THE ECONOMIC JOURNAL

The Economic Journal, 1–29 DOI: 10.1093/ej/ueaa022
Advance Access Publication Date: 30 May 2020

EDUCATION QUALITY AND TEACHING PRACTICES*

Marina Bassi, Costas Meghir and Ana Reynoso

Improving school quality with limited resources is a key issue of policy. This article uses a randomised controlled trial (RCT) to estimate the effectiveness of guided instruction methods as implemented in underperforming schools in Chile. The intervention improved performance substantially, and equally for boys and girls. However, the effect is mainly accounted for by children from relatively higher-income backgrounds. Basing our study on the Classroom Assessment Scoring System (CLASS) instrument, we document that the quality of teacher–student interactions is positively correlated with the performance of low-income students; however, the intervention did not affect these interactions. Guided instruction improves outcomes, but the challenge to reach the most deprived children remains.

Improving the quality of education for children from lower socio-economic backgrounds is key to offering equal opportunity and arresting the intergenerational transmission of poverty. However, achieving this can be challenging in practice. For example, observational studies, as well as studies with randomised assignment of students to teachers, have concluded that teachers can have a large impact on performance (Rivkin *et al.,* 2005, Chetty *et al.,* 2014 and Ecuador Araujo *et al.,* 2016). However this literature has been unable to identify what makes a good teacher; and, even if it did, turning around the quality of teachers to a sufficient extent will likely prove far too difficult and slow. So the natural question is whether we can improve outcomes by identifying and implementing innovative teaching practices, and relying on existing human resources; solving this problem can have major policy implications for most countries in the world.

An experiment in Chile provides a unique opportunity to address this question. The educational psychology literature focuses on the method of instruction as an approach to improve school performance. The basic principle is that it is possible to compensate for low teacher skills by providing them with specific prepackaged classroom material and directions for teaching to any group of students in standardised ways. These methods can be controversial, and there is an active debate on the extent to which prescriptive methods can be successful. While advocates

* Corresponding author: Costas Meghir, Department of Economics, Yale University, NBER, IFS, IZA and CEPR, 37 Hillhouse Av., New Haven, CT 06511, USA. Email: c.meghir@yale.edu

This paper was received on 29 September 2017 and accepted on 21 February 2020. The Editor was Frederic Vermeulen.

The authors were granted an exemption to publish their data because access to the data is restricted. However, the authors provided the Journal with temporary access to the data, which enabled the Journal to run their codes. The codes are available on the Journal website. The data and codes were checked for their ability to reproduce the results presented in the paper.

The authors would like to thank two anonymous referees and the Editor of the Journal for their comments. We are grateful to Daniel Alonso for his valuable assistance in the analysis of the data. We also thank Esteban M. Aucejo, Martha Bailey, Hoyt Bleakley, Eric Hanushek, Dean Yang and Paul Rodhe for helpful comments and suggestions. We would also like to recognise the support of the Ministry of Education of Chile (Division of General Education and Studies Department) in the different stages of this project. Costas Meghir benefited from funding by the Cowles foundation and the ISPS. This research uses data from *Agencia de Calidad de la Educación*. We are grateful to *Agencia de Calidad de la Educación* for granting access to the data. All results in this article are the sole responsibility of the authors and do not reflect the view of such institution. Fieldwork of the impact evaluation and Ana Reynoso were funded by the Inter-American Development Bank (IADB). The findings and conclusions expressed in this article are solely those of the authors and do not reflect the view of the IADB, its Executive Directors, or the countries they represent. All errors and opinions are our own.

of *minimal instructional guidance* argue that students learn best when they discover concepts by themselves, those who believe in *guided instruction* argue that the cognitive architecture of the human brain is such that students' learning is maximised when teachers directly explain the concepts that students are required to know (Kirschner *et al.*, 2006).

Guided instruction methods, in turn, come in many forms. They are distinguished by the degree of discretion that teachers have to adapt instruction according to the characteristics of the particular group of students they are facing (Ganimian and Murnane, 2014). These methods are usually complemented by training teachers to support them in the use of these instruction materials. This method is known in the literature as *scripted instruction* and became very popular ever since the launch of high-scale educational programmes such as Success for All and DISTAR in the United States (Slavin *et al.*, 2009).

In this article we contribute to the understanding of the effectiveness of direct-instruction approaches in schools that serve deprived populations, by analysing the impact of a large-scale guided instruction programme in Chile aimed at low-performing schools. We focus on the performance of students in the national standardised maths, language and science tests and our results are based on a school-level randomised trial.

The programme in question, known as *Plan Apoyo Compartido* (PAC), was implemented by the Chilean Ministry of Education in 2011. The main intervention of the programme was to support teachers through a modified method of instruction by adopting a more prescriptive model. Teachers in treated schools received detailed classroom guides and scripted material to follow in their lectures. The programme was intended to be implemented gradually, so only a group of eligible schools was invited to participate in the first year. Our measure of students' learning is their performance in the Chilean standardised Education Quality Measurement System evaluations (SIMCE evaluations, from their name in Spanish). We concentrate the analysis on students who were in their fourth grade of elementary school in years 2011 and 2012 and who attended eligible schools.

Our results suggest that the programme had positive and significant effects, particularly for kids from the most advantaged backgrounds within treated schools (students in schools with higher socio-economic status or from higher-income families within our lower-income population). Overall, the programme improves reading test scores by about 10% of a test score standard deviation the first year of implementation. Programme effects increase significantly in the second year: test scores improve in all subjects in between 9% and 13% of a test score standard deviation. All these effects are statistically significant. Moreover, kids in schools with high socio-economic status participating in the programme improved SIMCE scores by 20% of a test score standard deviation with respect to comparable kids in control schools. Finally, students from high-income families see the greatest benefits from the programme, their test scores improving by between 10% and 20% of a test score standard deviation. All these results are strongly robust to adjustments in our inference strategy to control for multiple testing.[1]

To better understand the impact of PAC on students' test scores we analyse the effects of the programme on the quality of teacher–student interactions based on the CLASS (Classroom Assessment Scoring System; see Pianta *et al.*, 2008).[2] A random subsample of treatment and control schools from the PAC programme were invited to participate in the CLASS experiment.

[1] A recent paper by Araujo *et al.* (2016) focuses on the relationship between the quality of teacher–student interactions and test scores in Ecuador. Their study finds that one standard deviation increase in the quality of teacher–student interaction results in approximately 10% of a standard deviation of higher students' tests scores.

[2] Araujo *et al.* (2016) also used the CLASS in their experiment.

The experiment involved filming several hours of classroom teaching and coding them to score teachers' interactions with their students based on very specific teachers' behaviours that coders look for. We first show that CLASS scores correlate positively and significantly with students' performance, and particularly for those from lower-income backgrounds. Then we show that PAC did not cause significant improvements in CLASS scores, which may explain why low-income students were more modestly impacted by the PAC.

Our study offers an important contribution to the literature on understanding and improving education quality. We are specifically testing a programme that is easily scalable and does not make inordinate demands on human resources, but which, according to a well-established literature, can offer real improvements in pupil performance. From a methodological point of view, the experimental design on a particularly large number of schools offers the power needed to detect even relatively small effect sizes. Secondly, we provide the first assessment of the interaction between a large-scale instruction intervention and the CLASS in producing learning outcomes. The use of CLASS as a tool for understanding the mechanisms through which the intervention works, by implementing it on both treatment and control groups, is new in the literature.[3]

The article is organised as follows. The next section describes the programme intervention, the experimental design and the data used in this article. Section 2 describes the identification and inference strategies. Section 3 presents the main results of the article. Section 4 studies the importance of teacher–student interactions to improve performance and the impact of PAC on these interactions. Finally, Section 5 concludes.

1. Experimental Design, Data and Randomisation Check

1.1. *Plan Apoyo Compartido (PAC)*

PAC was implemented by the Chilean Ministry of Education in 2011 as a targeted educational policy providing technical and paedagogical support to schools historically performing below average in the national standardised test, SIMCE. It aimed at improving students' learning outcomes in maths and language from pre-K to fourth grade (and, additionally, in natural and social sciences for students in third and fourth grades), changing practices inside the classroom and the school. The PAC targeted low-performing public and subsidised private schools nationwide.[4] Next, we describe the design and implementation of the PAC.

PAC design The design of PAC included five components and was implemented in six-week cycles (see Figure 1). The first component, called 'effective implementation of the national curriculum' (indicated [1] in the figure), consisted in the development of unified paedagogical material and planning tools distributed to teachers. These tools included an annual curricular programming, a series of teaching materials designed for each six-week cycle and a set of daily planning activities to be used by teachers in the classroom. The second component (indicated [2] in the figure) consisted of promoting a school culture and environment that encourages learning.

[3] He *et al.* (2009) also evaluates a scripted reading preschool programme in Mumbai, India and Albornoz *et al.* (2020) evaluates the effects of teacher training interventions in Buenos Aires, Argentina. Unlike theirs, our article focuses on fourth grade primary school students.

[4] The Chilean system of education includes three types of schools: public schools, subsidised private schools and private schools. Public schools are both financed and administered by the public sector; subsidised private schools are administered by private agencies but receive funding from the state in the form of vouchers per attending student; finally, the third group includes schools that are administered privately and tuition is paid by the students' families.

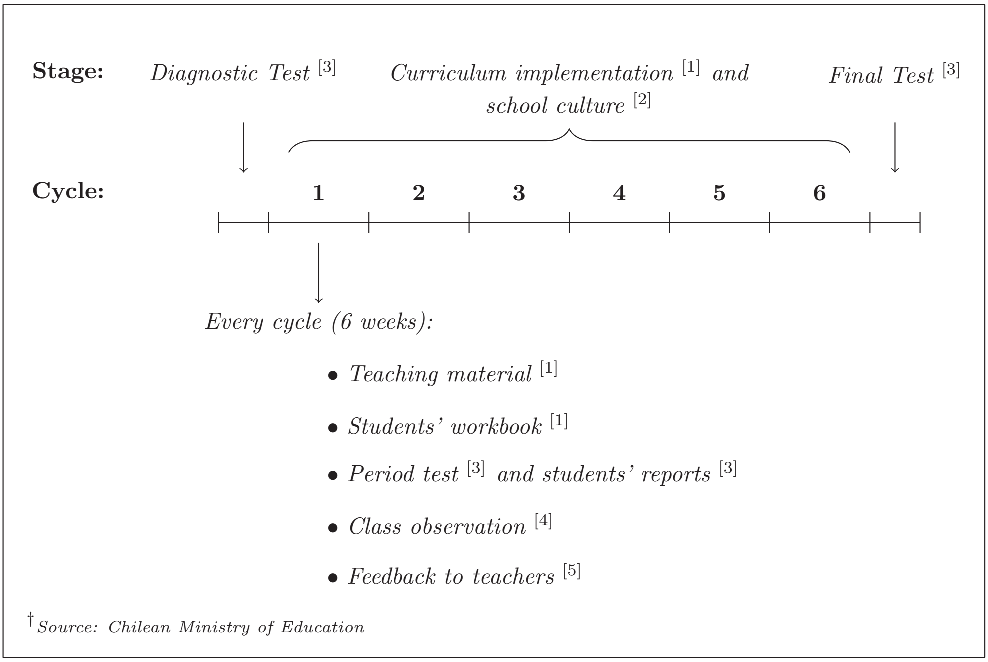

Fig. 1. *Design of Plan Apoyo Compartido*†.

A manual was developed and delivered to schools to guide the implementation of the ideas. The third component (indicated [3] in the figure) was the use of student evaluations as a tool for guiding teaching. This component included the development of four types of tests to monitor progress in students learning: a diagnostic test to determine the initial level of academic skills and knowledge administered at the beginning of the school year, intermediate and final tests to determine students' progress, and students' performance reports. Each of these testing instruments was applied in different moments of the semester to help to analyse students' performance in maths and language (MINEDUC, 2013). It is worth noting that, unlike the SIMCE tests, these instruments were not standardised tests and could be applied voluntarily by PAC schools. The fourth component (indicated [4] in the figure) was defined as the 'optimisation of the use of school time for learning in the classroom' and consisted in promoting class planning and frequent class observation in schools to provide feedback to teachers. Finally, the last component, known as 'promotion of teachers' professional development' (indicated [5] in the figure) aimed at promoting frequent internal school staff meetings to discuss students' progress.

A central feature of the programme's design to assist the implementation and monitoring of the PAC activities was the creation of two support teams—one internal and one external to schools—expected to work closely together. The first team, the Education Leadership Team (henceforth ELE, from its name in Spanish), consisted of the school principal, the head of the technical and paedagogic office of the school and two distinguished teachers. The second group, the Team of Technical and Pedagogic Advisors (ATP), comprised three representatives of the regional Department of Education (the DEPROV), aimed at providing external support to the ELE teams. Each ATP visited its assigned schools every six to seven weeks to advise the ELE on the use of the teaching material, on the development of a diagnosis of the school's strengths and weaknesses and on the analysis of the students' tests scores to study progress (MINEDUC, 2013).

Monitoring and assessment of PAC implementation at the school level The Ministry of Education collected information on the ATPs' visits to schools and its findings. Still, as part of the effort to assess the impact of the programme (which included this article's analysis), the Chilean Ministry of Education designed two instruments, mainly focused on gathering evidence on the extent to which teachers and schools used the paedagogic material provided by the programme, implemented PAC components and received the support of ELEs and ATPs.

The first instrument was a protocol specifically developed to observe and code class videotapes on key aspects of the programme.[5] Independent raters were asked to observe the videos and indicate whether the programme intervention activity is observed in the classroom, establish whether teachers in the classroom implemented the scripted instruction method, and determine whether teachers organised their classroom according to the programme guidelines. The key aspects that were observed in both PAC and control schools were class structure, encouragement of critical thinking, norms and schedule, and evaluation of students' performance. In Appendix A, Table A1, we present the correlation between participating in PAC and the average score in these relevant dimensions of the programme's implementation. The data indicate that PAC schools are observed to perform better in the PAC objectives within the classroom. The difference between treated and control schools is only significant when it comes to the enforcement of norms and schedule. PAC classrooms are 5.8 percentage points more likely to follow the norms and schedule relative to non-PAC schools. Within this dimension, a particularly interesting aspect concerns the use of prepacked material by the teachers. Coders were asked to indicate whether they observed teachers in the classroom using special workbooks as a paedagogic resource (the scripted class manuals, or 'PAC book' in the case of the treated schools). PAC schools are 23 percentage points more likely to use the scripted workbooks in class relative to non-PAC schools. The difference is significant at the 1% level.

The second instrument to assess the degree of implementation of the PAC at the school level was to interview the head of the technical and paedagogic office of the school (henceforth JUTP, from its name in Spanish) and a set of teachers within the schools that participated in the monitoring evaluation. These surveys gathered information about the key ingredients of the programme. For example, authorities were asked whether their school utilised annual planning in class, whether students were periodically evaluated, whether the teachers understood and explained norms of behaviour to students, and the frequency with which the school organised meetings between teachers and authorities to monitor the performance of students (see Figure 1).

Table A2 shows the correlation between PAC treatment and the responses of JUTPs. The first three columns are concerned with annual planning activities within the school. Some 93.38% of JUTPs in the sample respond that the school engages in annual planning of the curriculum, and we do not detect differences between PAC and control schools in annual planning. However, Columns (1) to (3) evidence that PAC schools are significantly more likely to have the annual planning performed by the PAC authorities or by the ELE team and significantly less likely to leave the curriculum design to teachers. This evidence may be an indication of one of the dimensions in which teachers in PAC schools were supported by the authorities. Columns (4) and (5) provide evidence of two additional dimensions of support to teachers. The evidence presented

[5] The protocol followed to code videos resembles the CLASS protocol described in Section 4.

in Column (4) indicates that PAC schools are observed about 34.2% more times by the ELE team relative to control schools. Moreover, Column (5) indicates that teachers in PAC schools receive feedback after these observations about 30.5% more times than those in control schools. Finally, teachers' responses are consistent with the JUTPs answers, although we do not detect significant differences between PAC schools and control schools in teachers' responses.[6] For example, consistent with the JUTPs reports, teachers in PAC schools are more likely to having their classrooms observed and to receive feedback after observations.

1.2. *Eligibility and Randomisation*

Among public and subsidised private schools in Chile, PAC considered two main eligibility criteria to define the target group of schools: first, the school's baseline average SIMCE score for the years between 2005 and 2009 in maths and language should be below the national average (252 points out of 500); and, secondly, there should be at least 20 students per level on average from pre-K to fourth grade.[7] Some 2,286 schools met these criteria and were ranked by their 2005–9 average SIMCE scores in language and maths. The bottom 1,000 schools were automatically considered eligible. Since participation in the programme was voluntary, refusal to participate was expected, so in order to reach a target of around 1,000 eligible schools in the first year of the programme, the Ministry increased the sample within each DEPROV by 50%, going up in the SIMCE ranking.[8] Of the resulting 1,480 eligible schools, 632 located in 'small' DEPROVs (DEPROVs with 40 schools or fewer) were allocated to the programme automatically and do not form part of the evaluation and analysis. The remaining 848 schools located in 'large' DEPROVs were randomly allocated to treatment and control groups. Five schools in the randomisation are excluded from the analysis because they show missing information on school and students' characteristics in both the 2011 and 2012 data sets. All in all, of the 843 schools considered in this analysis, 648 were randomly selected to the treatment group and 195 were randomly selected to the control group.[9]

1.3. *2012 CLASS Intervention*

The second part of this article analyses the relationship between teacher–student interactions and learning outcomes. To measure the quality of teacher–student interactions, we use the well-known *Classroom Assessment Scoring System* (CLASS, Pianta *et al.,* 2008). The CLASS is an instrument used in the education literature to measure the quality of teacher–student interactions, as a proxy to teachers' quality or effectiveness.

To produce the CLASS measures, a randomly selected group of 158 PAC schools (79 from the PAC treatment group and 79 from the PAC control group) were invited to have their fourth

[6] The vast majority of teachers in the sample respond positively to the questions. It is worth remarking that the fact that all teachers, including PAC teachers, implement the key programme instruments is reassuring of the implementation of PAC.

[7] The Ministry of Education also required that the schools' administrators should have no sanctions related to the voucher subsidies system in the previous three years.

[8] At this point some schools were excluded after consultation with DEPROV authorities either because of bad management or because they were already receiving technical and paedagogical assistance from well-known agencies of paedagogical support in Chile.

[9] Two of the 843 schools in the randomisation are missing from our 2011 data set. Therefore, we consider 841 schools in 2011. Four of the 843 schools in the randomisation drop out of our sample in 2012. Therefore, our analysis for 2012 is based on 839 schools. We discuss attrition in this section below.

grade classrooms videotaped for four full lessons. The CLASS intervention took place in 2012, the second year of implementation of the PAC programme. After class observations, thoroughly trained coders watch and analyse the videotapes and assign a score for teacher–student interactions in several dimensions (details will be presented in Section 4).

The CLASS experiment had extremely good compliance: in the end, 137 invited schools agreed to participate in the filming sessions and 185 classrooms within participating schools had lectures filmed. Non-participation is fairly well balanced between the treatment and control schools.[10] The sample of treated and control schools that participated in the CLASS experiment is also well balanced in school pre-treatment characteristics. These characteristics include the school income group, the past average SIMCE score of the school, the experience of fourth grade teachers, the experience of the school principal, and the tenure at the school of fourth grade teachers and the principal. For all these baseline characteristics we cannot reject the hypotheses that they are equal among PAC and non-PAC schools that participate in the CLASS experiment.

1.4. *Data*

The analysis in this article relies on administrative data provided by the Ministry of Education. This data set includes student-level information on treatment status, test scores and baseline demographic characteristics. Table B1 in Appendix B shows summary statistics of all the variables used in this article, namely: test scores and baseline characteristics, for the group of students who took each of the subject tests (post-attrition samples).

1.5. *Treatment–Control Balance and Attrition*

In our empirical results we exclude from the 2012 data those schools that implemented the CLASS observation system because of Hawthorne effects (Landsberger, 1959)—an issue to which we return in the results section. As we show there, the CLASS intervention had impacts of its own, thus contaminating the control schools that were part of it. CLASS was randomly allocated in both treatment and control schools and thus there is no bias in excluding these schools.

Table OA1 in the Online Appendix shows summary statistics and randomisation checks at the school level. The evidence shows that the randomisation at the school level was successful: all pre-treatment school characteristics are balanced across PAC and control schools. Moreover, the result of an F-test of joint significance of school baseline characteristics on random assignment indicated that we cannot reject the null hypothesis that no variable jointly predicts treatment.

Table OA2 in the Online Appendix, displays a set of randomisation checks for the entire population of fourth grade students (the pre-attrition sample) and for the three post-attrition samples (reading, maths and science test takers). The table is divided into three panels, corresponding to the 2011 cohort, the 2012 cohort that excludes schools participating in the CLASS intervention, and the whole 2012 sample. Each panel displays the results of a test of differences in means of attrition rates and baseline characteristics across treatment status, and a test of joint significance of the impact of baseline characteristics on treatment status.

In general, attrition rates in our sample are very low, and baseline characteristics are balanced in both the pre-attrition and post-attrition samples. In 2011 there is no student that missed

[10] Among these 185 classrooms, 94 were in control PAC schools and 91 were in treatment PAC schools. Among the 91 classrooms in PAC schools, in turn, 78 were participating in the PAC, while 13 were in schools invited to participate in PAC but did not accept.

all three subject tests in the sample. When analysing attrition rates by subject for this cohort (not reported in the table), only 2.06% of students missed the reading test, 2.08% missed the maths test and 1.97% missed the science test. Moreover, attrition rates are balanced between the treatment and control groups, as shown in the first three rows of the 2011 panel of Online Appendix Table OA2. There, the statistic reported is the difference in attrition rates between the treatment and control groups. These differences are very small: relative to the control group, 0.7% fewer students missed the reading test and 0.1% more students missed the maths and science tests in the treatment group. However, all p-values indicate that these differences are not significant.

The next set of rows show the results of a test of differences in means of baseline characteristics. Most baseline characteristics are balanced even among the students that did not drop out of the data. The exceptions are *low income* and mother and father *incomplete high school*: test takers in the treatment group are less likely to be from a low-income family and less likely to have a parent with incomplete high school. Even when the p-value indicates that these differences are individually significant, the magnitude of the economic effect is extremely small: around 2%. Moreover, the last row of the 2011 panel shows that, taken together, baseline characteristics do not significantly predict whether a student is in the treatment or the control group, even in the post-attrition samples. The statistic reported is the F-statistic of the joint test, and p-values indicate that we cannot reject the null hypothesis that baseline characteristics do not jointly determine the random allocation to the programme.

The conclusions from the 2012 cohort are similar. First, attrition rates are higher than in the 2011 cohort, but still low. In this cohort 15% of students missed the reading test, 15.26% missed the maths test and 15.36% missed the science test (statistics not reported in the table). However, differences in attrition rates between treatment and control groups for 2012 are small and insignificant. Being in the treatment group is associated with about 1% lower probability of sitting the reading, maths and science tests relative to the control group, but these differences are not significantly different from zero, which suggests that the higher overall attrition in 2012 does not bias our results of the impact of PAC on SIMCE.

As further corroborative evidence, Panel B in Table OA2 shows that the 2012 subsample excluding schools contaminated by the CLASS intervention is also balanced: all baseline characteristics are jointly insignificant in explain treatment status, as evidenced by the F-test. Individually, all baseline characteristics are balanced between treatment and control groups, with the exception of *Nbr years failed* (the number of grades a student had to retake prior to the fourth grade), which difference is economically negligible in magnitude and statistically only marginally significant.

In sum, we find no evidence that the experimental design was compromised in any way. In both cohorts the difference in the proportion of attritors is negligible in magnitude and not significant and the randomisation was successful in balancing baseline characteristics, even for the post-attrition samples.

2. Estimation and Inference

Our results explore overall effects as well as heterogeneous treatment effects by school and students' demographic characteristics. In our heterogeneity analysis we first analyse results by school socio-economic status. Secondly, we define four groups of students based on the

Table 1. *Randomisation and Implementation, School Level.*

		Implemented PAC					
		All schools		2011 sample		2012 sample	
		No	Yes	No	Yes	No	Yes
Randomised into PAC	No	185	10	194	0	115	10
	Yes	150	498	155	492	164	413

Notes: PAC stands for *Plan Apoyo Compartido*. *All schools* refers to the 843 schools that were considered eligible by the programme and considered in the programme evaluation at some point. The *2011 sample* consists of the 841 schools that were originally included in the programme evaluation sample. The *2012 sample* consists of the 702 schools that remained in the programme evaluation sample in the second year and were randomly excluded from the CLASS intervention (that is, 839 schools that continued in 2012 minus 137 schools that participated in CLASS). *Implemented PAC* is a dummy variable that takes value one if the school participated in PAC and zero otherwise. *Randomised into PAC* is a dummy variable that takes value one if the school that randomly assigned to the treatment group and zero otherwise.

interaction between the gender of the student and her household income (*Female—Low income*, *Female—Medium–High income*, *Male—Low income*, and *Male—Medium–High income*).

The focus on income is mainly motivated by the need to understand whether such programmes are particularly helpful for the most deprived or, by contrast, whether they reinforce resources provided by parents. In general there is ample evidence showing an association between income and wealth with child outcomes. Whether such association extends to responses to interventions is an open and important question. Gender is also important: girls tend to perform better than boys in reading and worse than boys in maths and science (OECD, 2015). These outcomes may be related to teachers' practices. Using the same sample of fourth grade teachers in Chile as this article, Bassi *et al.* (2018) show that teachers in fact pay more attention to boys than girls, and those differentiated behaviours are correlated with worse performance in SIMCE in maths and science among girls. It is thus important to understand whether there are substantial differences in the response to interventions.

The results that we present are obtained by a regression at the individual student level

$$SIMCE_{ijgk} = \beta_{gk} + \gamma_{gk} T_{ij} + \mathbf{X}_{ijg}\boldsymbol{\delta}_{gk} + \epsilon_{ijgk}, \tag{1}$$

where $SIMCE_{ijgk}$ is the test score of student i, in school j, in demographic group g, and in subject $k = \{Math, Language, Science\}$. This is measured in units of a standard deviation of the control group (which we will refer to as SD units henceforth). T_{ij} is is a dummy indicating whether the student attended a school j that was randomised into the programme (PAC); $\mathbf{X}_{ijg}$ is a vector of student–school characteristics that includes baseline characteristics;[11] and ϵ_{ijgk} is a random error term, which because of randomisation is uncorrelated with treatment assignment.

Not all schools assigned to the programme actually implemented it: there is non-compliance in both the 2011 and the 2012 cohorts. Table 1 shows take-up rates of schools considered eligible to the programme, for both years 2011 and 2012 and for the total schools that were ever considered eligible.

In 2011, about 23% of schools randomised into the programme did not implement it. In this case if we replace the randomisation indicator T_{ij} with whether treatment actually took place and

[11] The covariates include whether the student lives in a household with at least one parent and/or siblings; whether the student lives in a household with members of the extended family; the number of times the student failed a school year; mother's education: dummies for 'no education', 'incomplete primary', 'primary', 'incomplete high school', 'high school', 'some college', 'college +'; father's education (same dummies as mother education).

Table 2. *First Stage. Dependent Variable: Implemented PAC.*

	2011				
		Girls		Boys	
	All	Low income	High income	Low income	High income
Randomised into PAC	0.761***	0.774***	0.721***	0.769***	0.745***
	[0.734; 0.787]	[0.747; 0.802]	[0.684; 0.761]	[0.709; 0.783]	[0.741; 0.798]
Observations	31,384	10,938	2,330	11,492	2,581
	2012				
		Girls		Boys	
	All	Low income	High income	Low income	High income
Randomised into PAC	0.632***	0.653***	0.623***	0.651***	0.645***
	[0.582; 0.679]	[0.602; 0.702]	[0.56; 0.684]	[0.591; 0.698]	[0.595; 0.703]
Observations	29,905	8,740	2,225	8,894	2,622

Notes: PAC stands for *Plan Apoyo Compartido*. The dependent variable is *Implemented PAC*, a dummy variable that takes value one if the student attends a school that participated in PAC and zero otherwise. *Randomised into PAC* is a dummy variable that takes value one if the student attends a school that was randomly assigned to the treatment group and zero otherwise. *All* refers to all students pooled together. *Low income* is a dummy variable that takes value one if the student's family monthly income is less than 300,000 Chilean pesos, the minimum wage in such country. *High income* is a dummy that takes value one if *Low income* = 0. 95% bootstrapped confidence intervals are shown in brackets. ***Variable significant at the 1% level. Clustering at the school level. The 2012 sample excludes schools that implemented CLASS.

then use the randomisation indicator as an instrument we will identify the effect of treatment on the treated because non-compliance is only one-sided. In 2012 however, we have two-sided non-compliance, with 8% of schools not assigned to the programme by the randomisation actually getting it.[12] IV in this case identifies the LATE parameter under the additional monotonicity assumption that randomisation either does not change treatment status or induces the school to adopt the programme, but never the reverse.[13] In all cases using as treatment variable the original randomisation (T_{ij}) provides an unbiased estimate of the intention to treat parameter (ITT), namely the effect of having been offered the programme.

At the student level Table 2 shows that for the 2011 cohort 76% of students were exposed to it as a result of the school being assigned to receive PAC. No student in the control group was exposed. The percentage varies slightly by demographic groups because the composition of the schools is not uniform. For the 2012 cohort, the percent of exposed students as a result of being randomised into the programme is 63%; some students in the control group did, however, receive the treatment. Table OA5 in the Online Appendix shows similar conclusions for the whole sample of 2012 schools.

In deriving standard errors and carrying out inference we cluster at the school level, which is the randomisation unit. Since we will be splitting the sample by demographic characteristics and testing families of hypotheses, we adjust the p-values for multiple testing using the step-down procedure of Romano and Wolf (2005). The resulting p-value is the family wise error rate (FWE), namely the probability that we incorrectly identify one coefficient as significant in the entire group of hypotheses being tested. We use 1,000 bootstrap replications to compute all standard errors and confidence intervals.

[12] Table OA4 in the Online Appendix shows a similar picture when all 2012 schools are considered.
[13] See Imbens and Angrist (1994).

Table 3. *Impact of PAC on SIMCE.*

	Intention to treat effect (ITT)		
	Reading	Maths	Science
Randomised into PAC	0.108 [0.06; 0.15] (0.01)	0.087 [0.03; 0.14] (0.02)	0.055 [0.01; 0.1] (0.06)
Control group mean	243.83	236.12	235.18
Control group SD	50.38	47.17	44.83
Number of clusters	842	843	843
Observations	56,193	56,116	56,104

Notes: PAC stands for *Plan Apoyo Compartido*. *Randomised into PAC* is a dummy variable that takes value one if the student attends a school that was randomly assigned to the treatment group and zero otherwise. The effects shown are in units of the control group standard deviation. 95% bootstrapped confidence intervals are shown in brackets. Romano-Wolf step-down p-values from the two-sided test accounting for all three hypotheses are shown in parentheses. All regressions include cohort fixed effects. Clustering at the school level. The 2012 sample excludes schools that implemented CLASS.

3. Main Results

3.1. *CLASS, Hawthorne Effects and the 2012 Sample*

As mentioned earlier, in order to better understand how PAC works and how it may affect teacher practices it was decided to implement the CLASS observation system in a random subset of 137 PAC treatment and PAC control schools. The CLASS was effectively another intervention consisting of videotaping lectures in full knowledge of the teachers within randomly selected PAC control and PAC treated schools. The question is whether CLASS had an effect in itself, thus contaminating the control group. Exploiting the fact that CLASS was randomly allocated, we estimate the treatment effect of receiving CLASS by comparing the outcomes for children in whose schools CLASS was implemented with those in which it was not, among the schools that did not implement the PAC (PAC controls). The mere implementation of CLASS significantly improved SIMCE scores by 23%, 18% and 21% of a standard deviation for reading, maths and science respectively. However, CLASS had no additional effect on learning outcomes for children in the PAC treatment group.[14] Teachers observed through the CLASS protocol were not offered feedback on their classroom performace until *after* the SIMCE national examination that assesses student learning. Thus the effect of videotaping on achievement may be a characteristic example of the so-called *Hawthorne effects* (Landsberger, 1959; Levitt and List, 2011), which lead to productivity increases when people feel that they are being monitored.[15]

The implication is that by including schools that received CLASS in our evaluation sample we would blunt the estimated effects of the PAC; hence we exclude all 2012 schools in treatment and control that implemented CLASS. Importantly since CLASS was randomly allocated, this causes no bias, but instead produces results that correctly reflect the PAC intervention. The estimates that include the CLASS sample are presented in the Online Appendix for completeness.

3.2. *Overall Effects*

We start by showing in Table 3 the overall ITT effects of the experiment, pooling the data from the two cohorts and controlling for cohort fixed effects. In all tables that follow we report results

[14] Detailed results in Table C1 in Appendix C.

[15] In other contexts observation has been shown to reduce teacher absentees, which can also have an impact on performance (Duflo *et al.*, 2012).

Table 4. *Impact of PAC on SIMCE by Cohort.*

	Intention to treat effect (ITT)					
	2011			2012		
	Reading	Maths	Science	Reading	Maths	Science
Randomised into PAC	0.095	0.068	0.033	0.127	0.117	0.089
	[0.04; 0.15]	[0.01; 0.13]	[−0.02; 0.09]	[0.07; 0.19]	[0.04; 0.19]	[0.03; 0.15]
	(0.02)	(0.09)	(0.37)	(0.01)	(0.03)	(0.04)
	Instrumental variables (IV)					
Implemented PAC	0.125	0.089	0.044	0.2	0.184	0.139
	[0.05; 0.2]	[0.01; 0.17]	[−0.04; 0.12]	[0.11; 0.3]	[0.06; 0.31]	[0.06; 0.24]
	(0.02)	(0.09)	(0.37)	(0.01)	(0.04)	(0.04)
Control group mean	244.79	235.76	236.84	242.42	236.65	232.72
Control group SD	49.97	47.10	44.09	50.94	47.26	45.80
Number of clusters	840	841	841	702	702	702
Observations	30,736	30,731	30,765	25,457	25,385	25,339

Notes: PAC stands for *Plan Apoyo Compartido*. *Implemented PAC* is a dummy variable that takes value one if the student attends a school that participated in PAC and zero otherwise. *Randomised into PAC* is a dummy variable that takes value one if the student attends a school that was randomly assigned to the treatment group and zero otherwise. The effects shown are in units of the control group standard deviation. 95% bootstrapped confidence intervals are shown in brackets. Romano-Wolf step-down p-values from the two-sided test accounting for all six hypotheses in each panel (ITT and IV) are shown in parentheses. Clustering at the school level. In the second panel the instrument is *Randomised into PAC*. The 2012 sample excludes schools that implemented CLASS.

without covariates (other than cohort effects when we pool them). The Online Appendix reports the results when we include covariates. In this and all tables that follow we report 95% confidence intervals (CI) in square brackets and RW step-down p-values in parentheses, both computed using the bootstrap.

The impacts on reading and maths are both large, and significant even controlling for multiple testing. The results for science are smaller and the Romano Wolf (RW) step-down p-value is 0.06. Overall, the conclusion from this table is that the intervention was successful in improving learning standards. In what follows we first consider how the programme worked for separate cohorts and we then proceed with heterogeneity analysis.

3.3. *Effects by Cohort*

Table 4 shows the effects of the programme on SIMCE test scores for students in the 2011 and the 2012 cohorts separately. The top panel of the table shows the ITT estimate, while the bottom panel reports the corresponding instrumental variables (IV) results where the explanatory variable is actually receiving PAC and the instrument is being randomised into PAC; the parameter is interpreted as the effect of treatment on the treated for the 2011 cohort where all those randomised out were actually excluded from the programme (one-sided non-compliance), while for the 2012 cohort it is interpreted as the Local Average Treatment Effect (LATE) under the additional assumption of monotonicity, since there is two-sided non-compliance. The RW step-down p-values allows for all six hypotheses (reading, maths and science in each of the two years).

Considering the results for the ITT and the corresponding ones for IV, we find that reading improved by about 0.10 of a SD in 2011, giving an IV coefficient of 0.125 (RW p-value 0.02), revealing a large impact in the schools that actually received the PAC. There is also a 0.07

improvement in maths in the same year, that is individually significant (see CI) but not so once we account for multiple testing (see RW *p*-value). The remaining effects for 2011 are not significant.

The effects in the 2012 sample are strong: test scores in all subjects improve between 0.09 and 0.13 of SD units relative to control schools, and the effects are all significant, even accounting for multiple testing. The effects are larger than those in 2011, which is consistent with the programme maturing and being better embedded in the implementing schools. Indeed, the fraction of classrooms with teachers who are hired in 2012 or who are substitute teachers in the school is only 13.37%, implying improved experience levels with PAC implementation in the second year. So these results point to the persistent and even improving success for the programme overall. Table OA9 in the Online Appendix shows that these results are robust to including covariates in the specification.[16]

3.4. *Heterogeneity Analysis*

Given the overall impacts of the programme, we now investigate whether these differ across gender and socio-economic status, which are sources of disparities in performance. We will be focusing on three groups: boys and girls; children from low-income households versus higher-income ones; and low versus higher SES schools. We define low-income background as children from families with a monthly income less than 300,000 Chilean pesos (US$600 in 2011), which is the minimum wage. The SES status of the school is defined by the government, based on an index of parental education and income and a vulnerability index.[17] In our sample of PAC eligible schools lower SES schools are overrepresented, reflecting the fact that underperforming schools tend to serve lower SES students. Our *Low SES* group includes the schools classified by the government as belonging to the low and medium–low groups.

The heterogeneity analysis is important for targeting, programme improvement and understanding how to reduce important educational deficits: there are large disparities among the groups and a key question is whether the programme reduces such inequalities and more generally how it affects the outcomes of each category. Table 5 shows the difference in the SIMCE scores, in standard deviation units, for the control group in both cohorts between boys and girls and between children from lower- and higher-income families. We also show differences between children in low and higher SES schools.

Girls perform better in reading and worse in maths and science, while children from higher-income groups are performing uniformly better than lower-income ones. We also find differences between low and higher SES schools, although these are not significant. Nevertheless, it is still interesting to consider the impact of the programme across types of school, because policy

[16] The 2012 results including the CLASS subsample are shown in Online Appendix Table OA6. These show a decline in the impact of the programme. However, this is fully explained by the fact that CLASS raised the performance of the control schools (non-PAC) in which it was implemented, as we documented earlier. For further corroboration we also show in Section 1 that the PAC intervention was well implemented in 2012.

[17] The Education Quality Assurance Agency, responsible for the SIMCE, classifies schools into socio-economic categories based on four variables: mother's years of education, father's years of education, monthly income reported by parents, and a vulnerability index developed by the Ministry of Education. The first three variables are obtained from SIMCE parents' questionnaires. The average value among parents of the corresponding students grade is calculated for each of these variables. A cluster methodology is applied with these four variables to classify schools (for each grade level in which SIMCE is applied) into five socio-economic categories: low, medium–low, medium, medium–high, high. Appendix Table B2 shows the distribution of Chilean schools according to their SES along with the distribution in the PAC sample of eligible schools.

Table 5. *Differences in Performance in the Control Group, by Demographic Characteristics and School SES.*

	Reading	Maths	Science
Girls–Boys	0.219	−0.07	−0.094
	[0.19; 0.25]	[−0.11; −0.03]	[−0.13; −0.05]
Low income–Higher income student	−0.109	−0.145	−0.217
	[−0.16; −0.06]	[−0.21; −0.08]	[−0.28; −0.16]
Low SES–Higher SES school	−0.054	−0.011	−0.069
	[−0.15; 0.05]	[−0.13; 0.1]	[−0.17; 0.03]
Number of clusters (control schools)	194	194	194

Notes: Table shows pairwise differences in performance for students in the control group by demographic characteristics. *SES* variables refer to the school socio-economic status as described in Section 3. *Low income* is a dummy variable that takes value one if the student's family monthly income is less than 300,000 Chilean pesos, the minimum wage in such country. The effects shown are in SD units of students' test scores in the baseline group (Boys, Higher Income, and Higher SES, respectively). 95% bootstrapped confidence intervals are shown in brackets.

Table 6. *Impact of PAC on SIMCE by School Socio-economic Status, with Cohort Fixed Effects.*

	Low SES			Medium SES		
	Reading	Maths	Science	Reading	Maths	Science
Randomised into PAC	0.079	0.051	0.015	0.197	0.223	0.191
	[0.03; 0.13]	[−0.01; 0.11]	[−0.04; 0.07]	[0.1; 0.3]	[0.11; 0.34]	[0.09; 0.29]
	(0.02)	(0.24)	(0.67)	(0.01)	(0.01)	(0.01)
Control group mean	243.04	235.82	234.50	248.37	237.84	239.14
Control group SD	50.12	47.29	44.85	51.61	46.43	44.54
Number of clusters	706	707	707	194	194	194
Observations	43,637	43,544	43,537	12,480	12,497	12,492

Notes: PAC stands for *Plan Apoyo Compartido*. *Randomised into PAC* is a dummy variable that takes value one if the student attends a school that was randomly assigned to the treatment group and zero otherwise. *SES* variables refer to the school socio-economic status as described in Section 3. The effects shown are in units of the control group standard deviation (SD). 95% bootstrapped confidence intervals are shown in brackets. Romano-Wolf step-down p-values allowing for all six hypotheses from the two-sided tests are shown in parentheses. All regressions include cohort fixed effects. Clustering at the school level. Schools that implemented CLASS are excluded from the 2012 sample.

makers often target policies based on overall school composition and because the peer structure is different. We start by considering how the programme affected different SES type schools, and then we move to gender and family background differences.

Heterogeneity analysis can be particularly susceptible to false positives (i.e., finding significant results when there are none) because the number of hypotheses being tested is multiplied. Thus for each case we compute RW step-down p-values for the entire set of hypotheses involved as specified below. Moreover, for this analysis we pool the 2011 and the 2012 data (allowing for cohort effects) so as to increase statistical power.

3.4.1. *Effects by school socio-economic status*

Table 6 shows the effects of the PAC programme on test scores in each of the two SES school groups. The main conclusion from this table is that the programme was most successful in schools with higher socio-economic status among eligible schools: in the medium SES group the programme increased test scores in reading and science by about 0.20 of SD units and in maths by over 0.22 of SD units. All effects for the medium SES schools are highly significant with step-down p-values of at most 0.01 adjusting for all six hypotheses being considered. These are remarkable improvements. But even more remarkable from a policy perspective is the fact that the programme did not increase the performance in the low SES schools by nearly as much, despite the fact that the baseline performance is similar as shown above. In fact, we reject the hypothesis that the impact of PAC on SIMCE scores are equal for kids in low SES schools versus kids in medium SES schools. We aggregate the test scores of different subjects by obtaining the first principal component of test scores within each group, and we reject the hypothesis that PAC effects on aggregate test scores are equal for the low and high SES groups (p-value of 0.015). We also reject equality of coefficients when we compare each SIMCE subject separately instead of taking the principal component (p-values of 0.056 for reading, 0.022 for maths and 0.006 for science).[18]

3.4.2. *Effects by gender and family income*

We now turn to differences by gender and household income—both sources of disparities in performance. We are particularly interested in how impacts vary between children of different SES backgrounds, because it has been a challenge to improve outcomes for the most deprived populations. In addition, gender disparities in educational performance may partly explain male/female differences in labour market outcomes, including in wages and informality rates.

Appendix Tables D1 and D2 show differences between boys and girls and low- and higher-income background students respectively. We find that overall there are no significant differences in impacts between boys and girls and the estimates are almost the same. However, we find that the impacts for students from higher-income backgrounds are approximately twice those of the students from low-income families: the reading score improved by 0.167 SD units (p-value 0.01) for the higher-income group and only 0.089 SD units (p-value 0.01) for the lower-income group. The maths score improvements were 0.143 (p-value 0.01) and 0.074 (p-value 0.08) respectively.[19] We now look at this in greater detail by considering gender and income background differences jointly.

The results are shown in Table 7 that pools the two years together and includes cohort fixed effects in the regression.[20] The table presents 12 impacts and the RW step-down p-values provide significance levels accounting for the fact we are considering these multiple hypotheses. We also report 95% confidence intervals.

The main conclusion from Table 7 is that the programme produced significant impacts for the reading scores for boys from both income groups as well as for girls from a higher-income backround. It also improved significantly the maths performance of boys from the higher-income group. Specifically, reading scores for higher-income children improve by 0.135 SD units for boys (RW p-value 0.02) and 0.203 SD units for girls (RW p-value 0.01) relative to kids in control schools. The reading scores for boys in the lower-income families improved by 0.10 SD units (RW p-value 0.01). We also find that a 0.18 SD units improvement in the maths scores for boys from the higher-income group (RW p-value 0.01). Given the adjustment for multiple hypotheses testing these are particularly strong results. If we were using the conventional single hypothesis p-values many more of these 12 effects would have been classified as significant—for example, see the individually significant improvements, implied by the 95% confidence intervals, in maths

[18] See Online Appendix Table OA10 for results including covariates in the specification.

[19] p-values reported are RW step-down for six hypotheses.

[20] Once again, the 2012 sample excludes the contaminated CLASS sample. Table OA8 in the Online Appendix shows the 2012 results that include the sample contaminated by the CLASS intervention.

Table 7. *Impact of PAC on SIMCE by Students' Gender and Income, with Cohort Fixed Effects.*

	Girls					
	Low income			High income		
	Reading	Maths	Science	Reading	Maths	Science
Randomised into PAC	0.075	0.067	0.043	0.203	0.102	0.065
	[0.02; 0.13]	[0; 0.13]	[−0.01; 0.1]	[0.11; 0.3]	[0; 0.2]	[−0.03; 0.15]
	(0.10)	(0.32)	(0.38)	(0.01)	(0.32)	(0.38)
Control group mean	250.34	233.87	232.23	253.93	242.73	242.24
Control group SD	48.33	45.62	43.01	49.22	46.44	45.88
Number of clusters	835	835	835	761	761	760
Observations	19,222	19,245	19,201	4,494	4,484	4,493
	Boys					
	Low income			High income		
	Reading	Maths	Science	Reading	Maths	Science
Randomised into PAC	0.101	0.081	0.048	0.135	0.179	0.087
	[0.05; 0.15]	[0.02; 0.14]	[−0.01; 0.1]	[0.06; 0.21]	[0.1; 0.26]	[0.01; 0.17]
	(0.01)	(0.18)	(0.38)	(0.02)	(0.01)	(0.32)
Control group mean	239.1	238.10	237.45	246.75	243.96	246.09
Control group SD	51.18	48.04	45.42	52.93	48.68	46.59
Number of clusters	836	836	836	780	781	781
Observations	19,895	19,873	19,888	5,098	5,102	5,083

Notes: PAC stands for *Plan Apoyo Compartido*. *Randomised into PAC* is a dummy variable that takes value one if the student attends a school that was randomly assigned to the treatment group and zero otherwise. *Low income* is a dummy variable that takes value one if the student's family monthly income is less than 300,000 Chilean pesos, the minimum wage in such country. *High income* is a dummy that takes value one if *Low income* = 0. The effects shown are in units of the control group standard deviation (SD). 95% bootstrapped confidence intervals are shown in brackets. Romano-Wolf step-down p-values allowing for all 12 hypotheses from the two-sided tests are shown in parentheses. All regressions include cohort fixed effects. Clustering at the school level. Schools that implemented CLASS are excluded from the 2012 sample.

for lower-income boys and girls; however, the chance that these are false positives is quite high, given the adjusted p-values.[21]

3.5. *Discussion*

All in all, these results suggest that the PAC had a large and significant effect on the performance of fourth grade boys and girls from relatively higher-income backgrounds. This is clear evidence that the structured teaching intervention holds real promise. However, the effects on the lower-income children are much smaller and, when we break them down by gender, the only significant effect is confined to reading and to boys. The programme improved quality of education overall and improved the performance of boys and girls by the same amount, but did not reduce the disparities between socio-economic groups. Thus, we need further understanding on how to improve the learning outcomes of children from lower SES backgrounds, who tend to have worse levels of performance. The difficulty of intervening successfully for the most disadvantaged is consistent with much of the literature that shows complementarities between investments in children and earlier achievement (Cunha *et al.*, 2010; Attanasio *et al.*, 2019). It is also consistent with the results of other school interventions: for example, Machin *et al.* (2010) show that an

[21] Results from the estimation of the model with covariates are shown in Online Appendix table OA11 and are very similar.

inner-city school intervention in England improved most the performance of the children with higher achievement, although they also showed that the largest effects were observed in under-resourced schools, which is not surprising. This raises the urgency of how to design interventions for the most deprived populations and is likely to involve programmes specifically targeted at addressing developmental deficits of children in deprived populations from a very early age (Attanasio *et al.*, 2014; 2020; Gertler *et al.*, 2014).

We also documented that the CLASS intervention in 2012 had an impact on student outcomes. We interpret this impact as a Hawthorne effect caused by having cameras in the classroom. However, this impact was limited to control schools that did not receive PAC. There are two likely reasons for this: first, teachers in PAC-treated schools are already being subject to a great deal of monitoring as discussed in Section 1 so that an additional source of class observation does not significantly impact teachers' behaviour. Secondly, treated teachers are already following the tightly prescribed guidelines of the PAC, as we also show in Section 1, so that they have little room to significantly change their behaviour due to cameras in the classroom. This leaves the interesting question of whether class observation is itself a good substitute for prescribed teaching practices. While one can expect that observation without feedback is unlikely to have a sustained impact; we can only know this with longer term follow-up, and is therefore left for future research.

So far we have analysed how PAC affects students' achievement. In the next section we use the outcomes measured though the CLASS intervention to see whether the programme affected the way in which teachers and students interact.

4. The 2012 CLASS Experiment and Students' Learning

The small and growing literature that studies what characteristics of teachers matter the most for students' learning has recently started to focus on the quality of within classroom teacher–student interactions (Araujo *et al.*, 2016). In this section we study how important teacher–student interactions are to improving students' learning in our context, and whether the PAC had any positive impact on the quality of such interactions. As a preview of our results, we find that higher quality of teacher–student interactions is associated with better test scores of low-income students but are not correlated with test scores of high-income students. Moreover, we find that the PAC was not successful in improving teacher–student interactions by this measure.

4.1. *Measuring the Quality of Teacher–Student Interactions*

The main instrument used in this article to measure teacher–student interactions is the *Classroom Assessment Scoring System* or CLASS in its Upper Elementary version (fourth to sixth grade: see Pianta *et al.* (2008)). The CLASS is an instrument used in the education literature to measure the quality of teacher–student interactions, as a proxy to teachers' quality or effectiveness. To produce the CLASS measures, thoroughly trained coders watch and analyse videotaped classes and assign a score for teacher–student interactions in 11 dimensions. These dimensions can be grouped into three main domains: emotional support, classroom organisation and instructional support.[22] Coders look for very specific teachers' behaviours in each dimension, which are well described in the CLASS protocol that guides coders for their scoring.

[22] Emotional support includes the dimensions of Positive Climate, Negative Climate, Regard for Student Perspectives, and Teacher Sensitivity; Classroom Organisation includes the dimensions of Effective Behaviour Management, Instruc-

There are several studies that link better student outcomes (both in learning and in the development of socio-emotional skills) with teachers' scores in CLASS. Araujo *et al.* (2016) present a brief review of this literature for the United States and perform a study for kindergarten children in Ecuador. However, to the best of our knowledge, no study in the literature analyses the effect of CLASS on test scores for elementary school kids in developing countries.

In 2012, fourth grade teachers in participating schools were videotaped for four full lessons (see Subsection 1.3 for details on the random selection of PAC schools to participate in the CLASS intervention). A total of 185 teachers were filmed following the CLASS protocol.[23]

The coding was done by ten coders and a supervisor who were carefully trained and selected.[24] Each of the four school hours filmed per teacher were divided into 15-minute segments, and one segment per hour was coded (for a total of 760 segments) in each of the CLASS dimensions. Following the CLASS protocol, the score on each dimension was based on a scale from 1 to 7 ('low' for scores 1–2, 'medium' for scores 3–5, and 'high' for scores 6–7). The final CLASS scores for each domain consisted of the average across dimensions within the corresponding domain. For the coding, videos were randomly assigned to the ten certified coders. The coding process lasted five weeks. During the first week of coding, 100% of the videos were double coded. The double coding was expected to be gradually reduced in the following weeks if reliability rates remained above 80%.[25] Overall, 52% of the videos were double coded, with an average reliability rate of 84.2%.[26] This inter-coder reliability is comparable to that found in other studies. For example, Brown *et al.* (2010), Araujo *et al.* (2016) report an inter-coder reliability rate of 83% for the 12% of the classroom observations that were double coded.[27]

4.2. *CLASS, Teacher Performance and Programme Effects*

In Table 8 we report the association between CLASS and SIMCE scores for the 2012 cohort by school SES and in Table 9 by gender and student family income. The effects reported are in SD units of the SIMCE score for the corresponding demographic group and subject.

Table 8 shows that performance in CLASS is significantly and positively associated with the test scores of students in the most disadvantaged schools. Consistently, the most striking result from Table 9 is the association between better student–teacher interactions (reflected in a higher

tional Learning Formats, and Productivity; and Instructional Climate includes the dimensions of Language Modelling, Concept Development, Analysis and Inquiry, and Quality of Feedback.

[23] The fieldwork and coding according to CLASS was co-ordinated and implemented by a team of the Centro de Politicas Comparadas de Educacion from the Universidad Diego Portales, which had already applied CLASS for the evaluation of another programme in Chile, Un buen Comienzo (Leyva *et al*., 2015).

[24] The coders had to take a two-day training course provided by a Teachstone certified trainer, who also had the experience of applying CLASS to the Chilean context. After the course, coders took a four-hour online test (developed by Teachstone), which asks the candidate to watch and code five segments of model videos. The candidate is approved when achieving a reliability rate of at least 80% in all videos and at least in two of the videos the same reliability in all CLASS dimensions. Only the candidates that passed the test were certified to be CLASS coders in this evaluation. In addition, before starting the coding of the videos for the PAC evaluation, coders participated in another training course to adapt their knowledge of CLASS to the Chilean context. The training included watching and coding videos of Chilean teachers, which were previously coded by experienced CLASS coders.

[25] Coding is considered reliable if the difference between the two coders' score is less than two points for each CLASS dimension.

[26] When a coding was not considered not reliable, a supervisor did a third coding, which was the final score attributed to that teacher.

[27] Araujo *et al.* (2016) get a higher inter-coder reliability rate (93%) double-coding 100% of the videos.

Table 8. *Association between CLASS and SIMCE 2012, by School Socio-economic Status.*

	Low SES			Medium SES		
	Reading	Maths	Science	Reading	Maths	Science
CLASS first principal component	0.085 [0.06; 0.11] (0.01)	0.085 [0.06; 0.11] (0.01)	0.083 [0.06; 0.11] (0.01)	0.029 [−0.03; 0.08] (0.54)	0.011 [−0.09; 0.09] (0.86)	0.029 [−0.05; 0.09] (0.62)
SIMCE score mean	247.934	241.684	236.145	260.803	252.08	250.099
SIMCE score SD	51.984	48.961	47.098	48.858	45.899	44.11
Number of clusters	114	114	114	22	22	22
Observations	3,608	3,572	3,582	720	713	713

Notes: SES variables refer to the school socio-economic status as described in Section 3. The effects shown are in units of the corresponding SIMCE test score standard deviation (SD). 95% bootstrapped confidence intervals are shown in brackets. Romano-Wolf step-down *p*-values allowing for all six hypotheses from the two-sided tests are shown in parentheses. Clustering at the school level.

Table 9. *Association between CLASS and SIMCE 2012, by Students' Gender and Income.*

	Girls					
	Low income			High income		
	Reading	Maths	Science	Reading	Maths	Science
CLASS first principal component	0.074 [0.04; 0.11] (0.01)	0.075 [0.04; 0.11] (0.01)	0.067 [0.04; 0.1] (0.01)	0.075 [0.04; 0.12] (0.01)	0.013 [−.04;.06] (0.69)	0.031 [−0.01; 0.07] (0.32)
SIMCE score mean	253.862	239.124	234.441	265.009	252.042	246.767
SIMCE score SD	49.475	47.855	44.145	47.939	49.585	45.668
Number of clusters	128	128	128	102	102	102
Observations	1,415	1,404	1,403	297	296	298
	Boys					
	Low income			High income		
	Reading	Maths	Science	Reading	Maths	Science
CLASS first principal component	0.086 [0.06; 0.11] (0.01)	0.092 [0.06; 0.12] (0.01)	0.089 [0.06; 0.12] (0.01)	0.058 [0.01; 0.11] (0.19)	0.048 [0; 0.1] (0.32)	0.091 [0.04; 0.15] (0.03)
SIMCE score mean	244.621	245.274	239.259	254.037	252.437	250.899
SIMCE score SD	53.15	48.746	49.304	50.999	49.422	46.764
Number of clusters	129	129	129	109	109	109
Observations	1,472	1,461	1,461	365	360	361

Notes: Low income is a dummy variable that takes value one if the student's family monthly income is less than 300,000 Chilean pesos, the minimum wage in such country. *High income* is a dummy that takes value one if *Low income* = 0. The effects shown are in units of the corresponding SIMCE test score standard deviation (SD). 95% bootstrapped confidence intervals are shown in brackets. Romano-Wolf step-down *p*-values allowing for all 12 hypotheses from the two-sided tests are shown in parentheses. Clustering at the school level.

CLASS score) and the performance of low-income students. In effect, one additional standard deviation in the principal component of CLASS scores is associated with a higher SIMCE test score for low-income students of between 0.07 and 0.09 of SD units. For higher-income students, effects are smaller and in some cases insignificant. These results are potentially important and consistent with the finding that teachers have a large causal impact on student performance (see Rivkin *et al.*, 2005). Taken at face value, the results imply that moving a lower-income student from the bottom 2% of teachers to the top 2% can improve outcomes of low-income students

Table 10. *Impact of PAC on CLASS, Classroom Level.*

	Dependent variable: CLASS first principal component			
	(1) OLS	(2) IV	(3) OLS	(4) IV
Randomised into PAC	−0.5274		−0.2361	
	[−1.133; 0.112]		[−1.336; 0.133]	
Implemented PAC		−0.6153		−0.2697
		[−0.903; 0.33]		[−1.043; 0.384]
Covariates	No	No	Yes	Yes
Observations	185	185	184	184

Notes: PAC stands for *Plan Apoyo Compartido*. *Randomised into PAC* is a dummy variable that takes value one if the school was randomly assigned to the treatment group and zero otherwise. *Implemented PAC* is a dummy variable that takes value one if the school participated in PAC and zero otherwise. In Columns (2) and (4) we instrument actual participation in PAC with the random assignment to PAC. Covariates include an indicator of the income group the classroom belongs to, the type of administration of the school (private or public), average SIMCE scores of the school for the period 2005–9, general experience of the teacher and the school principal, and tenure of the teacher and the principal in the school. 95% bootstrapped confidence intervals are shown in brackets. Clustering at the school level.

by between 0.6 and 0.8 of a standard deviation.[28] There is no causality implied or presumed by these results, which may be entirely due to sorting of better low-income students to better teachers (say, because of more proactive parents). However, it does pose an interesting question as to whether improving interactions could actually lead to better performance for low-income students. We thus examine whether the CLASS score was affected by the programme.

The impact of PAC on CLASS Table 10 shows the result of regressing CLASS scores on treatment allocation and covariates. The results consistently suggest that the programme has no significant effect on teacher–classroom interactions in 2012. Given that CLASS is ranked based on interactions that may somehow be discouraged by the PAC intervention, it may well be that CLASS is not a particularly good way of understanding the mechanisms through which the PAC operated. It may also be that the improvements we observed relate to practices not captured by CLASS, namely the more structured approach to lesson planning and the monitoring of students. On the other hand, the loss in sample size has meant that these estimates are not as precise as we would desire. However, the association of CLASS scores with better performance of low-income students suggests that improving outcomes for deprived populations should focus more on how teachers interact with low-income students, as well as improving practices tested with this intervention. It is important to remember that, after all, the scripted instruction intervention was successful; moreover, while the impacts are concentrated among the relatively better off, the population we are studying is already lower-income and attending underperforming schools.

5. Discussion and Conclusions

Improving quality of education has proved to be a major policy challenge. While the quality of teachers seems to be of central importance, the policy question remains, particularly because it is not clear what constitutes a priori a good teacher. One possibility is to consider more structured teaching methods, which define carefully what teachers are supposed to do and monitor the progress of students throughout the year. This is the idea underlying PAC, the

[28] The second and 98th percentiles of the CLASS score principal component are −4.6 and 4.34, respectively.

programme we are analysing in this article, which was launched in 2011 in Chile. Through standardised teaching material (class preparation) and through the support of internal and external paedagogic teams, the programme aimed to reduce the gap, as measured by the standardised test SIMCE, between the poorest student population and the national average. The programme was designed with a gradual implementation, which implied that only half of eligible schools could be offered the programme. These were selected randomly, which forms the basis of our evaluation.

The results for the first 2011 cohort of implementation were encouraging implying overall improvements in reading. In the second year of the programme, the effects increase for all subjects and become significant also for maths and science. When we break down the impacts by school socio-economic status (SES) we find that the positive effects of the programme are concentrated among children in schools with higher SES. Moreover, heterogeneous effects by students' gender and family income reveal positive and significant effects particularly for students originating from relatively higher-income families. Importantly, the effects are the same for boys and girls. Overall, it seems that the programme can improve outcomes, but it mainly improves results for the relatively better off.[29]

In order to begin understanding what lies behind these results we used the CLASS system to record classroom sessions and score teacher–student interactions. CLASS is a well-documented instrument in the education literature that uses a very rigorous protocol to score the ways in which students and teachers interact along various dimensions (class organisation, instructional support and emotional support, measured in 11 different sub-dimensions). We find that CLASS scores are correlated with SIMCE results: a better CLASS score is associated with better-performing students, particularly among those from lower-income backgrounds. No causality should, of course, be attributed since it may well be the case that teachers interact better when they are interacting with better-performing students. We then examine whether the programme shifted the CLASS score, by improving teacher–student interactions, and we find no effect at all. However, since CLASS altered the performance of the control group it is hard to interpret this result: in other words, the group that was observed via CLASS is not comparable to the one that was not.

Despite the overall success of the programme, the urgent question of how to improve outcomes of children from the most deprived backgrounds remains. As such, research seems to show the answer may lie in early childhood development programmes, which attempt to ensure that children from the most deprived backgrounds have better cognitive development and access to improved opportunities from the earliest possible age, making them potentially better placed to benefit from schooling.

[29] PAC was discontinued in 2014 by the entering administration of the Ministry of Education.

Appendix A. Evidence on Implementation of PAC

Table A1. *Correlation between Participation in PAC and PAC Dimensions, Classroom Level Analysis.*

	(1) Class structure	(2) Reflective thinking	(3) Norms and schedule	(4) Evaluation	(5) Use of manual
Implemented PAC	0.0129	0.0369	0.0582**	0.0279	0.2300***
	(0.0238)	(0.0291)	(0.0225)	(0.0281)	(0.0733)
Constant	0.4589***	0.5397***	0.7390***	0.8730***	0.1619***
	(0.0171)	(0.0175)	(0.0153)	(0.0200)	(0.0354)
Observations	179	179	179	179	179

Notes: PAC stands for *Plan Apoyo Compartido*. *Implemented PAC* is a dummy variable that takes value one if the school participated in PAC and zero otherwise. **Variable significant at the 5% level. ***Variable significant at the 1% level. Clustering at the school level (clustered standard errors shown in parentheses).

Table A2. *Correlation between Participation in PAC and PAC Components, School Level Analysis.*

	(1)	(2)	(3)	(4)	(5)
	Annual planning performed by			Support to teachers	
	PAC team	ELE team	Teachers	ELE classroom observations	Feedback to teachers
Implemented PAC	0.2546***	0.1888**	−0.1830**	1.1593***	1.1176***
	(0.0641)	(0.0858)	(0.0810)	(0.3991)	(0.3826)
Constant	0.0385*	0.3974***	0.7692***	3.3924***	3.6582***
	(0.0219)	(0.0558)	(0.0481)	(0.2442)	(0.2439)
Observations	136	136	136	137	137

Notes: PAC stands for *Plan Apoyo Compartido*. ELE refers to the Education Leadership Team (consisting of the school principal, the head of the technical and pedagogic office of the school, and two distinguished teachers, as described in Section 1). *Implemented PAC* is a dummy variable that takes value one if the school participated in PAC and zero otherwise. *Variable significant at the 10% level. **Variable significant at the 5% level. ***Variable significant at the 1% level. Robust standard errors in parentheses.

Appendix B. Summary Statistics

The following table shows summary statistics. The names of columns indicate the set of students over which summary statistics are calculated.

Columns labelled *Reading*, *Math*s and *Science test takers* indicate the pool of students that took each of the corresponding subject tests. This corresponds to the *post-attrition* sample since, for each test, there is a small set of students who did not take the test (we discuss this in Subsection 1.5).

Sub-columns labelled *PAC=0* and *PAC=1* refer to treatment status. *PAC* is a dummy variable that takes value one if the student goes to a school that was invited to participate in the programme through the randomisation, and zero otherwise. In what follows, we refer to the set of students such that *PAC=0* as the *control group* and to the set of students such that *PAC=1* as the *treatment group*.

In turn, the table is divided in three panels (*2011*, *2012-excluding CLASS sample*, and *2012*), indicating the fourth grade cohorts considered in this article.

The names of rows indicate the variable for which we show summary statistics.

SIMCE scores (Reading, Maths and Science) refer to the grade obtained by students in the SIMCE subject tests.

Baseline characteristics indicate characteristics of the students that do not change because of treatment. They include student demographic characteristics and education of parents. Student demographics are *Female* (a dummy variable that takes value one if the student is a female and zero otherwise), *Low income* (a dummy variable that takes value one if the student's family monthly income is less than 300,000 Chilean pesos, or around 600 dollars at that time),[30] *Nuclear, Extended* and *Other* family (three dummies that indicate the family structure of the student), and *Nbr years failed* (a count variable that captures the number of primary school years the student had to retake previous to the forth grade). Mother's and father's education refer to the highest education level reached by the student's mother and father. These include *No education*, *Incomplete primary*, *Primary*, *Incomplete high school*, *High school*, *Incomplete college*, and *College*.

[30] SIMCE includes a 1 to 9 scale for the income reported by the parents in the questionnaire that they complete. We consider 'low-income' those reporting in categories 1 to 4. It is important to note, though, that students in our sample belong mainly to low- to middle-income families in Chile.

Table B1. *Summary Statistics: Post-Attrition Samples.*

	Reading test takers						Maths test takers						Science test takers					
	PAC=0			PAC=1			PAC=0			PAC=1			PAC=0			PAC=1		
	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD
Panel A: 2011																		
SIMCE scores:																		
Reading	6,886	245	50	23,850	248	50												
Maths							6,903	236	47	23,828	238	48						
Science													6,911	237	44	23,854	238	44
Baseline characteristics:																		
Students demographics																		
Female	6,554	0.472	0.499	22,861	0.484	0.5	6,653	0.473	0.499	23,152	0.484	0.5	6,660	0.473	0.499	23,162	0.484	0.5
Low income	6,219	0.841	0.366	21,470	0.813	0.39	6,235	0.841	0.366	21,490	0.813	0.39	6,244	0.841	0.366	21,502	0.813	0.39
Nuclear family	6,886	0.613	0.487	23,850	0.617	0.486	6,903	0.613	0.487	23,828	0.619	0.486	6,911	0.613	0.487	23,854	0.617	0.486
Extended family	6,886	0.244	0.43	23,850	0.239	0.426	6,903	0.244	0.43	23,828	0.239	0.426	6,911	0.244	0.43	23,854	0.239	0.427
Other family	6,886	0.143	0.35	23,850	0.144	0.352	6,903	0.143	0.35	23,828	0.143	0.35	6,911	0.143	0.35	23,854	0.143	0.35
Nbr years failed	6,187	0.238	0.527	21,353	0.232	0.531	6,202	0.239	0.528	21,367	0.232	0.531	6,211	0.239	0.528	21,387	0.232	0.531
Mother's education																		
No education	6,201	0.007	0.086	21,352	0.006	0.078	6,215	0.007	0.086	21,374	0.006	0.077	6,225	0.008	0.087	21,387	0.006	0.078
Inc. primary	6,201	0.179	0.383	21,352	0.174	0.379	6,215	0.178	0.383	21,374	0.174	0.379	6,225	0.178	0.383	21,387	0.174	0.379
Primary	6,201	0.17	0.375	21,352	0.164	0.37	6,215	0.17	0.375	21,374	0.164	0.37	6,225	0.169	0.375	21,387	0.164	0.37
Inc. high school	6,201	0.235	0.424	21,352	0.218	0.413	6,215	0.235	0.424	21,374	0.218	0.413	6,225	0.236	0.425	21,387	0.219	0.413
High school	6,201	0.324	0.468	21,352	0.341	0.474	6,215	0.324	0.468	21,374	0.341	0.474	6,225	0.324	0.468	21,387	0.341	0.474
Inc. college	6,201	0.034	0.182	21,352	0.038	0.191	6,215	0.034	0.182	21,374	0.038	0.191	6,225	0.034	0.181	21,387	0.037	0.19
College	6,201	0.051	0.22	21,352	0.059	0.235	6,215	0.051	0.22	21,374	0.059	0.235	6,225	0.051	0.22	21,387	0.059	0.235
Father's education																		
No education	5,992	0.008	0.089	20,577	0.008	0.089	6,006	0.008	0.09	20,597	0.008	0.089	6,013	0.008	0.089	20,611	0.008	0.09
Inc. primary	5,992	0.17	0.376	20,577	0.158	0.364	6,006	0.17	0.376	20,597	0.157	0.364	6,013	0.17	0.376	20,611	0.157	0.364
Inc. primary	5,992	0.16	0.366	20,577	0.162	0.368	6,006	0.16	0.367	20,597	0.161	0.368	6,013	0.16	0.366	20,611	0.162	0.368
Inc. high school	5,992	0.248	0.432	20,577	0.23	0.421	6,006	0.248	0.432	20,597	0.23	0.421	6,013	0.247	0.431	20,611	0.23	0.421
High school	5,992	0.329	0.47	20,577	0.345	0.475	6,006	0.328	0.469	20,597	0.344	0.475	6,013	0.328	0.47	20,611	0.344	0.475
Inc. college	5,992	0.036	0.187	20,577	0.043	0.202	6,006	0.036	0.187	20,597	0.043	0.203	6,013	0.036	0.187	20,611	0.043	0.203
College	5,992	0.05	0.217	20,577	0.056	0.229	6,006	0.05	0.219	20,597	0.056	0.23	6,013	0.05	0.218	20,611	0.056	0.229

Table B1. *Continued*

	Reading test takers						Maths test takers						Science test takers					
	PAC=0			PAC=1			PAC=0			PAC=1			PAC=0			PAC=1		
	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD
Panel B: 2012-excluding CLASS sample																		
SIMCE scores:																		
Reading	4,685	242	51	20,772	249	52												
Maths							4,658	237	47	20,727	242	49						
Science													4,654	233	46	20,685	237	46
Baseline characteristics:																		
Students demographics																		
Female	4,412	0.481	0.5	20,342	0.481	0.5	4,293	0.484	0.5	19,900	0.483	0.5	4,290	0.483	0.5	19,848	0.482	0.5
Low income	3,946	0.792	0.406	18,045	0.779	0.415	3,960	0.793	0.405	18,119	0.779	0.415	3,954	0.793	0.405	18,067	0.78	0.415
Nuclear family	4,685	0.563	0.496	20,772	0.572	0.495	4,658	0.569	0.495	20,727	0.576	0.494	4,654	0.569	0.495	20,685	0.576	0.494
Extended family	4,685	0.234	0.423	20,772	0.25	0.433	4,658	0.237	0.425	20,727	0.252	0.434	4,654	0.236	0.425	20,685	0.252	0.434
Other family	4,685	0.203	0.402	20,772	0.177	0.382	4,658	0.195	0.396	20,727	0.172	0.377	4,654	0.195	0.396	20,685	0.172	0.378
Nbr years failed	3,947	1.25	0.592	18,055	1.217	0.561	3,962	1.253	0.595	18,129	1.218	0.561	3,956	1.251	0.593	18,076	1.218	0.56
Mother's education																		
No education	3,836	0.007	0.08	17,474	0.006	0.075	3,852	0.007	0.082	17,554	0.006	0.075	3,847	0.006	0.08	17,505	0.006	0.075
Inc. primary	3,836	0.178	0.383	17,474	0.163	0.37	3,852	0.179	0.384	17,554	0.164	0.37	3,847	0.179	0.383	17,505	0.163	0.369
Primary	3,836	0.173	0.379	17,474	0.169	0.374	3,852	0.175	0.38	17,554	0.168	0.374	3,847	0.174	0.379	17,505	0.169	0.375
Inc. high school	3,836	0.217	0.412	17,474	0.212	0.408	3,852	0.217	0.412	17,554	0.211	0.408	3,847	0.217	0.412	17,505	0.211	0.408
High school	3,836	0.333	0.471	17,474	0.351	0.477	3,852	0.331	0.471	17,554	0.35	0.477	3,847	0.332	0.471	17,505	0.35	0.477
Inc. college	3,836	0.038	0.191	17,474	0.04	0.197	3,852	0.038	0.192	17,554	0.04	0.197	3,847	0.038	0.192	17,505	0.04	0.196
College	3,836	0.054	0.225	17,474	0.06	0.237	3,852	0.054	0.226	17,554	0.06	0.238	3,847	0.054	0.225	17,505	0.061	0.239
Father's education																		
No education	3,675	0.01	0.097	16,720	0.007	0.083	3,686	0.009	0.097	16,804	0.007	0.083	3,682	0.01	0.097	16,756	0.007	0.083
Inc. primary	3,675	0.155	0.362	16,720	0.158	0.364	3,686	0.155	0.362	16,804	0.158	0.364	3,682	0.155	0.362	16,756	0.157	0.364
Primary	3,675	0.176	0.381	16,720	0.162	0.368	3,686	0.176	0.381	16,804	0.162	0.369	3,682	0.174	0.379	16,756	0.162	0.369
Inc. high school	3,675	0.221	0.415	16,720	0.218	0.413	3,686	0.221	0.415	16,804	0.218	0.413	3,682	0.221	0.415	16,756	0.219	0.414
High school	3,675	0.351	0.477	16,720	0.352	0.478	3,686	0.35	0.477	16,804	0.351	0.477	3,682	0.351	0.477	16,756	0.35	0.477
Inc. college	3,675	0.037	0.189	16,720	0.044	0.206	3,686	0.037	0.189	16,804	0.045	0.206	3,682	0.037	0.189	16,756	0.045	0.207
College	3,675	0.051	0.22	16,720	0.06	0.237	3,686	0.051	0.221	16,804	0.059	0.236	3,682	0.051	0.22	16,756	0.06	0.237

Table B1. *Continued*

	Reading test takers						Maths test takers						Science test takers					
	PAC=0			PAC=1			PAC=0			PAC=1			PAC=0			PAC=1		
	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD
Panel C: 2012																		
SIMCE scores:																		
Reading	7,141	246	51	23,353	248	52												
Maths							7,095	239	47	23,273	242	49						
Social science													7,105	235	47	23,226	237	46
Baseline characteristics:																		
Students demographics																		
Female	6,773	0.481	0.5	22,810	0.481	0.5	6,572	0.483	0.5	22,303	0.482	0.5	6,580	0.483	0.5	22,245	0.482	0.5
Low income	6,048	0.8	0.4	20,229	0.783	0.412	6,065	0.801	0.399	20,292	0.783	0.412	6,063	0.801	0.399	20,236	0.783	0.412
Nuclear family	7,141	0.566	0.496	23,353	0.572	0.495	7,095	0.572	0.495	23,273	0.576	0.494	7,105	0.571	0.495	23,226	0.575	0.494
Extended family	7,141	0.235	0.424	23,353	0.248	0.432	7,095	0.237	0.425	23,273	0.25	0.433	7,105	0.236	0.425	23,226	0.249	0.433
Other family	7,141	0.2	0.4	23,353	0.18	0.385	7,095	0.191	0.393	23,273	0.175	0.38	7,105	0.193	0.395	23,226	0.175	0.38
Nbr years failed	6,050	1.243	0.588	20,242	1.219	0.565	6,068	1.245	0.593	20,305	1.22	0.564	6,066	1.243	0.592	20,248	1.219	0.564
Mother's education																		
No education	5,873	0.006	0.076	19,580	0.006	0.075	5,894	0.006	0.077	19,648	0.006	0.075	5,891	0.006	0.076	19,594	0.006	0.075
Inc. primary	5,873	0.178	0.383	19,580	0.166	0.372	5,894	0.18	0.384	19,648	0.167	0.373	5,891	0.179	0.384	19,594	0.166	0.372
Primary	5,873	0.173	0.378	19,580	0.169	0.375	5,894	0.173	0.378	19,648	0.169	0.375	5,891	0.172	0.378	19,594	0.17	0.376
Inc. high school	5,873	0.214	0.41	19,580	0.213	0.409	5,894	0.214	0.41	19,648	0.213	0.41	5,891	0.214	0.41	19,594	0.213	0.409
High school	5,873	0.338	0.473	19,580	0.348	0.476	5,894	0.336	0.472	19,648	0.348	0.476	5,891	0.337	0.473	19,594	0.348	0.476
Inc. college	5,873	0.036	0.187	19,580	0.039	0.193	5,894	0.036	0.187	19,648	0.039	0.193	5,891	0.036	0.187	19,594	0.039	0.192
College	5,873	0.055	0.228	19,580	0.058	0.235	5,894	0.055	0.228	19,648	0.059	0.235	5,891	0.055	0.228	19,594	0.059	0.236
Father's education																		
No education	5,630	0.009	0.092	18,745	0.007	0.085	5,647	0.009	0.092	18,820	0.007	0.085	5,644	0.009	0.092	18,768	0.007	0.085
Inc. primary	5,630	0.153	0.36	18,745	0.158	0.365	5,647	0.153	0.36	18,820	0.158	0.365	5,644	0.154	0.361	18,768	0.158	0.364
Inc. primary	5,630	0.172	0.377	18,745	0.165	0.371	5,647	0.172	0.377	18,820	0.165	0.371	5,644	0.17	0.376	18,768	0.165	0.372
Inc. high school	5,630	0.222	0.416	18,745	0.217	0.412	5,647	0.222	0.416	18,820	0.217	0.412	5,644	0.222	0.416	18,768	0.218	0.413
High school	5,630	0.357	0.479	18,745	0.351	0.477	5,647	0.356	0.479	18,820	0.35	0.477	5,644	0.357	0.479	18,768	0.35	0.477
Inc. college	5,630	0.035	0.183	18,745	0.043	0.203	5,647	0.035	0.184	18,820	0.043	0.204	5,644	0.035	0.184	18,768	0.043	0.204
College	5,630	0.054	0.225	18,745	0.058	0.234	5,647	0.054	0.226	18,820	0.058	0.234	5,644	0.054	0.225	18,768	0.059	0.235

Table B2. *Percent of Schools by Socio-economic Status (SES).*

SES	All Chilean schools 2011	PAC eligible schools Treatment	PAC eligible schools Control	PAC eligible schools excluding CLASS sample Treatment	PAC eligible schools excluding CLASS sample Control
Low	29	16.7	16.4	16.5	15.2
Medium-low	35	60.2	66.7	59.7	68
Medium	21	22.8	16.9	23.4	16.8
Medium-high	9	0.31	0	0.34	0
High	5	0	0	0	0
N	7,738	648	195	581	125

Notes: PAC stands for *Plan Apoyo Compartido*. *PAC eligible schools* refers to schools that were elgible to participate in the randomisation. *Treatment* refers to schools randomly assigned to receive PAC and *Control* refers to schools randomly assigned not to receive PAC. *SES* variables refer to the school socio-economic status as described in Section 3.

Appendix C. The Impact of Participating in CLASS on SIMCE

Table C1. *Impact of Participation in CLASS on SIMCE 2012.*

	PAC control group			PAC treatment group		
	Reading	Maths	Science	Reading	Maths	Science
In CLASS sample	0.229	0.18	0.206	−0.043	−0.014	−0.024
	[0.14; 0.33]	[0.08; 0.3]	[0.11; 0.32]	[−0.14; 0.06]	[−0.12; 0.1]	[−0.12; 0.07]
Not in CLASS mean	242.416	236.65	232.72	248.774	241.99	237.227
Not in CLASS SD	50.941	47.259	45.799	51.551	49.13	46.173
Number of clusters	195	195	195	644	644	644
Observations	7,141	7,095	7,105	23,353	23,273	23,226

Notes: PAC stands for *Plan Apoyo Compartido*. *In CLASS sample* is a dummy variable that takes value one if the student attends a school that was randomised into the CLASS intervention and zero otherwise. The effects shown are in units of the *Not in CLASS sample* standard deviation (SD). 95% bootstrapped confidence intervals are shown in brackets.

Appendix D. Results by Gender and Income (Separately)

Table D1. *Impact of PAC on SIMCE by Gender, with Cohort Fixed Effects.*

	Girls			Boys		
	Reading	Maths	Science	Reading	Maths	Science
Randomised into PAC	0.109	0.077	0.058	0.108	0.097	0.057
	[0.06; 0.16]	[0.02; 0.13]	[0; 0.11]	[0.06; 0.16]	[0.04; 0.16]	[0; 0.11]
	(0.01)	(0.08)	(0.12)	(0.01)	(0.02)	(0.12)
Control Group Mean	249.751	234.539	233.059	239.225	238.243	237.943
Control Group SD	48.659	45.635	43.42	51.38	48.142	45.678
Number of clusters	835	836	836	837	838	838
Observations	26,068	26,045	26,006	28,101	27,953	27,954

Notes: PAC stands for *Plan Apoyo Compartido*. *Randomised into PAC* is a dummy variable that takes value one if the student attends a school that was randomly assigned to the treatment group and zero otherwise. The effects shown are in units of the control group standard deviation (SD). 95% bootstrapped confidence intervals are shown in brackets. Romano-Wolf step-down p-values allowing for all six hypotheses from the two-sided tests are shown in parentheses. All regressions include cohort fixed effects. Clustering at the school level. Schools that implemented CLASS are excluded from the 2012 sample.

Table D2. *Impact of PAC on SIMCE by Family Income, with Cohort Fixed Effects.*

	Low income			High income		
	Reading	Maths	Science	Reading	Maths	Science
Randomised into PAC	0.089 [0.04; 0.14] (0.01)	0.074 [0.01; 0.13] (0.08)	0.044 [−0.01; 0.09] (0.18)	0.167 [0.11; 0.23] (0.01)	0.143 [0.07; 0.21] (0.01)	0.071 [0; 0.14] (0.18)
Control Group Mean	244.302	235.801	234.677	249.877	243.23	244.358
Control Group SD	50.17	46.92	44.319	51.423	47.994	46.75
Number of clusters	842	842	842	823	824	824
Observations	39,879	39,985	39,961	9,801	9,819	9,806

Notes: PAC stands for *Plan Apoyo Compartido*. *Randomised into PAC* is a dummy variable that takes value one if the student attends a school that was randomly assigned to the treatment group and zero otherwise. *Low income* is a dummy variable that takes value one if the student's family monthly income is less than 300,000 Chilean pesos, the minimum wage in such country. *High income* is a dummy that takes value one if *Low income* = 0. The effects shown are in units of the control group standard deviation (SD). 95% bootstrapped confidence intervals are shown in brackets. Romano-Wolf step-down *p*-values allowing for all six hypotheses from the two-sided tests are shown in parentheses. All regressions include cohort fixed effects. Clustering at the school level. Schools that implemented CLASS are excluded from the 2012 sample.

World Bank
Department of Economics, Yale University, NBER, IFS, IZA and CEPR
Department of Economics, University of Michigan

Additional Supporting Information may be found in the online version of this article:

Online Appendix
Replication Package

References

Albornoz, F., Anauati, M.V., Furman, M., Luzuriaga, M., Podestá, M.E. and Taylor, I. (2020). 'Training to teach science: experimental evidence from Argentina', *The World Bank Economic Review*, vol. 34(2), pp. 393–417.

Araujo, M.C., Carneiro, P., Cruz-Aguayo, Y. and Schady, N. (2016). 'A helping hand? Teacher quality and learning outcomes in kindergarten', *Quarterly Journal of Economics*, vol. 131(3), pp. 1415–53.

Attanasio, O., Cattan, S., Fitzsimmons, E., Meghir, C. and Codina, M.R. (2020). 'Estimating the production function for human capital: results from a randomized controlled trial in Colombia', *American Economic Review*, vol. 110(1), pp. 48–85.

Attanasio, O., Fernandez, C., Fitzsimons, E., Grantham-McGregor, S., Meghir, C. and Rubio-Codina, M. (2014). 'Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: a cluster randomised controlled trial', *BMJ—British Medical Journal*, vol. 349, p. g6126.

Attanasio, O., Meghir, C. and Nix, E. (2019). 'Human capital development and parental investment in India', NBER Working Paper No. 21740.

Bassi, M., Mateo Díaz, M., Blumberg, R.L. and Reynoso, A. (2018). 'Failing to notice? Uneven teachers' attention to boys and girls in the classroom', *IZA Journal of Labor Economics*, vol. 7.

Brown, J.L., Jones, S.M., LaRusso, M.D. and Aber, J.L. (2010). 'Improving classroom quality: teacher influences and experimental impacts of the 4Rs program', *Journal of Educational Psychology*, vol. 102(1), pp. 153–67.

Chetty, R., Friedman, J.N. and Rockoff, J.E. (2014). 'Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates', *American Economic Review*, vol. 104(9), pp. 2593–632.

Cunha, F., Heckman, J.J. and Schennach, S.M. (2010). 'Estimating the technology of cognitive and noncognitive skill formation', *Econometrica*, vol. 78(3), pp. 883–931.

Duflo, E., Hanna, R. and Ryan, S.P. (2012). 'Incentives work: getting teachers to come to school', *American Economic Review*, vol. 102(4), pp. 1241–78.

Ganimian, A.J. and Murnane, R.J. (2014). 'Improving educational outcomes in developing countries: lessons from rigorous impact evaluations', NBER Working Paper No. 20284.

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeerch, C., Walker, S., Chang, S. and Grantham-McGregor, S. (2014). 'Labor market returns to an early childhood stimulation intervention in Jamaica', *Science*, vol. 344, pp. 998–1001.

He, F., Linden, L.L. and MacLeod, M. (2009). 'A better way to teach children to read? Evidence from a randomized controlled trial', Unpublished manuscript. New York, NY: Columbia University.

Imbens, G. and Angrist, J. (1994). 'Identification and estimation of local average treatment effects', *Econometrica*, vol. 62(2), pp. 467–75.

Kirschner, P.A., Sweller, J. and Clark, R.E. (2006). 'Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching', *Educational Psychologist*, vol. 41(2), pp. 75–86.

Landsberger, H.A. (1959). 'Hawthorne revisited', *Soc. Forces*, vol. 37(1), pp. 361–4.

Levitt, S.D. and List, J.A. (2011). 'Was there really a Hawthorne effect at the Hawthorne plant? An analysis of the original illumination experiments', *American Economic Journal: Applied Economics*, vol. 3(1), pp. 224–38.

Leyva, D., Weiland, C., Barata, M., Yoshikawa, H., Snow, C., Treviño, E. and Rolla, A. (2015). 'Teacher–child interactions in Chile and their associations with prekindergarten outcomes', *Child development*, vol. 86, pp. 781–99.

Machin, S., McNally, S. and Meghir, C. (2010). 'Resources and standards in urban schools', *Journal of Human Capital*, vol. 4(4), pp. 365–93.

Ministerio de Education. (2013). 'Plan Apoyo Compartido', Santiago, Chile.

OECD. (2015). *The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence*, PISA, OECD Publishing.

Pianta, R.C., Mashburn, A.J., Downer, J.T., Hamre, B.K. and Justice, L. (2008). 'Effects of web mediated professional development resources on teacher child interactions in pre kindergarten classrooms', *Early Childhood Research Quarterly*, vol. 23(4), pp. 431–51.

Rivkin, S.G., Hanushek, E.A. and Kain, J.F. (2005). 'Teachers, schools, and academic achievement', *Econometrica*, vol. 73(2), pp. 417–58.

Romano, J.P. and Wolf, M. (2005). 'Stepwise multiple testing as formalized data snooping', *Econometrica*, vol. 73(4), pp. 1237–82.

Slavin, R.E., Lake, C., Chambers, B., Cheung, A. and Davis, S. (2009). 'Effective reading programs for the elementary grades: a best-evidence synthesis', *Review of Educational Research*, vol. 79(4), pp. 1397–466.