NBER WORKING PAPER SERIES

EDUCATION QUALITY AND TEACHING PRACTICES

Marina Bassi
Costas Meghir
Ana Reynoso

Working Paper 22719
http://www.nber.org/papers/w22719

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2016

The authors would like to thank Daniel Alonso for his valuable assistance in the analysis of the data. We would also like to recognize 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. Ana Reynoso was funded by the IADB. All errors and opinions are our own. Data collection for this project was financed by the Inter-American Development Bank. Administrative data was obtained from the Ministry of Education of Chile. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research, the Inter-American Development Bank, its Board of Directors, or the countries they represent.

Education Quality and Teaching Practices
Marina Bassi, Costas Meghir, and Ana Reynoso
NBER Working Paper No. 22719
October 2016
JEL No. I21,I24,I25,I3

ABSTRACT

This paper uses a RCT to estimate the effectiveness of guided instruction methods as implemented in under-performing schools in Chile. The intervention improved performance substantially for the first cohort of students, but not the second. The effect is mainly accounted for by children from relatively higher income backgrounds. Based on the CLASS instrument we document that 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 can improve outcomes, but it is a challenge to sustain the impacts and to reach the most deprived children.

Marina Bassi
Inter American Development Bank
1300 New York Avenue, N.W.
Washington, D.C. 20577
marinab@iadb.org

Ana Reynoso
Yale University
37 Hillhouse Avenue
New Haven, CT 06511
ana.reynoso@yale.edu

Costas Meghir
Department of Economics
Yale University
37 Hillhouse Avenue
New Haven, CT 06511
and IZA
and also NBER
c.meghir@yale.edu

1 Introduction

There is an urgent need to improve the quality of education in deprived areas. While it is well understood that good teachers can have a large impact (Rivkin *et al.* (2005a) and Chetty *et al.* (2014)) it is neither clear how to identify them, nor is it feasible to improve their stock soon enough for the current and upcoming cohorts of children. So the key question is whether we can improve outcomes by identifying and implementing innovative teaching practices. The educational psychology literature focusses on the method of instruction as an approach to improve outcomes. 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 standardized ways.

However, these methods can be controversial and there is an active debate on the extent to which prescriptive methods can be successful. While advocates 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 maximized 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 & 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 programs like Success for All and DISTAR in the United States Slavin *et al.* (2009).

In this paper we contribute to the understanding of the effectiveness of direct instruction approaches in schools that serve deprived populations, by analyzing the impact of a large-scale guided instruction program in Chile aimed low performing schools. We focus on the performance of students in the national standardized Math, Language, and Science tests and our results are based on a school-level randomized trial.

The program in question, known as *Plan Apoyo Compartido* (henceforth, PAC), was implemented by the Chilean Ministry of Education in 2011. The main intervention of the program 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 program 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 standardized Education Quality Measurement System evaluations (henceforth SIMCE evaluations, for its name in Spanish). We concentrate the analysis on students who were in their fourth grade of elementary school in years 2011 and 2012 and attended eligible schools.[1]

Our results suggest that only the most advantaged students within treated schools (students from higher

[1] A precedent to this study by He *et al.* (2009) evaluates a scripted reading preschool program in Mumbai, India. Unlike theirs, our paper focuses on fourth grade primary school students and, to the best of our knowledge, provides the first experimental evaluation of a teacher instruction intervention program for this age group.

income families within our lower income population) benefit from the program. For the 2011 cohort, middle-high income boys attending schools participating in the program improved SIMCE scores by almost 20% of a test score standard deviation with respect to comparable boys in control schools. For the 2012 cohort, middle-high income girls in treated schools improved SIMCE scores by more than 20% of a test score standard deviation relative to girls in control schools. These results are strongly robust to adjustments in our inference strategy to control for multiple testing.[2]

To better understand the impact of PAC on students' test scores we analyze the effects of the program on the quality of teacher-students interactions based on the CLASS (Classroom Assessment Scoring System) (Pianta et al., 2008), also used in Araujo *et al.* (2016). A random subsample of treatment and control schools from the PAC program where invited to participate in the CLASS 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' behaviors that coders look for (we provide details about the CLASS experiment in section 5). Consistent with Araujo *et al.* (2016), we first show that CLASS scores correlate positively and significantly with students' performance, and particularly for those from lower income background. Then, we show that PAC did not cause significant improvements in CLASS scores, which may explain why low income students were not impacted.

The paper is organized as follows. The next section describes the program intervention, the experimental design and the data used in this paper. Section 3 describes the identification and inference strategies. Section 4 presents the main results of the paper. Section 5 studies the importance of teacher-students' interactions to improve performance and the impact of PAC on these interactions. Finally, section 6 concludes.

2 Experimental design, data, and randomization check

2.1 Plan Apoyo Compartido (PAC)

PAC was implemented by the Ministry of Education of Chile in 2011 as a targeted educational policy providing technical and pedagogical support to schools historically performing below average in the national standardized test, SIMCE. It aimed at improving student' learning outcomes in Math 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 subsidized private schools nationwide.[3] We describe the eligibility criteria and the random assignment to the program below.

The program, consisting of five components described below, was implemented through two support teams (one internal and one external to schools) and expected to work together. The first team, the Education Leadership Team (henceforth ELE, from its name in Spanish), consisted of the school principal, the head of the

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

[3] The Chilean system of education includes three types of schools: public schools, subsidized private schools, and private schools. Public schools are both financed and administered by the public sector; subsidized 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.

technical and pedagogic office of the school, and two distinguished teachers. The second group, the Team of Technical and Pedagogic Advisors (henceforth ATP), was formed by three authorities of the regional Department of Education (the DEPROV), and aimed to provide external support to the ELE teams. Each ATP visited its assigned schools every 6 to 7 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).

The first component, called "effective implementation of the national curriculum" consisted in the development of unified pedagogical material and planning tools distributed to teachers. These tools included an annual curricular programming, a series of teaching materials designed for six-week periods, and a set of daily planning activities to be used by teachers in the classroom. The second component consisted of promoting a school culture and environment that encourages learning. A manual was developed and delivered to schools to guide the implementation of the ideas. The third component was the use of student evaluation 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 analyze students' performance in Math and Language (MINEDUC, 2013). It is worth noting that unlike the SIMCE tests, these instruments were not standardized tests and could be applied voluntarily by PAC schools. The fourth component was defined as the "optimization 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" aimed at promoting frequent internal school staff meetings to discuss students' progress.

2.2 Eligibility and Randomization

Among public and subsidized 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 Math and Language should be below the national average (252 points out of 500); and second, there should be at least 20 students per level on average from pre-K to fourth grade.[4] 2,286 schools met these criteria and were ranked by their 2005-2009 average SIMCE scores in Language and Math. The bottom 1,000 schools were automatically considered eligible. Since participation in the program 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 program, the Ministry increased the sample within each DEPROV by 50%, going up in the SIMCE ranking.[5] Of the resulting 1,480 eligible schools 632 located in "small" DEPROVs (DEPROVs with 40 schools or less) were allocated to the program automatically and do not form part of the evaluation and analysis. Among the remaining 848

[4]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.

[5]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 pedagogical assistance from well-known agencies of pedagogical support in Chile.

schools located in "large" DEPROVs 651 were randomly allocated to treatment and 197 to the control.

2.3 Data

This analysis in this paper 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 8 in appendix A shows summary statistics of all the variables used in this paper, namely, test scores and baseline characteristics, for the group of students that took each of the subject tests (post attrition samples).

2.4 Treatment- control balance and attrition

Table 9 in the Appendix, displays a set of randomization checks for the entire population of fourth grade students (the pre attrition sample) and for the three post attrition samples (Reading, Math, and Science test takers). The table is divided in two panels, corresponding to the 2011 and 2012 cohorts. 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 the post attrition samples. In 2011 there is no student that missed all three subject tests in the sample. When analyzing 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 Math 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 Table 9. 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, there is 0.7% less students missing the Reading test and 0.1% more students missing the Math 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 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 program.

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 Math 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 for the Reading, Math, and Science tests relative to the control group, but these differences are not statistically different from zero, which suggests that the higher overall attrition in 2012 does not bias our results of the impact of PAC on SIMCE.

Furthermore, with the exception of *father's incomplete primary*, all baseline characteristics are balanced between treatment and control groups, and they are jointly not significant to explain treatment status, as evidenced by the F-test.

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 randomization was successful in balancing baseline characteristics, even for the post attrition samples.

3 Estimation and inference

Our results explore overall effects as well as heterogeneous treatment effects by demographic characteristics. We define four groups based on the 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 programs are particularly helpful for the most deprived, or by contrast 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 Math and Science (OECD, 2015). These outcomes may be related to teachers' practices. Using the same sample of fourth grade teachers in Chile as this paper, Bassi *et al.* (2016) show that teachers in fact pay more attention to boys than girls, and those differentiated behaviors are correlated with worse performance in SIMCE in Math and Science among the girls. It is thus important to understand whether there are substantial differences in the response to interventions.

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

$$SIMCE_{ikg} = \beta + \gamma_g T_{ij} + X_{igj}\delta_g + \epsilon_{igj} \tag{1}$$

where $SIMCE_{ijkg}$ is the test score of student i, in school j, in subject $k = \{Math, Language, Science\}$ and in demographic group g. This is measure in units of 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 randomized into the program (PAC); X_{ijg} is a vector of student-school characteristics that includes baseline characteristics;[6] and ϵ_{ijg} is a random error term, which because of randomization is uncorrelated with treatment assignment.

Not all schools assigned to the program actually implemented it: there is non-compliance in both the 2011

[6]The covariates include, whether the students 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", "inc primary", "primary", "inc high school", "high school", "some college", "college +"; father's education (same dummies as mother education).

and the 2012 cohorts. Table 1 shows take up rates of schools invited to participate in the program, for both years 2011 and 2012.

Table 1: Randomization and implementation, school level

		2011		2012	
		Implemented PAC			
		No	Yes	No	Yes
Randomized into PAC	No	194	0	176	19
	Yes	155	492	179	465

In 2011 about 25% of schools randomized into the program did not implement it. In this case if we replace the randomization indicator T_{ij} with whether treatment actually took place and then use the randomization indicator as an instrument we will identify the effect of treatment on the treated. In 2012 however, we have two sided noncompliance, with 9% of schools not assigned to the program by the randomization actually getting it. IV in this case identifies the LATE parameter under the additional monotonicity assumption that randomization either does not change treatment status or induces the school to adopt the program.[7] In all cases using as treatment variable the original randomization (T_{ij}) provides an unbiased estimate of the intention to treat parameter (ITT), namely the effect of having been offered the program.

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 randomized into the program is 63%; some student in the control group did however receive the treatment.

Table 2: First stage. Dependent variable: Participated in PAC=1.

			2011		
	All	**Females**		**Males**	
		Low Income	**High Income**	**Low Income**	**High Income**
	(1)	(2)	(3)	(4)	(5)
Randomized into PAC	0.7606***	0.7745***	0.7207***	0.7691***	0.7451***
	[.734;.788]	[.746;.803]	[.68;.761]	[.707;.782]	[.74;.797]
Observations	31384	10938	2330	11492	2581
			2012		
	All	**Females**		**Males**	
		Low Income	**High Income**	**Low Income**	**High Income**
	(7)	(8)	(9)	(10)	(11)
Randomized into PAC	0.6238***	0.6342***	0.6108***	0.6360***	0.6291***
	[.577;.668]	[.586;.679]	[.555;.663]	[.575;.681]	[.585;.687]
Observations	35835	10479	2587	10709	3043

Notes: The dependent variable is *Participated in PAC*, a dummy variable that takes value one if the student attends a school that participated in PAC and zero otherwise. 95% bootstrapped confidence intervals are shown in brackets. ***Variable significant at the 1% level. Clustering at the school level.

In deriving standard errors and carrying out inference we cluster at the school level, which is the ran-

[7] See Imbens & Angrist (1994).

domization 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 & 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.

4 Main Results

4.1 Overall effects

Table 3: Impact of PAC on SIMCE 2011 and 2012

	Iintention to treat effect (ITT)					
		2011			**2012**	
	(1)	(2)	(3)	(4)	(5)	(6)
	Reading	Math	Science	Reading	Math	Science
Randomized into PAC	.095	.068	.033	.04	.051	.012
	[.04;.15]	[.01;.13]	[-.03;.09]	[-.01;.09]	[-.01;.11]	[-.04;.07]
	(.01)	(.13)	(.34)	(.32)	(.27)	(.72)
Control Group Mean	244.787	235.756	236.836	245.937	239.432	235.461
Control Group SD	49.97	47.103	44.09	51.055	47.326	46.527
Observations	30736	30731	30765	30494	30368	30331
	Instrumental Variables					
		2011			**2012**	
	(7)	(8)	(9)	(10)	(11)	(12)
	Reading	Math	Science	Reading	Math	Science
Received PAC	.125	.089	.044	.064	.081	.019
	[.05;.2]	[.01;.17]	[-.04;.12]	[-.02;.15]	[-.01;.18]	[-.06;.11]
	(.01)	(.13)	(.34)	(.32)	(.27)	(.72)
Control Group Mean	244.787	235.756	236.836	245.937	239.432	235.461
Control Group SD	49.97	47.103	44.09	51.055	47.326	46.527
Observations	30736	30731	30765	30494	30368	30331

Notes: The effects shown are in units of the control group standard deviation. 95% bootstrapped confidence intervals are shown in brackets. All 6 impacts in each panel are tested jointly to control for the Familywise Error Rate using the Romano-Wolf step down method. Romano-Wolf step down p-values from the two sided test are shown in parenthesis. Clustering at the school level. Inthe second panel instrument is original randomization into PAC.

In Table 3 we show the effects of the program on SIMCE test scores for all students pooled together. The top panel shows the ITT estimate while the bottom panel reports instrumental variables results where the explanatory variable is actually receiving PAC and the instrument is being randomized into PAC; the parameter is interpreted as the effect of treatment on the treated for the 2011 cohort where all those randomized out were actually excluded from the program (one sided noncompliance), while for the 2012 cohort it is interpreted as the Local Average Treatment Effect (LATE) since there is two sided noncompliance; as mentioned before in the latter case interpretation as a causal effect for compliers requires the monotonicity assumption, that the experiment did not induce any school to opt out of PAC, when they would have otherwise implemented it. Columns (1) to (3) and (7) to (9) show results for the 2011 fourth grade cohort while the rest of the columns show results for the 2012 cohort. In these Tables we report results without covariates. The appendix reports the results when we include them.

Overall the program has had significant positive effect on test scores. Specifically, reading improved for the 2011 cohort by about 10% of a standard deviation and this impact is significant at 1%. There was also an improvement of 7% of sd units for the 2011 Math score. However, this is only significant at a family wise error rate of 13%. All other impacts are not significant. The IV coefficients imply that the program itself improved reading for the treated by 12.5% of sd units. When we consider the 2012 cohort it is evident that the impacts have declined and are no longer significant. For example the reading impact falls by about a half, even when we use IV to allow for the greater degree of noncompliance. Given the data we have it is hard to know why this decline happened. One possibility is that the program was not applied with the same rigor; this generally raises the issue of how to maintain the momentum of interventions that seem to have the capability of producing positive results.

4.2 Effects by gender and family income

Much of the education debate relates to improving outcomes for more vulnerable, deprived or discriminated against populations. In what follows we thus consider the heterogeneity of these effects by gender and family income. We are particularly interested in family background because it has proved a challenge to improve outcomes for the most deprived populations. Gender is also important because it may define future imbalances, such as unequal pay or the increased informality among women. From a statistical point of view we avoid the inferential pitfalls of data mining by controlling for the overall Familywise error rate when splitting up the sample into groups by income and gender. The results are shown in Tables 4 and 5 separately for each of the two cohorts.[8]

The 2011 cohort The main conclusion from Table 4 is that the program benefited greatly medium-high income boys in the 2011 cohort, specially in Reading and Math - these have both a 2% p-value, controlling for FWE for all 12 coefficients; hence this is a particularly robust result and the impacts are large: for this demographic group and cohort being randomized into the program increases the Reading and Math test scores of high income boys by about 20% of sd units. Weaker effects are detected for medium-high income girls. The only individually significant effect is a 13.2% of sd units improvement in reading, but once we control for FWE the p-value is 28% and hence significance does not survive when we control for multiple hypotheses testing. Rescaling these effects by using IV implies that participating in the program causes an increase in Reading and Math tests scores for boys of more than 26% of sd units for students that were exposed to the program; this an effect of treatment on the treated. Finally, results from the estimation of the model with covariates are shown in tables 10 and 11 in the appendix and are very similar, although the effects are about three percentage points smaller.

[8]We carry out χ^2 joint test as a way of confirming that the program had an overall impact. We perform four sets of χ^2 joint tests. The first set considers the ITT effects of the program by gender and income for 2011 and 2012 reported in tables 4 and 5 and tests jointly the 24 null hypotheses that these effects are all zero. The second set replicates this exercise for the LATE effects. The third and fourth sets does the same for the effects estimated in specifications with covariates. Based on the χ^2 tests for the joint significance of all the effects reported, the program has an overall significant effect with p-values of zero in all joint tests.

Table 4: Impact of PAC on SIMCE 2011 (ITT parameter), by gender and income

				Females		
	Low Income			**Medium- high Income**		
	(1)	(2)	(3)	(4)	(5)	(6)
	Reading	Math	Science	Reading	Math	Science
Randomized into PAC	.048	.032	.01	.132	.058	.025
	[-.02;.11]	[-.04;.1]	[-.06;.08]	[.02;.25]	[-.06;.18]	[-.09;.15]
	(.69)	(.86)	(.98)	(.28)	(.84)	(.98)
Control Group Mean	251.092	233.751	233.211	255.152	242.041	243.128
Control Group SD	47.49	45.368	42.541	48.16	45.699	45.494
Observations	10854	10892	10877	2314	2313	2323
				Males		
	Low Income			**Medium- high Income**		
	(7)	(8)	(9)	(10)	(11)	(12)
	Reading	Math	Science	Reading	Math	Science
Randomized into PAC	.082	.053	.009	.197	.198	.13
	[.02;.15]	[-.02;.12]	[-.06;.08]	[.1;.3]	[.09;.3]	[.02;.24]
	(.19)	(.69)	(.98)	(.02)	(.02)	(.28)
Control Group Mean	241.078	238.484	240.424	244.988	242.971	246.545
Control Group SD	51.102	48.168	44.258	52.693	49.455	46.198
Observations	11391	11400	11425	2560	2562	2561

Notes: The effects shown are in units of the control group standard deviation. 95% bootstrapped confidence intervals are shown in brackets. All outcomes in this table are tested jointly to control for the Familywise Error Rate using the Romano-Wolf step down method. Romano-Wolf step down p-values from the two sided test are shown in parenthesis. Clustering at the school level.

Table 5: Impact of PAC on SIMCE 2012 (ITT parameter), by gender and income

				Females		
	Low Income			**Medium- high Income**		
	(1)	(2)	(3)	(4)	(5)	(6)
	Reading	Math	Science	Reading	Math	Science
Randomized into PAC	.039	.068	.025	.208	.117	.075
	[-.03;.11]	[0;.15]	[-.04;.1]	[.09;.32]	[0;.24]	[-.03;.18]
	(.80)	(.57)	(.87)	(.01)	(.45)	(.74)
Control Group Mean	251.989	236.495	232.941	256.406	244.927	242.931
Control Group SD	49.769	46.37	44.465	49.098	47.692	46.727
Observations	10030	10005	9973	2533	2522	2520
				Males		
	Low Income			**Medium- high Income**		
	(7)	(8)	(9)	(10)	(11)	(12)
	Reading	Math	Science	Reading	Math	Science
Randomized into PAC	.046	.072	.031	.026	.107	.013
	[-.02;.11]	[0;.15]	[-.03;.1]	[-.06;.12]	[.01;.2]	[-.08;.1]
	(.74)	(.47)	(.87)	(.87)	(.45)	(.87)
Control Group Mean	239.601	240.548	235.503	250.521	247.172	246.618
Control Group SD	51.678	47.936	48.001	51.718	47.433	46.927
Observations	10235	10191	10182	2950	2946	2930

Notes: The effects shown are in units of the control group standard deviation. 95% bootstrapped confidence intervals are shown in brackets. All outputs in this table are tested jointly to control for the Familywise Error Rate using the Romano-Wolf step down method. Romano-Wolf step down p-values from the two sided test are shown in parenthesis.

The 2012 cohort For the 2012 cohort (Table 5) we are only able to find a significant effect (based on the strict criterion of controlling for the FWE for all 12 hypotheses) for medium-high income girls in Reading scores, while the effects for high income boys disappear .[9] Being randomized into the program causes an increase in the Reading scores of higher income girls of about 21% of sd units. The effect is significant at a one percent level even after controlling for the FWE. The estimation of the IV parameter suggests that the LATE program effect on Reading scores for high income girls is 34% of sd units with a p-value of less than two percent. When estimating the program effect in the specifications that include covariates (shown in tables 12 and 13 in the appendix), the treatment effects remain significant at the 5% level when testing all hypothesis independently, but are not longer significant after controlling for the FWE.

All in all, the main results of this paper suggest that the PAC when first implemented in 2011 had a large and significant effect on the performance of four grader high income boys and girls, although the effects weaken during the second year of the program. There was never an effect for the children from lower income groups. However, this structured teaching intervention holds real promise, particularly if we are able to understand how to sustain the effects over time and how to impact on the poorer children. In the next section we use the *Classroom Assessment Scoring System* (CLASS) to see whether the program affected the way teachers and students interact.

5 The 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-students interactions Araujo *et al.* (2016). The aim of this section is, therefore, to study how important are teacher-students interactions to improve students' learning in our context, and whether the PAC had any positive impact on the quality of teacher-students interactions. As a preview of our results, we find that higher quality of teacher-students interactions are 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-students interactions by this measure.

5.1 Background

The main instrument used in this paper to measure teacher-students interactions is the 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 teacher-student interactions, as a proxy to teachers' quality or effectiveness. To produce the CLASS measures, thoroughly trained coders watch and analyze videotaped classes and assign a score for teacher-students interactions in 11 dimensions. These dimensions can be grouped into three main domains: Emotional Support, Classroom Organization, and Instructional Support.[10] Coders

[9] It is interesting to point out that based on the standard criteria of individual 5% significance levels many more effects are significant in the 2012 cohort. See for example the effect on math for both boys and girls in both income groups.

[10] Emotional support includes the dimensions of Positive Climate, Negative Climate, Regard for Student Perspectives, and Teacher Sensitivity; Classroom Organization includes the dimensions of Effective Behavior Management, Instructional Learning Formats,

look for very specific teachers' behaviors in each dimension, which are well described in the CLASS protocol that guides coders for their scoring.

There are several studies that link better student outcomes (both in learning and in the development of socioemotional skills) with teachers' scores in CLASS. Araujo *et al.* (2016) present a brief review of this literature for the US 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.

5.2 The CLASS experiment

In 2012, among the entire PAC experimental sample, a subsample of 210 schools (105 from the PAC treatment group and 105 from the PAC control group) was randomly selected to also participate in the CLASS experiment. Selected schools were asked if they would agree to have some classroom lectures videotaped and analyzed afterwards. The CLASS experiment had some non compliance: in the end, 185 invited schools agreed to participate in the filming sessions. Nonparticipation is fairly well balanced between the treatment and control schools.[11] 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.

Fourth grade teachers in the participating schools were videotaped for four full lessons (mostly Language classes). A total of 185 teachers were filmed following the CLASS protocol.[12] The coding was done by 10 coders and a supervisor carefully trained and selected.[13] Each of the four school hours filmed per teacher was 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 1 to 7 scale ("low" for scores 1-2, "medium" for scores 3-5, and "high" for scores 6-7). The final CLASS scores for each domain consisted on the average across dimensions within the corresponding domain. For the coding, videos were randomly assigned to the 10 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

and Productivity; and Instructional Climate includes the dimensions of Language Modeling, Concept Development, Analysis and Inquiry, and Quality of Feedback.

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

[12] The fieldwork and coding according to CLASS was coordinated 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 program in Chile, Un buen Comienzo (Yoshikawa, et al. 2013).

[13] 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), that 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.

the following weeks if reliability rates remained above 80%.[14] Overall, 52% of the videos were double coded, with an average reliability rate of 84.2%.[15] This inter-coder reliability is comparable to that found in other studies. For example, as cited in Araujo *et al.* (2016), Brown *et al.* (2010) report an inter-coder reliability rate of 83% for the 12% of the classroom observations which were double-coded.[16]

5.3 CLASS, Teacher Performance and program effects

Table 6: Association between CLASS and SIMCE 2012, by gender and income

			Females			
	Low Income			**Medium- high Income**		
	(1)	(2)	(3)	(4)	(5)	(6)
	Reading	Math	Science	Reading	Math	Science
CLASS first principal component	.158	.161	.143	.159	.028	.066
	[.09; .24]	[.09;.24]	[.06;.21]	[.08;.23]	[-.08;.13]	[-.02;.15]
	(.01)	(.01)	(.01)	(.01)	(.69)	(.33)
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
Observations	1415	1404	1403	297	296	298
			Males			
	Low Income			**Medium- high Income**		
	(7)	(8)	(9)	(10)	(11)	(12)
	Reading	Math	Science	Reading	Math	Science
CLASS first principal component	.187	.201	.194	.126	.105	.198
	[.13;.24]	[.13;.26]	[.13;.26]	[.02;.24]	[0;.22]	[.09;.33]
	(.01)	(.01)	(.01)	(.2)	(.33)	(.02)
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
Observations	1472	1461	1461	365	360	361

Notes: The effects shown are in units of the corresponding test standard deviation. 95% bootstrapped confidence intervals are shown in brackets. All outcomes in this table are tested jointly to control for the FWE Rate using the Romano-Wolf step down method. Step down p-values from the two sided test are shown in parenthesis. Clustering at the school level.

In Table 6 we report the association between CLASS and SIMCE scores for the 2012 cohort, at the student level using a regression of SIMCE on the standardized CLASS score as well as covariates. The effects reported are in units of a standard deviation of the SIMCE score for the corresponding demographic group and subject. The most striking result from the table is the association between better student teacher interactions (reflected in a higher 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 15% and 20% of sd units. These results are robust to adjustments in p-values to control for the FWE rate. 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.* (2005b)). Taken at face value these results imply that moving a lower income student from a bottom 2% of teachers to the top 2% can improve outcomes of low income students by

[14] Coding is considered reliable if the difference between the two coders' score is less than 2 points for each CLASS dimension.

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

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

close to one standard deviation.

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 pro-active 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 program.

Table 7: Impact of PAC on CLASS, classroom level

	Dependent variable: CLASS first principal component			
	(1)	(2)	(3)	(4)
	OLS	IV	OLS	IV
PAC = 1	-0.5274		-0.2361	
	[-1.128;.093]		[-1.318;.107]	
Participated in PAC = 1		-0.6153		-0.2697
		[-.878;.338]		[-.992;.38]
Covariates	No	No	Yes	Yes
Observations	185	185	184	184

Notes: 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 2005-2009, 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.

The impact of PAC on CLASS Table 7 shows the result of regressing CLASS on treatment allocation and covariates. The results consistently suggest that the program has no significant effect on teacher-classroom interactions in 2012. The absence of an effect of the program on CLASS scores may be the reason why PAC had a much weaker effect in the 2012 cohort. 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. It is regretful that we have no CLASS scores for the first 2011 cohort, where the effects of the program were much stronger. 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 the intervention was successful in the first implementation cohort; moreover, while the impacts are concentrated among the relatively better off, the population we are studying is already lower-income and attending underperforming schools.

6 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, that define carefully what teachers are supposed to do and monitor the progress of students throughout the year. This is the idea

underlying PAC (Plan Apoyo Compartido) the program we are analyzing in this paper and which was launched in 2011 in Chile. Through standardized teaching material (class preparation) and through the support of internal and external pedagogic teams, the program aimed at reducing the gap, as measured by the standardized test SIMCE, between the poorest student population and the national average. The program was designed with a gradual implementation, which implied that only half of eligible schools could be offered the program. 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. However, there are no overall effects for the 2012 cohort of students. When we break down the effects by gender and family income of the students we find positive and significant effects (controlling for multiple hypotheses testing) in both cohorts for students originating from higher income families, and particularly for boys in 2011. So it seems that the program can improve outcomes, but it mainly improves results for the relatively better off; moreover the effects of the program seem to fade, which poses the urgent question of how to sustain the impact of what seems to be a successful program. Our statistical analysis indicates that the success observed for the 2011 cohort is extremely unlikely to be down to luck: we believe the effects were real and as shown quite substantial for some groups. The policy question is making sure practices and implementation are sustained so that all cohorts of children can benefit.[17]

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 students and teachers interact along various dimensions (class organization, 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 program shifted the CLASS score, by improving teacher-student interactions and we find no effect at all. There is an open question here, whether this is down to CLASS not capturing the dimensions of the program that led to the improvements we observe or whether practices had reverted to pre-policy ones in the 2012 cohort, i.e. whether the PAC was no longer implemented as effectively as it was for the 2011 cohort in the treatment schools.

However, even at the most successful point the PAC program only benefited those from better-off family backgrounds. Thus the urgent question of how to improve outcomes of children from the most deprived backgrounds remains. As much research seems to show the answer may lie in Early Childhood Development Programs, which attempt to ensure that children from the most deprived backgrounds have better cognitive development and access to improved opportunities (see Attanasio *et al.* (2014)).

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

References

Araujo, Maria Caridad, Carneiro, Pedro, Cruz-Aguayo, Yyannu, & Schady, Norbert. 2016. A helping hand? Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics*, **131**(3), 1415–1453.

Attanasio, O., Fernandez, C., Fitzsimons, E., Grantham-McGregor, S., Meghir, C., & Rubio-Codina, M. 2014. Using the infrastructure of a conditional cash transfer programme to deliver a scalable integrated early child development programme in colombia: a cluster randomised controlled trial. *BMJ - British Medical Journal*, **349, g5785**.

Bassi, Marina, Blumberg, Rae Lesser, & Diaz, Mercedes Mateo. 2016. Under the "Cloak of Invisibility": Gender Bias in Teaching Practices and Learning Outcomes. *IDB WORKING PAPER SERIES NÂº IDB-WP-696.*

Brown, Joshua L, Jones, Stephanie M, LaRusso, Maria D, & Aber, J Lawrence. 2010. Improving classroom quality: Teacher influences and experimental impacts of the 4rs program. *Journal of Educational Psychology*, **102**(1), 153.

Chetty, Raj, Friedman, John N., & Rockoff, Jonah E. 2014. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, **104**(9), 2593–2632.

Ganimian, Alejandro J., & Murnane, Richard J. 2014 (July). *Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Impact Evaluations.* Working Paper 20284. National Bureau of Economic Research.

He, Fang, Linden, Leigh L, & MacLeod, Margaret. 2009. A better way to teach children to read? Evidence from a randomized controlled trial. *Unpublished manuscript. New York, NY: Columbia University.*

Imbens, Guido, & Angrist, Joshua. 1994. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, **62**(2), 467–475.

Kirschner, Paul A., Sweller, John, & Clark, Richard 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*, **41**(2), 75–86.

OECD. 2015. The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence. *Paris, France: OECD Publishing.*

Pianta, Robert C., Mashburn, Andrew J., Downer, Jason T., Hamre, Bridget K., & Justice, Laura. 2008. Effects of web mediated professional development resources on teacher child interactions in pre kindergarten classrooms. *Early Childhood Research Quarterly*, **23**(4), 431–451.

Rivkin, Steven G., Hanushek, Eric A., & Kain, John F. 2005a. Teachers, Schools, and Academic Achievement. *Econometrica*, **73**(2), 417–458.

Rivkin, Steven G., A., Hanushek Eric, & F., Kain John. 2005b. Teachers, Schools, and Academic Achievement. *Econometrica*, **73.2**, 417–58.

Romano, Joseph P., & Wolf, Michael. 2005. Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica*, **73**(4), 1237–1282.

Slavin, Robert E, Lake, Cynthia, Chambers, Bette, Cheung, Alan, & Davis, Susan. 2009. Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research*, **79**(4), 1397–1466.

Appendix A Summary Statistics

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

Columns labeled *Reading*, *Math*, 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 that did not take the test (we discuss the issue of attrition in the next subsection).

Sub-columns labeled *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 program through the randomization, 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 two panels, *2011* and *2012*, indicating the fourth grade cohorts considered in this paper.

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

SIMCE scores (Reading, Math, 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),[18] *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*.

[18]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, tough, that students in our sample belong mainly to low-middle income families in Chile.

Table 8: Summary statistics - post attrition samples

	Reading test takers						Math 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	6886	245	50	23850	248	50												
Math							6903	236	47	23828	238	48						
Science													6911	237	44	23854	238	44
Baseline characteristics:																		
Students demographics																		
Female	6554	.472	.499	22861	.484	.5	6653	.473	.499	23152	.484	.5	6660	.473	.499	23162	.484	.5
Low income	6219	.841	.366	21470	.813	.39	6235	.841	.366	21490	.813	.39	6244	.841	.366	21502	.813	.39
Nuclear family	6886	.613	.487	23850	.617	.486	6903	.613	.487	23828	.619	.486	6911	.613	.487	23854	.617	.486
Extended family	6886	.244	.43	23850	.239	.426	6903	.244	.43	23828	.239	.426	6911	.244	.43	23854	.239	.427
Other family	6886	.143	.35	23850	.144	.352	6903	.143	.35	23828	.143	.35	6911	.143	.35	23854	.143	.35
Nbr years failed	6187	.238	.527	21353	.232	.531	6202	.239	.528	21367	.232	.531	6211	.239	.528	21387	.232	.531
Mother's education																		
No education	6201	.007	.086	21352	.006	.078	6215	.007	.086	21374	.006	.077	6225	.008	.087	21387	.006	.078
Inc. primary	6201	.179	.383	21352	.174	.379	6215	.178	.383	21374	.174	.379	6225	.178	.383	21387	.174	.379
Primary	6201	.17	.375	21352	.164	.37	6215	.17	.375	21374	.164	.37	6225	.169	.375	21387	.164	.37
Inc. high school	6201	.235	.424	21352	.218	.413	6215	.235	.424	21374	.218	.413	6225	.236	.425	21387	.219	.413
High school	6201	.324	.468	21352	.341	.474	6215	.324	.468	21374	.341	.474	6225	.324	.468	21387	.341	.474
Inc. college	6201	.034	.182	21352	.038	.191	6215	.034	.182	21374	.038	.191	6225	.034	.181	21387	.037	.19
College	6201	.051	.22	21352	.059	.235	6215	.051	.22	21374	.059	.235	6225	.051	.22	21387	.059	.235
Father's education																		
No education	5992	.008	.089	20577	.008	.089	6006	.008	.09	20597	.008	.089	6013	.008	.089	20611	.008	.09
Inc. primary	5992	.17	.376	20577	.158	.364	6006	.17	.376	20597	.157	.364	6013	.17	.376	20611	.157	.364
Inc. primary	5992	.16	.366	20577	.162	.368	6006	.16	.367	20597	.161	.368	6013	.16	.366	20611	.162	.368
Inc. high school	5992	.248	.432	20577	.23	.421	6006	.248	.432	20597	.23	.421	6013	.247	.431	20611	.23	.421
High school	5992	.329	.47	20577	.345	.475	6006	.328	.469	20597	.344	.475	6013	.328	.47	20611	.344	.475
Inc. college	5992	.036	.187	20577	.043	.202	6006	.036	.187	20597	.043	.203	6013	.036	.187	20611	.043	.203
College	5992	.05	.217	20577	.056	.229	6006	.05	.219	20597	.056	.23	6013	.05	.218	20611	.056	.229
Panel B: 2012																		
SIMCE scores:																		
Reading	7141	246	51	23353	248	52												
Math							7095	239	47	23273	242	49						
Social Science													7105	235	47	23226	237	46
Baseline characteristics:																		

Table 8: Summary statistics - post attrition samples (continued)

	Reading test takers						Math 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
Students demographics																		
Female	6773	.481	.5	22810	.481	.5	6572	.483	.5	22303	.482	.5	6580	.483	.5	22245	.482	.5
Low income	6048	.8	.4	20229	.783	.412	6065	.801	.399	20292	.783	.412	6063	.801	.399	20236	.783	.412
Nuclear family	7141	.566	.496	23353	.572	.495	7095	.572	.495	23273	.576	.494	7105	.571	.495	23226	.575	.494
Extended family	7141	.235	.424	23353	.248	.432	7095	.237	.425	23273	.25	.433	7105	.236	.425	23226	.249	.433
Other family	7141	.2	.4	23353	.18	.385	7095	.191	.393	23273	.175	.38	7105	.193	.395	23226	.175	.38
Nbr years failed	6050	1.243	.588	20242	1.219	.565	6068	1.245	.593	20305	1.22	.564	6066	1.243	.592	20248	1.219	.564
Mother's education																		
No education	5873	.006	.076	19580	.006	.075	5894	.006	.077	19648	.006	.075	5891	.006	.076	19594	.006	.075
Inc. primary	5873	.178	.383	19580	.166	.372	5894	.18	.384	19648	.167	.373	5891	.179	.384	19594	.166	.372
Primary	5873	.173	.378	19580	.169	.375	5894	.173	.378	19648	.169	.375	5891	.172	.378	19594	.17	.376
Inc. high school	5873	.214	.41	19580	.213	.409	5894	.214	.41	19648	.213	.41	5891	.214	.41	19594	.213	.409
High school	5873	.338	.473	19580	.348	.476	5894	.336	.472	19648	.348	.476	5891	.337	.473	19594	.348	.476
Inc. college	5873	.036	.187	19580	.039	.193	5894	.036	.187	19648	.039	.193	5891	.036	.187	19594	.039	.192
College	5873	.055	.228	19580	.058	.235	5894	.055	.228	19648	.059	.235	5891	.055	.228	19594	.059	.236
Father's education																		
No education	5630	.009	.092	18745	.007	.085	5647	.009	.092	18820	.007	.085	5644	.009	.092	18768	.007	.085
Inc. primary	5630	.153	.36	18745	.158	.365	5647	.153	.36	18820	.158	.365	5644	.154	.361	18768	.158	.364
Inc. primary	5630	.172	.377	18745	.165	.371	5647	.172	.377	18820	.165	.371	5644	.17	.376	18768	.165	.372
Inc. high school	5630	.222	.416	18745	.217	.412	5647	.222	.416	18820	.217	.412	5644	.222	.416	18768	.218	.413
High school	5630	.357	.479	18745	.351	.477	5647	.356	.479	18820	.35	.477	5644	.357	.479	18768	.35	.477
Inc. college	5630	.035	.183	18745	.043	.203	5647	.035	.184	18820	.043	.204	5644	.035	.184	18768	.043	.204
College	5630	.054	.225	18745	.058	.234	5647	.054	.226	18820	.058	.234	5644	.054	.225	18768	.059	.235

Table 9: Randomization check

	Pre attrition sample		Post attrition samples					
			Reading test takers		Math test takers		Science test takers	
	Stat.	*p-val*	*Stat.*	*p-val*	*Stat.*	*p-val*	*Stat.*	*p-val*
Panel A: 2011								
Balancing of attrition rates and baseline characteristics (E(PAC=1) - E(PAC=0))								
Proportion of attritors								
Reading	-.007	.258	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Math	.001	.67	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Science	.001	.784	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Baseline characteristics:								
Students demographics								
Female	.003	.758	.003	.758	.003	.758	.003	.758
Low income	-.02	.048	-.02	.048	-.02	.048	-.02	.048
Nuclear family	.012	.324	.012	.324	.012	.324	.012	.324
Extended family	-.009	.284	-.009	.284	-.009	.284	-.009	.284
Other family	-.003	.786	-.003	.786	-.003	.786	-.003	.786
Nbr years failed	-.01	.497	-.01	.497	-.01	.497	-.01	.497
Mother's education								
No education	-.001	.578	-.001	.578	-.001	.578	-.001	.578
Inc. primary	.005	.567	.005	.567	.005	.567	.005	.567
Primary	-.004	.597	-.004	.597	-.004	.597	-.004	.597
Inc. high school	-.022	.007	-.022	.007	-.022	.007	-.022	.007
High school	.017	.131	.017	.131	.017	.131	.017	.131
Inc. college	0	.923	0	.923	0	.923	0	.923
Father's education								
College	.004	.363	.004	.363	.004	.363	.004	.363
No education	-.001	.661	-.001	.661	-.001	.661	-.001	.661
Inc. primary	-.001	.92	-.001	.92	-.001	.92	-.001	.92
Inc. primary	.006	.446	.006	.446	.006	.446	.006	.446
Inc. high school	-.02	.015	-.02	.015	-.02	.015	-.02	.015
High school	.009	.372	.009	.372	.009	.372	.009	.372
Inc. college	.005	.186	.005	.186	.005	.186	.005	.186
College	.001	.854	.001	.854	.001	.854	.001	.854
Test of joint significance of baseline characteristics (F-statistic)								
	1.18	0.277	1.18	0.272	1.17	0.282	1.15	0.305
Panel B: 2012								
Balancing of attrition rates and baseline characteristics (E(PAC=1) - E(PAC=0))								
Proportion of attritors								
Reading	-.01	.213	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Math	-.012	.125	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Science	-.01	.231	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Baseline characteristics:								
Students demographics								
Female	-.001	.909	-.001	.909	-.001	.909	-.001	.909
Low income	-.005	.652	-.005	.652	-.005	.652	-.005	.652
Nuclear family	.014	.299	.014	.299	.014	.299	.014	.299
Extended family	.003	.687	.003	.687	.003	.687	.003	.687
Other family	-.017	.251	-.017	.251	-.017	.251	-.017	.251
Nbr years failed	-.024	.123	-.024	.123	-.024	.123	-.024	.123
Mother's education								
No education	.001	.441	.001	.441	.001	.441	.001	.441
Inc. primary	-.003	.783	-.003	.783	-.003	.783	-.003	.783
Primary	.001	.844	.001	.844	.001	.844	.001	.844
Inc. high school	-.001	.934	-.001	.934	-.001	.934	-.001	.934
High school	.004	.675	.004	.675	.004	.675	.004	.675
Inc. college	-.004	.3	-.004	.3	-.004	.3	-.004	.3
College	.001	.922	.001	.922	.001	.922	.001	.922
Father's education								

Table 9: Randomization check (continued)

	Pre attrition sample		Post attrition samples					
			Reading test takers		Math test takers		Science test takers	
	Stat.	*p-val*	*Stat.*	*p-val*	*Stat.*	*p-val*	*Stat.*	*p-val*
No education	-.002	.338	-.002	.338	-.002	.338	-.002	.338
Inc. primary	.015	.061	.015	.061	.015	.061	.015	.061
Inc. primary	-.003	.692	-.003	.692	-.003	.692	-.003	.692
Inc. high school	-.006	.447	-.006	.447	-.006	.447	-.006	.447
High school	-.009	.355	-.009	.355	-.009	.355	-.009	.355
Inc. college	.005	.21	.005	.21	.005	.21	.005	.21
College	0	.945	0	.945	0	.945	0	.945
Test of joint significance of baseline characteristics (F-statistic)								
	1.19	0.268	1.17	0.282	1.15	0.300	1.12	0.332

Notes: Pre attrition sample refers to the universe of students in the fourth grade. Post attrition sample refers to the sub sample of students that took each of the subject SIMCE tests. The statistic (*Stat.*) reported in the balancing exercises is E(PAC=1) - E(PAC=0), that is, the difference in means between the treatment and the control groups. The statistic (*Stat.*) reported in the test of joint significance exercises is the F-test. Baseline characteristics include student demographics and Mother's and father's education. 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, the minimum wage in such country), *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*.

Appendix B Results with covariates

Table 10: Impact of PAC on SIMCE 2011 (ITT parameter), by gender and income

			Females			
	Low Income			**Medium- high Income**		
	(1)	(2)	(3)	(4)	(5)	(6)
	Reading	Math	Science	Reading	Math	Science
Randomized into PAC	.058	.045	.01	.089	.018	-.01
	[-.01;.12]	[-.03;.11]	[-.06;.08]	[-.02;.2]	[-.09;.13]	[-.12;.1]
	(.6)	(.72)	(1)	(.62)	(1)	(1)
Control Group Mean	251.092	233.751	233.211	255.152	242.041	243.128
Control Group SD	47.49	45.368	42.541	48.16	45.699	45.494
Observations	10035	10064	10052	2195	2195	2204
			Males			
	Low Income			**Medium- high Income**		
	(7)	(8)	(9)	(10)	(11)	(12)
	Reading	Math	Science	Reading	Math	Science
Randomized into PAC	.084	.048	.014	.165	.16	.091
	[.02;.15]	[-.02;.12]	[-.06;.08]	[.06;.26]	[.05;.26]	[-.01;.2]
	(.16)	(.71)	(1)	(.04)	(.11)	(.6)
Control Group Mean	241.078	238.484	240.424	244.988	242.971	246.545
Control Group SD	51.102	48.168	44.258	52.693	49.455	46.198
Observations	10560	10566	10593	2423	2427	2424

Notes: The effects shown are in units of the control group standard deviation. 95% bootstrapped confidence intervals are shown in brackets. All outputs in this table are tested jointly to control for the Familywise Error Rate using the Romano-Wolf step down method. Romano-Wolf step down p-values from the two sided test are shown in parenthesis. All regressions include covariates: whether the students 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", "inc primary", "primary", "inc high school", "high school", "some college", "college +"; father's education (same dummies as mother education).

Table 11: Impact of PAC on SIMCE 2011 (Instrumental Variables), by gender and income

	Females					
	Low Income			**Medium- high Income**		
	(1)	(2)	(3)	(4)	(5)	(6)
	Reading	Math	Science	Reading	Math	Science
Received PAC	.075	.058	.013	.123	.025	-.013
	[-.01;.16]	[-.03;.15]	[-.07;.1]	[-.02;.28]	[-.13;.18]	[-.16;.14]
	(.6)	(.72)	(1)	(.61)	(1)	(1)
Control Group Mean	251.092	233.751	233.211	255.152	242.041	243.128
Control Group SD	47.49	45.368	42.541	48.16	45.699	45.494
Observations	10035	10064	10052	2195	2195	2204
	Males					
	Low Income			**Medium- high Income**		
	(7)	(8)	(9)	(10)	(11)	(12)
	Reading	Math	Science	Reading	Math	Science
Received PAC	.108	.062	.018	.222	.215	.122
	[.02;.19]	[-.03;.15]	[-.07;.1]	[.09;.35]	[.07;.35]	[-.02;.26]
	(.16)	(.72)	(1)	(.05)	(.11)	(.61)
Control Group Mean	241.078	238.484	240.424	244.988	242.971	246.545
Control Group SD	51.102	48.168	44.258	52.693	49.455	46.198
Observations	10560	10566	10593	2423	2427	2424

Notes: The effects shown are in units of the control group standard deviation. 95% bootstrapped confidence intervals are shown in brackets. All outputs in this table are tested jointly to control for the Familywise Error Rate using the Romano-Wolf step down method. Romano-Wolf step down p-values from the two sided test are shown in parenthesis. All regressions include covariates: whether the students 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", "inc primary", "primary", "inc high school", "high school", "some college", "college +"; father's education (same dummies as mother education). Instrument is original randomization into PAC.

Table 12: Impact of PAC on SIMCE 2012 (ITT parameter), by gender and income

	Females					
	Low Income			**Medium- high Income**		
	(1)	(2)	(3)	(4)	(5)	(6)
	Reading	Math	Science	Reading	Math	Science
Randomized into PAC	.038	.05	.022	.161	.067	.021
	[-.03;.1]	[-.02;.13]	[-.04;.09]	[.05;.27]	[-.04;.18]	[-.08;.12]
	(.91)	(.87)	(.99)	(.12)	(.91)	(.99)
Control Group Mean	251.989	236.495	232.941	256.406	244.927	242.931
Control Group SD	49.769	46.37	44.465	49.098	47.692	46.727
Observations	9064	9049	9022	2396	2384	2382
	Males					
	Low Income			**Medium- high Income**		
	(7)	(8)	(9)	(10)	(11)	(12)
	Reading	Math	Science	Reading	Math	Science
Randomized into PAC	.035	.059	.022	.029	.106	.004
	[-.03;.1]	[-.01;.13]	[-.04;.09]	[-.06;.12]	[0;.21]	[-.09;.1]
	(.91)	(.73)	(.99)	(.99)	(.49)	(.99)
Control Group Mean	239.601	240.548	235.503	250.521	247.172	246.618
Control Group SD	51.678	47.936	48.001	51.718	47.433	46.927
Observations	9234	9206	9194	2785	2781	2767

Notes: The effects shown are in units of the control group standard deviation. 95% bootstrapped confidence intervals are shown in brackets. All outputs in this table are tested jointly to control for the Familywise Error Rate using the Romano-Wolf step down method. Romano-Wolf step down p-values from the two sided test are shown in parenthesis. All regressions include covariates: whether the students 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", "inc primary", "primary", "inc high school", "high school", "some college", "college +"; father's education (same dummies as mother education).

Table 13: Impact of PAC on SIMCE 2012 (Instrumental Variables), by gender and income

	Females					
	Low Income			**Medium- high Income**		
	(1)	(2)	(3)	(4)	(5)	(6)
	Reading	Math	Science	Reading	Math	Science
Received PAC	.059	.078	.035	.258	.108	.034
	[-.04;.16]	[-.03;.21]	[-.06;.15]	[.08;.44]	[-.07;.3]	[-.13;.2]
	(.9)	(.88)	(.99)	(.14)	(.9)	(.99)
Control Group Mean	251.989	236.495	232.941	256.406	244.927	242.931
Control Group SD	49.769	46.37	44.465	49.098	47.692	46.727
Observations	9064	9049	9022	2396	2384	2382
	Males					
	Low Income			**Medium- high Income**		
	(7)	(8)	(9)	(10)	(11)	(12)
	Reading	Math	Science	Reading	Math	Science
Received PAC	.055	.092	.034	.047	.169	.006
	[-.04;.16]	[-.02;.21]	[-.06;.14]	[-.09;.19]	[.01;.33]	[-.14;.15]
	(.9)	(.75)	(.99)	(.99)	(.49)	(.99)
Control Group Mean	239.601	240.548	235.503	250.521	247.172	246.618
Control Group SD	51.678	47.936	48.001	51.718	47.433	46.927
Observations	9234	9206	9194	2785	2781	2767

Notes: The effects shown are in units of the control group standard deviation. 95% bootstrapped confidence intervals are shown in brackets. All outputs in this table are tested jointly to control for the Familywise Error Rate using the Romano-Wolf step down method. Romano-Wolf step down p-values from the two sided test are shown in parenthesis. All regressions include covariates. All regressions include covariates: whether the students 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", "inc primary", "primary", "inc high school", "high school", "some college", "college +"; father's education (same dummies as mother education). Instrument is original randomization into PAC.