The direct and spillover effects of a nationwide mental health program for disruptive students.*

Clément de Chaisemartin† Nicolás Navarrete H.‡

March 8, 2019

Abstract

A large literature finds that cognitive behavioral therapy programs for disruptive students can reduce their disruptiveness and improve their academic outcomes. However, the literature has mostly considered small-scale programs, and has not studied spillover effects on ineligible students. We use a randomized controlled trial to estimate the effects of one such program, implemented as a nationwide policy in Chile. The program has no effect on eligible students' disruptiveness and academic outcomes. It worsens the studying conditions in treated classes, and it may increase the disruptiveness of ineligible students. More research is needed to determine how to successfully scale-up such programs.

Keywords: disruptive students, spillover effects, peer effects, cognitive behavioral therapy.

JEL Codes: I21, I24, I28, D62.

*We thank the *Junta Nacional de Auxilio Escolar y Becas* (JUNAEB) for allowing us to randomize the timing of the *Habilidades para la vida* program, and for providing us with some of the data used in this research. The authors gratefully acknowledge financial support from the Centre for Competitive Advantage in the Global Economy at Warwick University, and from the Economics department at Warwick University. We would also like to thank Romina Guzmán, Juan Pablo Arias, Antonio Figueroa and Gerardo Alvarez for outstanding research assistance. Finally, we also thank Miya Barnett, Peter Kuhn, Zoe Liberman, Shelly Lundberg, Kyle Ratner, Fabian Waldinger, and seminar participants at INSEAD, UC Santa Barbara, University of Bologna, and University of Virginia for their useful comments. This research has been approved by the University of Warwick Research Ethics Committee (approval date: 2014-11-03, approval number: 111/13-14), and has been registered on the social science registry website (RCT ID AEARCTR-0001080). † Clément de Chaisemartin: UC Santa Barbara; J-Pal. ‡Nicolas Navarrete H: Paris School of Economics. For questions please email: `clementdechaisemartin@ucsb.edu` or `nicolas.navarrete@psemail.eu`.

1 Introduction

Lazear [2001] has proposed that classroom learning is a public good suffering from congestion effects, which are negative externalities created when one student is disruptive and impedes the learning of her classmates. Those externalities are important: Carrell and Hoekstra [2010] and Carrell et al. [2018] find that being exposed to one peer experiencing domestic violence at home, a good proxy for a disruptive peer, reduces classmates' test scores by 0.07 standard deviation (σ), and reduces their earnings at age 26 by 3 to 4 percent. Figlio [2007] also finds that being exposed to disruptive peers reduces classmates test scores. Betts and Shkolnik [1999] find that US middle and high schools teachers devote 6.1% of instruction time to discipline, and that this fraction is higher in disadvantaged schools. Therefore, programs effective at reducing troubled students' disruptiveness may generate large positive spillover on their classmates, and may have large social benefits.

School-based psychosocial programs are a commonly used strategy to reduce students' disruptiveness. For instance, more than three fourths of schools in the U.S. offer mental health, social service, and prevention service options for students and their families (see [Brener et al., 2001]). Those programs can be divided into three categories. First, *universal* programs are delivered in classroom settings to all the students in the classroom. Second, *selected* programs are provided to specific students identified by teachers as having conduct problems, during the school day and outside of their classroom. Third, *mixed* programs include both actions targeted at all students and actions targeted at selected students. For instance, school counselors in the U.S. often use a mixed approach (see [Carrell and Carrell, 2006] and [Carrell and Hoekstra, 2014]).

In this paper, we study the effects of "Skills for Life" (SFL), a nationwide *selected* program that provides cognitive and behavioral therapy (CBT) to disruptive second graders in Chile. SFL is one of the largest school-based mental health programs in the world, screening and treating more than 1,000,000 students over the past decade (see [Guzmán et al., 2015]). To identify eligible students, SFL teams use a psychometric scale measuring students' disruptiveness, and students above some cut-off are eligible. We randomly assigned 172 classes to either receive SFL in the first or second semester of the 2015 school year, and we measured outcomes at the start of the second semester, after the treatment group had received the treatment but before the control group received it. By comparing eligible students in the treatment and control groups, we can estimate the direct effects of the program, and by comparing ineligible students in the two groups we can estimate its spillover effects.

A number of studies conducted in high-income countries have shown that selected programs

similar in spirit and intensity to SFL can improve the behavior and academic performance of disruptive children. A meta-analysis conducted by Wilson and Lipsey [2007] includes 108 studies of such programs. The authors find that on average, selected programs reduce treated children's disruptive behavior by 0.29σ, and increase their academic performance by 0.23σ.[1] There are much fewer studies of selected programs similar to SFL in middle- and low-income countries. In a recent meta-analysis of psychosocial interventions for disruptive children in low- and middle-income countries (see [Burkey et al., 2018]), only two selected programs specifically target disruptive children (see [Ellas et al., 2003] and [Bratton, 2011]). Consistent with the evidence from high-income countries, both studies find large positive effects.

However, two key features distinguish our study from earlier work. First, Ellas et al. [2003], Bratton [2011], and most of the studies reviewed in Wilson and Lipsey [2007] consider demonstration programs, implemented with significant researcher involvement. However, meta-analyses reveal substantial falloff in effect size when interventions move from research to practice contexts, and when they are implemented away from the developer's control (see e.g. [Curtis et al., 2004]). This phenomenon is referred to as the "implementation cliff" in clinical psychology (see [Weisz et al., 2014]), and it has also been documented in economics (see e.g. [Banerjee et al., 2016]). Therefore, these studies may not be informative about the potential of selected programs for disruptive children if they were to be scaled-up as nationwide policies. Our paper, on the other hand, can shed some light on that question, as we study a program implemented as a nationwide policy.

Second, to the best of our knowledge, we are the first to study the spillover effects of school-based CBT programs for disruptive children on their non-disruptive peers. Neither the articles considered by Wilson and Lipsey [2007] in their review nor the 785 articles citing Wilson and Lipsey [2007] on Google Scholar as of December 2018 estimate such spillover effects, thus suggesting that this question remains unaddressed in the literature. Relatedly, Aizer [2008] shows that when a student is treated for ADHD, her disruptiveness diminishes and her peers' academic performance increases. However, ADHD treatment is usually pharmacological, not school based, and individually administered. Its spillover effects may then differ from those created by the CBT, school-based, and group-level treatment we study.

We do not find any direct effect of the SFL program on eligible students' disruptiveness, academic achievement, and mental health. The difference between our results and those in the litera-

[1] In a more recent review, Sandler et al. [2014] find similar results to those in Wilson and Lipsey [2007].

ture does not come from a lack of statistical precision in our study: based on our estimates, we can reject effects much smaller than the average effects found by Wilson and Lipsey [2007]. Later in the paper, we hypothesize that this difference may be another manifestation of the "implementation cliff" phenomenon: SFL is a nationwide program run by the government in thousands of schools, without any researcher or NGO involvement.

Moreover, the programs worsens the studying conditions in treated classes. We asked teachers and the enumerators we sent to observe the classes to rate classes' overall disruptiveness. Both teachers' and enumerators' ratings are significantly higher in the treatment than in the control group. The standardized index formed using those ratings and other measures of the disruptiveness of the class is also higher in the treatment than in the control group, and the difference is highly significant. This worsening of studying conditions seems to come from an increase in the disruptiveness of ineligible students. As per their teachers, ineligible students in treated classes are more disruptive than those in control classes. However, this effect is no longer significant after accounting for multiple testing. An exploratory analysis suggests that, if real, this increase in ineligible students' disruptiveness may come from the fact they perceive the treatment as a reward: eligible students get to skip classes during the sessions, and the sessions mostly consist in games and role play. Then, ineligible students may increase their disruptiveness, in the hope of joining the program, or because they find it unfair that disruptive students get rewarded and not them.

Finally, ineligible students have more friends in treated than in control classes. This difference comes from a larger number of friendship ties with other ineligible students, rather than with eligible ones. This may come from a "minimal group" effect that has been extensively documented in the psychological literature: the mere fact of labelling people into groups leads them to favor members of their own group (see [Tajfel et al., 1971]). Accordingly, we find that the treatment increases the segregation of friendship networks between eligible and ineligible students.

Overall, our results suggest that when implemented as nationwide policies, school-based mental health programs for disruptive students may not produce the strong positive effects found in demonstration studies. Unfortunately, we cannot disentangle which part of the scaling-up process should be improved. Our results call for further research in that area.

The remainder of the paper is organized as follows. In Section 2, we present the SFL program. In Section 3, we present the randomization, the data we use, and the population under study. In Section 4, we present compliance with randomization, the balancing checks, and attrition. In

Section 5, we present the main results. In Section 6, we interpret the results and present some exploratory analysis. Section 7 concludes.

2 The SFL program

SFL is a school-based mental health program whose goal is to equip second graders suffering from conduct disorders with the soft skills necessary to adapt to the school environment. It is managed by JUNAEB (*Junta Nacional de Auxilio Escolar y Becas*), the division of the Chilean Department of Education in charge of most of the non-teaching programs implemented in Chilean schools. The program started as a pilot in 1998. Over the next 3 years, members of JUNAEB collaborated with researchers from Chile and other countries to review the screening measures and intervention programs available at that time and adapt them for use in Chile. The program became a nationwide policy in 2001, and it is currently implemented in 1,637 publicly-funded elementary schools in Chile (see [Guzmán et al., 2015]). These schools account for 20% of all elementary schools in Chile, and they are the most disadvantaged. As the Chilean public school system is administrated at the municipal level, SFL teams implementing the program are also organized at this administrative level.

The municipalities participating in our study have been implementing SFL for 11.5 years, so their teams have a fair amount of experience in the program. A college degree with a psychology major is considered sufficient for being part of an SFL team (see [Guzmán et al., 2015]).

To identify eligible students, SFL teams use a psychometric scale, the Teacher Observation of Classroom Adaptation (TOCA, see [Kellam et al., 1977], and [Werthamer-Larsson et al., 1990]). In the end of each academic year, first grade teachers fill the TOCA questionnaire for each of their student. Based on this questionnaire, students receive scores on the following six scales: authority acceptance (AA), attention and focus (AF), activity levels (AL), social contact (SC), motivation for schooling (MS), and emotional maturity (EM). The TOCA questionnaire concludes with two summary questions, where teachers have to give ratings of the overall disruptiveness and academic ability of each of their student.

Then, the three following groups of students are eligible for the program:

- Students above the 75th percentile of the AA scale, above the 85th percentile of the AF and AL scales, and below the 25th percentile of the MS scale;
- Students below the 25th percentile of the SC scale, and either above the 75th percentile of

the AA scale or above the 85th percentile of the AL scale;

- Students below the 25th percentile of the SC, MS, and EM scales, and below the 50th percentile of either the AA or AL scale.

The cut-offs are gender specific, to ensure that not only males are eligible. Students in the third eligibility group are not disruptive, but they only account for 7% of eligible students, while the first two groups respectively account for 40% and 53% of eligible students. Depending on the year, eligible students account for 15 to 20% of first-grade students whose teachers fill the TOCA questionnaire.

In second grade, SFL teams ask eligible students' parents the authorization to enroll their child in the program. If their parents accept, eligible students are enrolled in a workshop consisting in 10 two-hours cognitive and behavioral therapy (CBT) group sessions, implemented by two SFL psychologists. The sessions take place weekly, during the class day, over the course of one semester. During sessions, enrolled students leave the classroom, while their classmates remain there and continue with their normal schedule. The time of the group sessions is set in coordination with teachers, to avoid that enrolled students lose key instruction time.

As per the SFL manual, the program is divided into five parts. The goals of the first part are to welcome children and build a group identity, for instance by having children choose a group name. The goals of the second part are to improve children's self-esteem, and their respect of others. Then, during the third part, the psychologists help students put words on their and others' emotions, and help them share their emotions with others. Then, the fourth part is dedicated to self-control techniques, and to strategies to find non-violent solutions to conflicts. Finally, the last part is dedicated to a review of what has been learnt during the workshop. Sessions are activity based, involve games and role play, and are rooted in positive psychology. For instance, if they behave well during a session, students are praised and receive rewards like cakes or candies.

To ensure that the program is properly delivered, SFL teams attend every year a two-days retraining, in which they review how to conduct the workshops. Moreover, SFL implementers attend "good practices" meetings every six months, in which they share what seems to work in the sessions.

As per the SFL guidelines, six to 12 students should participate in a workshop. If there are less than six eligible students in a school, no workshop takes place, and if a school has more than 12 eligible students, two workshops take place in that school. In the next section, we explain how we exploit these features in our randomization. Finally, the parents of enrolled children are invited to

three information sessions, whose goals are to ensure that they are aware of the activities conducted in the workshop, and that they encourage their child to attend.

The 108 selected programs reviewed by Wilson and Lipsey [2007] in their meta-analysis are similar in spirit and in intensity to SFL. Like SFL, many of these programs teach students CBT techniques to help them acknowledge their emotions and deal with anger. Their median duration (13 weeks) is close to that of SFL (10 weeks). Finally, the majority of those programs treat students in groups, and a large fraction of them target elementary school students.[2] However, the personnel delivering the programs reviewed by Wilson and Lipsey [2007] is very different from that delivering SFL. 90% of those programs are "mounted by a researcher for research or demonstration purposes with the researcher often being the program developer and heavily involved in the program implementation", and only 10% are routine practice programs implemented without researcher involvement.[3] Among the research or demonstration programs, a third are fully delivered by researchers (often in psychology or education), a fifth are delivered both by researchers and by psychologist or teachers, while the remainder are delivered by psychologists or teachers under the close supervision of researchers. On the other hand, SFL is a routine practice program delivered by psychologists with college-level degrees, without any researcher involvement.

3 Randomization, data, and study population

3.1 Sample selection and randomization

Our sample consists of 172 classes. All municipal teams conducting the SFL program in the Santiago and Valparaiso regions, the two most populated regions in Chile, were invited to join the study. 32 out of 39 accepted our invitation. In March 2015, these teams visited the schools covered by the program in their municipalities, and collected data on the number of students eligible for the program enrolled in each second grade class. 172 classes with four or more eligible students and in schools with six or more eligible students were included in the study. The second criterion ensured that group sessions would indeed take place in the school, while the first criterion ensured that there were enough treated students per class to potentially generate spillover effects. About 450 classes participate in a SFL workshop each year in the Santiago and Valparaiso regions, so our

[2]Wilson and Lipsey [2007] do not find that programs durations and the age of participants are correlated with programs treatment effects. Individual programs seem to produce slightly larger effects than group programs, though the difference becomes insignificant when controlling for other study characteristics.

[3]Wilson and Lipsey [2007] do not report the average effect of the routine practice programs in their meta-analysis. They note that this average effect is smaller than, but not significantly different from, the average effect of research or demonstration programs, but they do not report the confidence interval of the difference between these two averages.

sample covers about 40% of the classes covered by the program in those regions.

Randomization took place both within schools and within municipalities. There were 29 schools with two classes included in our sample and where it was possible to form two groups of six students or more without grouping students of the two classes together. In such instances, we conducted a lottery within the school, to assign one of the two classes to receive the treatment in the first semester of 2015, and the second class to receive it in the second semester. For the remaining 114 classes in our sample, randomization took place within municipalities. Overall, we conducted 56 lotteries (29 within schools, and 27 within municipalities) and we assigned 89 classes to receive the treatment in the first semester, from April to June 2015, and 83 to receive it in the second semester, from September to December 2015.

3.2 Data

In our analysis, we use data produced by JUNAEB. First, we use the six first-grade TOCA scores that determine students' eligibility to SFL, as well as the teachers' ratings of students' disruptiveness and academic ability in the TOCA questionnaire. Then, we also use another psychometric scale collected by JUNAEB and measuring students' disruptiveness, the pediatric symptom checklist (PSC, see [Jellinek et al., 1988]), which is filled by students' parents. We also use JUNAEB's data on treatment implementation. Specifically, for each class in our sample we know how many SFL group sessions were conducted in the first semester of 2015. For each student, we know how many sessions she attended, and how many sessions her parents attended. Finally, JUNAEB also provided us data on students' socio-economic background, as well as their monthly school attendance from March 2015 to June 2015.

We also use baseline data collected in March 2015, before the treatment started in the treatment group classes, and endline data collected in August 2015, after the treatment ended in the treatment group classes and before it started in the control group classes. Both at baseline and endline, two enumerators visited each of the 172 classes included in the experiment during a half day. Enumerators were undergraduate students, mostly psychology and education majors. Every person who applied to become an enumerator first had to attend a half-day training, during which he/she was taught how to administer our questionnaires. Candidates also had to take a test at the end of the training, and only those who scored above some threshold became enumerators.

Our questionnaires slightly changed from baseline to endline. Below, we describe our endline questionnaires, and we explain the difference between our baseline and endline questionnaires when

needed later in the paper.

The enumerators first administered a non-cognitive questionnaire to the students. That questionnaire aimed at measuring:

- Students' happiness in school, using a question from the student SIMCE questionnaire.[4]
- Students' self-control, using items of the child self-control psychometric scale (see [Rorhbeck et al., 1991]) that we translated into Spanish.
- Students' self-esteem, using items of the self-perception for children psychometric scale (see [Harter, 1985]) translated and validated into Spanish (see [Molina et al., 2011]).

Second, the enumerators administered a Spanish and mathematics test to the students. Third, the enumerators interviewed individually each student and asked her to name up to three students that she likes to play with during breaks, hereafter referred to as the student's friends. Fourth, the enumerators observed a one-hour lecture. During that observation, they observed the behaviour of each student during five seconds, and assessed whether the student was studying, not studying, or being disruptive. They repeated that process five times, and then rated the overall disruptiveness of each student by answering the summary question from the TOCA questionnaire. During that one-hour lecture, the enumerators also recorded the decibel levels in the class using a smartphone app, and wrote down the time at which the lecture was supposed to start and the time when it effectively started. Fifth, the enumerators filled a short questionnaire aimed at assessing the overall disruptiveness in the class, using questions taken from the PISA (Program for International Student Assessment) questionnaire, asking them their agreement with statements such as: "There is noise and disorder in this class," or "The teacher has to wait for a long time before students calm down and he/she can start teaching".

Finally, the enumerators also administered a questionnaire to the teachers. That questionnaire aimed at collecting: teachers' socio-demographic characteristics; teachers' ratings of the overall disruptiveness of the class, using the same PISA questions as those asked to enumerators; teachers' rating of the prevalence of bullying in the class; teachers' motivation, taste for their job, and mental health levels. The questionnaire was for the most part composed of questions from the SIMCE teacher questionnaire. Teachers also rated the overall disruptiveness of each of their student by answering the summary question from the TOCA questionnaire.

[4]The SIMCE (*Sistema de Medición de la Calidad de la Educación*) questionnaires are the nationwide standardized cognitive and non cognitive questionnaires administered to students and teachers in Chile.

The list of the outcome variables we consider in the paper was pre-specified in a pre-analysis plan (PAP) available at `https://www.socialscienceregistry.org/trials/1080`. That plan was time-stamped on 04/28/2017, before JUNAEB sent us students' first grade TOCA scores, as a letter from JUNAEB officials also available on the social science registry website testifies. Students' first-grade TOCA scores are necessary to distinguish eligible and ineligible students in our data, a distinction that underlies most of our analysis. The analysis presented in Sections 4, 5.1, 5.2, and 5.3 follows our pre-analysis plan, except for a few exceptions described below. On the other hand, the analysis presented in Sections 5.5 and 6 was not pre-specified in our PAP.

The student-level outcome measures listed in our PAP are:

- the student's happiness in school, self-control, self-esteem, Spanish, and mathematics scores,
- the percentage of school days missed by the student from April to June 2015,
- the rating of the student's disruptiveness by her teacher,
- the average rating of the student's disruptiveness across the two enumerators,
- the percentage of the student's classmates that nominate her as one of their friends,
- an indicator for whether the student is not nominated as a friend by any other student,
- the average disruptiveness at baseline of the student's endline friends,
- the average baseline Spanish and mathematics scores of the student's endline friends.

The class-level outcome measures listed in our PAP are:

- the teacher's rating of the class's disruptiveness, constructed using teachers' answers to the PISA questions measuring the disruptiveness in the class,
- the teacher's rating of the prevalence of bullying in the class,
- the average rating of the class's disruptiveness across the two enumerators, constructed using enumerators' answers to the PISA questions measuring the disruptiveness in the class,
- the number of minutes between the moment the class was supposed to start and the moment it effectively started according to the enumerators,
- the average decibel levels during the class across the two enumerators' recordings.

We standardize the school happiness, self-control, self-esteem, disruptiveness and test score measures to have a mean of 0 and a σ of 1 in the sample.

3.3 Assessing data quality

Some of the dimensions we are trying to measure are hard to observe. To get a sense of the reliability of our measures, Table A1 shows their baseline-endline correlation in the control group. Students' Spanish and mathematics test scores have high positive baseline-endline correlations, around 0.5. Those correlations are still far from one, probably because students in our study are young and their cognitive ability is not fixed yet. Our measure of students' popularity also has a fairly high baseline-endline correlation, around 0.3. On the other hand, our school happiness, self-esteem, and self-control measures have lower baseline-endline correlations, around 0.1-0.2.

Turning to disruptiveness measures, the rating of students' disruptiveness by teachers has a high positive baseline-endline correlation, equal to 0.42, which is almost as high as the baseline-endline correlation of test scores. This is all the more remarkable as we use first grade teachers' answer to the TOCA summary question as our baseline measure,[5] so our baseline and endline measures were not made by the same teacher. This suggests that disruptiveness is a relatively stable characteristic of students, and that different teachers tend to agree in their ratings of students' disruptiveness. Then, Table A2 shows that this measure is negatively correlated with students' academic ability: at baseline, its correlation with students' average test score in Spanish and mathematics is equal to -0.28. Finally, the bottom panel of Table A1 shows that teachers' rating of the disruptiveness of the class also has a high baseline-endline correlation, equal to 0.50.

Enumerators' ratings of students' disruptiveness has a baseline-endline correlation close to, and insignificantly different from, zero. This could be due to the fact that endline and baseline observations are made by different enumerators, who may have a different interpretation of what it means to be disruptive. Still, Table A2 shows that at baseline, the ratings made by the two enumerators are highly correlated, thus suggesting that different enumerators observing the same one-hour lecture agree on students' disruptiveness. Then, this lack of correlation could be due to the fact that the enumerators only observe students during an hour, and students' behavior during that hour may differ from their average behavior. Still, Table A2 shows that enumerators' ratings correlate reasonably well with teachers', and with students' academic ability. Overall, enumerators' ratings

[5] We decided to include the summary TOCA question in our baseline teacher questionnaire after having collected more than half of the baseline data, so that variable is missing for many classes at baseline.

of students' disruptiveness seem to be noisier than that of teacher. On the other hand, enumerators' ratings of classes' disruptiveness has a relatively high baseline-endline correlation, around 0.25, and Table A2 shows that this measure correlates well with that of teachers.

The decibel measure constructed following our PAP also has a very low baseline-endline correlation, and it does not correlate at all with teachers' and enumerators' ratings of classes' disruptiveness. The app's measurement does not seem very precise: enumerators recording the same lecture sometimes end up with average noise levels differing by more than 10 decibels. This measurement also seems to depend on the make of the phone and on idiosyncratic factors specific to the enumerator's phone. Therefore, we depart from our PAP and use a slightly different measure. We start by regressing the average decibels measured by enumerator i in class j on enumerator fixed effects, in the sample of control group classes. Then, we compute the residuals from that regression both for treatment and control group classes, and we define our measure for class j as the average of the residuals of the two enumerators for that class. Our measure can therefore be interpreted as the difference between the average decibels recorded in class j and the average of the recordings made by the same enumerators in the control group. This measure has a higher baseline-endline correlation than the measure described in our PAP, though Table A1 shows that this correlation is still not significant. But it also has a much larger correlation with enumerators' ratings of the class disruptiveness, and that correlation is significant as shown in Table A2. Throughout the paper, we use that measure instead of that described in our PAP.

3.4 Study population

The 172 classes included in our sample bear 5,704 students, meaning that classes have an average of 33.2 students. 4,466 students are ineligible to the program (26.0 per class), while 1,238 students are eligible (7.2 per class). Column (1) in Table 1 below presents the baseline characteristics of ineligible students. 33.8% of them are born to teenage mothers, which is more than twice the corresponding proportion in Chile.[6] 75.2% of them live in households below the 20th percentile of the social security score. Being below this threshold opens eligibility for 22 social programs and is usually considered as a proxy for poverty. 44.4% of them live in households below the 5th percentile of the social security score. Being below this threshold opens eligibility for 3 more social programs and is usually considered as a proxy for extreme poverty. Overall, the students included in our study live in households disproportionately coming from the bottom of the Chilean income distribution.

[6] See `http://web.minsal.cl/portal/url/item/c908a2010f2e7dafe040010164010db3.pdf`.

Column (2) in Table 1 presents the baseline characteristics of eligible students, and Column (3) reports the p-value of tests that the baseline characteristics of eligible and ineligible students are equal. Panel A shows that eligible students are more likely to be males and less likely to live with their father. Their parents are also less educated than that of ineligible students. Panel B shows that eligible students's self-control and self-esteem scores are about 0.2σ lower than that of ineligible students. But differences between the two groups are even more pronounced when one considers their disruptiveness and academic ability. Eligible students score 1.2σ higher than ineligible students on first-grade teachers' disruptiveness ratings, and 0.4σ higher on enumerators' baseline ratings. They also score 0.4σ lower on the Spanish and mathematics tests. Eligible students are also less popular than ineligible ones: 7.6% of the students in the class nominate them as friends, against 8.8% for ineligible students. The average disruptiveness of their friends is also about 0.2σ higher than that of ineligible's friends, thus suggesting some assortative matching along the disruptiveness dimension, though the difference between the disruptiveness of the two groups is much larger than that between their friends.

Finally, Table A4 shows some characteristics of the teachers in our sample. 96.3% of teachers are females. Their average age is 42.8 years old, they have an average of 16.5 years of experience as a teacher, and 8.6 years of experience in the school where they currently teach, and 86.3% percent of them have a university degree.

4 Compliance, internal validity, and estimation methods

4.1 Compliance with randomization and fidelity of treatment assignment

In this section, we show that the SFL teams followed the randomization, and implemented the treatment as per the program's rules: in the treatment group classes, very few ineligible students received the program.

To do so, we estimate the effect of being assigned to treatment on actual exposure to treatment during the first semester of 2015. Let Y_{ijk} be a measure of exposure to treatment for student i in class j and lottery k. We estimate the following regression:

$$Y_{ijk} = \gamma_k + \beta D_{jk} + u_{ijk}, \tag{1}$$

where the γ_ks are lottery fixed effects, and where D_{jk} is an indicator variable equal to 1 if lottery k assigned class j to the treatment group and to 0 otherwise. $\widehat{\beta}$ estimates a weighted average across lotteries of the within-lottery difference between the average of Y_{ijk} in treatment and control group

Table 1: Characteristics of eligible and ineligible students

	Ineligible (1)	Eligible (2)	P-value (3)	N (4)
Panel A: demographic characteristics				
Male	0.498	0.582	0.000	5704
Teen mother	0.338	0.36	0.199	4440
Student lives with father	0.635	0.554	0.000	3765
$\leq$ p20 social security score	0.752	0.77	0.198	5068
$\leq$ p5 social security score	0.444	0.456	0.469	5068
Mother's education	9.131	8.564	0.000	4727
Father's education	9.163	8.439	0.000	4117
Panel B: baseline measures				
School happiness score	0.023	-0.063	0.022	4431
Self-control score	0.048	-0.166	0.000	4594
Self-esteem score	0.041	-0.146	0.000	4610
Overall disruptiveness TOCA	-0.293	0.873	0.000	4850
Disruptiveness, enumerator	-0.089	0.322	0.000	4646
Spanish test score	0.095	-0.335	0.000	4758
Math test score	0.082	-0.289	0.000	4758
% class friends with student	0.088	0.076	0.000	4721
Friends' average disruptiveness	-0.051	0.188	0.000	3931

Notes: This table reports descriptive statistics for students in the sample. Column (1) reports the mean of the outcome variable for ineligible students and Column (2) reports the mean of the outcome variable for eligible students. Column (3) reports the p-value of a test that the two means are equal. Column (4) reports the number of observations used in the comparison.

classes. To account for the fact that the treatments of classes participating in the same lottery are correlated, we cluster the standard errors at the lottery level.

To estimate the effect of assignment to treatment on class-level measures of exposure, we estimate Regression (1), except that we use propensity score reweighting instead of lottery fixed effects in the regression. With propensity score reweighting, β is also identified out of comparisons of treatment and control group classes in the same lottery (see [Rosenbaum and Rubin, 1983] and [Hirano et al., 2003]). Using propensity score reweighting ensures that the regression does not have too many independent variables with respect to its number of observations (with lottery fixed effects, Regression (1) would have 57 independent variables and at most 172 observations). In any case, as the share of treated classes is equal to 0.5 in more than 80% of the lotteries (46 out of 56), using lottery fixed effects or propensity score reweighting does not make a large difference in this paper.

Column (1) of Table 2 below shows the mean value of eight measures of exposure to the treatment in the control group. Column (2) shows estimates of β for these eight measures. Column (3) shows estimates of the standard error of $\widehat{\beta}$. Column (4) shows the p-value of a t-test of the null hypothesis $\beta = 0$. To account for the fact that we consider several measures of exposure to the treatment, Column (5) shows the p-value controlling the False Discovery Rate (FDR) across the eight tests we conduct (see [Benjamini and Hochberg, 1995]). Finally, Column (6) shows the number of observations used in the estimation.

Panel A of the table shows that in the first semester of 2015, SFL sessions were conducted in 8.4% of the control group classes and in 98.1% of the treatment group classes. On average, 0.6 sessions were conducted in the control group classes against 9.5 in the treatment group classes. Throughout the paper, we estimate intention to treat (ITT) effects of assigning a class to the treatment. Given that less than 10% of treatment group classes received the treatment, while almost 100% of treatment group classes received it, this ITT effect "almost" estimates the effect of implementing the treatment in a class. This parameter is probably the policy-relevant one in our context: policy-makers can control whether the treatment is implemented in a class, but they cannot control whether students actually attend.

Moving to attendance, panel A also shows that 4.8% of eligible students in the control group attended at least one session, against 84.9% in the treatment group. Discussions with the SFL team suggest that the most common reason why some eligible students did not attend any group session is that their parents refused that they participate. Table A3 compares the characteristics of the

"takers", eligible students in the treatment group that attended at least one session, to those of the "non takers" that did not attend any session. The main difference between the two groups is that the takers seem to be less disruptive at baseline: their average first-grade teacher disruptiveness rating is 0.3σ below that of non takers. On average, eligible students attended 0.4 sessions in the control group, against 7.4 in the treatment group. This number is 8% lower than $9.5 \times 0.849 = 8.1$, the number we would have observed if students attending at least one session had attended all the sessions conducted in their class. This small difference may for instance arise from the fact that some of those students have missed class on a few workshop days.

Finally, Panel A shows that the fidelity with the program's assignment rules was very high: in the treatment group, only 1% of ineligible students attended at least one session, and they attended an average of 0.089 sessions. The fact that ineligible students were almost not exposed at all to the treatment is crucial for us to be able to estimate the spillover effects of the treatment on them. If in the treatment group, a non-negligible proportion of ineligible students had received the treatment, the comparisons of ineligible students in the treatment and control groups would have estimated a mixture of direct and spillover effects of the program.

Panel B of the table shows that compliance with randomization was lower for the parents' than for the students' workshops: 53.5% of eligible parents in the treatment group attended at least one session, and eligible parents attend on average 1.0 sessions out of 3. There again, fidelity with the program's assignment rules was almost perfect, with very few ineligible parents attending a session.

4.2 Internal validity

Balancing checks

We test for baseline differences between the treatment and control groups by estimating Regression (1) with class-, teacher-, and student-level baseline measures as the dependent variables. First, Table A7 compares eligible students in the treatment and control groups on 29 baseline characteristics. Only two differences are significant at the 10% level: treatment group students are more disruptive as per enumerators' ratings, and they are more likely not to be nominated as a friend by any other student in the class. Those differences are not significant at the 5% level, and they become insignificant when p-values are adjusted for the fact we conduct 29 tests in the table.

Second, Table A10 compares ineligible students in the treatment and control groups on the same 29 baseline characteristics. Four differences are significant at the 10% level, one of which

Table 2: Compliance with randomization

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: students' workshops						
$\geq$1 session conducted in class	0.084	0.897	0.035	0.000	0.000	172
Sessions conducted in class	0.602	8.942	0.337	0.000	0.000	172
Eligible students attended $\geq$1 session	0.048	0.801	0.029	0.000	0.000	1238
Sessions attended by eligible students	0.37	6.992	0.304	0.000	0.000	1238
Ineligible students attended $\geq$1 session	0.000	0.01	0.004	0.011	0.016	4466
Sessions attended by ineligible students	0.000	0.089	0.038	0.022	0.028	4466
Panel B: parents' workshops						
Eligible parents attended $\geq$1 ses.	0.048	0.487	0.039	0.000	0.000	1238
Sessions attended by eligible parents	0.099	0.933	0.107	0.000	0.000	1238
Ineligible parents attended $\geq$1 ses.	0.000	0.008	0.004	0.039	0.043	4466
Sessions attended by ineligible parents	0.000	0.016	0.008	0.062	0.062	4466

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator. For student-level dependent variables, the regression includes lottery fixed effects. For class-level dependent variables, the regression is computed with propensity score weights. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression. All the dependent variables come from JUNAEB's program implementation data sets.

is also significant at the 5% level. Treatment group students have slightly worse social contact, attention and focus, activity level, and disruptiveness TOCA scores. Here again, those differences become insignificant when p-values are adjusted for multiple testing.

Third, Table A13 compares teachers in the treatment and control groups on 12 characteristics. One difference is significant at the 5% level: treatment group teachers have worse happiness scores than control group ones. Again, this difference becomes insignificant when p-values are adjusted for multiple testing.

Finally, Table A14 compares six class-level characteristics in the treatment and control groups. Three differences are significant at the 10% level, one of which is significant at the 5% level. Treated classes are more disruptive than control ones according to teachers and enumerators, and have higher decibel levels. Again, these differences become insignificant when p-values are adjusted for multiple testing.

Overall, we conduct 76 balancing checks in Tables A7, A10, A13, and A14. We find 10 significant differences between the treatment and control groups at the 10% level, three significant differences at the 5% level, and no significant difference at the 1% level. In Section 5, we show that our results do not change when we account for these differences in our statistical analysis.

Attrition

In this section, we document the percentage of students in our sample for which endline measures are not available, and the most common reasons for such attrition. We also provide evidence suggesting that the treatment and the control groups do not present differential levels of attrition, and that the characteristics of treatment and control group students for which endline measures are available are still balanced.

Table A5 considers attrition among eligible students. Column (1) shows the levels of attrition in the control group. Endline measures collected by the enumerators are missing for 25.3% of students. For 5.9% of them this is because they have left the class between baseline and endline, for instance because their parents have moved to a different neighborhood. For the most part, the remaining 19.4% correspond to students that were absent on the day when the enumerators visited the class.[7] The teacher's endline disruptiveness rating is missing for 24.8% of students. Again, for some of them

[7]There are also a couple of classes that enumerators could not visit at endline, because the school principal did not want to sacrifice again a half day of instruction for the purpose of the study.

this is because they have left the class at endline. But for the majority of students, this is because their teachers refused to rate their students' disruptiveness, or only rated, say, the first half of the class and then stopped because they thought the task was too time-consuming. Column (2) of Table A5 shows tests of differential attrition between the treatment and control groups. Those tests are conducted by estimating Regression (1) with variables measuring whether students are still in the sample at endline as the dependent variables. Attrition does not seem differential: of the five measures we consider, only one is significantly different between the treatment and control groups at the 10% level, and this difference becomes insignificant when p-values are adjusted for multiple testing.

Table A6 considers attrition among ineligible students. Columns (1) and (2) respectively show the levels of attrition in the control group, as well as tests for differential attrition between the treatment and control groups. The attrition levels in the control group are similar to those observed among eligible students. Here again, attrition is not differential: of the five measures we consider, only one is significantly different between the treatment and control groups at the 10% level, and this difference becomes insignificant when p-values are adjusted for multiple testing.

Finally, we conduct balancing checks again, among the students whose endline measures are available. Table A8 (resp. Table A9) considers the same 29 baseline characteristics as in Table A7, and compares their mean in the treatment and control groups, among the eligible students for which enumerators' endline measures (resp. the teacher's endline disruptiveness rating) are (resp. is) available. As in Table A7, few differences are significant. Table A11 repeats the same exercise, among ineligible students for which enumerators' endline measures are available. Again, few differences are significant. Finally, Table A12 compares ineligible students for which the teacher's endline disruptiveness rating is available in the treatment and control groups. More differences are significant, but most become insignificant once p-values are adjusted for multiple testing. Overall, the post-attrition treatment and control group students whose outcomes are compared in Section 5 have balanced baseline characteristics.

Turning to class-level outcomes, while we have teachers' and enumerators' ratings of classes' disruptiveness for more than 90% of classes in our sample, we have some differential attrition for teachers' questionnaires: none is missing in the control group, while 8% are missing in the treatment group, and the difference is statistically significant. In Table A15, we conduct again the balancing checks on the baseline class-level measures in Table A14.[8] For measures made by teachers, we

[8]Table A15 was not pre-specified in our PAP, because we had not anticipated the possibility of differential attrition for the class-level measures.

restrict the sample to classes for which all class-level endline teacher measures are available, while for measures made by enumerators we restrict the sample to classes for which all class-level endline enumerators measures are available. As in Table A14, three differences are significant at the 10% level, but none is significant at the 5% level, and these differences become insignificant when p-values are adjusted for multiple testing.

4.3 Estimation methods

In this section, we discuss the methods we use to estimate the effect of the treatment. For all the student-level outcomes, we estimate the following regression:

$$Y_{ijk} = \gamma_k + X'_{ijk}\theta_1 + Z'_{jk}\theta_2 + \beta D_{jk} + u_{ijk}, \tag{2}$$

where Y_{ijk} is the outcome of student i in class j and lottery k, the γ_ks are lottery fixed effects, X_{ijk} and Z_{jk} respectively denote student- and class-level baseline variables used as statistical controls, and D_{jk} is an indicator variable equal to 1 if class j in lottery k was assigned to the treatment group. $\widehat{\beta}$ estimates the ITT effect of being assigned to the treatment on the outcome. As in Regression (1), we cluster the standard errors at the lottery level. To select the controls, we follow Belloni et al. [2014]. For the student-level controls, we run a Lasso regression of the outcome on all the student-level baseline variables in Table A7, and we pick the variables selected by the Lasso.[9] For the class-level controls, we run a Lasso regression of the class average of the outcome on the class average of all the student-level baseline variables in Table A7, and all the class-level baseline variables in Tables A13 and A14, and we pick the variables selected by the Lasso.

For all the class-level outcomes, we estimate the following regression:

$$Y_{jk} = \alpha + Z'_{jk}\theta + \beta D_{jk} + u_{jk}, \tag{3}$$

where Y_{jk} is the outcome of class j in lottery k, Z_{jk} denotes class-level baseline variables used as statistical controls, and D_{jk} is the treatment indicator. The regression is weighted by propensity score weights, and as in Regression (1), we cluster the standard errors at the lottery level. To select the controls, we follow again Belloni et al. [2014], and we run a Lasso regression of the outcome on the class average of all the student-level baseline variables in Table A7, and all the class-level baseline variables in Tables A13 and A14, and we pick the variables selected by the Lasso.

[9] In a randomized experiment, the treatment is by construction uncorrelated with the controls, so it is not necessary to run a Lasso regression of the treatment on the controls.

To account for multiple testing, we follow the same approach as Finkelstein et al. [2010]. First, we group related outcomes into hypothesis. For instance, students' happiness, self-esteem, and self-control scores are grouped together into an "emotional stability" hypothesis. Then, for each outcome, we report both the unadjusted p-value of the estimated effect, and the adjusted p-value controlling the FDR within the hypothesis the outcome belongs to. Each panel in Tables 3, 4, and 5 corresponds to a set of related outcomes grouped into an hypothesis. Finally, for each hypothesis we also report the effect of the treatment on a weighted average of the outcomes in that hypothesis, using the weights proposed in Anderson [2008]. We refer to the effect of the treatment on this weighted average as the standardized treatment effect.

5 Treatment Effects

5.1 Effects on eligible students

In this section, we assess whether the SFL workshops generate positive effects on eligible students.

Panel A of Table 3 suggests that being assigned to the SFL workshops may have conflicting effects on eligible students' emotional stability, though the estimated effects are not very significant. The average school happiness score is 0.123σ higher in the treatment than in the control group, but this difference is not very significant (p-value=0.101), and becomes insignificant after adjusting for multiple testing. The average self-esteem score is 0.106σ lower in the treatment group, but this difference is insignificant even before adjusting for multiple testing (p-value=0.176). Students' self-esteem scores can be decomposed into general, academic, and social self-esteem scores, so in an exploratory analysis we look at the treatment effect separately on these three scores. The average academic self-esteem score is 0.155σ lower in the treatment group (p-value=0.061), while the differences are much smaller for general and social self-esteem. This may indicate that being assigned to the program is stigmatizing, and reduces students' beliefs in their academic potential. The average self-control score is very close in the treatment and control groups. Finally, the average standardized score is also very close in the treatment and control groups, because the positive effect on school happiness is canceled out by the negative effect on self-esteem.

Panel B shows that SFL does not have any strong effect on eligible students' disruptiveness. At endline, the average teachers' disruptiveness rating is 0.089σ higher in the treatment than in the control group. This difference is not statistically significant at conventional levels, but based on its estimated standard error, we can rule out at the 5% level that SFL reduces teachers' disruptiveness

Table 3: Treatment effect on eligible students

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: emotional stability						
School happiness score	-0.107	0.123	0.075	0.101	0.304	876
Self-control score	-0.184	-0.04	0.087	0.648	0.648	880
Self-esteem score	-0.17	-0.106	0.079	0.176	0.264	903
Standardized Treatment Effect	0.015	-0.002	0.08	0.977		915
Panel B: disruptiveness						
Disruptiveness, teacher	0.353	0.089	0.101	0.377	0.755	904
Disruptiveness, enumerator	0.153	0.07	0.104	0.501	0.501	955
Standardized Treatment Effect	-0.036	0.053	0.088	0.547		1111
Panel C: academic outcomes						
% school days missed	12.82	1.055	1.016	0.299	0.896	1236
Spanish test score	-0.308	-0.045	0.067	0.497	0.745	956
Math test score	-0.274	-0.002	0.078	0.976	0.976	956
Standardized Treatment Effect	0.011	-0.031	0.07	0.662		1238
Panel D: integration in the class network						
No friends in the class	0.27	-0.028	0.027	0.307	0.613	1147
% class friends with student	0.07	0.007	0.005	0.145	0.581	1147
Friends' average ability	-0.061	0.007	0.069	0.919	0.919	829
Friends' average disruptiveness	0.177	0.085	0.084	0.307	0.41	787
Standardized Treatment Effect	-0.008	0.038	0.063	0.54		1148

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator, lottery fixed effects, and control variables for eligible students. The control variables are selected by a Lasso regression of the dependent variable on all potential controls, following the methodology proposed by Belloni et al. [2014]. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression. All the dependent variables, except for *% school days missed*, were collected by the authors at endline.

ratings by more than 0.109σ. This is around 1/3 of the treatment effect on students' disruptiveness found by Wilson and Lipsey [2007] in their meta-analysis of 108 programs similar to SFL. Similarly, the average enumerators' disruptiveness rating is 0.070σ higher in the treatment than in the control group, and the difference is again insignificant.

Panel C shows that SFL also does not have any strong effect on the academic outcomes of eligible students. The percentage of missed school days is approximately the same in the treatment and control groups, and the average Spanish and mathematics scores are also very close in the two groups. We can reject at the 5% level that SFL increases eligible students' Spanish and mathematics scores by more than 0.086σ and 0.151σ, respectively. Again, these effects are much smaller than those found in the meta-analysis by Wilson and Lipsey [2007].

Finally, Panel D shows that SFL may slightly improve eligible students' integration in the class network, though the estimated effects are not significant. For instance, the proportion of students not nominated as a friend by any other student in the class is 2.8 percentage points lower in the treatment than in the control group, but this difference is insignificant. Similarly, eligible students are nominated as friends by 7.7% of their classmates in the treatment group, against 7.0% in the control group, but again the difference is insignificant.

Overall, we do not find strong evidence of a positive effect of SFL on any of the dimensions we consider. We can also rule out much smaller effects on students' disruptiveness and academic achievement than those previously found for similar programs.

5.2 Effects on ineligible students

In this section, we explore whether the SFL workshops have spillover effects on ineligible students. Panel A of Table 4 shows that these workshops do not generate strong spillover effects on the emotional stability of ineligible students. The average school happiness, self-control, and self-esteem scores are very close and do not significantly differ in the treatment and control groups.

Panel B suggests that the SFL workshops may generate negative spillover effects on ineligible students' disruptiveness. At endline, the average teachers' disruptiveness rating is 0.208σ higher in the treatment than in the control group. This difference is significant (p-value=0.056), but it becomes marginally insignificant after adjusting for multiple testing (adjusted p-value=0.112). As shown in Table A12 in the Appendix, some baseline characteristics of the ineligible students with a teacher endline rating are imbalanced in the treatment and control groups. Not all of these

Table 4: Treatment effect on ineligible students

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: emotional stability						
School happiness score	0.026	0.016	0.037	0.666	0.666	3360
Self-control score	0.097	-0.05	0.043	0.25	0.751	3404
Self-esteem score	0.084	-0.043	0.047	0.36	0.54	3446
Standardized Treatment Effect	0.027	-0.023	0.042	0.577		3476
Panel B: disruptiveness						
Disruptiveness, teacher	-0.212	0.208	0.109	0.056	0.112	3203
Disruptiveness, enumerator	-0.068	-0.013	0.061	0.835	0.835	3530
Standardized Treatment Effect	-0.049	0.05	0.075	0.505		4034
Panel C: academic outcomes						
% school days missed	13.089	0.282	0.596	0.636	0.954	4427
Spanish test score	0.128	-0.054	0.054	0.318	0.954	3517
Math test score	0.08	0.004	0.056	0.945	0.945	3517
Standardized Treatment Effect	0.018	-0.004	0.046	0.929		4452
Panel D: integration in the class network						
No friends in the class	0.197	-0.035	0.013	0.008	0.033	4168
% class friends with student	0.087	0.005	0.003	0.081	0.163	4168
Friends' average ability	0.027	-0.006	0.055	0.91	0.91	3342
Friends' average disruptiveness	-0.11	-0.037	0.041	0.366	0.488	3176
Standardized Treatment Effect	0.003	0.081	0.036	0.026		4171

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator, lottery fixed effects, and control variables for ineligible students. The control variables are selected by a Lasso regression of the dependent variable on all potential controls, following the methodology proposed by Belloni et al. [2014]. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression. All the dependent variables, except for *% school days missed*, were collected by the authors at endline.

characteristics are selected as controls by the Lasso in the estimation in Table 4, so we reestimate the treatment effect on this measure, controlling for all the variables in Table A12 imbalanced at the 10% level. The estimated treatment effect is now equal to 0.220σ, with a p-value of 0.045. This suggests that the significant estimated treatment effect on teachers' disruptiveness rating is not due to some attrition bias. On the other hand, the average enumerators' disruptiveness rating is very similar in the treatment and control groups, but as explained in Section 3.3, the enumerators' rating seems noisier than the teacher's rating. A one-hour observation may not be sufficient for enumerators to precisely assess students' disruptiveness.

Panel C shows that the SFL workshops do not have strong spillover effects on ineligible students' academic outcomes. The percentage of missed school days and the average Spanish and mathematics scores are close and do not significantly differ in the treatment and control groups.

Panel D shows that SFL improves the integration of ineligible students in the class network. The proportion of students not nominated as a friend by any other student in the class is 3.5 percentage points lower in the treatment than in the control group, which represents a 17.7% reduction in the fraction of ineligible students who have no friends. This difference is significant (p-value=0.008), and it remains significant after accounting for multiple testing (adjusted p-value=0.033). Similarly, ineligible students are nominated as friends by 9.2% of their classmates in the treatment group, against 8.7% in the control group. This difference is significant before but not after adjusting for multiple testing. On the other hand, the treatment does not significantly alter the academic ability and disruptiveness of ineligible students' friends. Finally, the average standardized score constructed from these four outcomes is significantly higher in the treatment than in the control group (p-value=0.026), thus confirming that SFL significantly improves the integration of ineligible students in the class network.

Overall, we find suggestive evidence of a negative spillover effect on the disruptiveness of ineligible students, though this effect becomes insignificant after adjusting for multiple testing. We also find a positive spillover effect on the integration of ineligible students in the class network, that remains significant after adjusting for multiple testing.

5.3 Effects on the classroom environment

In this section, we study how the SFL workshops affect different measures of classrooms' environment at endline. Table 5 shows that according to teachers, treated classes are 0.232σ more

Table 5: Treatment effect on classroom environment

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Disruptiveness, teacher	-0.187	0.232	0.137	0.091	0.226	160
Bullying in class, teacher	-0.038	0.105	0.153	0.492	0.492	160
Disruptiveness, enumerator	-0.186	0.389	0.148	0.009	0.043	167
Delay in class's start (minutes)	9.938	1.204	1.046	0.25	0.312	160
Average decibels during class	0.022	0.681	0.487	0.162	0.27	169
Standardized Treatment Effect	-0.215	0.424	0.131	0.001		169

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator and control variables, computed with propensity score weights. The control variables are selected by a Lasso regression of the dependent variable on all potential controls, following the methodology proposed by Belloni et al. [2014]. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression. All the dependent variables were collected by the authors at endline.

disruptive than control ones. This difference is statistically significant before adjusting for multiple testing (p-value=0.091), but it becomes insignificant after adjusting for it (adjusted p-value=0.226). Enumerators agree with teachers: according to them, treated classes are 0.389σ more disruptive than control ones. This difference is statistically significant before and after adjusting for multiple testing (p-value=0.009, adjusted p-value=0.043). While enumerators' assessments of students' disruptiveness seem to be noisy, their assessments of classes' disruptiveness seem quite reliable (see Section 3.3), so this result is informative. Moreover, enumerators presumably do not know if the class they observe has been treated or not, contrary to teachers. The fact that they also find that treated classes are more disruptive suggests that teachers' worse perception of the treatment-group classes is not a mere placebo effect.

Table A14 in the Appendix shows that treated and control classes are imbalanced on these two measures at baseline, so we reestimate these two effects controlling for these two measures.[10] The estimated treatment effects on teachers' and enumerators' ratings are now respectively equal to 0.247σ (p-value=0.084) and 0.282σ (p-value=0.066). This suggests that the significant estimated treatment effects on these two measures are not due to imbalances already existing at baseline.

Table 5 also shows that treated classes have higher levels of bullying, that their lectures start

[10]In the estimation of the treatment effect on teachers' ratings, the Lasso selects teachers' baseline ratings as a control, but it does not select enumerators' ratings. In the estimation of the treatment effect on enumerators' ratings, the Lasso does not select any control.

1.2 more minutes after the scheduled time than in control classes, and that they have higher levels of decibels. Even though these results are not statistically significant, they go in the same direction as the results on the disruptiveness measures.

Finally, the average standardized score constructed from the five outcomes in the table is 0.424σ higher in the treatment than in the control group. This difference is highly significant (p-value=0.001), and it remains highly significant even accounting for the fact that in Tables 3, 4, and 5 we estimate the effect of the treatment on nine standardized scores (adjusted p-value=0.009). The Lasso does not select any control in the estimation of the treatment effect on this measure, so to ensure that this significant effect is not due to imbalances already existing at baseline, we reestimate it controlling for the five measures in the table at baseline. The estimated treatment effects is now equal to 0.365σ and is still very significant (p-value=0.009). Again, this effect remains significant after accounting for the fact we estimate the effect of the treatment on nine standardized scores (adjusted p-value=0.081).

Overall, we find that SFL significantly worsens teachers' and enumerators' perception of the classroom environment.

5.4 Effects on teachers and robustness checks

In our PAP, we had indicated that we would study the effect of the treatment on some teachers' outcomes, such as teachers' motivation for their job or their mental health. Those estimated effects are all non significant. They are not reported here, because teachers are not the main target of the SFL program, and to preserve space. Those supplementary results are available upon request.

As a robustness check, we reestimate all the regressions in Tables 3, 4, and 5 without controls. The results of that exercise can be found in Tables B1, B2, and B3. Results with and without controls are pretty similar. The effects on ineligible students' and classrooms' disruptiveness are more significant without controls, while the effects on ineligible students' integration in the class network are no longer significant without controls. In our PAP, we had indicated that as a further robustness check, we would recompute all the unadjusted p-values in Tables 3, 4, and 5 using randomization inference. Doing so does not change our main findings so the results of that exercise are not reported here but are available upon request.

5.5 Heterogeneous treatment effects

In our PAP we had indicated that to investigate treatment effect heterogeneity, we would reestimate the treatment effect in various subgroups of students (e.g.: boys and girls, students with a baseline disruptiveness below and above the median, etc.). However, when estimating the treatment effect in many subgroups, one may spuriously find heterogeneous treatment effects due to type 1 error if p-values are not adjusted for multiple testing, while one may fail to detect truly heterogeneous effects due to type 2 error if p-values are adjusted. Instead, we prefer to use one of the machine-learning based methods that has been proposed since we wrote our PAP.

Specifically, we use the method proposed by Chernozhukov et al. [2018]. Omitting a few technical details,[11] the method amounts to repeating the following steps, say 100 times:

1. Randomly split the sample into a training and a validation sample.
2. Train a machine-learning model to predict the outcome of the control-group training-sample observations, based on some baseline covariates. Then, train the same machine-learning model to predict the outcome of the treatment-group training-sample observations.
3. Use those two models to predict the treatment effect of the validation-sample observations, and divide the validation sample into, say, quartiles of the predicted treatment effect.
4. Regress the outcome of validation-sample observations on their predicted outcome without the treatment, their predicted treatment effect, and on indicators of predicted-treatment-effect quartiles interacted with the treatment. Let $\widehat{\theta}$ denote the difference between the coefficients of the fourth and first quartiles interacted with the treatment.

To estimate the amount of treatment effect heterogeneity along the covariates used in step 2, Chernozhukov et al. [2018] show that one can use med($\widehat{\theta}$), the median of $\widehat{\theta}$ across the 100 replications. To compute the p-value of this estimator, the authors show that one can use the median p-value of $\widehat{\theta}$ multiplied by two.

We use this method to investigate treatment effect heterogeneity along seven students' baseline characteristics: their gender; the social security score of the household they live in; their mother's education; their average Spanish and mathematics score; the average of their authority acceptance,

[11] For instance, in step 4 below, the treatment has to be demeaned, and the regression has to be weighted. See Chernozhukov et al. [2018] for a comprehensive description of the method.

attention and focus, activity levels, and disruptiveness first-grade TOCA scores; their school happiness score; the percentage of their classmates that nominate them as a friend. The machine-learning method we use is the elastic net, as this is the model that performs the best in the application in Chernozhukov et al. [2018]. Our elastic net regressions include the seven variables listed above, their square, and the 42 products between the variables.

In Table 6 below, we investigate treatment effect heterogeneity for our two main outcomes: teachers' endline disruptiveness ratings, and the average of students' endline Spanish and mathematics scores. Across the split-sample replications, the median difference between the treatment effect of eligible students predicted to be in the top and bottom quartiles of the treatment effect by the elastic net is equal to 0.459σ for teachers' ratings of disruptiveness, and 0.320σ for students' Spanish and mathematics scores. These differences are both insignificant: their p-values are respectively equal to 0.206 and 0.468. For ineligible students, these median differences are also both insignificant. Overall, we do not find any evidence of heterogeneous treatment effects, at least with respect to students' gender, family background, disruptiveness, academic ability, mental health, and popularity.

Table 6: Heterogeneous treatment effects

Variables	$me(\widehat{\theta})$ (1)	Unadj. P (2)
Panel A: eligible students		
Disruptiveness, teacher	0.459	0.206
Average Spanish and math test scores	0.320	0.468
Panel B: eligible students		
Disruptiveness, teacher	0.163	0.435
Average Spanish and math test scores	0.121	0.552

Notes: This table investigates treatment effect heterogeneity for teachers' disruptiveness ratings and for students' Spanish and mathematics tests scores, using a method proposed by Chernozhukov et al. [2018]. Column (1) reports the median of the difference between the treatment effect of students predicted to be in the highest and lowest quartiles of treatment effect according to elastic net regressions of the outcome on students' baseline characteristics, across 100 split-sample replications where the elastic net regressions are estimated on the first half of the sample while the treatment effect is estimated on the second half. Column (2) reports the p-value of that median. Students' baseline characteristics that are used to predict treatment effect heterogeneity are: their gender; the social security score of the household they live in; their mother's education; their average Spanish and mathematics score; the average of their authority acceptance, attention and focus, activity levels, and disruptiveness first-grade TOCA scores; their school happiness score; the percentage of their classmates that nominate them as a friend.

6 Interpretation, and exploratory analysis

6.1 Why SFL does not improve eligible students' outcomes?

Panels B and C of Table 3 show that SFL does not have any positive effect on eligible students' disruptiveness and academic ability. This is surprising, as an extensive literature has shown that programs similar to SFL usually produce fairly large positive effects on these dimensions (see the meta-analysis by Wilson and Lipsey [2007]). The difference between our results and those in the literature does not come from a lack of statistical precision in our study: based on our estimates, we can reject effects much smaller than the average effects found by Wilson and Lipsey [2007].

A first potential explanation for this discrepancy is that our study and those in Wilson and Lipsey [2007] may measure the treatment effect at different times after the end of the program. As Wilson and Lipsey [2007] do not report how long after the end of the program the endline measures were collected in the 108 studies they consider, we randomly selected and reviewed one third of those studies. The median study collected its endline measures two weeks after the end of the program. We collected our measures three weeks after the end of SFL, so our study is actually pretty similar to those in Wilson and Lipsey [2007] on that dimension.

A second potential explanation is that our study design may have created some spillovers between the treatment and control groups. For instance, eligible students in control group classes may be friends with eligible students in treatment group classes, and may then indirectly benefit from the treatment. Assuming that students only form friendships with students in the same school as them, such spillover effects can only happen in schools that have both a treated and a control class. To test the plausibility of this explanation, we therefore estimate the treatment effect in schools where only one class was included in our experiment: in those schools, spillover effects cannot affect control students as nobody is treated in their school. This subsample still has 114 classes, so this test is decently powered. In this subsample, teachers' rating of eligible students' disruptiveness is 0.2σ higher in the treatment than in the control group. This difference is not statistically significant (p-value=0.12), but we can rule out at the 5% level that SFL reduces eligible students' disruptiveness by more than 0.06σ. Results are similar when we consider other outcomes, such as students' test scores. Overall, spillover effects are unlikely to account for the lack of effect of SFL on eligible students.

A third potential explanation for this discrepancy is that all the studies in Wilson and Lipsey [2007] consider programs implemented in high-income countries, while the program we study is implemented in a middle-income country. However, the few studies of interventions similar to SFL in

low- and middle-income countries also find large effects, and we are not aware of a well-documented pattern whereby interventions effective in richer countries are less effective in poorer countries.

A final potential explanation is that most of the existing studies consider research or demonstration programs, implemented with significant researcher involvement, while we study a nationwide policy run by the government. Researchers in clinical psychology and economics have recently started documenting that interventions often produce smaller effects when they move from research to practice contexts and when they are implemented away from the developer's control, a phenomenon referred to as the "implementation cliff" (see [Weisz et al., 2014]). Various causes of this phenomenon have been suggested. First, participants in demonstration studies may be more motivated to receive the program than participants in routine practice studies (see [Muralidharan and Niehaus, 2017]). In our study, the motivation of children's parents indeed seems quite low, with around half of them not attending any of their sessions. Students' take up is also not perfect. Second, when scaling up a labor intensive intervention, it may not be possible to maintain its quality, because the skilled labor that implemented it at high quality in the demonstration study is scarce (see [Davis et al., 2017]). This mechanism may also be relevant here: SFL psychologists only hold a college degree, while program implementers in the studies reviewed by Wilson and Lipsey [2007] often have PhDs. In our context, the "implementation cliff" phenomenon may be aggravated by the fact that SFL is implemented without any researcher or NGO involvement. On the other hand, Banerjee et al. [2016] show that a small-scale program can be successfully scaled-up in a government school system, provided the NGO that designed it is strongly involved in the scale-up.

Overall, the "implementation cliff" may be a plausible explanation of why SFL produces much smaller effects than those found in the literature for similar programs. However, this claim remains speculative. SFL was not shown to be effective in a demonstration study with significant researcher involvement. It resembles other interventions that prove effective in demonstration studies, but it is not exactly identical to any of them, so it could be the case that SFL would also prove ineffective in a demonstration study.

6.2 More integrated or more segregated classrooms?

Panel D of Table 4 shows that SFL increases the integration of ineligible students in the class network, while Panel D of Table 3 suggests that SFL may also increase the integration of eligible students. In an exploratory analysis, we study whether SFL creates within- or between-groups friendship links. Do eligible (resp. ineligible) students become more friends with each other, or do

they become more friends with ineligible (resp. eligible) students? To answer those questions, we estimate the treatment effect on two student-level outcomes, the proportion of a student's eligible classmates that nominate her as a friend, and the proportion of a student's ineligible classmates that nominate her. We also estimate the treatment effect on a class-level measure of the segregation of friendships between eligible and ineligible students, the spectral homophily measure proposed by Golub and Jackson [2012]. That measure is included between 0 (no segregation) and 1 (full segregation). To estimate the treatment effect on the two student-level outcomes, we use the same estimation method as in Tables 3 and 4, while for the class-level outcome we use the same method as in Table 5. P-values are not adjusted for multiple testing because this analysis is exploratory.

We find that SFL only increases within-group links, and increases the segregation between eligible and ineligible students. Panel A of Table 7 shows that eligible students are nominated as friends by 9.6% of their eligible classmates in treated classes, against 7.2% in control classes. The difference is statistically significant (p-value=0.008), and represents a 33% increase with respect to the level of the variable in the control group. On the other hand, eligible students are nominated by 7.1% of their ineligible classmates in treated classes, against 7.0% in control classes, and the difference is insignificant. Conversely, Panel B of Table 7 shows that ineligible students are nominated as friends by 8.3% of their eligible classmates in treated classes, against 8.4% in control classes, and the difference is insignificant. On the other hand, ineligible students are nominated by 9.3% of their ineligible classmates in treated classes, against 8.6% in control classes. The difference is statistically significant (p-value=0.013), and represents a 8% increase with respect to the level of the variable in the control group.[12] Finally, the average spectral homophily is equal to 0.141 in treated classes, against 0.111 in control classes. The difference is significant (p-value=0.056), and represents a 27% increase with respect to the level of the variable in the control group.

It may not be surprising that SFL increases the connections between eligible students. During the SFL sessions, eligible students get to spend time together in a small-group setting, doing various games and non-academic activities. Moreover, the program explicitly aims at creating a group identity, for instance by having students choose a group name. On the other hand, it may be more surprising that SFL also increases the connections between ineligible students. The SFL sessions take place during regular class hours, while our definition of friendships is based on students' non-academic interactions during breaks. It could still be the case that more academic interactions take

[12]In Panels A and B, the number of observations is slightly lower for the "% eligibles friends with student" outcome than for the other outcome, because of a few classes where no or only one eligible student answered the friendship questions at endline.

Table 7: Exploratory analysis of the effect of SFL on friendship links

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	N (5)
Panel A: eligible students					
% eligibles friends with student	0.072	0.024	0.009	0.008	1145
% ineligible friends with student	0.07	0.001	0.006	0.824	1147
Panel B: ineligible students					
% eligibles friends with student	0.085	-0.001	0.004	0.798	4098
% ineligible friends with student	0.087	0.007	0.003	0.013	4168
Panel C: class level					
Spectral homophily	0.111	0.03	0.016	0.057	165

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator and control variables. The control variables are selected by a Lasso regression of the dependent variable on all potential controls, following the methodology proposed by Belloni et al. [2014]. For student-level dependent variables, the regression includes lottery fixed effects. For class-level dependent variables, the regression is computed with propensity score weights. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient. Finally, Column (5) reports the number of observations used in the regression. All the dependent variables were collected by the authors at endline.

place between ineligible students during the sessions, which may in turn lead to more non-academic interactions during breaks. It could also be the case that some non-academic interactions take place between ineligible students during the sessions, despite the fact those take place during regular class hours. Finally, the mere fact of dividing the class into groups may lead students to develop more in-group connections. A large social-psychology literature shows that assigning group labels to individuals leads them to favour in-group members, even if those labels are arbitrary, the so-called minimal group effect (see [Tajfel et al., 1971]). Bigler et al. [2001] document this effect in an elementary-school setting, with groups sharing some features with those created by the SFL program.[13]

6.3 Why does SFL worsen the studying conditions in treated classes?

Table 5 shows that SFL worsens teachers' and enumerators' perceived disruptiveness of the treated classes. This effect is strongly significant, even after accounting for multiple testing. Panel B of Table 3 suggests that this increase does not come from an increase in the disruptiveness of eligible students. On the other hand, Panel B of Table 4 suggests this increase may come from an increase

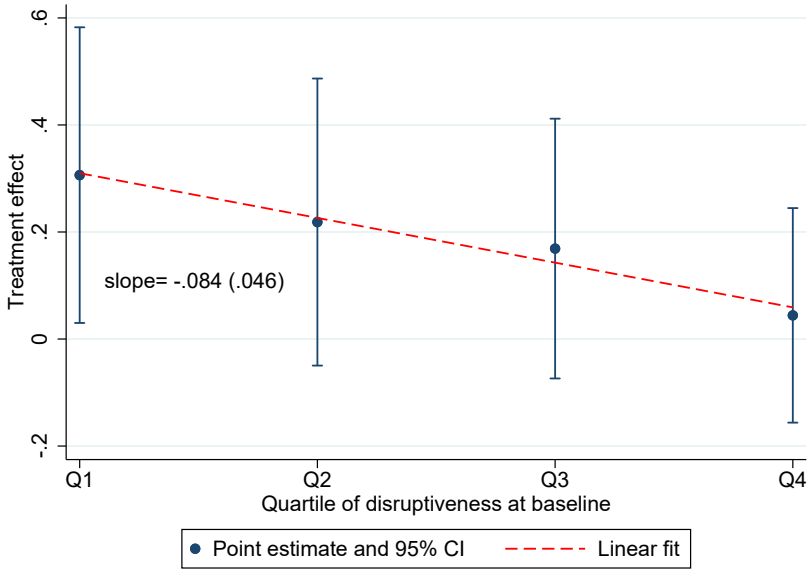
[13]Bigler et al. [2001] randomly allocate blue and yellow t-shirts to children, implicitly assign a low/high social status to one of the two colors, and organize some color-based group activities during classes. They find that after a few weeks students tend to have more positive opinions of students in their group.

in the disruptiveness of ineligible students, even though this effect is no longer significant after accounting for multiple testing. In this section, we explore various mechanisms that could explain this increase in ineligible students' disruptiveness, if one is ready to believe that this effect is real.

First, this effect could come from an increase in friendship ties between eligible and ineligible students, which would then lead ineligible students to be more disruptive to imitate their friends' behaviour. However, in the previous section we did not find any evidence that SFL increases friendships between eligible and ineligible students, so this mechanism is unlikely to be at play.

Second, when they leave the classroom to attend the SFL sessions, eligible students may create a "disruptiveness vacuum" that gets filled by ineligible students, and idea that can be rationalized by the following toy model of disruption. Assume that a class bears three students, that respectively derive utilities λ_H, λ_M, and λ_L from being disruptive. If the proportion of students that are being disruptive exceeds a threshold d^*, the teacher punishes the students, in which case all students experience a disutility γ. For each $i \in \{H, M, L\}$, let d_i be an indicator equal to 1 if student i chooses to be disruptive. Let SFL be an indicator for whether student H is out of the class for an SFL group session. Students simultaneously choose to be disruptive, and student i's payoff is $\lambda_i d_i - \gamma 1\{(d_H(1 - SFL) + d_M + d_L)/(3 - SFL) > d^*\}$. If $\lambda_L \leq 0 < \lambda_M < \gamma < \lambda_H$ and $d^* \in (1/2, 2/3)$, when $SFL = 0$ only student H chooses to be disruptive, while when $SFL = 1$ only student M chooses to be disruptive. However, this second potential mechanism is again not supported by the data. A prediction of the toy model is that the program should increase more the disruptiveness of the most disruptive ineligible students. If anything, we actually find the opposite. Figure 1 below shows that the effect of SFL on ineligible students' disruptiveness is higher for students in the bottom quartiles of disruptiveness at baseline. The differences are not significant, but the slope of the line connecting the four dots is negative and marginally significant.

Figure 1: Treatment effect on ineligibles' disruptiveness, by quartiles of baseline disruptiveness



Notes: The figure plots the coefficients $\widehat{\beta}$ in Regression (2), estimated separately for ineligible students in each quartile of disruptiveness at baseline, with teachers' ratings of students' disruptiveness as the dependent variable. The controls used in each regression are those picked by a Lasso regression of ineligible students' disruptiveness on all potential controls, using the entire sample and following the methodology proposed by Belloni et al. [2014]. The figure also shows the 95% confidence interval attached to each coefficient $\widehat{\beta}$, using standard errors clustered at the lottery level. Finally, the figure shows the line arising from an OLS regression of the estimated treatment effect in each quartile on the quartile numbers attached to it. The standard error of the regression's slope is computed by bootstrapping 200 times 56 lottery groups, estimating the treatment effects for each quartile of disruptiveness in the bootstrapped sample, computing an OLS regression of the estimated treatment effects on the quartiles, and computing the standard error of these 200 coefficients.

Third, we interviewed some of the SFL psychologists, and they told us that ineligible students experience their exclusion from the program as a punishment and ask them to be included. From their perspective, the program often appears as a reward given to disruptive students: they get to leave the classroom to play games during the SFL sessions, and they may come back with cakes or candies if they have behaved well during the session. This could lead ineligible students to increase their disruptiveness for at least two reasons. First, SFL may give them an incentive to be disruptive, if they believe that they will get included in the program if they become as disruptive as the eligible students. This hypothesis is consistent with the pattern in Figure 1: the least disruptive students are those who need to increase their disruptiveness the most to reach disruptiveness levels comparable to that of eligible students. Second, they may consider it unfair that disruptive students get rewarded by being included in the program, and they may increase their disruptiveness levels to

protest against that injustice. This hypothesis is again consistent with the pattern in Figure 1: the least disruptive students may be those who find it the most unfair that eligible students get rewarded while they do not. These two mechanisms appear to us as plausible explanations of the increase in ineligible students' disruptiveness, though we can certainly not rule out other explanations.

7 Conclusion

We explore the effects of a nationwide school-based CBT program for disruptive second graders in Chile. Eligibility to the program is based on first-grade teachers' ratings of students' disruptiveness, and the program consists in 10 two-hours sessions during which psychologists teach students CBT techniques to help them improve their behavior. We randomly assigned 172 classes to either receive the treatment in the first or in the second semester of the 2015 school year, and we measured outcomes between the two semesters.

Eligible students in treated classes see no improvement in their level of disruptiveness or test scores, when compared to eligible students in control classes. A large literature has found that school-based mental health programs for disruptive students can be very successful, but the literature has mostly considered small-scale programs (see [Wilson and Lipsey, 2007]). Our results suggest that school-based CBT programs may not be as successful when implemented as nationwide policies, at least in the context we study. We hypothesize that this finding arises from the fact that, as documented in earlier work, programs often become less effective when they move from research to practice (see [Weisz et al., 2014] and [Banerjee et al., 2016]). Unfortunately, we cannot disentangle which part of the scaling up process should be improved to produce effects similar to those found in small-scale studies. Our results call for further research in that area.

We also find that SFL worsens teachers' and enumerators' perceived disruptiveness of the treated classes. This negative effect seems to come from an increase of the disruptiveness of ineligible students in the treatment group. We conjecture that this increase may be due to the fact that ineligible students perceive the treatment as a reward: eligible students get to skip classes during the sessions, and the sessions mostly consist in games and role play. Ineligible students may then increase their disruptiveness to be included in the program, or because they find it unfair that disruptive students get rewarded and not them. Should this conjecture be correct, one could then mitigate the negative unintended consequences of the program by making it less salient to ineligible students, for instance by conducting the sessions after the school day.

References

Anna Aizer. Peer effects and human capital accumulation: The externalities of add. Technical report, National Bureau of Economic Research, 2008.

Michael L Anderson. Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American statistical Association*, 103(484):1481–1495, 2008.

Abhijit Banerjee, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton. Mainstreaming an effective intervention: Evidence from randomized evaluations of "teaching at the right level" in india. Technical report, National Bureau of Economic Research, 2016.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50, 2014.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

Julian R Betts and Jamie L Shkolnik. The behavioral effects of variations in class size: The case of math teachers. *Educational Evaluation and Policy Analysis*, 21(2):193–213, 1999.

Rebecca S Bigler, Christia Spears Brown, and Marc Markell. When groups are not created equal: Effects of group status on the formation of intergroup attitudes in children. *Child Development*, 72(4):1151–1162, 2001.

Sue C Bratton. *Effectiveness of group activity play therapy on internalizing and externalizing behavior problems of preadolescent orphans in Uganda*. PhD thesis, University of North Texas, 2011.

Nancy D Brener, Jim Martindale, and Mark D Weist. Mental health and social services: Results from the school health policies and programs study 2000. *Journal of School Health*, 71(7): 305–312, 2001.

Matthew D Burkey, Megan Hosein, Isabella Morton, Marianna Purgato, Ahmad Adi, Mark Kurzrok, Brandon A Kohrt, and Wietse A Tol. Psychosocial interventions for disruptive behaviour problems in children in low-and middle-income countries: a systematic review and meta-analysis. *Journal of Child Psychology and Psychiatry*, 2018.

Scott E Carrell and Susan A Carrell. Do lower student to counselor ratios reduce school disciplinary problems? *The BE Journal of Economic Analysis & Policy*, 5(1), 2006.

Scott E Carrell and Mark Hoekstra. Are school counselors an effective education input? *Economics Letters*, 125(1):66–69, 2014.

Scott E Carrell and Mark L Hoekstra. Externalities in the classroom: How children exposed to domestic violence affect everyone's kids. *American Economic Journal: Applied Economics*, 2(1): 211–28, 2010.

Scott E Carrell, Mark Hoekstra, and Elira Kuka. The long-run effects of disruptive peers. *American Economic Review*, 108(11):3377–3415, 2018.

Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. Generic machine learning inference on heterogenous treatment effects in randomized experiments. Technical report, National Bureau of Economic Research, 2018.

Nicola M Curtis, Kevin R Ronan, and Charles M Borduin. Multisystemic treatment: a meta-analysis of outcome studies. *Journal of family psychology*, 18(3):411, 2004.

Jonathan Davis, Jonathan Guryan, Kelly Hallberg, and Jens Ludwig. The economics of scale-up. Technical report, National Bureau of Economic Research, 2017.

Luciana Carla Dos Santos Ellas, Edna Maria Marturano, Ana Maria De Almeida Motta, and Alessandra Gaspar Giurlani. Treating boys with low school achievement and behavior problems: Comparison of two kinds of intervention. *Psychological Reports*, 92(1):105–116, 2003.

David N Figlio. Boys named sue: Disruptive children and their peers. *Education finance and policy*, 2(4):376–394, 2007.

A Finkelstein, S Taubman, H Allen, J Gruber, JP Newhouse, B Wright, and K Baicker. The short-run impact of extending public health insurance to low-income adults: Evidence from the first year of the oregon medicaid experiment. *Analysis plan*, 2010.

Benjamin Golub and Matthew O Jackson. How homophily affects the speed of learning and best-response dynamics. *The Quarterly Journal of Economics*, 127(3):1287–1338, 2012.

Javier Guzmán, Ronald C Kessler, Ana Maria Squicciarini, Myriam George, Lee Baer, Katia M Canenguez, Madelaine R Abel, Alyssa McCarthy, Michael S Jellinek, and J Michael Murphy.

Evidence for the effectiveness of a national school-based mental health program in chile. *Journal of the American Academy of Child & Adolescent Psychiatry*, 54(10):799–807, 2015.

Susan Harter. *Manual for the self-perception profile for children:(revision of the perceived competence scale for children)*. University of Denver, 1985.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Michael S Jellinek, J Michael Murphy, John Robinson, Anita Feins, Sharon Lamb, and Terrence Fenton. Pediatric symptom checklist: screening school-age children for psychosocial dysfunction. *The Journal of pediatrics*, 112(2):201–209, 1988.

Sheppard G Kellam, Margaret E Ensminger, and R Jay Turner. Family structure and the mental health of children: Concurrent and longitudinal community-wide studies. *Archives of General Psychiatry*, 34(9):1012–1022, 1977.

Edward P Lazear. Educational production. *The Quarterly Journal of Economics*, 116(3):777–803, 2001.

María Fernanda Molina, María Julia Raimundi, Carolina López, Silvana Cataldi, and Lucia Bugallo. Adaptación del perfil de autopercepciones para niños para su uso en la ciudad de buenos aires. *Revista Iberoamericana de Diagnóstico y Evaluación-e Avaliação Psicológica*, 2(32), 2011.

Karthik Muralidharan and Paul Niehaus. Experimentation at scale. *Journal of Economic Perspectives*, 31(4):103–24, 2017.

Cynthia A Rorhbeck, Sandra T Azar, and Patricia E Wagner. Child self-control rating scale: Validation of a child self-report measure. *Journal of Clinical Child and Adolescent Psychology*, 20(2):179–183, 1991.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Irwin Sandler, Sharlene A Wolchik, Gracelyn Cruden, Nicole E Mahrer, Soyeon Ahn, Ahnalee Brincks, and C Hendricks Brown. Overview of meta-analyses of the prevention of mental health, substance use, and conduct problems. *Annual review of clinical psychology*, 10:243–273, 2014.

Henri Tajfel, Michael G Billig, Robert P Bundy, and Claude Flament. Social categorization and intergroup behaviour. *European journal of social psychology*, 1(2):149–178, 1971.

John R Weisz, Mei Yi Ng, and Sarah Kate Bearman. Odd couple? reenvisioning the relation between science and practice in the dissemination-implementation era. *Clinical Psychological Science*, 2(1):58–74, 2014.

L Werthamer-Larsson, SG Kellam, and KE Ovesen-McGregor. Teacher interview: Teacher observation of classroom adaptation—revised (toca-r). *Johns Hopkins Prevention Center training manual. Baltimore, MD: Johns Hopkins University*, 1990.

Sandra Jo Wilson and Mark W Lipsey. School-based interventions for aggressive and disruptive behavior: Update of a meta-analysis. *American journal of preventive medicine*, 33(2):S130–S143, 2007.

For Online Publication

Appendix A Tables

Table A1: Baseline - endline correlations in the control group

	Correlation (1)	P-value (2)	N (3)
Panel A: student-level measures			
School happiness score	0.221	0.000	1753
Self-control score	0.143	0.000	1836
Self-esteem score	0.135	0.000	1862
Disruptiveness, teacher	0.419	0.000	1785
Disruptiveness, enumerator	0.031	0.171	1895
% school days missed	0.034	0.078	2756
Spanish test score	0.522	0.000	1916
Math test score	0.509	0.000	1916
% class friends with student	0.323	0.000	2271
Friends' average ability	0.409	0.000	1662
Friends' average disruptiveness	0.343	0.000	1518
No friends in the class	0.096	0.000	2271
Panel B: class-level measures			
Disruptiveness, teacher	0.5	0.000	78
Bullying in class, teacher	0.392	0.000	76
Disruptiveness, enumerator	0.254	0.024	79
Average decibels during class	0.152	0.18	79
Delay in class's start (minutes)	0.031	0.788	79

Notes: This table reports the correlation, in control classes, of several covariates between baseline and endline. Column (1) reports the baseline - endline correlation of the covariates. Column (2) reports the p-value of the significance of the correlation. Column (3) reports the number of observations used to compute the correlation.

Table A2: Correlations between baseline disruptiveness measures

	Correlation (1)	P-value (2)	N (3)
Panel A: student-level measures			
Enumerator 1 - enumerator 2	0.545	0.000	4482
Teacher - enumerator	0.28	0.000	4041
Teacher dis. - avg. test score	-0.277	0.000	4144
Enumerator dis. - avg. test score	-0.153	0.000	4705
Panel B: class-level measures			
Enumerator 1 - Enumerator 2	0.618	0.000	157
Enumerator - Teacher	0.337	0.000	159
Enumerator - decibels	0.2	0.011	163
Teacher - decibels	-0.018	0.82	157

Notes: This table reports the correlation, in control classes, between several baseline measures of disruption. Column (1) reports the correlation between the measures. Column (2) reports the p-value of the significance of the correlation. Column (3) reports the number of observations used to compute the correlation.

Table A3: Characteristics of takers and non-takers

	Non-takers (1)	Takers (2)	P-value (3)	N (4)
Panel A: demographic characteristics				
Male	0.667	0.567	0.05	655
Teen mother	0.415	0.368	0.43	525
Student lives with father	0.515	0.551	0.577	478
$\leq$ p20 social security score	0.842	0.741	0.016	596
$\leq$ p5 social security score	0.463	0.441	0.693	596
Mother's education	8.448	8.327	0.798	576
Father's education	8.014	8.198	0.727	485
Panel B: baseline measures				
School happiness score	0.08	-0.034	0.41	477
Self-control score	-0.27	-0.172	0.493	511
Self-esteem score	-0.233	-0.176	0.708	513
Overall disruptiveness TOCA	1.128	0.81	0.011	645
Disruptiveness, enumerator	0.723	0.406	0.072	517
Spanish test score	-0.496	-0.326	0.22	548
Math test score	-0.489	-0.248	0.085	548
% class friends with student	0.069	0.079	0.168	539
Friends' average disruptiveness	0.324	0.241	0.604	422

Notes: This table reports descriptive statistics for eligible students, comparing those who attended and did not attend the workshops. Column (1) reports the mean of the outcome variable for eligible students who did not attend any session. Column (2) reports the mean of the variable for eligible students who attended at least one session. Column (3) reports the p-value of a test that the two means are equal. Column (4) reports the number of observations used in the comparison.

Table A4: Characteristics of teachers

	Mean (1)	N (2)
Female	0.963	160
Age	42.78	159
University degree	0.863	160
Years of experience	16.547	161
Years of experience, school	8.568	162

Notes: This table reports descriptive statistics for teachers in the sample. Column (1) reports the mean of the variables and Column (2) reports the number of observations used to compute that mean.

Table A5: Test of differential attrition for eligible students

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Eligible students per class at endline	6.651	0.473	0.386	0.22	0.55	169
Join class btw baseline and endline	0.023	0.004	0.008	0.649	0.649	1229
In class at baseline and endline	0.941	0.024	0.014	0.078	0.311	1178
With all enumerators' measures	0.748	-0.035	0.03	0.247	0.329	1238
With teacher's disruption measure	0.768	-0.084	0.071	0.235	0.47	1238

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator. For student-level dependent variables, the regression includes lottery fixed effects. For class-level dependent variables, the regression is computed with propensity score weights. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression. All the dependent variables were collected by the authors at endline.

Table A6: Test of differential attrition for ineligible students

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Ineligible students per class at endline	25.518	-1.009	0.853	0.237	0.592	169
Join class btw baseline and endline	0.045	-0.005	0.008	0.553	0.737	4433
In class at baseline and endline	0.962	-0.001	0.007	0.842	0.842	4159
With all enumerators' measures	0.783	-0.048	0.027	0.074	0.297	4466
With teacher's disruption measure	0.753	-0.059	0.067	0.383	0.766	4466

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator. For student-level dependent variables, the regression includes lottery fixed effects. For class-level dependent variables, the regression is computed with propensity score weights. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression. All the dependent variables were collected by the authors at endline.

Table A7: Balancing tests of eligible students' baseline characteristics

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: demographic characteristics						
Male	0.581	-0.004	0.046	0.937	0.97	1238
Teen mother	0.343	0.018	0.031	0.549	0.885	991
Student lives with father	0.563	-0.012	0.034	0.726	1	899
Social security score	5564.943	137.239	173.203	0.428	0.828	1124
Payment rate in health services	2.879	0.327	0.361	0.365	0.963	1122
Mother's education	8.813	-0.292	0.32	0.362	1	1080
Father's education	8.743	-0.565	0.38	0.137	0.995	913
Panel B: TOCA and PSC scores						
Authority Acceptance TOCA	1.027	-0.084	0.063	0.181	0.751	1223
Social Contact TOCA	0.842	-0.025	0.072	0.723	1	1223
Motiv. for Schooling TOCA	0.842	-0.036	0.06	0.543	0.985	1223
Emotional Maturity TOCA	0.563	-0.12	0.076	0.117	1	1223
Atention and Focus TOCA	0.834	-0.054	0.063	0.391	0.873	1223
Activity Level TOCA	0.831	-0.054	0.064	0.404	0.837	1223
Academic ability TOCA	0.667	-0.016	0.071	0.82	0.951	1222
Overall disruptiveness TOCA	0.891	-0.046	0.076	0.548	0.935	1220
PSC	0.477	-0.011	0.08	0.889	0.955	903
Panel C: baseline measures						
School happiness score	-0.107	0.082	0.083	0.323	1	929
Self-control score	-0.148	-0.057	0.063	0.371	0.897	986
Self-esteem score	-0.107	-0.105	0.076	0.168	0.811	991
Disruptiveness, teacher	0.396	0.087	0.276	0.753	0.993	253
Disruptiveness, enumerator	0.192	0.205	0.112	0.068	0.993	1007
Spanish test score	-0.321	-0.021	0.086	0.806	0.973	1036
Math test score	-0.301	0.021	0.099	0.829	0.924	1036
% class friends with student	0.075	0.002	0.006	0.769	0.97	1030
Friends' average ability	-0.09	-0.002	0.114	0.988	0.988	863
Friends' average disruptiveness	0.122	0.099	0.103	0.333	1	822
No friends in the class	0.128	0.047	0.026	0.065	1	1030
Distance to teacher	4.361	-0.079	0.18	0.66	1	863
% school days missed, March	36.971	-4.809	3.421	0.16	0.927	1236

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator and lottery fixed effects for eligible students. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression.

Table A8: Balancing tests of eligible students' baseline characteristics, for those with all enumerators' endline measures.

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: demographic characteristics						
Male	0.56	0.016	0.054	0.767	1	906
Teen mother	0.324	0.081	0.04	0.044	1	731
Student lives with father	0.58	-0.051	0.038	0.183	0.883	665
Social security score	5640.612	-62.531	227.803	0.784	1	819
Payment rate in health services	3.005	0.122	0.472	0.795	1	824
Mother's education	8.836	-0.218	0.404	0.589	1	794
Father's education	8.768	-0.197	0.396	0.619	1	667
Panel B: TOCA and PSC scores						
Authority Acceptance TOCA	1	-0.038	0.063	0.548	1	894
Social Contact TOCA	0.785	0.008	0.077	0.919	0.987	894
Motiv. for Schooling TOCA	0.809	-0.009	0.065	0.893	1	894
Emotional Maturity TOCA	0.591	-0.128	0.083	0.123	0.895	894
Atention and Focus TOCA	0.798	0.013	0.064	0.845	1	894
Activity Level TOCA	0.821	-0.026	0.07	0.713	1	894
Academic ability TOCA	0.626	-0.014	0.079	0.859	1	894
Overall disruptiveness TOCA	0.801	0.034	0.092	0.712	1	893
PSC	0.441	-0.005	0.089	0.957	0.991	669
Panel C: baseline measures						
School happiness score	-0.077	0.069	0.091	0.445	1	700
Self-control score	-0.136	-0.018	0.079	0.824	1	745
Self-esteem score	-0.13	-0.043	0.093	0.643	1	744
Disruptiveness, teacher	0.341	0.061	0.215	0.776	1	192
Disruptiveness, enumerator	0.129	0.219	0.124	0.077	0.744	742
Spanish test score	-0.264	-0.01	0.084	0.908	1	769
Math test score	-0.22	0.037	0.11	0.736	1	769
% class friends with student	0.077	0.006	0.006	0.353	1	765
Friends' average ability	-0.071	0.000	0.126	0.997	0.997	656
Friends' average disruptiveness	0.094	0.162	0.118	0.17	0.987	623
No friends in the class	0.111	0.048	0.026	0.068	0.985	765
Distance to teacher	4.377	-0.203	0.225	0.366	1	630
% school days missed, March	37.887	-4.312	3.658	0.238	0.988	904

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator and lottery fixed effects for eligible students with all enumerators' endline measures. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression.

Table A9: Balancing tests of eligible students' baseline characteristics, for those with teacher's endline disruptiveness measure.

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: demographic characteristics						
Male	0.574	-0.01	0.053	0.848	0.984	901
Teen mother	0.337	0.033	0.038	0.394	1	724
Student lives with father	0.564	-0.006	0.045	0.89	0.922	659
Social security score	5533.873	205.674	236.641	0.385	1	814
Payment rate in health services	3.144	-0.045	0.506	0.929	0.929	816
Mother's education	8.897	-0.594	0.415	0.152	0.883	798
Father's education	8.771	-0.483	0.511	0.345	1	673
Panel B: TOCA and PSC scores						
Authority Acceptance TOCA	0.983	-0.12	0.08	0.136	0.983	889
Social Contact TOCA	0.829	0.041	0.096	0.666	1	889
Motiv. for Schooling TOCA	0.852	-0.018	0.081	0.821	0.992	889
Emotional Maturity TOCA	0.597	-0.123	0.1	0.219	1	889
Atention and Focus TOCA	0.842	-0.046	0.082	0.572	0.922	889
Activity Level TOCA	0.821	-0.124	0.081	0.124	1	889
Academic ability TOCA	0.676	-0.052	0.091	0.563	0.961	888
Overall disruptiveness TOCA	0.877	-0.069	0.099	0.482	1	887
PSC	0.434	-0.017	0.103	0.869	0.933	662
Panel C: baseline measures						
School happiness score	-0.064	-0.064	0.096	0.503	0.972	680
Self-control score	-0.128	-0.165	0.085	0.053	0.762	718
Self-esteem score	-0.078	-0.106	0.088	0.23	0.952	720
Disruptiveness, teacher	0.275	0.057	0.245	0.815	1	190
Disruptiveness, enumerator	0.167	0.099	0.151	0.513	0.93	743
Spanish test score	-0.34	0.03	0.088	0.736	1	758
Math test score	-0.28	0.036	0.133	0.786	1	758
% class friends with student	0.075	0.006	0.008	0.451	1	751
Friends' average ability	-0.138	0.129	0.143	0.367	1	635
Friends' average disruptiveness	0.129	0.026	0.14	0.853	0.951	611
No friends in the class	0.102	0.088	0.035	0.011	0.31	751
Distance to teacher	4.441	0.061	0.178	0.732	1	643
% school days missed, March	37.204	-2.795	4.143	0.5	1	899

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator and lottery fixed effects for eligible students with teacher's endline disruptiveness measure. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression.

Table A10: Balancing tests of ineligible students' baseline characteristics.

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: demographic characteristics						
Male	0.486	0.026	0.027	0.327	0.678	4466
Teen mother	0.328	0.016	0.02	0.434	0.662	3449
Student lives with father	0.639	-0.012	0.017	0.501	0.727	2866
Social security score	5965.036	-108.938	107.006	0.309	0.746	3944
Payment rate in health services	4.132	-0.019	0.313	0.951	0.951	3927
Mother's education	9.239	-0.19	0.2	0.341	0.582	3647
Father's education	9.181	-0.017	0.177	0.925	0.958	3204
Panel B: TOCA and PSC scores						
Authority Acceptance TOCA	-0.356	0.059	0.054	0.278	1	3654
Social Contact TOCA	-0.346	0.14	0.055	0.01	0.297	3654
Motiv. for Schooling TOCA	-0.312	0.071	0.047	0.132	0.765	3654
Emotional Maturity TOCA	-0.171	0.024	0.092	0.795	0.922	3654
Atention and Focus TOCA	-0.32	0.092	0.053	0.086	0.624	3654
Activity Level TOCA	-0.33	0.124	0.066	0.059	0.86	3645
Academic ability TOCA	-0.244	0.043	0.041	0.292	0.847	3633
Overall disruptiveness TOCA	-0.335	0.075	0.041	0.068	0.66	3630
PSC	-0.171	0.043	0.044	0.333	0.603	2882
Panel C: baseline measures						
School happiness score	0.039	-0.015	0.039	0.697	0.879	3502
Self-control score	0.05	-0.005	0.045	0.917	0.985	3608
Self-esteem score	0.066	-0.051	0.043	0.234	0.971	3619
Disruptiveness, teacher	-0.132	0.052	0.181	0.772	0.933	804
Disruptiveness, enumerator	-0.151	0.111	0.092	0.23	1	3639
Spanish test score	0.139	-0.065	0.076	0.393	0.632	3722
Math test score	0.083	0.033	0.079	0.676	0.891	3722
% class friends with student	0.09	-0.003	0.005	0.523	0.722	3691
Friends' average ability	0.055	0.017	0.099	0.86	0.959	3260
Friends' average disruptiveness	-0.094	0.075	0.073	0.305	0.804	3109
No friends in the class	0.097	0.02	0.02	0.328	0.635	3691
Distance to teacher	4.519	0.168	0.158	0.286	0.923	3129
% school days missed, March	38.922	-2.992	2.969	0.314	0.699	4427

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator and lottery fixed effects for ineligible students. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression.

Table A11: Balancing tests of ineligible students' baseline characteristics, for those with all enumerators' endline measures.

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: demographic characteristics						
Male	0.473	0.038	0.027	0.154	0.64	3376
Teen mother	0.322	0.015	0.021	0.481	0.734	2646
Student lives with father	0.647	-0.008	0.021	0.702	0.783	2203
Social security score	5982.408	-99.568	119.473	0.405	0.838	2989
Payment rate in health services	4.305	-0.181	0.376	0.63	0.795	2974
Mother's education	9.239	-0.184	0.223	0.409	0.791	2788
Father's education	9.189	0.022	0.19	0.908	0.941	2454
Panel B: TOCA and PSC scores						
Authority Acceptance TOCA	-0.365	0.054	0.05	0.282	0.745	2768
Social Contact TOCA	-0.39	0.173	0.061	0.005	0.138	2768
Motiv. for Schooling TOCA	-0.351	0.074	0.05	0.137	0.661	2768
Emotional Maturity TOCA	-0.182	0.079	0.103	0.44	0.751	2768
Atention and Focus TOCA	-0.346	0.095	0.052	0.069	0.501	2768
Activity Level TOCA	-0.331	0.164	0.061	0.007	0.108	2762
Academic ability TOCA	-0.28	0.05	0.045	0.264	0.766	2759
Overall disruptiveness TOCA	-0.363	0.075	0.038	0.045	0.436	2756
PSC	-0.195	0.047	0.058	0.417	0.756	2210
Panel C: baseline measures						
School happiness score	0.045	-0.018	0.045	0.688	0.798	2715
Self-control score	0.07	-0.021	0.05	0.673	0.813	2789
Self-esteem score	0.102	-0.081	0.051	0.112	0.651	2797
Disruptiveness, teacher	-0.208	0.101	0.167	0.545	0.752	641
Disruptiveness, enumerator	-0.152	0.099	0.095	0.294	0.71	2806
Spanish test score	0.171	-0.067	0.071	0.347	0.774	2870
Math test score	0.106	0.042	0.08	0.598	0.789	2870
% class friends with student	0.09	0.000	0.006	0.95	0.95	2852
Friends' average ability	0.073	0.012	0.099	0.904	0.971	2524
Friends' average disruptiveness	-0.095	0.101	0.075	0.176	0.636	2402
No friends in the class	0.098	0.014	0.022	0.515	0.746	2852
Distance to teacher	4.522	0.124	0.163	0.446	0.718	2416
% school days missed, March	38.416	-3.897	3.252	0.231	0.744	3353

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator and lottery fixed effects for ineligible students with all enumerators' endline measures. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression.

Table A12: Balancing tests of ineligible students' baseline characteristics, for those with teacher's endline disruptiveness measure.

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: demographic characteristics						
Male	0.486	0.061	0.03	0.043	0.248	3202
Teen mother	0.319	0.04	0.025	0.118	0.381	2490
Student lives with father	0.641	0.012	0.023	0.61	0.804	2071
Social security score	5966.787	18.269	149.837	0.903	1	2838
Payment rate in health services	4.271	-0.156	0.42	0.71	0.823	2826
Mother's education	9.281	-0.293	0.281	0.296	0.537	2637
Father's education	9.276	-0.151	0.272	0.579	0.8	2310
Panel B: TOCA and PSC scores						
Authority Acceptance TOCA	-0.347	0.056	0.067	0.405	0.652	2645
Social Contact TOCA	-0.378	0.24	0.075	0.001	0.041	2645
Motiv. for Schooling TOCA	-0.323	0.122	0.055	0.028	0.267	2645
Emotional Maturity TOCA	-0.136	0.012	0.116	0.915	0.948	2645
Atention and Focus TOCA	-0.329	0.121	0.055	0.027	0.393	2645
Activity Level TOCA	-0.308	0.082	0.074	0.27	0.603	2637
Academic ability TOCA	-0.245	0.06	0.049	0.222	0.585	2632
Overall disruptiveness TOCA	-0.328	0.104	0.048	0.032	0.229	2630
PSC	-0.172	0.075	0.06	0.212	0.614	2084
Panel C: baseline measures						
School happiness score	0.047	-0.061	0.05	0.227	0.548	2531
Self-control score	0.106	-0.117	0.058	0.046	0.22	2592
Self-esteem score	0.09	-0.107	0.063	0.089	0.367	2604
Disruptiveness, teacher	-0.268	0.285	0.172	0.097	0.353	634
Disruptiveness, enumerator	-0.175	0.122	0.122	0.316	0.54	2638
Spanish test score	0.118	0.009	0.078	0.906	0.973	2689
Math test score	0.094	0.059	0.101	0.56	0.813	2689
% class friends with student	0.091	0.000	0.005	0.937	0.937	2659
Friends' average ability	0.045	0.058	0.123	0.635	0.767	2366
Friends' average disruptiveness	-0.073	0.096	0.091	0.289	0.558	2259
No friends in the class	0.088	0.023	0.021	0.277	0.575	2659
Distance to teacher	4.565	0.158	0.226	0.483	0.738	2355
% school days missed, March	39.314	-1.844	3.663	0.615	0.775	3178

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator and lottery fixed effects for ineligible students with teacher's endline disruptiveness measure. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression.

Table A13: Balancing tests of teachers' baseline characteristics

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: demographic characteristics						
Age	43.013	-0.256	1.763	0.885	0.965	159
University degree	0.872	-0.019	0.06	0.748	1	160
Years of experience	16.367	0.508	2.108	0.809	1	161
Years of experience in the school	8.139	0.729	1.331	0.584	1	162
Absenteeism	0.646	-0.101	0.547	0.853	1	162
Panel B: motivation and taste for their job						
Taste for her job	0.007	0.031	0.144	0.827	1	161
Confident to improve students' life	0.076	-0.146	0.172	0.395	1	161
Effort to prepare lectures	0.497	0.023	0.042	0.588	1	143
Diverse methods used in class	-0.005	0.016	0.161	0.919	0.919	161
Panel C: mental health						
Stress score	0.073	-0.138	0.156	0.377	1	160
Happiness score	0.148	-0.317	0.15	0.034	0.41	161
Control on life score	0.054	-0.115	0.151	0.447	1	158

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator for teachers. The regression is estimated with propensity score weights. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression. All the dependent variables were collected by the authors at baseline.

Table A14: Balancing tests of classes' baseline characteristics

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Academic level of the class, teacher	0.059	-0.086	0.14	0.538	0.538	162
Disruptiveness, teacher	-0.143	0.286	0.16	0.074	0.148	161
Bullying in class, teacher	0.033	-0.094	0.147	0.519	0.623	160
Disruptiveness, enumerator	-0.131	0.275	0.153	0.072	0.217	168
Delay in class's start (minutes)	8.802	1.122	1.253	0.37	0.555	166
Average decibels during class	0.053	1.796	0.745	0.016	0.095	165

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator. The regression is estimated with propensity score weights. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression. All the dependent variables were collected by the authors at baseline.

Table A15: Balancing tests of classes' baseline characteristics, for classes with all teacher's or enumerators' endline measures.

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: classes with all teacher's measures						
Academic level of the class, teacher	0.052	-0.095	0.143	0.509	0.611	150
Disruptiveness, teacher	-0.145	0.326	0.17	0.055	0.332	149
Bullying in class, teacher	0.036	-0.099	0.158	0.532	0.532	148
Panel B: classes with all enumerators' measures						
Disruptiveness, enumerator	-0.136	0.277	0.152	0.068	0.205	155
Average decibels during class	-0.108	1.391	0.815	0.088	0.176	153
Delay in class's start (minutes)	8.885	1.424	1.412	0.313	0.469	153

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator for classes with all teacher's or enumerators' measures. The regression is estimated with propensity score weights. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression. All the dependent variables were collected by the authors at baseline.

Appendix B Results without controls

Table B1: Treatment effect on eligible students

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: emotional stability						
School happiness score	-0.107	0.136	0.082	0.097	0.292	876
Self-control score	-0.184	-0.04	0.09	0.654	0.654	880
Self-esteem score	-0.17	-0.107	0.081	0.183	0.275	903
Standardized Treatment Effect	0.015	-0.002	0.08	0.977		915
Panel B: disruptiveness						
Disruptiveness, teacher	0.353	0.057	0.099	0.562	0.562	904
Disruptiveness, enumerator	0.153	0.063	0.104	0.545	1	955
Standardized Treatment Effect	-0.036	0.05	0.086	0.563		1111
Panel C: academic outcomes						
% school days missed	12.82	1.055	1.016	0.299	0.896	1236
Spanish test score	-0.308	-0.044	0.082	0.59	0.886	956
Math test score	-0.274	-0.006	0.081	0.946	0.946	956
Standardized Treatment Effect	0.011	-0.049	0.083	0.555		1238
Panel D: integration in the class network						
% class friends with student	0.07	0.008	0.005	0.118	0.472	1147
Friends' average ability	-0.061	-0.022	0.096	0.816	0.816	829
Friends' average disruptiveness	0.177	0.146	0.096	0.13	0.259	787
No friends in the class	0.27	-0.025	0.027	0.348	0.464	1147
Standardized Treatment Effect	-0.008	0.035	0.066	0.592		1148

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator and lottery fixed effects for eligible students. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression. All the dependent variables, except for *% school days missed*, were collected by the authors at endline.

Table B2: Treatment effect on ineligible students

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: emotional stability						
School happiness score	0.026	-0.009	0.04	0.828	0.828	3360
Self-control score	0.097	-0.067	0.044	0.126	0.377	3404
Self-esteem score	0.084	-0.066	0.047	0.161	0.241	3446
Standardized Treatment Effect	0.027	-0.062	0.046	0.183		3476
Panel B: disruptiveness						
Disruptiveness, teacher	-0.212	0.258	0.104	0.014	0.027	3203
Disruptiveness, enumerator	-0.068	0.01	0.057	0.856	0.856	3530
Standardized Treatment Effect	-0.049	0.089	0.073	0.221		4034
Panel C: academic outcomes						
% school days missed	13.089	0.331	0.742	0.656	0.656	4427
Spanish test score	0.128	-0.097	0.07	0.167	0.5	3517
Math test score	0.08	-0.035	0.065	0.589	0.884	3517
Standardized Treatment Effect	0.018	-0.038	0.058	0.515		4452
Panel D: integration in the class network						
% class friends with student	0.087	0.002	0.003	0.538	0.718	4168
Friends' average ability	0.027	-0.033	0.1	0.745	0.745	3342
Friends' average disruptiveness	-0.11	0.097	0.07	0.163	0.652	3176
No friends in the class	0.197	-0.018	0.013	0.175	0.349	4168
Standardized Treatment Effect	0.003	0.001	0.051	0.992		4171

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator and lottery fixed effects for ineligible students. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression. All the dependent variables, except for *% school days missed*, were collected by the authors at endline.

Table B3: Treatment effect on classroom environment

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Disruptiveness, teacher	-0.187	0.39	0.131	0.003	0.015	160
Bullying in class, teacher	-0.038	0.062	0.159	0.698	0.698	160
Disruptiveness, enumerator	-0.186	0.389	0.148	0.009	0.021	167
Delay in class's start (minutes)	9.938	1.204	1.046	0.25	0.312	160
Average decibels during class	0.022	0.681	0.487	0.162	0.27	169
Standardized Treatment Effect	-0.215	0.424	0.131	0.001		169

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator. The regression is estimated with propensity score weights. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment indicator. Column (3) reports the standard error of this coefficient, clustered at the lottery level. Column (4) reports the unadjusted p-value of this coefficient, while Column (5) reports its p-value adjusted for multiple testing, following the method proposed in Benjamini and Hochberg [1995]. Finally, Column (6) reports the number of observations used in the regression. All the dependent variables were collected by the authors at endline.