics with respect to the material and non-material assets. In this way, households are ranked according to their degree of vulnerability. Households that have the highest vulnerability assume the value of 1 in the index while those that have the lowest vulnerability assume the value of 0 in the index.

which are, most commonly, divided in two classrooms. Repetition rate in 7th grade is 3% on average, and school over-aged and promotion rate are on average 15% and 97%, respectively.

We also report the FEPBA score in Language, which presents an average of 448.[20] We do not discuss here the meaning of this score, but use this variable with the sole intention of comparing student academic performance across our groups of CABA schools.

Are the experimental groups similar with each other and representative? Table 3 displays the differences in the means along with p-values from two-tailed t-tests of equality of means across experimental groups. As we can see, the treatment and Control groups do not differ significantly in any observable dimension. The only variable with a statistical difference at the 95% level of confidence is student age, where Control Group students are slightly younger than those in the Sequence Group. However, this difference is very small and vanishes when considering 7th grade repetition rate, which is balanced across the three experimental groups (see Appendix 6 for further differences in the means of other variables).

[20] It is important to note that there is no school-specific measure of Science knowledge of 6th or 7th grade student available in Argentina or CABA. Therefore, to the best of our knowledge, FEPBA score in Language is the best approximation, based on administrative data, which we can make.

**Table 2. Pre-treatment characteristics**

| | | All simple | | Control Group | | Sequence Group | | Coaches Group | |
|---|---|---|---|---|---|---|---|---|---|
| | N | Mean | Sd | Mean | Sd | Mean | Sd | Mean | Sd |
| ***Student-level variables*** | | | | | | | | | |
| Percent female | 2359 | 0.49 | 0.50 | 0.51 | 0.50 | 0.48 | 0.50 | 0.49 | 0.50 |
| Age | 2346 | 12.19 | 0.52 | 12.17 | 0.49 | 12.22 | 0.55 | 12.18 | 0.52 |
| Percent of Argentines | 2341 | 0.86 | 0.34 | 0.86 | 0.35 | 0.87 | 0.34 | 0.86 | 0.35 |
| Mother or father education (secondary) | 1858 | 0.71 | 0.46 | 0.72 | 0.45 | 0.71 | 0.45 | 0.69 | 0.46 |
| Have internet in their home | 2279 | 0.90 | 0.31 | 0.91 | 0.29 | 0.89 | 0.32 | 0.89 | 0.31 |
| Have air conditioning in their home | 2130 | 0.59 | 0.49 | 0.60 | 0.49 | 0.60 | 0.49 | 0.58 | 0.49 |
| Have at least one car in their home | 2162 | 0.64 | 0.48 | 0.64 | 0.48 | 0.66 | 0.48 | 0.62 | 0.49 |
| At most, missed one class per month | 2288 | 0.65 | 0.48 | 0.66 | 0.48 | 0.64 | 0.48 | 0.64 | 0.48 |
| ***Teacher-level variables*** | | | | | | | | | |
| Percent female | 91 | 0.88 | 0.33 | 0.85 | 0.36 | 0.89 | 0.32 | 0.90 | 0.31 |
| Age | 90 | 41.52 | 8.75 | 39.59 | 8.69 | 42.64 | 9.46 | 42.75 | 7.94 |
| Percent with Post-Graduate Certificate | 91 | 0.43 | 0.50 | 10.42 | 6.67 | 12.68 | 6.49 | 12.26 | 8.36 |
| Percent with University degree | 91 | 0.10 | 0.30 | 5.81 | 5.58 | 7.04 | 5.62 | 6.75 | 7.79 |
| Seniority in teaching (in years) | 91 | 11.70 | 7.19 | 3.52 | 3.28 | 4.09 | 3.49 | 3.23 | 4.29 |
| Seniority in teaching Science (in years) | 88 | 6.48 | 6.33 | 0.35 | 0.49 | 0.50 | 0.51 | 0.45 | 0.51 |
| Percent of teachers that participated in trainings | 91 | 0.90 | 0.30 | 0.91 | 0.29 | 0.96 | 0.19 | 0.83 | 0.38 |
| Percent of teachers that used a teaching sequence | 91 | 0.55 | 0.50 | 0.62 | 0.49 | 0.46 | 0.51 | 0.55 | 0.51 |
| ***School-level variables*** | | | | | | | | | |
| Students per school | 70 | 301.20 | 132.10 | 316.50 | 139.60 | 289.60 | 127.70 | 296.80 | 132.80 |
| Students of 7th grade | 70 | 42.36 | 19.42 | 45.25 | 22.16 | 39.87 | 16.85 | 41.83 | 19.21 |
| School promotion rate (%) | 70 | 0.97 | 0.02 | 0.98 | 0.03 | 0.97 | 0.02 | 0.97 | 0.02 |
| School drop-out rate (%) | 70 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 |
| School over-aged rate (%) | 70 | 0.15 | 0.07 | 0.15 | 0.08 | 0.16 | 0.08 | 0.14 | 0.06 |
| FEPBA score in Language | 70 | 488.16 | 18.05 | 487.03 | 18.17 | 487.86 | 20.02 | 489.66 | 16.49 |

Note: N means number of observation in the full sample and Sd means standard deviation.

**Table 3: Balance across Treatments**

| | Sequence<br>vs.<br>Control | Coaches<br>vs.<br>Control | Coaches<br>vs.<br>Sequence |
|---|---|---|---|
| ***Student-level variables*** | | | |
| Percent female | -0.03 | -0.02 | 0.01 |
| Age | 0.06** | 0.01 | -0.04 |
| Percent of Argentines | 0.01 | 0.00 | -0.01 |
| Mother or father education (secondary) | -0.01 | -0.04 | -0.03 |
| Have internet in their home | -0.03 | -0.02 | 0.00 |
| Have air conditioning in their home | 0.00 | -0.02 | -0.02 |
| Have at least one car in their home | 0.02 | -0.02 | -0.04 |
| At most, missed one class per month | -0.02 | -0.02 | -0.01 |
| *Teacher-level variables* | | | |
| Percent female | 0.04 | 0.04 | 0.00 |
| Age | 3.06 | 3.16 | 0.11 |
| Percent with Post-Graduate Certificate | 0.15 | 0.10 | -0.05 |
| Percent with University degree | 0.13 | 0.11 | 0.10 |
| Seniority in teaching (in years) | 2.26 | 1.84 | -0.42 |
| Seniority in teaching Science (in years) | 1.23 | 0.94 | -0.29 |
| Percent of teachers that participated in trainings | 0.05 | 0.08 | -0.14* |
| Percent of teachers that used a teaching sequence | 0.15 | 0.07 | 0.09 |
| ***School-level variables*** | | | |
| Students per school | -26.98 | -19.72 | 7.26 |
| Students of 7th grade | -5.38 | -3.42 | 1.96 |
| School promotion rate (%) | -0.01 | 0.00 | 0.01 |
| School drop-out rate (%) | 0.00 | 0.00 | 0.00 |
| School over-aged rate (%) | 0.01 | 0.00 | -0.01 |
| 7th student’s repetition rate (%) | 0.02 | 0.00 | -0.02 |
| FEPBA score in Language | 0.83 | 2.63 | 1.80 |

Note: Each entry indicates the mean difference between the two experimental groups in the column for the corresponding variable in each line. * indicates that the difference of means test is significant at 10%; ** significant at 5%; *** significant at 1%.

Only 14.5%, 15.3% and 17% of students in the Control, Sequence and Coaches group, respectively, did not complete the Human Body test. These are relatively low non-response rates and, as it is shown in Table 4, there is no statistically significant difference in the number of students who missed or omitted the test across the experimental groups. Finally, only 7 out of the 136 classrooms failed to teach lessons on the Human Body unit, which implies an attrition rate of 5%. Appendix 7 shows that excluding classrooms not completing the Human Body unit carries no effect on the balance across the experimental groups (see Table A7.1).

**Table 4. Differences in the non-Response Rates**

| | Sequence<br>vs.<br>Control<br>(1) | Coaches<br>vs.<br>Control<br>(2) | Coaches<br>vs<br>Sequence<br>(3) |
|---|---|---|---|
| Missing (or omitting) student test | 0.008 | 0.025 | 0.017 |

Note: Each entry indicates the mean difference between the two experimental groups in the column for the students who missed or did not complete the student test. * indicates that the difference of means test is significant at 10%; ** significant at 5%; *** significant at 1%.

## 6. Identification Strategy

Our goal is to understand how using a structured curriculum or receiving, as a complement, weekly coaching meetings can influence learning outcomes in a randomized controlled experimental setting. In this setting, the Control Group estimates what would have happened to the treated groups in the absence of the intervention. The validity of the Control Group is evaluated by examining the exogeneity of treatment status with respect to the potential outcomes, and by testing whether pre-intervention characteristics of the treatment and Control Groups are reasonably similar. As discussed in Section 5, we find a strong similarity across the three experimental groups. The similarity across pre-treatment characteristics is consistent with the exogeneity in the allocation of schools in each treatment.

When the treatment status is exogenous, estimating the average treatment effects is straightforward. The random assignment of schools to treatment/control groups allows us to identify the average treatment effect by simply comparing the means of each of the two treatment groups with respect to the Control Group. Operationally, we estimate by Ordinary Least Squares a set of models of the following form:

$$Y_{ij} = \alpha + \beta DT_j + \gamma X_{ij} + \theta_j + \varepsilon_{ij} \tag{1}$$

where $i$ indexes students and $j$ indexes schools. $Y_{ij}$ is the outcome of interest (e.g., student performance in the Human Body test) of student $i$ in school $j$. $DT_j$ is a dummy variable indicating treatment status. We also include control variables ($X$). Specifically, we control for students characteristics (gender, age, nationality, parent's education, if the student missed at most one class per month, if the student has internet in his home), teacher characteristics (gender, age, years of experience in teaching, if she/he has post-graduate certificates) and school characteristics (school size, 7th size, 7th repetition rate, FEPBA score in Language, school district –or location). The parameter of interest $\beta$ is the average treatment effect (e.g., the average effect on student performance in the Human Body test of being in the treatment group versus the status quo). Finally, $\varepsilon_{ij}$ is the error term.

The specifications of the model stated in equation (1) take into account the potential correlation between students' and teachers' performance and behavior by clustering the standard errors at the school level (i.e. the unit of randomization). However, the standard error estimates are typically not sensitive to the level of clustering.

## 7. Results

In section 7.1 we discuss our main results regarding the effect of the treatments on student learning. As their effects may depend on how teachers respond to the intervention, section 7.2 explores the average treatment effect conditioned to teacher experiences. Finally, section 7.3 extends our analysis beyond the test results and explores how the different interventions affected students' and teachers' perceptions about learning and teaching Science.

### 7.1. Student Learning

Table 5 shows the mean and standard deviation of the standardized score in the Science test, which was calculated using the mean and standard deviation of the Control Group, the score according to different levels of skills (Basic, Medium and High-order skills) as well as the percentage of correct, incorrect and omitted answers across experimental groups. We can see that the average score for the Sequence Group is 0.36 standard deviations higher than the average score for the Control Group, whereas the average score of the Coaches Group is 0.53 standard deviations higher that of the Control Group. It seems that both treatments were more effective in promoting middle and higher order skill development in students than basic skills. In addition to this, the percentage of correct answers increased in both treatments, while the percentage of incorrect and omitted answers decreased.

**Table 5. Mean and Standard deviation of learning outcomes**

| | Control Group | | Sequence Group | | Coaches Group | |
|---|---|---|---|---|---|---|
| | Mean | Sd | Mean | Sd | Mean | Sd |
| Science score | 0 | 1 | 0.356 | 1.115 | 0.530 | 1.152 |
| Science score (Basic skills) | 0 | 1 | 0.212 | 0.957 | 0.210 | 0.980 |
| Science score (Medium skills) | 0 | 1 | 0.605 | 1.055 | 0.515 | 1.084 |
| Science score (Higher-order skills) | 0 | 1 | 0.406 | 1.102 | 0.568 | 1.157 |
| Percent of correct answers | 0.326 | 0.202 | 0.409 | 0.240 | 0.436 | 0.243 |
| Percent of incorrect answers | 0.217 | 0.160 | 0.166 | 0.158 | 0.161 | 0.153 |
| Percent of omitted answers | 0.203 | 0.203 | 0.145 | 0.171 | 0.113 | 0.166 |

Table 6 presents the results on student learning. The dependent variable is the standardized score in the Science test, which was calculated using the mean and standard deviation of the Control Group. This allows interpreting the coefficient as the treatment effects in terms of the standard deviation. Columns (1) and (2) of panel A show the effect of the structured curriculum unit (Sequence Group) and the coaches (Coaches Group) in comparison with the Control Group. The estimated coefficients are all statistically significant and present a positive sign. The average treatment effect of the structured curriculum is an increase of 55%

of a standard deviation in Science scores and the average treatment effect of the coaches is an increase of 64% of a standard deviation in Science scores. Thus, students in the Sequence and Coaches Group learned between 55% and 64% of a standard deviation more than those students in the Control Group. This is equivalent to an average increase in student achievement from the 50th to the 66th percentile in the case of the Sequence Group whereas, if they were treated in the Coaches Group, the improvement goes from the 50th to the 70th percentile approximately. These effects are considered to be rather large for interventions with similar characteristics (see for example Allen et al (2011); Campbell and Malkus (2011); Sailors and Price (2010 and 2015); Matsumura, Garnier, Correnti, Junker, and DiPrima Bickel (2010); Bassi, Meghir, and Reynoso (2016)).

Although the estimated coefficient of the Coaches Group (column 2) is higher than that of the Sequence Group (column 1), we find its marginal effect is not statistically significant. This is shown in column 3 that presents the result of the Wald test, which evaluates the difference in the coefficient between column 1 and 2. This finding is relevant in terms of policy as it would suggest that just the implementation of a structured curriculum would be sufficient to improve average results in learning outcomes in Science in the short term.

Panel B of Table 6 reports the same analysis but splitting the score according to different levels of skills (Basic, Medium and High-order skills). The findings are similar: although both treatments improve learning, we do not find a significant difference in their effects. Interestingly, the average treatment effect of the Coaches Group increases as the content evaluated (or items) becomes more complex (see column 2). This implies that the Coach treatment was more effective in promoting higher order skill development in students than either the 4-hour training session or the provision of the structured curriculum unit.

The mechanism through which both the Sequence and Coaches treatments appear to increase Science test-scores involves an increase in the percentage of correct answers. Students in the Sequence and Coaches Group exhibit 10% more correct answers than students in the Control Group (panel C). But we also observe that the treatments reduce the number of omitted and incorrect answers. This suggests that the interventions did not only increase Science learning (which is shown in both the increase of correct answers and the reduction of incorrect ones), but also motivated students to answer more questions.

Our findings are especially important given the big difference in treatments' costs. Whilst the cost per student for the Control Group was 1.4 dollars, for the Sequence Group the cost per student was 4.6 dollars and for the Coaches Group it was 14.7 dollars[21]. This includes the costs of hiring and training the tutors, teacher seminar materials as well as curriculum unit design and printing. The estimates of Table 7 allow us calculating the cost-effectiveness of the program. Providing teachers with a short-term training complemented with a structured curriculum costs (per student) 0.84 dollars per 0.1 standard deviations. Whereas, providing teachers with the same short-term training complemented with on-going coaching through

[21] Our calculations correspond to 2016 US dollar.

the use of the same structured curriculum costs (per student) 2.28 dollars per 0.1 standard deviations. In other words, it costs 0.84 dollars to move a child from the 50th to the 53th percentile, approximately, in the first intervention, and 2.28 dollars in the second intervention. Therefore, providing teachers with a structured curriculum is 2.7 times more cost-effective for the total score than complementing it with on-going coaching. Even though cost effectiveness calculations might not be perfectly comparable across program, in general terms, our calculations are in line with other interventions based on teacher training programs (see for example Banerjee, Cole, Duflo and Linden (2007)).

**Table 6. Results on Science learning**

| ***Dependent variable*** | **Sequence vs. Control (1)** | **Coaches vs. Control (2)** | **Wald Test (3)** |
|---|---|---|---|
| *Panel A* | | | |
| Science score | 0.548*** | 0.644*** | 0.45 |
| | (0.125) | (0.137) | (0.502) |
| *Panel B* | | | |
| Science score (Basic skills) | 0.285*** | 0.290*** | 0.01 |
| | (0.083) | (0.083) | (0.951) |
| Science score (Medium skills) | 0.577*** | 0.540*** | 0.07 |
| | (0.131) | (0.132) | 0.791 |
| Science score (Higher-order skills) | 0.405*** | 0.612*** | 2.39 |
| | (0.121) | (0.135) | 0.122 |
| *Panel C* | | | |
| Percent of correct answers | 0.103*** | 0.110*** | 0.06 |
| | (0.023) | (0.026) | (0.800) |
| Percent of incorrect answers | -0.057*** | -0.050*** | 0.17 |
| | (0.014) | (0.015) | (0.680) |
| Percent of omitted answers | -0.053** | -0.079*** | 1.62 |
| | (0.020) | (0.020) | (0.203) |

Note: ** significant at 5%; *** significant at 1%. Robust standard errors in parentheses clustered at school level. Controls: (i) students characteristics (gender, age, nationality, parent's education, if the student missed, at most, one class per month, if the student has internet in his home), (ii) teacher characteristics (gender, age, years of experience, if she/he has post-graduate certificates) and (iii) school characteristics (school size, 7th size, 7th repetition rate, FEPBA score in Language, school district –or location). The number of observations for regression in column (1) is 1105 and in column (2) 1100. All the regressions exclude the classrooms where the Human Body Unit was not taught.

In order to deepen our analysis, we investigated whether any group of students experienced more gains in test-score results. Table A8.1 in Appendix 8 displays separate estimates for students below (first panel, "high performance") and above (second panel, "low performance") the mean test-score for each group. We can see that, in general, both the use of the structured curriculum unit and the coaches seem to benefit more high performing than low performing students: in both treatment groups, high performing students obtained a significantly higher percentage of correct answers and lower percentage of incorrect and omitted ones. In particular, the gain of high performing students in the Coaches Group is almost twice the gain of low performing students in the same group, while the gain of high performing students in the Sequence Group is 19% higher than the gain of low performing students in the same group. However, although the Coaches Group has a slightly higher impact in increasing test-scores than the Sequence Group for the high performing students, we still observe no statistically significant difference between these two treatments (see column 3).

Another factor of interest is to analyze whether learning results differ according student gender. Table A8.2 in Appendix 8 reports separate estimates of the average treatment effects for female and male students. As it can be observed in column (1), there is no difference in the average treatment effect within the Sequence Group between girls and boys. However, column (2) shows that the average treatment effect of the coaches is higher for girls than boys. In particular, test-scores for girls in this group are almost 33% higher than for boys, while the percentage of correct answers and omitted answers are near 25% higher and 39% lower for girls than boys respectively. Nevertheless, we still find no difference in the average treatment effect of the coaches in comparison with the Sequence Group either for males or females (see column 3).

## 7.2. The role of teachers' experience

An important message conveyed in our previous results is that there is no statistical difference between supporting teacher training with a structured curriculum unit with or without pedagogical coaches, which suggests that the additional learning gain from coaching in Science is weak. We explore in this section whether this result is conditional on teaching experience. To do so with estimate the following model using OLS:

$$Y_{ij} = \alpha + \beta DT_j + \gamma X_{ij} + \mu(DT_j \times E_i) + \upsilon E_i + \theta_j + \varepsilon_{ij} \qquad (2)$$

where, as in equation (1), $i$ indexes students and $j$ indexes schools. $Y_{ij}$ is the outcome of interest, i.e. student scores in the Human Body test. $DT_j$ is a dummy variable indicating treatment status. $X_{ij}$ represents a set of control variables (students characteristics: gender, age, nationality, parent's education, if the student missed at most one class per month, if he has internet in his home; teacher characteristics: gender, age, general teaching experience (in years), if she/he has post-graduate certificates; and school characteristics: school size, 7th size, 7th repetition rate, FEPBA score in Language, school district –or location). $DT_j \times E_i$ represents an interaction term between treatment status ($DT$) and a dichotomous variable ($E$) that equals to 1 if the teacher has less than two years of experience in teaching Science (first quartile in our sample) and zero otherwise, we call this variable "low experience". Now, our parameters of interest are $\beta$ (the average treatment effect) and $\mu$ (the marginal effect of teaching experience). Finally, $\varepsilon_{ij}$ is the error term.

The following table shows that the average treatment effect of the Coaches Group versus the Sequence Group is conditioned by teacher experience in Science. Specifically, the average treatment effect of the coaches is an increase of 82% of a standard deviation in Science scores in comparison with the teaching sequence when we considered the least experienced teachers (column 3). This increase in test-scores is considerable; therefore, coaches add value for the teachers who were relatively inexperienced in teaching Science.

**Table 7. Results on Science learning according to teacher experience**

| ***Independent variable*** | **(1)** | **(2)** | **(3)** |
|---|---|---|---|
| Sequence vs. Control | 0.572*** | | |
| | (0.131) | | |
| Low experience | -0.300 | -0.112 | -0.749*** |
| | (0.208) | (0.255) | (0.199) |
| (Sequence vs. Control)*Low experience | -0.332 | | |
| | (0.281) | | |
| Coaches vs. Control | | 0.568*** | |
| | | (0.146) | |
| (Coaches vs. Control)*Low experience | | 0.287 | |
| | | (0.333) | |
| Coaches vs. Sequence | | | -0.174 |
| | | | (0.140) |
| (Coaches vs. Sequence)*Low experience | | | 0.820** |
| | | | (0.331) |

Note: ** significant at 5%; *** significant at 1%. Robust standard errors in parentheses clustered at school level. Controls: (i) students characteristics (gender, age, nationality, parent's education, if the student missed, at most, one class per month, if the student has internet in his home), (ii) teacher characteristics (gender, age, general teaching experience (in years), if she/he has post-graduate certificates) and (iii) school characteristics (school size, 7th size, 7th repetition rate, FEPBA score in Language, school district –or location). Low experience represents a dummy variable equals to 1 if the teacher has less than two years of experience in teaching Science and zero otherwise. An interaction term between treatment status and the dichotomous variable of low experience is included. The number of observations for regression in column (1) is 1042, in column (2) 1072, and in column (3) 1102. All the regressions exclude the classrooms where the Human Body Unit was not taught.

Furthermore, we explore if Science teaching experience conditions the average treatment effect for the higher-order skills that we expect students to develop. This is particularly relevant as they underlie the development of both complex reasoning and scientific competencies that could be more challenging for teachers to enhance in students. Table 8 shows the results of estimating equation (2) on the test-scores for higher-order skills. We confirm the finding that coaches add value in comparison with the teaching sequence for the least experienced teachers, and this is particularly true when we considered the higher-order skills, which often require more intensive teaching (column 3).

**Table 8. Results on Science test-scores in higher-order items according to teacher experience**

| *Independent variable* | (1) | (2) | (3) |
|---|---|---|---|
| Sequence vs. Control | 0.423*** | | |
| | (0.127) | | |
| Low experience | -0.233 | -0.033 | -0.602*** |
| | (0.184) | (0.227) | (0.151) |
| (Sequence vs. Control)* Low experience | -0.246 | | |
| | (0.246) | | |
| Coaches vs. Control | | 0.523*** | |
| | | (0.163) | |
| (Coaches vs. Control)* Low experience | | 0.308 | |
| | | (0.292) | |
| Coaches vs. Sequence | | | -0.050 |
| | | | (0.138) |
| (Coaches vs. Sequence)* Low experience | | | 0.732** |
| | | | (0.293) |

Note: ** significant at 5%; *** significant at 1%. Robust standard errors in parentheses clustered at school level. Controls: (i) students characteristics (gender, age, nationality, parent's education, if the student missed, at most, one class per month, if the student has internet in his home), (ii) teacher characteristics (gender, age, general teaching experience (in years), if she/he has post-graduate certificates) and (iii) school characteristics (school size, 7th size, 7th repetition rate, FEPBA score in Language, school district –or location). Low experience represents a dummy variable equals to 1 if the teacher has less than two years of experience in teaching Science and zero otherwise. An interaction term between treatment status and the dichotomous variable of low experience is included. The number of observations for regression in column (1) is 1042, in column (2) 1072, and in column (3) 1102. All the regressions exclude the classrooms where the Human Body Unit was not taught.

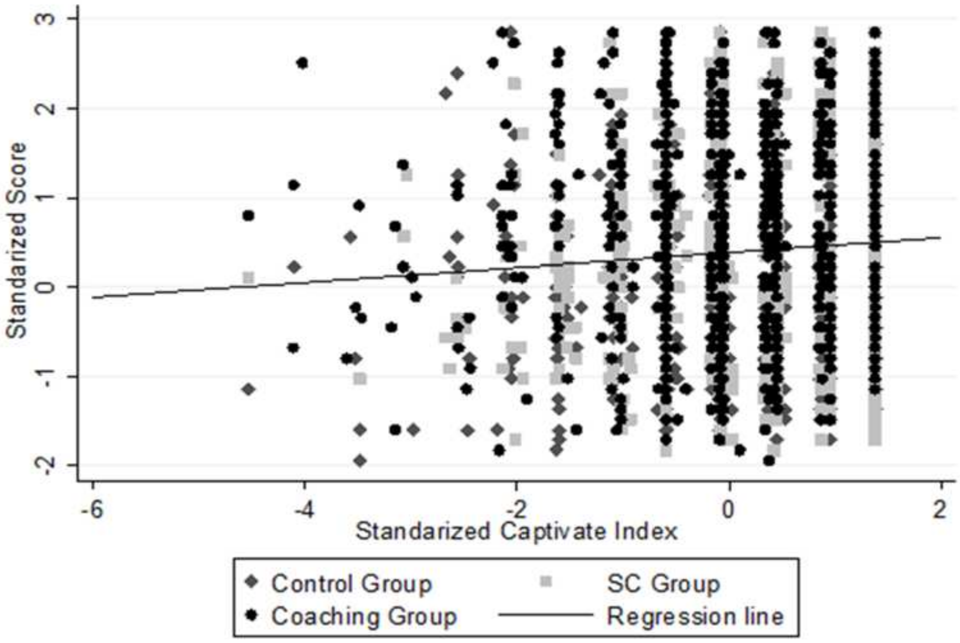
### 7.3. Effects on Perceptions

Beyond student performance, we explore whether the different treatments have any effects on students and teachers´ perceptions on Science teaching and learning. This is an important issue since research has shown that both teacher and student motivation and perceptions are associated with learning outcomes (Christophel, 1990; Bill and Melinda Gates Foundation, 2012). The next subsections explore these issues from the perspectives of the students (Section 7.3.1) and teachers (Section 7.3.2).

#### *7.3.1 Student Perceptions*

Students can be a primary source of information on the quality of teaching and the learning environment in individual classrooms (Bill and Melinda Gates Foundation, 2012). To explore whether our treatments affect student perceptions about learning Science, we constructed an index of captivate to evaluate whether teaching practices inspired curiosity and interest, and whether teachers were able to hold the student's attention in class and provide the basis for continuing interest. The construction of this index is explained in Appendix 4.

**Figure 1. Science scores and Captivate Index**



Note: Controls include: (i) students characteristics (gender, age, nationality, parent's education, if the student missed, at most, one class per month, if the student has internet in his home), (ii) teacher characteristics (gender, age, years of experience, if she/he has post-graduate certificates), and (iii) school characteristics (school size, 7th size, 7th repetition rate, FEPBA score in Language, school district –or location). Both the captivate index and the Science cores are standardized in terms of the Control Group. The captivate index is combined scale, whose construction is described in Appendix 4.

Figure 1 relates the captivate index (which ranges between -6 to 2)[22] to the Science test-scores (both variables are standardized in term of the Control Group). The figure shows that classrooms in which students rated their teachers higher on the captivate index tended also to produce greater average achievement gains. The black line, which shows the statistically significant partial correlation (0.08) of the scores and the captivate index controlling for students, teachers and schools' characteristics, confirms this relation.

Columns (1) and (2) of Table 9 reports the results of estimating equation (1) on the captivate index as well as on each separate question (with a 4-point scale) that conforms it, while column (3) shows the results of the Wald test, which evaluates the difference between the average treatment effect of the Sequence and Coaches Group . The results suggest that the sequence treatment seemed to be an effective instrument to enhance curiosity and interest among students.

[22] The value of -6 indicates that the student strongly disagrees in all the questions included in the index (see Appendix 4). This means that, for that student, teacher practices do not inspire curiosity and interest at all, or fails to keep his attention in class. In contrast, the value of 2 indicates that the student strongly agrees in all the questions that make up the index suggesting that, for that student, teacher practices do inspire curiosity and interest or are successful in keeping his full attention in class.

**Table 9. Effect on students' perception**

| ***Dependent variable*** | **Sequence vs. Control (1)** | **Coaches vs. Control (2)** | **Wald Test (3)** |
|---|---|---|---|
| Captivate index (A+B+C+D) | 0.202** | 0.051 | 1.91 |
| | (0.096) | (0.117) | (0.166) |
| A. This class (of Science) keeps my attention | 0.100** | 0.118** | 0.13 |
| | (0.044) | (0.046) | (0.722) |
| B. My teacher (of Science) makes learning enjoyable | 0.064 | -0.049 | 2.82 |
| | (0.063) | (0.067) | (0.093) |
| C. My teacher (of Science) makes lessons interesting | 0.118 | -0.012 | 3.10 |
| | (0.073) | (0.078) | (0.078) |
| D. I like the way we learn in this class (of Science) | 0.127*** | 0.062 | 1.27 |
| | (0.043) | (0.064) | (0.260) |
| E. I like the class of Science | 0.153*** | 0.151** | 0.00 |
| | (0.055) | (0.070) | (0.970) |

Note: ** significant at 5%; *** significant at 1%. Robust standard errors in parentheses clustered at school level. Controls: (i) students characteristics (gender, age, nationality, parent's education, if the student missed, at most, one class per month, if the student has internet in his home and if he has a car), (ii) teacher characteristics (gender, age, years of experience, if she/he has post-graduate certificates), and (iii) school characteristics (school size, 7th size, 7th repetition rate, FEPBA score in Language, school district –or location). The dependent variable in A-E represents a 4-point-scale, where 1 means strongly disagree, 2 disagree, 3 agree and 4 strongly agree. The captivate index is combined scale, which construction is described in Appendix 4. The number of observations for regression in column (1) is 974, and in column (2) 964. All the regressions exclude the classrooms where the Human Body Unit was not taught.

Picking up on the gender perspective previously considered, we explore if girls experienced more gains in the captivate index than boys. Table A9.1 in Appendix 9 presents the estimation of the equation (1) using as dependent variable the captivate index for female and male students. Results show that female students from the Sequence Group were more interested over their male classmates in comparison with the Control Group. Indeed, the captivate index for female students is 35% of a standard deviation higher than that of the Control Group. In contrast, the Coaches treatment seems to reduce the captivate index for females in comparison with the Sequence Group, which does not happen when we restrict the sample to male students.

### *7.3.2 Teacher Perceptions*

Finally, this subsection explores the effect of the intervention on how teachers perceived their experience and the effects they observed on their students. For that, we constructed a 4-point-scale[23] to measure to what extent did teachers agree with the following statements: A. I feel that the way I teach Science changed a lot; B. I liked or enjoyed more teaching Science than

[23] In the 4-point-scale, 1 represents "strongly disagree", 2 "disagree", 3 "agree" and 4 "strongly agree".

previous years; C. I feel that by implementing the ideas of this training my students learned more in comparison with other groups and/or subjects; D. I feel my students developed more skills than in previous years; and E. I taught more hours of Science classes.

Column (1) and (2) of Table 10 displays the results of estimating the effect of the treatments on these variables, controlling for teacher characteristics. Our findings suggest that both the sequence and coaches treatments favorably changed teacher perceptions on their practices and their expectations of student learning. Compared to the Control Group, teachers in the Sequence Group (Coaches Group) present a scale 86% (95%) higher in their perception than their teaching practices meaningfully changed. In a similar vein, teachers in the Sequence Group (Coaches Group) present a scale 72% (97%) higher in their perception that they enjoyed more teaching Science in comparison with teachers of the Control Group. Furthermore, teachers in the Sequence Group (Coaches Group) present a scale between 63% and 67% (77% and 88%) higher in their perception that students learned more and developed more skills than teachers in the Control Group. Finally, according to column (3), which shows the results of a test that evaluates the difference between the coefficients in columns (1) and (2), teachers in the Coaches Group expressed that they taught more hours of Science than teachers in the Control and Sequence groups. This is important since evidence shows that more hours of class are associated with more learning (OECD, 2016). All these differences are statistically significant.

**Table 10. Teacher perceptions**

| ***Dependent variable*** | **Sequence vs. Control (1)** | **Coaches vs. Control (2)** | **Wald test (3)** |
|---|---|---|---|
| A. I feel it the way I teach Science changes a lot | 0.864*** | 0.953*** | 0.14 |
| | (0.312) | (0.247) | (0.708) |
| B. I like and/or enjoy more teaching Science than in previous years | 0.717** | 0.974*** | 0.93 |
| | (0.345) | (0.307) | (0.334) |
| C. I feel that by implementing the ideas of this training, my students learned more in comparison with other groups and/or subjects | 0.632 | 0.769** | 0.30 |
| | (0.333) | (0.324) | (0.586) |
| D. I feel my students develop more skills than in previous years | 0.672** | 0.876*** | 0.86 |
| | (0.293) | (0.306) | (0.354) |
| E. I taught more hours of Science classes | 0.260 | 1.107*** | 18.33 |
| | (0.247) | (0.275) | (0.000) |

Note: ** significant at 5%; *** significant at 1%. Robust standard errors in parentheses clustered at school level. Controls: teacher characteristics (gender, age, years of experience, if she/he has post-graduate certificates). The dependent variables are 4-point-scale, where 1 means strongly disagree, 2 disagree, 3 agree and 4 strongly agree.

## 8. Conclusion

This study used a randomized controlled trial to assess the impact of different CPD approaches on student Science learning. We randomly assigned the 70 participating schools to one of three conditions: (1) short-term teacher training (Control Group), (2) short-term teacher training complemented with a structured curriculum unit (Sequence Group), and (3) short-term teacher training complemented with both a structured curriculum unit and tutoring of pedagogical coaches (Coaches Group). The study included 2965 students and 99 teachers in the seventh grade of public schools in CABA, Argentina. The experiment was internally valid and performed on a representative sample of schools of CABA.

We find that providing teachers with a structured curriculum unit increased student performance by 55% of a standard deviation compared to a short-term training session. This finding is consistent with the literature that shows that developing high quality structured curriculum units are valid instruments for assisting teachers and promoting more effective teaching and learning (Brown, 2009). The structured curriculum unit also sparked interest and curiosity amongst students. Using an index that measures if learning was interesting and relevant as well as if teaching practices inspired curiosity, we find that students in the Sequence Group presented a scale 20% of a standard deviation higher than those in the Control Group. This finding is in line with research that shows that motivation is an essential factor to generate and sustain student learning (Ercan, Ural and Ates, 2016).

We also find that student in the Coaches Group learned significantly more than those in the Control Group. Specifically, students whose teachers had a coach learned about 64% of a standard deviation more than those in the Control Group. However, according to our results, there is no general additional benefit in terms of student learning between the structured curriculum unit with and without on-going coaching. Nevertheless, we find that the marginal effect of coaches is statistically significant for relatively inexperienced teachers in Science education. Specifically, student in the Coaches Group learned 82% of a standard deviation more than students in the Sequence Group when we consider the least experienced teachers. And this is particularly true when we focus on higher-order skills, which may require more specific teaching. This points that coaches should be targeted to teachers who have little prior experience in teaching science and focus their support on getting teachers to master the teaching of cognitively demanding activities, rather than simply implement basic active learning strategies (which teachers seem to be able to pick up alone by just working with a structured curriculum unit).

Additionally, our results show that the average treatment effect of the coaches is higher for girls than boys. This is an encouraging finding as not only girls´ participation and achievement in school science (Liben and Coyle, 2014) but also the choice of careers related to Science shows a noticeable gender gap (Beede, Julian, Langdon, McKittrick, Khan and Doms, 2011). This result may indicate that coaches help teachers create more gender inclusive Science lessons, supporting them in strategies that better cater to active participation of girls. In this sense, one important value of working with coaches would be promoting more female engagement in Science.

Both the structured curriculum unit and the pedagogical coaching favorably change teacher satisfaction with their practice and their perceptions of student learning. Compared to the Control Group, teachers in the Sequence Group and Coaches Group present a scale between 63% and 100% higher in their perception that their teaching practices meaningfully changed and that students learned more and developed more skills, that they enjoyed more teaching Science and that they taught more hours of Science.

The first policy advice that emerges from our study is that a short-term teacher training complemented with a structured curriculum is a cost-effective CPD intervention to increase student learning on Science. Specifically, complementing training sessions with a structured curriculum unit costs (per student) 0.84 dollars per 0.1 standard deviations, this means that it costs 0.84 dollars to move a child from the 50th to the 53th percentile approximately.

The second policy advice is that additional coaching does make a difference in student scores, but only for relatively inexperienced teachers in Science. This finding suggests that experienced teachers already have the pedagogical toolkit that enables them to confidently implement the lessons outlined in the curriculum unit, at least up to a basic level. For less experienced teachers, coaching can bridge the gap between structured lesson plans and the complex world of the actual science classroom.

This suggests that improving teachers' practice in Science is not a matter of choosing the best ("one size fits all") CPD strategy, but selecting the strategy that suits best the specific population of teachers and student learning goals being targeted. These are relevant contributions for public policies focused on CPD interventions since hiring, training and providing coaches is an expensive and human-resource intensive approach. Our study shows that providing teachers with a structured curriculum is 2.7 times more cost-effective for the total score than complementing it with on-going coaching).

In all, our study speaks to the need to tailor CPD interventions in order to maximize their effects based on evidence of what works better and taking into account the cost-effectiveness of each strategy. We believe that evidence of this nature is urgent and necessary for the development of effective public policies aimed at promoting students scientific literacy and thus effective participation in the global knowledge economy.

## References

Abeberese, A. B., Kumler, T. J., and Linden, L. L. (2014). Improving reading skills by encouraging children to read in school: A randomized evaluation of the Sa Aklat Sisikat reading program in the Philippines. *Journal of Human Resources*, 49(3), 611-633.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., and Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, 333, 1034-1037.

Angrist, J. and Lavy, V. (2001). Does Teacher Training Effect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools. *Journal of Labor Economics*, 19(2), 343-369.

Angrist, J., and Lavy, V. (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, 114 (2), 533_575.

Angrist, J., Bettinger, E., Bloom, E., King, E., and Kremer, M. (2002). Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. *American Economic Review*, 92, 1535–1558. doi:10.1257/000282802762024629

Arancibia, V., Popova, A., and Evans, D. K. (2016). Training Teachers on the Job: What Works and How to Measure it. Policy Research Working Paper, No. 7834. World Bank.

Argentine Ministry of Education (2007). *Mejorar la enseñanza de las ciencias y las matemáticas - Una prioridad nacional*. Buenos Aires: Argentine Ministry of Education. Retrieved from: http://repositorio.educacion.gov.ar/dspace/handle/123456789/95085

Argentine Ministry of Education and Sports (2017). *Aprender 2016. Primer informe de resultados*. Buenos Aires: Argentine Ministry of Education and Sports.

Argentine Ministry of Education (2015). *Presentación Nuestra Escuela - Programa Nacional de Formación Docente.* Retrieved from: http://nuestraescuela.educacion.gov.ar/pdf/presentacionnuestraescuela.pdf

Argentine National Institute of Teacher Training. (2016). *Plan Nacional de Formación Docente 2016-2021*. Buenos Aires: Argentine National Institute of Teacher Training. Retrieved from: http://cedoc.infd.edu.ar/upload/Plan_Nacional_de_Formacion_Docente1.pdf

Arias, A. M., Davis, E. A., Marino, J. C., Kademian, S. M., and Palincsar, A. S. (2016). Teachers' use of educative curriculum materials to engage students in science practices. *International Journal of Science Education*, 38(9), 1504-1526.

Banerjee, A., Cole, S., Duflo, E., and Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *Quarterly Journal of Economics*, 122, 1235–1264. doi:10.1162/qjec.122.3.1235.

Barrera-Osorio, F., and Linden, L. L. (2009). The use and misuse of computers in education: Evidence from a randomized controlled trial of a language arts program. Unpublished manuscript, Columbia University, New York, NY.

Bassi, M., Meghir, C., and Reynoso, A. (2016). Education Quality and Teaching Practices. National Bureau of Economic Research, No. w22719.

Beede, D. N., Julian, T. A., Langdon, D., McKittrick, G., Khan, B., and Doms, M. E. (2011). Women in STEM: A gender gap to innovation. *Economics and Statistics Administration, Issue Brief No. 04-11* Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1964782

Berlinski, S., and Busso, M. (2017). Challenges in educational reform: An experiment on active learning in mathematics. *Economics Letters, 156,* 172-175.

Bill and Melinda Gates Foundation. (2012). Asking Students about Teaching. Student Perception Surveys and Their Implementation. Met Project. Policy and Practice Brief.

Brown, M. W. (2009). The teacher-tool relationship: Theorizing the Design and Use of Curriculum Materials. In Remillard, J. T., Herbel-Eisenmann, B. A., and Lloyd, G. M. (Eds.). (2011). *Mathematics teachers at work: Connecting curriculum materials and classroom instruction*. New York: Routledge.

Bruns, B., & Luque, J. (2014). *Profesores excelentes. Docentes excelentes: Cómo mejorar el aprendizaje en América Latina y el Caribe.* Washington, DC: Banco Mundial. Retrieved from http://www. bancomundial. org/content/dam/Worldbank/Highlights% 20&% 20Features/lac/LC5/Spanish-excellent-teachers-report. pdf.

Campbell, P. F., and Malkus, N. N. (2011). The impact of elementary mathematics coaches on student achievement. *The Elementary School Journal*, 111(3), 430-454.

Christophel, D. M. (1990). The relationships among teacher immediacy behaviors, student motivation, and learning. *Communication education*, *39*(4), 323-340.

Cristia, J. P., Ibarrarán, P., Cueto, S., Santiago, A., and Severín, E. (2012). *Technology and child development: Evidence from the One Laptop per Child program* (Working Paper No. IDB-WP-304). Washington, DC: Inter-American Development Bank.

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., and Orphanos, S. (2009). *Professional learning in the learning profession*. Washington, DC: National Staff Development Council. Retrieved from http://www.ostrc.org/docs/document_library/ppd/Professionalism/Professional%20Learning%20in%20the%20Learning%20Profession.pdf

Das, J., S. Dercon, J. Habyarimana, P. Krishnan, K. Muralidharan, and V. Sundararaman (2013).When can school inputs improve test scores? *American Economic Journal: Applied Economics*, 5(2), 29-57.

Davis, E. A., Janssen, F. J., and Van Driel, J. H. (2016). Teachers and science curriculum materials: where we are and where we need to go. *Studies in Science Education*, 52(2), 127-160.

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

DINIECE. (2004). *Programme for International Student Assessment. Informe Nacional República Argentina*. Buenos Aires: DiNIECE. Retrieved from http://repositorio.educacion.gov.ar/dspace/handle/123456789/55289

DINIECE. (2015). *Anuario Estadístico Educativo. Relevamientos anuales 2007-2014.* Buenos Aires: DiNIECE. Retrieved from http://portales.educacion.gov.ar/diniece/

Duflo, E, Hanna, R., and Ryan, S. (2012). Incentives work: Getting teachers to come to school. *American Economic Review,* 102(4), 1241-1278.

Duflo, E., Dupas, P., and Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review*, 101(5), 1739-1774.

Educ.ar. (2005). *Anuario-En cifras.* Retrieved from http://portal.educ.ar/acercade/anuarios/2006/cifras.html

Ercan, O., Ural, E., and Ateş, D. (2016). The Effect of Educational Software Based on Ausubel's Expository Learning on Students' Academic Achievement, Science and Computer Attitudes: “Human and Environment” Unit Example. *British Journal of Education, Society & Behavioural Science*, 14(1): 1-10.

Forbes, C. T., and Davis, E. A. (2010). Curriculum design for inquiry: Preservice elementary teachers' mobilization and adaptation of science curriculum materials. *Journal of research in science teaching*, 47(7), 820-839.

Fredriksson, P., Ockert, B., and Oosterbeek, H. (2012). Long term effects of class size. *Quarterly Journal of Economics*, 249-285.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P., Sepanik, S. and Doolittle, F. (2011). Middle School Mathematics Professional Development Impact Study: Findings after the Second Year of Implementation. NCEE 2011-4024. National Center for Education Evaluation and Regional Assistance.

Glewwe, P., Kremer, M. and Moulin, S. (2009). Many Children Left Behind? Textbooks and Test Scores in Kenya. *American Economic Journal: Applied Economics*, 1(1), 112-35.

Glewwe, P., Kremer, M., Moulin, S., and Zitzewitz, E. (2004). Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. *Journal of development Economics*, 74(1), 251-268.

Glewwe, P., N. Ilian, and M. Kremer (2010). Teacher incentives. *American Economic Journal: Applied Economics*, 2(3), 205-227.

Gulamhussein, A. (2013). Teaching the teachers: Effective professional development in an era of high stakes accountability. *Center for Public Education*, 1-47. Retrieved from goo.gl/DYNuct

Harris, C. J., Penuel, W. R., DeBarger, A., D'Angelo, C., and Gallagher, L. P. (2014) *Curriculum materials make a difference for Next Generation Science Learning: Results from year 1 of a randomized controlled trial*. Menlo Park, CA: SRI International. Retrieved from https://www.sri.com/sites/default/files/publications/pbis-efficacy-study-y1-outcomes-report-2014_0.pdf

He, F., L. Linden, and M. Margaret (2009). A better way to teach children to read? evidence from a randomized control trial. *Unpublished manuscript*, Columbia University.

He, F., Linden, L. L., and MacLeod, M. (2008). How to teach English in India: Testing the relative productivity of instruction methods with Pratham English Language Education Program. *Unpublished manuscript*, Columbia University, New York, NY.

Jacob, B. A., and Lefgren, L. (2004a). Remedial education and student achievement: A regressiondiscontinuity approach. *Review of Economics and Statistics*, 86(1), 226-244.

Jacob, B. and Lefgren, L. (2004b) The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from School Reform Efforts in Chicago. *Journal of Human Resources*, 39(1), 50-79.

Kraft, M. A., and Blazar, D. (2016). Individualized coaching to improve teacher practice across grades and subjects: New experimental evidence. *Educational Policy* (advance online publication). Retrieved from http://journals.sagepub.com/doi/abs/10.1177/0895904816631099

Kraft, M. A., Blazar, D., and Hogan, D. (2016). The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence. Brown University Working Paper.

Kretlow, A. G., and Bartholomew, C. C. (2010). Using coaching to improve the fidelity of evidence-based practices: A review of studies. *Teacher Education and Special*

*Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, 33(4). 279 – 299.

Krueger, A. and Whitmore, D. (2002). “Would Smaller Classes Help Close the Black-White Achievement Gap?” In John E. Chubb and Tom Loveless, (eds.), *Bridging the Achievement Gap*. Washington: Brookings Institution Press.

Liben, L. S., and Coyle, E. F. (2014). Chapter three-developmental interventions to address the STEM gender gap: exploring intended and unintended consequences. *Advances in child development and behavior*, 47, 77-115.

Linden, L. L. (2008). Complement or substitute? The effect of technology on student achievement in India. Unpublished manuscript, Columbia University, New York, NY.

Martin, M. O., Mullis, I. V. S., Foy, P., and Hooper, M. (2016). *TIMSS 2015 International Results in Science*. Boston College: TIMSS & PIRLS International Study Center. Retrieved from http://timssandpirls.bc.edu/timss2015/international-results/

Matsumura, L. C., Garnier, H. E., Correnti, R., Junker, B., and DiPrima Bickel, D. (2010). Investigating the effectiveness of a comprehensive literacy coaching program in schools with high teacher mobility. *The Elementary School Journal*, 111(1), 35-62.

McEwan, P. J. (2015). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of Educational Research*, 85(3), 353-394.

Minner, D. D., Levy, A. J., and Century, J. (2010). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of research in science teaching*, 47(4), 474-496.

Mo, D., Zhang, L., Luo, R., Qu, Q., Huang, W., Wang, J., Qiao, Y., Boswell, M., and Rozelle, S. (2014). Integrating computer-assisted learning into a regular curriculum: evidence from a randomised experiment in rural schools in Shaanxi. *Journal of development effectiveness*, 6(3), 300-323.

Muralidharan, K., Singh, A., and Ganimian, A. J. (2016). Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India. National Bureau of Economic Research, No. w22923.

Näslund-Hadley, E., and Bando, R. (Eds.) (2016). *Todos los niños cuentan: enseñanza temprana de las matemáticas y ciencias en América Latina y el Caribe*. Washington, DC: Banco Interamericano de Desarrollo. Retrieved from http://doi.org/http://dx.doi.org/10.18235/0000226#sthash.7MqRqw2c.dpuf

Novak, J. D. (2005). Results and implications of a 12-year longitudinal study of science concept learning. *Research in Science Education*, 35(1), 23-40

OECD (2014), PISA 2012 *Results: What Students Know and Can Do (Volume I, Revised edition) Student Performance in Mathematics, Reading and Science*. Paris: OECD Publishing.

OECD (2016). *PISA 2015 Results (Volume 1): Excellence and Equity in Education.* Paris: OECD Publishing.

Organización de las Naciones Unidas para la Educación la Ciencia y la Cultura (UNESCO) (2009). *Aportes para la enseñanza de las Ciencias Naturales: Segundo estudio Regional Comparativo y Explicativo (SERCE)*. Santiago de Chile: OREALC/UNESCO.

Organización de las Naciones Unidas para la Educación la Ciencia y la Cultura (UNESO). (2016). *Informe de resultados TERCE. Logros De Aprendizaje.* Santiago de Chile: OREALC/UNESCO. Retrieved from http://unesdoc.unesco.org/images/0024/002435/243532S.pdfUNESCO 2015

Sailors, M., and Price, L. (2010). Professional Development that supports the teaching of cognitive reading strategy instruction. *Elementary School Journal*, 110, 301e323.

Sailors, M., and Price, L. (2015). Support for the Improvement of Practices through Intensive Coaching (SIPIC): A model of coaching for improving reading instruction and reading achievement. *Teaching and Teacher Education*, 45, 115-127.

Serra, J. C. (2001) "La política de capacitación docente en Argentina: La Red Federal de Formación Docente Continua (1994-1999)." Ministry of Education, Argentina

Sloan, H. A. (1993). Direct instruction in fourth and fifth grade classrooms. *Dissertation Abstracts International*, 54(08), 2837A.

Stanger-Hall, K. F. (2012). Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. *CBE-Life Sciences Education*, *11*(3), 294-306.

Urquiola, M. (2006). Identifying class size effects in developing countries: Evidence from rural bolivia. *Review of Economics and Statistics*, 88 (1), 171-177.

Valverde, G. and Näslund-Hadley, E. (2010). *La condición de la educación en matemáticas y ciencias naturales en América Latina y el Caribe*. Technical Notes, IDB-TN-211. Washington, DC: InterAmerican Developmnet Bank (IADB).

Vegas, E., Ganimian, A. and Bos, M. S. (2014). *América Latina en PISA 2012: ¿Cuántos estudiantes tienen bajo desempeño?* Washington, DC: InterAmerican Developmnet Bank (IADB). Retrieved from https://publications.iadb.org/handle/11319/701

World Bank. (2014). *Argentina—Country partnership strategy for the period of FY2015-18.* Washington, DC: World Bank Group. Retrieved from http://documents.worldbank.org/curated/en/846861468210572315/Argentina-Country-partnership-strategy-for-the-period-of-FY2015-18

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., and Shapley, K. L. (2007). *Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement.* (Issues & Answers Report, REL 2007-No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation an Regional Assistance, Regional Educational Laboratory Southwest.

# Appendixes

## Appendix 1: Timeline

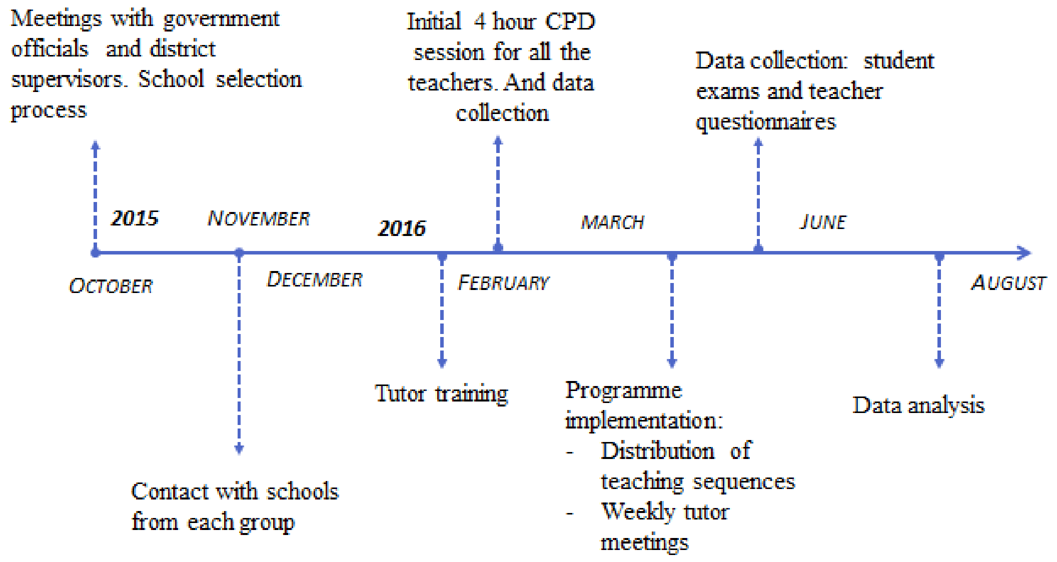
We started the school selection process in October 2015 (see Figure A3.1). We selected 75 primary state schools from six school districts of CABA. We notified the Ministry of Education of Ciudad de Buenos Aires of the lottery results the first days of November, and then the Ministry communicated the results to the schools. We invited those schools to participate in the experiment during November 2015; 70 schools agreed to participate and were effectively included in the experiment. We organized meetings with the schools supervisors to inform them how the experiment would be implemented in December 2015. During these meeting, supervisors received a letter with the following information: description of the project, main objectives, grade and topic to be covered, teacher training, a calendar for the experiment implementation, and contact details.

Pedagogical Coaches were recruited by the Education School of Universidad de San Andrés to carry out weekly sessions with teachers from Group C. The main aim of these sessions was to guide and coach them on how to implement the structured curriculum as well as reflect on their practice at the end of each week. They received an initial training session during February 2016 where key aspects of Science inquiry were discussed and revising the curriculum unit. Throughout the intervention they also attended regular training sessions every fortnight and had access to an extensive library of guiding documents and videos.

The intervention started at the end of February 2016 when all the teachers of the sample received an initial 4 hour CPD session. During this session, teachers in the Sequence and Coaches Group received the structured curriculum unit. Meanwhile, Coaches coordinated with teachers the meeting agenda. Coaches visited the teachers at their school during free periods over 12 weeks. The first Coaches meeting was carried out in mid-March and the last in mid-June.

The data collection process was carried out in two stages. First, at the beginning of the intervention (late February) we collected information about schools, teachers and students, which served as our baseline. Then, at the end of the intervention (late June) we evaluated students as well as collected information about teachers and schools.

## Figure A3.1 Intervention timeline


Meetings with government officials and district supervisors. School selection process
Initial 4 hour CPD session for all the teachers. And data collection
Data collection: student exams and teacher questionnaires
2015
NOVEMBER
2016
MARCH
JUNE
OCTOBER
DECEMBER
FEBRUARY
AUGUST
Tutor training
Programme implementation:
- Distribution of teaching sequences
- Weekly tutor meetings
Data analysis
Contact with schools from each group

## Appendix 2: Descriptive Statistics of Coaches

### Table A2.1 Descriptive Statistics of Coaches

| | Age | Maximum educational level | Field | Seniority (in years) |
|---|---|---|---|---|
| Coach 1 | 45 - 49 | Graduate Degree | Science Education | 20 |
| Coach 2 | 40 - 44 | B.A. | Biology and Chemistry Education | 14 |
| Coach 3 | 40 - 44 | Ph.D. | Biology | 10 |
| Coach 4 | 30 - 34 | B.A. | Health Science | 7 |
| Coach 5 | 30 - 34 | M.A. | Science Education | 7 |
| Coach 6 | 35 - 39 | M.A. | Science Education | 8 |
| Coach 7 | 30 - 34 | M.A | Science Education | 6 |
| Coach 8 | 25 - 29 | B.Sc. | Biology | 5 |
| Coach 9 | 25 - 29 | M.A. | Science Education | 3 |
| Coach 10 | 25 - 29 | B.A. | Education | 4 |

## Appendix 3: Tests Items

### Table A3.1 Description of the levels of skills addressed in the Science Test

| Basic skills | Medium skills | Higher-order skills |
|---|---|---|
| Recall of scientific knowledge | Describe scientific phenomena. | Identify research questions. |
| Read information from tables. | Interpret conclusions from Science experiments. | Design Science experiments to test hypothesis. |
| | | Explain more complex Science phenomena (such as the integration of the systems in the Human Body). |

## Appendix 4. Student Perception: Captivate Index

Throughout the student questionnaire we asked students a series of questions about specific teachers and their practices in order to construct an index that aims to measure if teaching practices inspire curiosity and interest and whether teachers are able to hold the student's attention in class and provide the basis for continuing interest. The questions were based on the Student Tripod Survey, a pre-specified well-known questionnaire validated by the Measures of Effective Teaching (MET) project sponsored by the Bill and Melinda Gates Foundation.[24]

Column (1) of Table A4.1 lists the questions used to build this index, which we named Captivate Index. These questions were randomly mixed in the student survey instrument. Students gave categorical answers of the type "strongly agree", "agree", "disagree" and "strongly disagree". We aggregated these answers into an index using a maximum likelihood principal components estimator. According to our estimation, only one factor is retained because it has an eigenvalue over one. Specifically, the Eigenvalue is 2.327 and the Cronbach's alpha reliability coefficient for our sample is 0.759. Column (2) shows the loading associated with each variable.

After the prediction was computed to produce the index, we standardized it using the mean and standard deviation of the Control Group. The index ranges between -6 to 2, where -6 indicates that the student strongly disagrees in all the questions included in the index. This means that, for that student, teacher practices do not inspire curiosity and interest or fails to keep his attention in class. In contrast 2 indicates that the student strongly agrees in all the questions that make up the index suggesting that, for that student, teacher practices do inspire curiosity and interest or are successful in keeping his full attention in class.

**Table A4.1 Captivate Index**

| Question | Factor Loadings |
|---|---|
| A. I feel it the way I teach Science changes a lot. | 0.666 |
| B. I like and/or enjoy more teaching Science than in previous years. | 0.793 |
| C. I feel that by implementing the ideas of this training, my students learned more in comparison with other groups and/or subjects. | 0.802 |
| D. I feel my students develop more skills than in previous years. | 0.782 |

[24] Tripod survey is the US's leading provider of classroom-level survey assessments for K-12 education. Tripod surveys are in their 18th generation, refined over more than a decade of field experience and in response to valuable feedback from educators. For further information see http://tripoded.com.

## Appendix 5: Sample representativeness

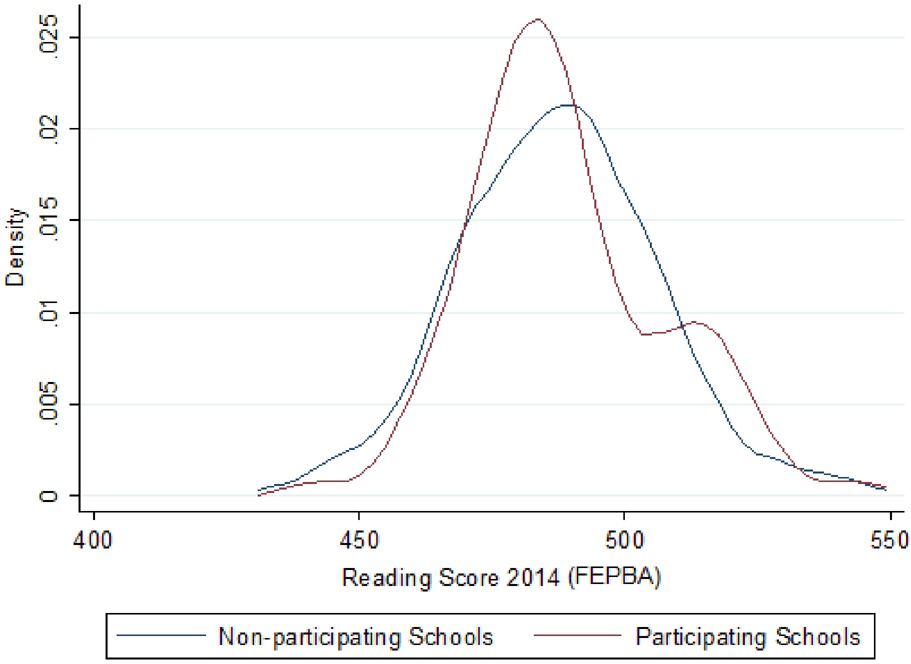
The following table compares the average characteristics of the 70 participating schools with the characteristics of the non-participating primary state schools in CABA. Then, Figure A5.1 shows the distribution of study participants and non-study participants' scores in the FEBPA test.

**Table A5.1 Mean test between non-participating and participating schools**

| Variable | Non-participating schools | | Participating schools | | Difference |
|---|---|---|---|---|---|
| | N | Average | N | Average | |
| School enrollment | 385 | 321.51 | 70 | 298.50 | |
| 7th grade enrollment | 385 | 45.30 | 70 | 40.67 | |
| Number of 7th classrooms | 385 | 2.12 | 70 | 1.97 | |
| School promotion rate (%) | 385 | 97.56 | 70 | 97.32 | |
| School drop-out rate (%) | 385 | 0.17 | 70 | 0.17 | |
| School repetition rate (%) | 385 | 2.06 | 70 | 2.47 | * |
| School over-aged rate (%) | 385 | 14.28 | 70 | 15.02 | |
| 7th grade score in language FEBPA test | 341 | 487.45 | 70 | 488.74 | |
| Social Vulnerability Index | 371 | 0.18 | 70 | 0.17 | |

Note: N indicates the number of schools. * indicates that the difference of means test is significant at 10%.

**Figure A4.1 Distribution of FEBPA scores**



## Appendix 6: Balance across experimental groups

**Table A1.6. Pre-treatment characteristics (*continuation*)**

| | **All simple** | | | **Control Group** | | **Sequence Group** | | **Coaches Group** | |
|---|---|---|---|---|---|---|---|---|---|
| | **N** | **Mean** | **Sd** | **Mean** | **Sd** | **Mean** | **Sd** | **Mean** | **Sd** |
| ***Student-level variables*** | | | | | | | | | |
| Took holidays in the last two years | 2297 | 0.8 | 0.4 | 0.78 | 0.41 | 0.81 | 0.4 | 0.8 | 0.4 |
| Scored "Very Good" in Language | 2249 | 0.45 | 0.5 | 0.46 | 0.5 | 0.45 | 0.5 | 0.43 | 0.5 |
| Scored "Very Good" in Mathematics | 2243 | 0.44 | 0.5 | 0.46 | 0.5 | 0.44 | 0.5 | 0.43 | 0.5 |
| ***Teacher-level variables*** | | | | | | | | | |
| Seniority in teaching 7th grade (in years) | 88 | 4.39 | 5.1 | 4.57 | 5.11 | 4.35 | 3.36 | 4.23 | 6.48 |
| Seniority in teaching at the current school (in years) | 87 | 3.6 | 3.65 | 0.06 | 0.24 | 0.07 | 0.26 | 0.17 | 0.38 |
| ***School-level variables*** | | | | | | | | | |
| Percent of schools with double school day | 70 | 0.39 | 0.49 | 0.46 | 0.51 | 0.35 | 0.49 | 0.35 | 0.49 |
| Classrooms of 7th grade | 70 | 1.99 | 0.81 | 2.08 | 0.93 | 1.96 | 0.71 | 1.91 | 0.79 |
| Percent of schools in school-district: | | | | | | | | | |
| 2 | 70 | 0.17 | 0.38 | 0.21 | 0.42 | 0.09 | 0.29 | 0.22 | 0.42 |
| 3 | 70 | 0.19 | 0.39 | 0.21 | 0.42 | 0.17 | 0.39 | 0.17 | 0.39 |
| 4 | 70 | 0.16 | 0.37 | 0.21 | 0.42 | 0.13 | 0.34 | 0.13 | 0.34 |
| 5 | 70 | 0.14 | 0.35 | 0.08 | 0.28 | 0.17 | 0.39 | 0.17 | 0.39 |
| 7 | 70 | 0.17 | 0.38 | 0.13 | 0.34 | 0.3 | 0.47 | 0.09 | 0.29 |
| 8 | 70 | 0.17 | 0.38 | 0.17 | 0.38 | 0.13 | 0.34 | 0.22 | 0.42 |

Note: N means number of observation in the full sample and Sd means standard deviation

**Table A6.2. Balance across Treatments (*continuation)***

| | Sequence vs. Control | Coaches vs. Control | Coaches vs. Sequence |
|---|---|---|---|
| *Student-level variables* | | | |
| Took holidays in the last two years | 0.03 | 0.02 | 0.00 |
| Scored "Very Good" in Language | -0.01 | -0.02 | -0.01 |
| Scored "Very Good" in Mathematics | -0.02 | -0.03 | -0.01 |
| *Teacher-level variables* | | | |
| Seniority in teaching 7th grade (in years) | -0.22 | -0.34 | -0.12 |
| Seniority in teaching at the current school (in years) | 0.58 | -0.29 | -0.87 |
| *School-level variables* | | | |
| Percent of schools with double school day | -0.11 | -0.11 | 0.00 |
| Classrooms of 7th grade | -0.13 | -0.17 | -0.04 |
| Percent of schools of district: | | | |
| 2 | -0.12 | 0.01 | 0.13 |
| 3 | -0.03 | -0.03 | 0.00 |
| 4 | -0.08 | -0.08 | 0.00 |
| 5 | 0.09 | 0.09 | 0.00 |
| 7 | 0.18 | -0.04 | -0.22* |
| 8 | -0.04 | 0.05 | 0.09 |

Note: Each entry indicates the mean difference between the two experimental groups in the column for the corresponding variable in each line. * indicates that the difference of means test is significant at 10%; ** significant at 5%; *** significant at 1%.

## Appendix 7: Attrition

This Appendix explores the attrition of the sample. If individuals move in or out of the sample randomly, then the design would have a change in power, but there is no need to make further adjustments besides documenting. One way to check whether individuals move in or out of the sample randomly is to test if individuals who move out of the sample are different from those that did not migrate (Card, Ibarraran and Villa, 2011).

As mentioned in Section 5, the attrition rate is 5% of classrooms that belonged only to the Control Group. In what follows we inspect the differences in the basic characteristics between students, teachers and schools that remained in the sample and those who left. The first column of Table A7.1 shows the differences between those who left the sample in the Control Group and those who remained, while the second and third columns show the differences within treatment and controls in terms of the realized sample (i.e. the sample that includes the 67 schools, or the 129 classrooms of 7th grade). Regarding the first column, we can see that there are no practically differences, at 95% of confidence, between those who left the sample in the Control Group and those who remained. The only significant difference is observed in student nationality, at 99% of confidence. However, this difference vanishes when we compare the control with the treatment groups –both the Sequence and Coaches Group –suggesting that the balance between groups remains the same (columns 2 and 3). Although there are some small statistical differences (at 90% of confidence), they are substantially very small, therefore, the balance is maintained overall.

As in Table 3, the realized sample of treatments and Control Groups do not differ significantly in any observable dimension except in age. This is the only variable with a statistical difference at the 99% level of confidence for the control and Sequence Groups (students in the former group are slightly younger than those in the latter). Nevertheless, this difference is very small and vanishes when we consider 7th grade repetition rates, which are balanced across the three experimental groups.

## Table A7.1. Mean differences

| | Left the sample vs. Control Group | Realized Sample | |
|---|---|---|---|
| | | Sequence vs. Control | Coaches vs. Control |
| ***Student-level variables*** | | | |
| Percent female | -0.001 | 0.026 | 0.021 |
| Age | 0.081* | -0.07*** | -0.023 |
| Percent of Argentines | -0.141*** | 0.015 | 0.022 |
| Mother or father education (secondary) | -0.082* | 0.024 | 0.051* |
| Have internet in their home | -0.02 | 0.028* | 0.025 |
| Have air conditioning in their home | -0.03 | 0.007 | 0.023 |
| Have at least one car in their home | -0.025 | -0.015 | 0.02 |
| Took holidays in the last two years | -0.016 | -0.024 | -0.02 |
| At most, missed one class per month | -0.01 | 0.018 | 0.024 |
| Scored "Very Good" in Language | -0.015 | 0.015 | 0.027 |
| Scored "Very Good" in Mathematics | 0.023 | 0.017 | 0.024 |
| ***Teacher-level variables*** | | | |
| Percent female | 0.172 | 0.061 | 0.069 |
| Age | -3.503 | 2.304 | 2.647 |
| Percent with Post-Graduate Certificate | -0.179 | 0.14 | 0.069 |
| Percent with Universitary degree | -1.572 | 1.604 | 1.607 |
| Seniority in teaching (in years) | 0.166 | 0.003 | 0.138* |
| Seniority in teaching 7th grade (in years) | -2.222 | -0.424 | -0.676 |
| Seniority in teaching Science (in years) | 1.27 | 1.656 | 1.137 |
| Seniority in teaching at the current school (in years) | 0.434 | 0.416 | -0.224 |
| Percent of teachers that used a teaching sequence | -0.021 | -0.14 | -0.069 |
| ***School-level variables*** | | | |
| Students per school | 19.202 | -27.387 | -20.126 |
| Students of 7th grade | -0.119 | -5.511 | -3.555 |
| Classrooms of 7th grade | 0.095 | -0.138 | -0.182 |
| Percent of schools with double school day | -0.024 | -0.128 | -0.128 |
| School promotion rate (%) | 1.74 | -1.181* | -0.512 |
| School drop-out rate (%) | 0.033 | 0.176 | -0.046 |
| School over-aged rate (%) | -1.049 | 1.185 | -0.246 |
| 7th student's repetition rate (%) | -0.692 | 0.997* | 0.58 |
| Percent of schools of district: 2 | -0.012 | -0.151 | -0.021 |
| Percent of schools of district: 3 | 0.238 | -0.064 | -0.064 |
| Percent of schools of district: 4 | -0.357 | -0.013 | -0.013 |
| Percent of schools of district: 5 | 0.095 | 0.079 | 0.079 |
| Percent of schools of district: 7 | 0.143 | 0.161 | -0.056 |
| Percent of schools of district: 8 | -0.107 | -0.013 | 0.074 |

Note: * significant at 10%; ** significant at 5%; *** significant at 1%.

## Appendix 8: The role of student ability and gender in student learning

This Appendix explores whether any experimental group presents more gains in test-score results. The following table presents separate estimates for students below (panel A) and above (panel B) the mean test-score, while Table A8.2 shows separate estimates of the average treatment effects for female and male students.

**Table A8.1. Results on Science learning according to student ability**

| ***Dependent variable*** | **Sequence vs. Control (1)** | **Coaches vs. Control (2)** | **Wald Test (3)** |
|---|---|---|---|
| *Panel A: Low ability students* | | | |
| Science score | 0.154** | 0.118* | 0.31 |
| | (0.066) | (0.069) | (0.575) |
| Percent of correct answers | 0.013 | 0.001 | 0.64 |
| | (0.012) | (0.014) | (0.425) |
| Percent of incorrect answers | -0.011 | 0.023 | 2.31 |
| | (0.019) | (0.016) | (0.128) |
| Percent of omitted answers | -0.035 | -0.059** | 0.91 |
| | (0.024) | (0.024) | (0.339) |
| *Panel B: High ability students* | | | |
| Science score | 0.183** | 0.236*** | 0.58 |
| | (0.078) | (0.073) | (0.448) |
| Percent of correct answers | 0.047*** | 0.037** | 0.47 |
| | (0.016) | (0.017) | (0.493) |
| Percent of incorrect answers | -0.033*** | -0.026*** | 0.48 |
| | (0.009) | (0.009) | (0.490) |
| Percent of omitted answers | 0.004 | -0.017** | 6.27 |
| | (0.011) | (0.007) | (0.012) |

Note: ** significant at 5%; *** significant at 1%. Robust standard errors in parentheses clustered at school level. Controls: (i) students characteristics (gender, age, nationality, parent's education, if the student missed, at most, one class per month, if the student has internet in his home), (ii) teacher characteristics (gender, age, years of experience, if she/he has post-graduate certificates), and (iii) school characteristics (school size, 7th size, 7th repetition rate, FEPBA score in Language, school district –or location). The number of observations for regression for the sub-sample of students with low skills (high skills) in column (1) is 598 (507), in column (2) 586 (524), and in column (3) 476 (651). All the regressions exclude the classrooms where the Human Body Unit was not taught.

### Table A8.2. Results on Science learning according to student gender

| ***Dependent variable*** | **Sequence vs. Control (1)** | **Coaches vs. Control (2)** | **Wald Test (3)** |
|---|---|---|---|
| *Panel A: Female* | | | |
| Science score | 0.516*** | 0.730*** | 1.93 |
| | (0.145) | (0.138) | (0.165) |
| Percent of correct answers | 0.089*** | 0.121*** | 1.06 |
| | (0.030) | (0.025) | (0.304) |
| Percent of incorrect answers | -0.035 | -0.049*** | 0.49 |
| | (0.021) | (0.016) | (0.486) |
| Percent of omitted answers | -0.062*** | -0.092*** | 2.14 |
| | (0.022) | (0.019) | (0.143) |
| *Panel B: Male* | | | |
| Science score | 0.539*** | 0.545*** | 0.00 |
| | (0.140) | (0.168) | (0.970) |
| Percent of correct answers | 0.108*** | 0.097*** | 0.13 |
| | (0.024) | (0.032) | (0.720) |
| Percent of incorrect answers | -0.071*** | -0.048** | 1.02 |
| | (0.016) | (0.021) | (0.313) |
| Percent of omitted answers | -0.045** | -0.066*** | 0.81 |
| | (0.021) | (0.023) | (0.367) |

Note: ** significant at 5%; *** significant at 1%. Robust standard errors in parentheses clustered at school level. Controls: (i) students characteristics (gender, age, nationality, parent's education, if the student missed, at most, one class per month, if the student has internet in his home), (ii) teacher characteristics (gender, age, years of experience, if she/he has post-graduate certificates), and (iii) school characteristics (school size, 7th size, 7th repetition rate, FEPBA score in Language, school district –or location). The number of observations for regression for the sub-sample of students with low skills (high skills) in column (1) is 544 (561), in column 2 is 545 (565), and in column (3) is 543 (584). All the regressions exclude the classrooms where the Human Body Unit was not taught.

## Appendix 9: The role of student gender in the captivate index

The following table presents separate estimates of the average treatment effects for female and male students.

### Table A9.1 Results on the captivate index according to student gender

| ***Dependent variable*** | **Sequence vs. Control (1)** | **Coaches vs. Control (2)** | **Wald Test (3)** |
|---|---|---|---|
| *Female students* | | | |
| Captivate index (A+B+C+D) | 0.353*** | -0.013 | 6.65 |
| | (0.130) | (0.149) | (0.010) |
| *Male students* | | | |
| Captivate index (A+B+C+D) | 0.096 | 0.126 | 0.07 |
| | (0.097) | (0.104) | (0.785) |

Note: ** significant at 5%; *** significant at 1%. Robust standard errors in parentheses clustered at school level. Controls: (i) students characteristics (gender, age, nationality, parent's education, if the student missed, at most, one class per month, if the student has internet in his home), (ii) teacher characteristics (gender, age, years of experience, if she/he has post-graduate certificates), and (iii) school characteristics (school size, 7th size, 7th repetition rate, FEPBA score in Language, school district –or location). The dependent variable in A-E represents a 4-point-scale, where 1 means strongly disagree, 2 disagree, 3 agree and 4 strongly agree. The captivate index is combined scale, which construction is described in Appendix 5. All the regressions exclude the classrooms where the Human Body Unit was not taught.