The Direct and Spillover Effects of a Nationwide Socio-Emotional Learning Program for Disruptive Students

Clément de Chaisemartin, Nicolás Navarrete

The direct and spillover effects of a nationwide socio-emotional learning program for disruptive students.*

Clément de Chaisemartin† Nicolás Navarrete H.‡

January 21, 2022

Abstract

Social and emotional learning (SEL) programs that target disruptive students aim to improve their classroom behavior. Small-scale programs in high-income countries have demonstrated positive effects. Using a randomized experiment, we show that a nationwide SEL program in Chile has no effect. Very disruptive students seem to reduce the program's effectiveness. ADHD being more prevalent in middle- than high-income countries, very disruptive students may be more present there, which could diminish the effectiveness of SEL programs. Moreover, implementation fidelity seems lower in this program than in the small-scale ones considered earlier, which could also explain the program's null effect.

Keywords: disruptive students, spillover effects, peer effects, social and emotional learning, implementation fidelity.

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, for providing us with some of the data used in this research, and for answering all our questions about the program. 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 Juan Pablo Arias, Antonio Figueroa and Gerardo Alvarez for outstanding research assistance. Finally, we also thank Miya Barnett, Myriam George, Peter Kuhn, Loreto Leiva-Bahamondes, Zoe Liberman, Shelly Lundberg, Michael Murphy, 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: Sciences Po; J-Pal; NBER. ‡Nicolás Navarrete H: City, University of London. For questions please email: `clement.dechaisemartin@sciencespo.fr` or `nicolas.navarrete-hernandez@city.ac.uk`.

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. In the US, 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, on top of their direct effects.

Epidemiological studies show that the prevalence of ADHD, a disorder correlated with conduct problems, is higher in some low- and middle-income countries than in high-income countries. Then, addressing students' conduct problems may be an even more pressing issue in those countries. In Chile, the country where the intervention we study takes place, 15.5% of primary school children have ADHD (see [de la Barra et al., 2013]). Primary school children also have be found to have high ADHD rates in Colombia (16.9%, see [Cornejo et al., 2005]), or in Iran (17.3%, see [Safavi et al., 2016]). On the other hand, the ADHD prevalence rate among school-age children is estimated at 8.7% in the US (see [Froehlich et al., 2007]), between 3.5 and 5.6% in France (see [Lecendreux et al., 2011]), and 3% in Italy (see [Bianchini et al., 2013]).[1]

School-based mental health programs are often used to reduce students' disruptiveness. Some programs are universal, meaning that they are delivered in classroom settings to all the students in the class. Other programs are selected, meaning that they are provided to students identified by teachers as having conduct problems, during the school day and outside of the classroom. Many school-based mental health programs are social and emotional learning (SEL) programs (see [Wilson and Lipsey, 2007]), that teach children to recognize and manage their emotions, and to handle interpersonal situations effectively, using cognitive and behavioral therapy (CBT).

In this paper, we study the effects of Skills for Life" (SFL), a selected SEL program for disruptive second graders in Chile. Since its creation in 1998, SFL has screened and treated around

[1] In all those studies, ADHD diagnosis is based on structured interviews conducted by trained health professionals, and on criteria set in the DSM-IV that do not allow clinical input by the interviewer. Accordingly, these cross-country differences are unlikely to be driven by measurement differences.

1,000,000 children, making it the fifth largest school-based mental health program in the world (see [Murphy et al., 2017]). To identify eligible students, SFL teams use a psychometric scale measuring students' disruptiveness, and students above some cut-off are eligible. Eligible students then follow 10 two-hours SEL sessions with a psychologist and a social worker. SFL is a costly program: we estimate that its cost per student is equivalent to at least 15% of the expenditure per primary school student in Chile.

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.

We find that a few weeks after the end of the intervention, SFL does not have short-run effects on eligible students' disruptiveness, mental health, and academic achievement. The effects we can rule out are fairly small. For instance, we can rule out at the 5% level that the program increases students' Spanish scores by more than 0.09σ, or that it reduces teachers' assessment of students' disruptiveness by more than 0.10σ. Not surprisingly, as we do not find that SFL impacts eligible students, we also do not find spillover effects on ineligible ones. We even find that the program has a strong negative effect on teachers' and enumerators' ratings of the overall disruptiveness of treated classes.

Even though control students were supposed to receive the treatment in the second semester of 2015, a significant proportion of them did not, in part because they changed school during the year, and in part because some control classes did not receive the intervention. Accordingly, being randomly assigned to receive the intervention in the first semester of 2015 increases students' probability of receiving the intervention at any point by 17 percentage points. We leverage this first-stage to study the medium-term effects of the program, and do not find any effect one and two years after the end of the intervention, on attendance, dropout, and students' performance in national exams. To increase power, we also look at the subsample of towns where this medium-term first stage is the highest. There, being randomly assigned to the treatment group increases students' probability of receiving the intervention at any point by 36.5 percentage points. There again, we do not find any effect, and we can reject moderate effects of the intervention.

Our results are at odds with a vast literature in psychology, that has found SEL programs to be successful. In a meta-analysis of 80 selected SEL interventions, Payton et al. [2008] find that

they reduce conduct problems by 0.47σ, and respectively improve mental health and academic performance by 0.50 and 0.43σ.[2] To account for the discrepancy between our results and the literature, we compared SFL to the selected SEL interventions reviewed in Payton et al. [2008]. Three conclusions emerge. First, SFL's intensity (number of sessions, duration, etc.) is comparable to that of the meta-analysis's interventions, so it is not the case that SFL is not intensive enough to produce an effect. SFL's curriculum is also similar to that of the PATHS intervention, another SEL intervention that has been shown to be successful Sorrenti et al. [2020].

Second, SFL may be implemented with lower fidelity than the meta-analysis's interventions. The interventions in the meta-analysis are demonstration programs mounted by researchers, that typically treat a few dozens children in a handful of schools. Half are delivered by the researchers, while the other half are delivered by psychologists or teachers under researchers' close supervision: typically, researchers review their delivery of the intervention every week. Accordingly, those interventions are probably implemented with very high fidelity. On the other hand, SFL is a large-scale governmental program, delivered by psychologists without any researcher involvement. The turnover rate among the psychologists implementing the program is fairly high, and more than a third of implementers have less than a year of experience working for SFL. The governmental agency in charge of the program loosely monitors the program implementers, and very rarely audits their workshops. Without sufficient experience and monitoring, teams may not implement the program with high-enough fidelity, which could explain why SFL does not produce an effect. Consistent with that hypothesis, Rojas-Andrade [2018] finds that slightly less than half of SFL teams implement the program with high fidelity. However, while low implementation fidelity may account for SFL's lack of effect, it cannot explain the negative effects we find on our classroom-level measures of disruption.

Third, all the interventions reviewed by Payton et al. [2008] take place in high-income countries, where ADHD is much less prevalent than in Chile. Moreover, a substantial fraction of those interventions exclude very disruptive students, unlike SFL. Accordingly, SFL is faced with a harder-to-treat population than those interventions. We find evidence that SFL's effectiveness may be hampered by the presence of very disruptive students. In classes with at least one very disruptive eligible student, defined as students above eligible students' 90th percentile of baseline disruptiveness,[3] the program increases the disruptiveness of other eligible students, of ineligible students, and it worsens

[2]Meta-analyses of universal SEL interventions find smaller but still large effects on those dimensions, around -0.25σ for conduct problems, and $+0.55\sigma$ and $+0.30\sigma$ for mental health and test scores ([Durlak et al., 2011], [Sklad et al., 2012], [Wigelsworth et al., 2016], [Taylor et al., 2017], and [Corcoran et al., 2018]).

[3]This definition ensures that about 50% of classes have at least one very disruptive student.

teachers' and enumerators' ratings of the overall class disruptiveness. The program also strongly increases the friendship ties between very disruptive and other eligible students. The former may then have a negative influence on the latter, which would explain the negative effects we observe.

Overall, our paper is the first to study the effects of a nationwide SEL program, implemented in a middle-income country with a high-ADHD prevalence rate. We do not find any positive effect of the program, unlike what previous research looking at small-scale programs in high-income countries with relatively low ADHD prevalence rates had found.

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.

2 The SFL intervention

SFL is a Chilean school-based SEL program for second graders suffering from conduct disorders. SFL, as other SEL programs, teaches children the necessary skills to recognize and manage their emotions, set and achieve positive goals, and handle interpersonal situations effectively. SFL tries to enhance children's self-awareness (accurately assessing one's feelings and maintaining a sense of self-confidence), self-management (regulating one's emotions and controlling impulses), and social awareness (being able to take the perspective of others, preventing, managing, and resolving interpersonal conflict).

SFL is managed by JUNAEB (*Junta Nacional de Auxilio Escolar y Becas*), a division of the Chilean Department of Education that administers non-teaching programs for socially vulnerable students, to help them succeed in school. The program started as a pilot in 1998. Over the next 3 years, JUNAEB collaborated with psychologists from the University of Chile and renowned international psychiatrists (e.g. Drs. Sheppard Kellam and Thomas Anders) to review the screening measures and successful interventions available at that time and adapt them for use in Chile, where the prevalence of ADHD is particularly high among children (see [de la Barra et al., 2013]). 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. Since 1998, the SFL program has screened and treated around 1,000,000 children, making it the fifth largest school-based mental

health program in the world (see [Murphy et al., 2017]).

To identify eligible students, SFL uses a psychometric scale, the Teacher Observation of Classroom Adaptation (TOCA, see [Kellam et al., 1977], and [Werthamer-Larsson et al., 1990]), adapted to the Chilean context by George et al. [1994]. In the end of each academic year, first-grade teachers answer 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:

- Hyperactive and aggressive 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;
- Disobedient 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 with poor social skills and motivation for learning: below the 25th percentile of the SC, MS, and EM scales, and below the 50th percentile of either the AA or AL scale.

The percentiles are gender specific, to ensure that not only males are eligible, and were computed using a representative sample of the 2nd grade population in Chile (see [George et al., 1994, De La Barra et al., 2005]). Students in the third eligibility group are not disruptive, but they only account for 7% of eligible students. The first two groups account for the bulk of eligible students, and they overlap: half of hyperactive and aggressive students are also considered as disobedient. Depending on the year, eligible students account for 15 to 20% of first-grade students whose teachers answer the TOCA questionnaire.

In second grade, SFL asks eligible students' parents the authorization to enroll their child in the program. If their parents accept, eligible students are enrolled in a workshop implemented by a team of two SFL employees. A survey conducted in 2015 (see [Rojas-Andrade and Leiva, 2018]) shows that half of SFL employees are psychologists. In Chile, this title can be obtained after a college degree with a psychology major (see [Guzmán et al., 2015]). The other half of employees are social workers and former teachers, titles that can also be obtained after a college degree. Usually, an SFL team consists of a psychologist and a social worker or teacher. 77% of SFL employees

are women, their average age is 31 years. They have on average 2.6 years of experience into the program, and 36% have less than one year of experience, indicating a high rate of turnover. During their first year, SFL employees receive three eight-hours-long days of training. They also attend "good practices" meetings every six months, in which they share with other teams what works in their workshops. As the Chilean public school system is administrated at the municipal level, SFL teams are also organized at this administrative level.

SFL workshops consist in 10 two-hours group sessions, taking place weekly, during the class day, over the course of one semester. During sessions, enrolled students leave the classroom, while their classmates stay 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.[4] The workshop takes place over two school periods, and eligible students come back to their classroom before the break. Sessions are divided into three blocks. The goal of the first block is to improve children's self-esteem, and their respect of others. The goal of the second block is to help students put words on their and others' emotions, and help them share their emotions with others. The goal of the third block is to teach students' self-control techniques, and strategies to find non-violent solutions to conflicts. Sessions are activity based, involve games, story-telling, and role play, and make use of CBT techniques. If they behave well during a session, students sometimes receive rewards like cakes or candies. SFL employees are provided with a 114-pages-long manual describing the goal and the content of each session, and suggesting games and activities. But they are also encouraged to tailor the content of their sessions to the specific needs of the students enrolled. Examples of activities in the SFL manual are presented in Appendix E. Several SFL activities are remarkably similar to activities proposed to students in the Promoting Alternative Thinking Strategies (PATHS) program, another SEL intervention [Sorrenti et al., 2020].

SFL rests on the idea that both student's mental health and academic achievement depend on their school adaptation. Whether students will thrive or not depends on whether they can meet the requirements imposed by the school around learning activities, relationship among peers, and behavioral regulation and autonomy [Perry and Weinstein, 1998, Kellam et al., 2011, Vargas and Peña, 2016]. SFL aims to help students that have difficulties to meet those requirements, using a socio-emotional curriculum incorporating evidence-based CBT techniques [Vargas and Peña, 2016].

[4] In a survey of 197 professionals implementing SFL in 127 municipalities (76% of all municipalities implementing SFL in Chile), the majority of professionals reported that during SFL sessions, teachers teach subjects deemed less crucial than Spanish or mathematics, like religion (a mandatory subject in Chile) or music, to the ineligible students that stay with them in the classroom (see [Rojas-Andrade and Leiva, 2018]).

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 training sessions, whose goal is to encourage them to reproduce the workshop's activities at home.

Though estimates of the cost of the program per student are not directly available, using two indirect methods we find that SFL costs between 200 and 458 USD per treated student (see Appendix F). We also estimate that the government spends 1,316 USD on instruction per student and per year in the schools in our sample.[5] Therefore, the program's cost represents a sizeable 15-35% increase of the expenditure per student.[6]

Previous research has found that from first to third grade, the disruptiveness of students that attend seven to 10 SFL sessions in second grade decreases more than that of students attending six sessions or less (see e.g. [Guzmán et al., 2015]). However, SFL attendance is driven by students' school attendance, and students who attend school less may do so because they experience negative shocks, which could explain why their disruptiveness decreases less. To avoid that type of endogeneity bias, our paper relies on an experimental control group to measure the effect of SFL.

A vast psychology literature has found that selected SEL interventions that target disruptive students tend to produce large effects. In a meta-analysis of 80 such interventions, Payton et al. [2008] find that they reduce conduct problems by 0.47σ, and respectively improve emotional stability and academic performance by 0.50σ and 0.43σ.[7] In Appendix D, we conduct a thorough and systematic

[5]The government funds public schools by giving them a voucher per student, whose amount depends on the student's attendance (`https://www.oecd-ilibrary.org/docserver/9789264287112-6-es.pdf?expires=1586606397&id=id&accname=guest&checksum=17AF0B3C9CF0863F8300FAA082FE969D`). For public primary schools, the school voucher is worth 754 USD for an attendance of 84%, the average attendance that is observed in our sample. Then, the government gives schools another voucher which is worth 721 USD for every very disadvantaged student in the school (`https://ate.mineduc.cl/usuarios/admin3/doc/2015020312570909985.Manual_Apoyo_a_la_Gestion.pdf`), and 78% of students are very disadvantaged in the schools we study, thus leading to our $754+0.78\times721=1{,}316$ USD estimate.

[6]SFL's cost is higher than that of the PATHS program, estimated at 74 USD per student in 2015 USD Sorrenti et al. [2020]. PATHS's cost does not include implementers' salary or transportation costs: PATHS is implemented by teachers, and the PATHS curriculum replaced an existing subject so it did not generate an increase in teachers' working hours. On the other hand, implementers' salary and transportation costs account for 75% of SFL's cost in the first method we use to estimate its cost. This explains the difference between the two programs' costs.

[7]Several other meta-analyses of SEL interventions have been peer-reviewed, unlike Payton et al. [2008], and are more recent. However, they either focus on universal interventions delivered to the whole class rather than to a selected group of students (see [Durlak et al., 2011], [Sklad et al., 2012], [Wigelsworth et al., 2016], [Taylor et al., 2017], and [Corcoran et al., 2018]), or they include both universal and selected interventions but do not report effects separately for both types of interventions (see [Dymnicki et al., 2012]). To our knowledge, Payton et al. [2008] is the only meta-analysis reporting effects separately for selected SEL interventions comparable to SFL. In any case,

comparison of SFL and the SEL interventions reviewed in Payton et al. [2008].[8] Three main findings emerge. First, SFL is no less intensive than the meta-analysis's interventions, in terms of number of sessions, sessions' duration, or number of students per workshop (see Panel A of Table D1). Many of the meta-analysis's interventions do not have a parental training, unlike SFL, but those that have one tend to have a more intensive parental training than SFL. Payton et al. [2008] do not mention heterogeneous effects across interventions with/without a parental training. Second, while many of the metanalysis's interventions also target primary school students with conduct problems, students receiving SFL may be harder to treat than those in the meta-analysis's interventions. All the meta-analysis's interventions take place in high-income countries, where the prevalence of ADHD among children is much lower than in Chile. Moreover, many of the meta-analysis's interventions exclude students with a psychological disorder or very disruptive students (see Panels C and D of Table D1): designers of SEL programs seem to consider they may be less effective with very disruptive students. On the other hand, SFL does not exclude such students. Third, SFL strikingly differs from the meta-analysis's interventions in terms of delivery. All of the meta-analysis's interventions are small-scale programs mounted by researchers, that are either delivered by researchers, or by personnel trained and supervised by researchers, with implementers' delivery typically reviewed every week by the research team (see Panel E of Table D1). On the other hand, JUNAEB loosely monitors delivery. We interviewed three SFL teams, and only one had a workshop observed over the last two years.

In the economics literature, while few articles have looked at SEL interventions specifically, several articles have studied interventions that intend to improve students' non-cognitive skills, and have also found large effects. Table D2 in the Appendix reviews some of those interventions. Many of them are more intensive than SFL, in terms of duration and numbers of sessions. While several of them are implemented at a much larger scale than the interventions in Payton et al. [2008], most are still implemented either by researchers, or by personnel trained and supervised by researchers, or by personnel employed by the NGO that created the program.

those six other meta-analyses also find pretty large effects, even though they are slightly lower than those in Payton et al. [2008]. The effects they find on conduct problems range from -0.14 to -0.47σ, with an average equal to -0.25σ. Similarly, effects on emotional stability range from 0.23 to 0.74σ (average=0.55σ). Finally, effects on academic performance range from 0.26 to 0.53σ (average=0.28σ).

[8] Several features of those interventions are readily available from Table 7 in Payton et al. [2008], but we also manually collected features that seemed important to us but were not reported in the paper. Thus, Appendix D also contains new findings on those interventions that may be of independent interest.

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 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. The remaining 114 schools each only had one class included in our sample, so 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 each semester of 2015. For each student, we know how many sessions she attended, and how many sessions her parents attended. Finally, JUNAEB also pro-

vided 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.[9] 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.[10]
- 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

[9] In a few towns, a teachers' strike interrupted the intervention for a couple of weeks, so that the last sessions of the workshop had to take place at the beginning of the second semester. Accordingly, in those towns, endline took place a few weeks after what had been originally planned. This issue arose in less than 15% of classes in our sample, and our point estimates of the program's effects remain similar to those in the tables below if we exclude those classes. Teachers' strikes are common in Chile. From 2005 to 2015, 7 strikes led to classes' interruptions for 2 or more weeks, 3 of which were national strikes affecting all towns in Chile, unlike the strike that took place in 2015 Villalobos Dintrans [2019]. Accordingly, SFL teams routinely need to adjust their delivery of the intervention to teachers' strikes.

[10] 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.

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".

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 similar 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.

Finally, in July 2019 we also conducted qualitative interviews to shed light on the mechanisms underlying our results. We interviewed three of the SFL municipal teams that had participated in our experiment, and that account for 12% of our sample.

The student-level outcomes we consider are:[11]

- 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 (netted out of enumerators' fixed effects),
- the percentage of the student's classmates that nominate her as one of their friends,

[11] We pre-specified a list of outcome variables in a pre-analysis plan (PAP) registered on the socialscienceregistry website. The analysis presented in Sections 4 and 5.1-5.3 mostly follows our PAP, except for a few exceptions described in Appendix G. On the other hand, the analysis presented in Sections 5.4-5.5 and 6 was not pre-specified in our PAP.

- 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 outcomes we consider 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,[12]
- 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 (netted out of enumerators' fixed effects).

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. Appendix G shows that most of our measures have good baseline-endline correlations.

3.3 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.[13] 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

[12]This measure is not netted out of enumerators' fixed effects (FEs), because netting out those FEs was not pre-specified in our PAP, and this measure has a good baseline-endline correlation even without netting out enumerators' FEs. On the other hand, enumerators' ratings of students' disruptiveness and decibel measurements have good baseline-endline correlations only after netting out enumerators' FEs, see Appendix G for further details.

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

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.

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. Differences are even more pronounced when one considers students' 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.

Finally, Column (5) of Table 1 shows some of those variables for the 2014 population of Chilean first graders. Students in our sample have lower school attendance and GPA, are more likely to attend a public school, and attend schools that do worse in the Spanish and math fourth grade Chilean national test. Students in our sample are also enrolled in classes that are slightly larger than the national average.

Table A2 in the Appendix shows some characteristics of the teachers in our sample. 96.3% of teachers are females. Their average age is 42.8 years, 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.

Table 1: Characteristics of eligible and ineligible students

	Ineligible (1)	Eligible (2)	P-value (3)	N (4)	National sample (5)
Panel A: demographic characteristics					
Male	0.498	0.582	0.000	5704	
Student's age at the end of 1st grade	6.842	6.835	0.747	5683	
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	
Male	0.501	0.587	0.000	5609	0.515
Panel B: cognitive measures					
Spanish test score	0.095	-0.335	0.000	4758	
Math test score	0.082	-0.289	0.000	4758	
1st grade GPA	-0.133	-0.459	0.000	5609	0.005
School's Spanish test score, 4th grade	-0.427				0.009
School's math test score, 4th grade	-0.244				0.005
Panel C: non-cognitive 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	-0.03	0.341	0.000	4645	
% class friends with student	0.088	0.076	0.000	4721	
Friends' average disruptiveness	-0.051	0.188	0.000	3931	
Attendance, 2014	88.863	87.71	0.000	5609	91.161
Class size, 2014	34.48	33.894	0.01	5609	31.334
Public school, 2014	0.677	0.755	0.000	5609	0.362

Notes: This table reports descriptive statistics for students in the sample and the population of Chilean students not in our sample who attended 1st grade in 2014. Column (1) reports the mean of the variable for ineligible students and Column (2) reports the mean of the 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 this comparison. Column (5) reports the mean of the variable for the population of Chilean students not in our sample enrolled in 1st grade in 2014.

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 fixed effects for the 56 lotteries we conducted to assign the treatment, and where D_{jk} is 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 classes. As our lotteries have few classes, the treatments of classes in the same lottery are strongly negatively correlated. Therefore, we cluster standard errors at the lottery level, following the recommendation of [de Chaisemartin and Ramirez-Cuellar, 2019]), who show that clustering at the class level could lead to substantial over-rejection of the null hypothesis.

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. With propensity score reweighting, β is also identified out of comparisons of treatment and control group classes in the same lottery (see [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 46 of the 56 lotteries, using lottery fixed effects or propensity score reweighting does not make a large difference.

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 $\beta = 0$. 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 (see [Benjamini and Hochberg, 1995]). Finally, Column (6) shows the number of observations used in the estimation.

Panel A of the table shows that 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 the control group classes received the treatment, while almost 100% of treatment group classes received it, this ITT effect "almost" estimates the effect of delivering the treatment in a class.

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. Some eligible students did not attend any group session, either because their parents refused that they participate, or forgot to send back the document they had to sign to authorize their child's participation. Table A1 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 are less disruptive at baseline. 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 is due to the fact that those students sometimes miss school on a workshop day, but school absenteeism does not seem to reduce students' exposure to the program very much. 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.

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.

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						
≥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 ≥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 ≥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 ≥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 ≥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.

4.2 Internal validity

Balancing checks in the original sample

We compare our outcome measures at baseline in the treatment and control groups, by estimating Regression (1) with those baseline measures as the dependent variables. First, Table 3 shows that among eligible students, only two of our twelve outcomes are significantly different at baseline in the treatment and control groups, at the 10% level. Treatment group students are more disruptive as per enumerators' ratings, and they are more likely to not be nominated as a friend by any other student in the class. Only the first of those two differences is significant at the 5% level. Second, Table 4 shows that among ineligible students, none of our twelve outcome is significantly different at baseline in the two groups. Finally, Table 5 compares our five class-level outcomes at baseline 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. Overall, we conduct 29 balancing checks in Tables 3, 4, and 5. We find five significant differences between the treatment and control groups at the

10% level, two significant differences at the 5% level, and no significant difference at the 1% level.

Table 3: Balancing of eligible students' outcomes at baseline

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	0.082	0.083	0.323	0.485	929
Self-control score	-0.148	-0.057	0.063	0.371	0.371	986
Self-esteem score	-0.107	-0.105	0.076	0.168	0.504	991
Panel B: disruptiveness						
Disruptiveness, teacher	0.396	0.087	0.276	0.753	0.753	253
Disruptiveness, enumerator	0.242	0.204	0.085	0.017	0.033	1007
Panel C: academic outcomes						
% school days missed, March	36.971	-4.809	3.421	0.16	0.479	1236
Spanish test score	-0.321	-0.021	0.086	0.806	1.000	1036
Math test score	-0.301	0.021	0.099	0.829	0.829	1036
Panel D: integration in the class network						
No friends in the class	0.128	0.047	0.026	0.065	0.259	1030
% class friends with student	0.075	0.002	0.006	0.769	1.000	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	0.667	822

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.

More balancing tests are shown in the Appendix. Table A5 (resp. A8) compares the demographic characteristics and TOCA scores of eligible (resp. ineligible) students in the treatment and control groups. Table A11 compares teachers in the treatment and control groups. Overall, our treatment and control groups appear to be well balanced: in those three tables, only 5 out of 46 differences are significant at the 10% level.

Attrition, and post-attrition balancing checks

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 show that the treatment and the control groups do not present differential levels of attrition, and that the characteristics of

Table 4: Balancing of ineligible students' outcomes at baseline.

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: emotional stability						
School happiness score	0.039	-0.015	0.039	0.697	1	3502
Self-control score	0.05	-0.005	0.045	0.917	0.917	3608
Self-esteem score	0.066	-0.051	0.043	0.234	0.703	3619
Panel B: disruptiveness						
Disruptiveness, teacher	-0.132	0.052	0.181	0.772	0.772	804
Disruptiveness, enumerator	-0.067	0.078	0.06	0.193	0.386	3638
Panel C: academic outcomes						
% school days missed, March	38.922	-2.992	2.969	0.314	0.941	4427
Spanish test score	0.139	-0.065	0.076	0.393	0.589	3722
Math test score	0.083	0.033	0.079	0.676	0.676	3722
Panel D: integration in the class network						
No friends in the class	0.097	0.02	0.02	0.328	0.656	3691
% class friends with student	0.09	-0.003	0.005	0.523	0.697	3691
Friends' average ability	0.055	0.017	0.099	0.86	0.86	3260
Friends' average disruptiveness	-0.094	0.075	0.073	0.305	1	3109

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 5: Balancing of classes' outcomes at baseline

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Disruptiveness, teacher	-0.143	0.286	0.16	0.074	0.123	161
Bullying in class, teacher	0.033	-0.094	0.147	0.519	0.519	160
Disruptiveness, enumerator	-0.131	0.275	0.153	0.072	0.181	168
Delay in class's start (minutes)	8.802	1.122	1.253	0.37	0.463	166
Average decibels during class	0.053	1.796	0.745	0.016	0.079	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.

treatment and control group students for which endline measures are available are still balanced.

Table A3 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 25.2% 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. Our data shows that for the most part, the remaining 19.3% are students who were absent on the day when the enumerators visited the class.[14] The teacher's endline disruptiveness rating is missing for 23.2% of students. Again, for some of them this is because they have left the class at endline. But our data shows that for the majority of students, this is because their teachers refused to rate 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 A3 shows tests of differential attrition between the treatment and control groups, conducted by estimating Regression (1) with measures of attrition 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.

Table A4 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.

Finally, we conduct balancing checks again, among the students whose endline measures are available. Table A6 (resp. Table A7) considers the same 29 baseline characteristics as in Tables 3 and A5, 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. In Table A6 (resp. Table A7), three (resp. two) differences out of 29 are significant at the 10% level. Table A9 repeats the same exercise, among ineligible students for which enumerators' endline measures are available. Four differences out of 29 are significant at the 10% level. Finally, Table A10 compares ineligible students for which the teacher's endline disruptiveness rating is available in the treatment and control groups. Eight differences out of 29 are significant at the 10% level, but most become insignificant once p-values are adjusted for multiple testing. Overall, the post-attrition treatment and control groups whose outcomes are compared in Section

[14]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.

5 seem to have balanced baseline characteristics.

Turning to class-level attrition, 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 A12, we conduct again the balancing checks on the baseline class-level measures in Table 5. For measures made by teachers, we 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 5, three differences are significant at the 10% level.

4.3 Estimation methods

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

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

where Y_{ijk} is the outcome of student i in class j and lottery k, the γ_ks are lottery fixed effects, X_{ijk} denotes student-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]. We run a Lasso regression of the outcome on all the student-level baseline variables in Tables 3 and A5, and we pick the variables selected by the Lasso.[15]

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 Tables 3 and A5, and all the class-level baseline variables in Tables 5 and A11, and we pick the variables selected by the Lasso.

[15] 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 hypotheses. 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 6, 7, and 8 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

Table 6 below shows the effect of the SFL workshops on eligible students' outcomes.

Panel A shows that the SFL workshops do not have large effects on eligible students' emotional stability. The average school happiness score is 0.123σ higher in the treatment than in the control group, but this difference is marginally 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). 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.

Panel B shows that SFL does not have a large effect on eligible students' disruptiveness. At endline, the average teachers' disruptiveness rating is 0.1σ 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 ratings by more than 0.1σ. This is around 1/5 of the treatment effect on students' disruptiveness found by Payton et al. [2008] in their meta-analysis of selected SEL programs. Enumerators' disruptiveness ratings also do not significantly differ in the treatment and control groups.

Panel C shows that SFL also does not have large effects on the academic outcomes of eligible students. For instance, students' Spanish and mathematics scores are 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

Table 6: 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	0.1	0.102	0.327	0.654	904
Disruptiveness, enumerator	0.157	0.02	0.083	0.805	0.805	948
Standardized Treatment Effect	-0.025	0.062	0.089	0.489		1110
Panel C: academic outcomes						
% school days missed	12.82	1.055	1.016	0.299	0.896	1236
Spanish test score	-0.308	-0.049	0.069	0.482	0.723	956
Math test score	-0.274	-0.006	0.08	0.945	0.945	956
Standardized Treatment Effect	0.011	-0.035	0.071	0.622		1238
Panel D: integration in the class network						
No friends in the class	0.27	-0.028	0.027	0.307	0.409	1147
% class friends with student	0.07	0.007	0.005	0.145	0.291	1147
Friends' average ability	-0.061	-0.011	0.077	0.883	0.883	829
Friends' average disruptiveness	0.177	0.132	0.087	0.131	0.525	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 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.

found in the meta-analysis by Payton et al. [2008].

Finally, Panel D shows that SFL does not have large effects on eligible students' friendship ties. 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.

Overall, we do not find evidence of a positive effect of SFL on any of the dimensions we consider, and we can also rule out much smaller effects 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 7 below 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.

Table 7: 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	0.106	0.05	0.101	3203
Disruptiveness, enumerator	-0.046	-0.003	0.046	0.954	0.954	3518
Standardized Treatment Effect	-0.051	0.063	0.072	0.384		4033
Panel C: academic outcomes						
% school days missed	13.089	0.382	0.634	0.547	0.82	4427
Spanish test score	0.128	-0.055	0.055	0.316	0.948	3517
Math test score	0.08	-0.013	0.056	0.821	0.821	3517
Standardized Treatment Effect	0.018	-0.019	0.044	0.66		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	0.004	0.003	0.156	0.312	4168
Friends' average ability	0.027	-0.011	0.077	0.884	0.884	3342
Friends' average disruptiveness	-0.11	0.051	0.053	0.338	0.45	3176
Standardized Treatment Effect	0.003	0.066	0.037	0.076		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 potential controls, following 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.

Panel B suggests that the SFL workshops may generate negative spillover effects on ineligible students' disruptiveness. At endline, the average of teachers' disruptiveness ratings is 0.208σ higher in the treatment than in the control group. This difference is significant (p-value=0.05), but becomes marginally insignificant after adjusting for multiple testing (adjusted p-value=0.101).

Then, Panel C shows that the SFL workshops do not have large spillover effects on ineligible students' academic outcomes. Finally, 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, a 17.8% 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 9.1% of their classmates in the treatment group, against 8.7% in the control group, but this difference is not significant. 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=0.076).

5.3 Effects on the classroom environment

Table 8: 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	-0.100	0.308	0.095	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 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.

In this section, we study how the SFL workshops affect different measures of classrooms' environment at endline. Table 8 above shows that SFL worsens teachers' and enumerators' disruptiveness ratings of the classes. Those ratings are based on teachers' and enumerators' agreement with statements like "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". According to teachers, treated classes are 0.232σ more 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). According to enumerators, treated classes are 0.389σ more disruptive. This difference is statistically significant before and after adjusting for multiple testing (p-value=0.009,

adjusted p-value=0.043). Enumerators 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 5 shows that treated and control classes are imbalanced on these two measures at baseline, so Table A13 reestimates these two effects controlling for these two measures at baseline.[16] 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), so the treatment effects on these two measures do not seem due to imbalances already existing at baseline. Table A13 also shows the results we obtain without any controls. The estimated effects are even more significant without than with controls.

It may be surprising that the treatment significantly worsens enumerators' ratings of classes' overall disruptiveness, without affecting their ratings of eligible and ineligible students' disruptiveness, as shown in Panel B of Tables 6 and 7. While the limited amount of time they spend in each classroom may be enough for them to observe that there is more disorder in the treated classes, it may not be sufficient for them to pinpoint the students responsible for that disorder.

Table 8 also shows that treated classes have higher levels of bullying, that their lectures start 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 Table 8 is 0.308σ 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 6, 7, and 8 we estimate the effect of the treatment on nine standardized scores (adjusted p-value=0.009). Therefore, we can conclude that SFL significantly worsens the studying conditions in treated classes.

5.4 Heterogeneous treatment effects

In this section, we investigate treatment effect heterogeneity, along six student and class variables: students' gender; the social security score of their family, a good proxy for socio-economic status; the average of students' Spanish and mathematics scores; class size; the average of students' authority acceptance, attention and focus, activity levels, and overall disruptiveness TOCA scores;

[16] 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.

an indicator for whether the class has a very disruptive eligible student, defined as a student above the 90th percentile[17] of the average of those four TOCA scores among eligible students. Investigating treatment effect heterogeneity along the first four dimensions seems relatively uncontroversial. The inclusion of the last two dimensions in our heterogeneity analysis is motivated by the fact that primary school students with serious behavioral problems are more present in Chile than in high-income countries, and are not excluded from SFL while they are often excluded from SEL interventions in high-income countries. Accordingly, SFL may face a harder-to-treat population, which could explain why it does not produce the large positive effects typically produced by SEL interventions in high-income countries Payton et al. [2008]. Moreover, each SFL workshop is more likely to comprise some very disruptive students than an SEL workshop in a high-income country, and the presence of those hard-to-treat students may lower the workshop's effectiveness for every student, including the less disruptive ones.

In our heterogeneity analysis, we use the machine-learning method proposed by Chernozhukov et al. [2018]. To predict our outcomes, we use elastic net regressions including the six variables listed above, their square, and the products between the variables. We investigate treatment effect heterogeneity for our two main student-level outcomes (teachers' endline disruptiveness ratings, and the average of students' endline Spanish and mathematics scores), and for our two main class-level outcomes (teachers' and enumerators' endline disruptiveness ratings).

Results are shown in Table 9 below. Panels A and B suggest that no subgroup of students is strongly affected by the treatment. Across the split-sample replications, the median treatment effects of eligible students predicted to be in the top and bottom quartiles of the treatment effect by the elastic net are respectively equal to 0.293σ and -0.118σ for teachers' ratings of disruptiveness, and 0.150σ and -0.198σ for students' test scores. These effects are all insignificant, and they do not significantly differ between the top and bottom quartiles. Results are similar for ineligible students, though we find some evidence that SFL may increase the disruptiveness of some ineligible students. Finally, Panels C and D may suggest that the treatment effect is actually very negative for some classes, though we lack power to make definitive conclusions. For both outcomes, the median treatment effect of classes predicted to be in the effect's top quartile is large, around 0.75σ for teachers' disruptiveness ratings, and above 0.80σ for enumerators' ratings. This second effect

[17]The choice of the 90th percentile was guided by the fact that $0.9^7 = 48\%$, so assuming that students' disruptiveness levels are independent within a class and that all classes have 7 eligible students, 52% of classes should have at least one student above that percentile. In practice, the proportion of classes that have at least one eligible student above that percentile is slightly lower (46%), but still close to 50%.

is significant at the 10% level. This is despite the fact that the approach to compute p-values proposed in Chernozhukov et al. [2018], which we follow, is conservative.

Table 9: Heterogeneous treatment effects

	Estimate (1)	P-value (2)
Panel A: teachers' ratings of students' disruptiveness at endline		
Eligible students		
Median effect in bottom quartile of predicted effect	-0.118	1.000
Median effect in top quartile of predicted effect	0.293	0.378
Median difference between effects in top and bottom quartiles	0.404	0.276
Ineligible students		
Median effect in bottom quartile of predicted effect	0.106	0.841
Median effect in top quartile of predicted effect	0.328	0.098
Median difference between effects in top and bottom quartiles	0.231	0.399
Panel B: average of students' endline Spanish and mathematics scores		
Eligible students		
Median effect in bottom quartile of predicted effect	-0.198	0.582
Median effect in top quartile of predicted effect	0.150	0.842
Median difference between effects in top and bottom quartiles	0.368	0.350
Ineligible students		
Median effect in bottom quartile of predicted effect	-0.105	0.504
Median effect in top quartile of predicted effect	0.049	1.000
Median difference between effects in top and bottom quartiles	0.146	0.470
Panel C: teachers' disruptiveness ratings of the classes		
Median effect in bottom quartile of predicted effect	-0.051	1.000
Median effect in top quartile of predicted effect	0.749	0.164
Median difference between effects in top and bottom quartiles	0.800	0.344
Panel D: enumerators' disruptiveness ratings of the classes		
Median effect in bottom quartile of predicted effect	-0.034	1.000
Median effect in top quartile of predicted effect	0.817	0.099
Median difference between effects in top and bottom quartiles	0.857	0.326

Notes: This table uses the method proposed by Chernozhukov et al. [2018] to investigate treatment effect heterogeneity for teachers' disruptiveness ratings of students, for the average of students' Spanish and mathematics tests scores, and for teachers' and enumerators' disruptiveness ratings of classes. Column (1) reports the median treatment effect of students or classes in the bottom and top quartiles of the predicted effect, as well as the median difference of the treatment effect between students or classes in those quartiles, across 100 split-sample replications. The treatment effect is predicted using elastic net regressions. Column (2) reports the p-value of those estimates, computed according to the method proposed by Chernozhukov et al. [2018]. The variables used in the elastic net regressions are: students' gender; the social security score of their family; the average of their Spanish and mathematics score; class size; the average of students' authority acceptance, attention and focus, activity levels, and overall disruptiveness TOCA scores; whether the class has a very disruptive eligible student; the squares and interactions of those variables.

Table A14 shows how different are classes in the bottom and top quartiles of the predicted effect in Table 9, for the two class-level outcomes where we find some evidence of very negative effects in some classes. Panel A shows that classes predicted to be in the top quartile of the treatment effect on teachers' assessment of classes' disruptiveness are 9.2 percentage points more likely to have at least one very disruptive student than classes predicted to be in the bottom quartile. This difference represents 21% of the average of that variable in the control group, and 17% of its standard deviation in the control group. Per these metrics, the difference between classes predicted to be in the top and bottom quartiles is substantially larger for that variable than for the other variables in the elastic net regressions, even though none of these differences are significant. Panel B shows that the same holds when one looks at the differences between classes predicted to be in the top and bottom quartiles of the treatment effect on enumerators' assessment of classes' disruptiveness: none of the differences are significant, and the one that's quantitatively the largest is for the proportion of classes with at least one very disruptive student. On the other hand, students' TOCA scores are not very different in classes in the top and bottom quartiles of the predicted effect, in both panels.

5.5 SFL effects on eligible students' medium-term outcomes.

The previous results show SFL's effect on outcomes measured around three weeks after the workshops ended. As improvements in socio-emotional skills might take more time to impact student's behavior and academic outcomes, we merge our data with administrative data from the Ministry of Education and look at outcomes up to two years after SFL workshops were implemented.[18] The administrative data contains the following outcomes: whether the student was promoted to the next grade; student's attendance and dropout; student's Spanish score in the 2^{nd} and 4^{th} grade Chilean national tests; student's math score in the 4^{th} grade Chilean national test. To preserve students' anonymity, we could only include a limited set of control variables in the merge. In all specifications we use the following student-level controls: an indicator variable for whether the student is female, lottery fixed effects, and student's 2014 school GPA and attendance. We cannot control for our baseline Spanish, maths, and disruptiveness scores, because those variables could not be included in the merge with the administrative data. Instead, we control for class $\times$ gender average, among eligible students, of those scores. When the outcome is one of the 4^{th} grade test score, we also control for the year in which the student took the test.

Column (2) of Panel A in Table 10 shows that being randomly assigned to receive the SFL

[18]There are 10 eligible students that we cannot identify in the administrative data due to the anonymization process carried out by the Ministry of Education.

workshops in the first semester of 2015 increases the probability to attend at least one session during the 2015 school year, either in the first or in the second semester, by 17%. Then, this random assignment can be used as an instrument to study the medium-term effects of SFL workshops. There are two reasons why students assigned to SFL workshops in the second semester of 2015 are less likely to attend the SFL workshops at any point. First, a small number of teams did not implement the second-semester workshops as planned. Second, some students in schools due to receive an SFL workshop in the second semester left between the two semesters. Overall, this 17 percentage points difference is the first-stage we leverage to study SFL's medium-term effects.

We do not have data on whether students attend an SFL workshop in 2016, 2017, or 2018, the years when we measure our medium-term outcomes. There is no SFL workshop in 3^{rd}, 4^{th}, and 5^{th} grade, so only students that repeat 2^{nd} grade in 2015 may have benefited of the intervention in those years. As shown in Table 10 below, only 12% of eligible students in the control group do not pass in 3^{rd} grade in 2015. Moreover, the 2016 2^{nd} grade class of those students may not have benefited from the intervention, and treatment group students that repeated 2^{nd} grade in 2015 and that had not benefited from the intervention in 2015 are equally likely to benefit from it in 2016. Accordingly, our 17 percentage points difference is probably very close from the first-stage effect of our random assignment on the probability of receiving the intervention at any point.

Some of our medium-term outcomes are observed for almost all students. On the other hand, scores in the Chilean national tests are only observed for students that take those tests, so differential attrition may bias our estimated effects for those outcomes. Moreover, treatment and control group students could take those tests in different years, which could also bias the results. Panel B of Table 10 evacuates such concerns: treatment and control group students are equally likely to take those tests, and take them in the same year on average.

Panel C shows the ITT effects of being assigned to SFL workshops in the first semester on different academic outcomes. Being promoted to the next grade, attendance, and school dropout do not significantly differ in the treatment and control groups two years after the SFL workshops were implemented. Being in the treatment group increases 2^{nd} grade Spanish test score by 0.060σ (p-value=0.303), and decreases 4^{th} grade Spanish test score by 0.079σ (p-value=0.355) and maths test score by 0.074σ (p-value=0.314). As the first stage effect in Panel B is far from one, Panel D shows results from two stage least square (2SLS) regressions, where attending at least one SFL session in 2015 is instrumented by the assignment to SFL workshops in the first semester. Estimates

are imprecise, but we can reject at the 5% level that SFL workshops increase 4^{th} grade Spanish and math test score by more than 0.452σ and 0.353σ, respectively. Those effects are respectively comparable to and smaller than the 0.43σ effect on academic performance found in the meta-analysis by Payton et al. [2008].

Table 10: Treatment effect on eligible student's medium-term outcomes

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	N (5)
Panel A: Workshop attendance in 2015, in the first or second semester					
Eligible students attended ≥1 session, 2015	0.699	0.171	0.037	0.000	1228
Panel B: Attrition in cognitive tests					
Took 2nd grade test, 2015	0.760	0.007	0.022	0.731	1228
Took 4th grade test, 2017	0.692	0.022	0.022	0.323	1228
Took 4th grade test, 2018	0.105	0.011	0.020	0.583	1228
Panel C: ITT estimates					
Student passed grade, 2015	0.876	0.011	0.022	0.611	1228
Student passed grade, 2016	0.934	–0.019	0.014	0.188	1228
Student passed grade, 2017	0.916	–0.002	0.019	0.914	1228
Attendance, 2015	88.046	–0.309	0.648	0.634	1212
Attendance, 2016	88.326	1.010	0.615	0.101	1207
Attendance, 2017	89.751	–0.128	0.536	0.811	1192
School dropout, 2015	0.010	0.002	0.007	0.715	1228
School dropout, 2016	0.012	0.011	0.008	0.187	1228
School dropout, 2017	0.028	–0.002	0.010	0.861	1228
Spanish test score, 2nd grade	–0.755	0.060	0.058	0.303	930
Spanish test score, 4th grade	–0.523	–0.079	0.086	0.355	952
Math test score, 4th grade	–0.570	–0.074	0.073	0.314	957
Panel D: 2SLS estimates					
Student passed grade, 2015	0.876	0.066	0.125	0.597	1228
Student passed grade, 2016	0.934	–0.109	0.088	0.215	1228
Student passed grade, 2017	0.916	–0.012	0.106	0.910	1228
Attendance, 2015	88.046	–1.787	3.649	0.624	1212
Attendance, 2016	88.326	5.906	3.732	0.113	1207
Attendance, 2017	89.751	–0.748	3.012	0.804	1192
School dropout, 2015	0.010	0.014	0.038	0.709	1228
School dropout, 2016	0.012	0.065	0.052	0.211	1228
School dropout, 2017	0.028	–0.01	0.057	0.857	1228
Spanish test score, 2nd grade	–0.755	0.331	0.331	0.317	930
Spanish test score, 4th grade	–0.523	–0.451	0.461	0.328	952
Math test score, 4th grade	–0.57	–0.421	0.395	0.286	957

Notes: Panels A, B, and C of this table report results from OLS regressions of several dependent variables on an indicator for being assigned to the SFL workshop in the first semester of 2015, lottery fixed effects and control variables. Panel D reports results from 2SLS regressions, where participation in the SFL workshop in 2015 is instrumented by the assignment indicator. Individual-level control variables are 2014 school GPA and attendance, as well as an indicator variable for whether the student is female. We also control for class × gender average, among eligible students, of Spanish and math baseline test score, and for class × gender average, among eligible students, of teacher's and enumerator's baseline disruptiveness assessment. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment. 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.

To increase the first stage and obtain more precise estimates, in Table A15 we replicate the analysis in Table 10 using only municipalities where being in the treatment group increases the probability of attending an SFL workshop in 2015 by more than 15%, the median first-stage across municipalities in our sample. Panel A shows that in this subsample, being in the treatment group increases the probability of attending a workshop in 2015 by 36.5%. Panel B shows no differential attrition between treatment and control group students. Panel C shows that based on 2SLS estimates in that subsample, we can reject at the 5% level that being in the treatment group increases 4^{th} grade Spanish and math test score by more than 0.134σ and 0.256σ, respectively. To preserve space, ITT effects and effects on being promoted to the next grade, attendance, and school dropout are not shown: those effects are insignificant.

Overall, in spite of our imprecise estimates, we can rule out that participating in the SFL program in second grade has large and lasting impacts on students' academic outcomes.

5.6 Robustness checks

As a robustness check, we reestimate the regressions in Tables 6, 7, 8, 10, and A15 without controls. Results can be found in Tables B1, B2, B3, B4, and B5. Results with and without controls are similar, except that the effects on ineligible students' friendships are no longer significant without controls, while the effect on teachers' assessment of their disruptiveness is more significant. We also recompute all the unadjusted p-values in Tables 6, 7, and 8 using randomization inference. The results, in Tables B6, B7, and B8, show that our main findings do not change.

Another methodological concern is that our control group may have benefited from the treatment, as we have some schools that have both treated and control classes, and treated students may interact with students from control classes in their school. To assess if this is a serious concern, we estimate SFL's effect in schools where only one class was included in our experiment. In this subsample, which still has 114 classes, we find that teachers' ratings of eligible students' disruptiveness is 0.2σ higher in the treatment than in the control group (p-value=0.12), and 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, control-group contamination seems unlikely to account for SFL's lack of effect.

6 Interpretation

In a meta-analysis of 80 selected SEL interventions, Payton et al. [2008] find that they reduce conduct problems by 0.47σ, and respectively improve emotional stability and academic performance by 0.50 and 0.43σ.[19] We can reject effects much smaller than those found in Payton et al. [2008]. In Section 2, we identified two important dimensions on which SFL differs from those interventions. First, SFL faces a harder-to-treat population. Second, SFL is delivered by government employees without any monitoring, while the interventions reviewed by Payton et al. [2008] are small-scale programs either implemented by researchers or implemented by personnel trained and closely monitored by researchers. We now investigate if those differences can account for SFL's lack of effect.

6.1 Very disruptive students may hamper SFL's effectiveness

Though our estimates are imprecise, our analysis of SFL's heterogeneous effects using machine-learning methods (see Section 5.6) suggests that classes whose overall disruptiveness is actually increased by the program are more likely to have at last one very disruptive student. We now investigate if very disruptive students hamper SFL's effectiveness. To do so, in Table 11 below we estimate SFL's effects in the 79 classes with at least one very disruptive student, as defined in Section 5.6. This subgroup analysis was not pre-specified in our pre-analysis plan, and remains exploratory. Those classes have 123 very disruptive eligible students, 534 other eligible students, and 2,064 ineligible students. We estimate SFL's effects separately for each group of students, focusing on disruptiveness ratings and test scores, and on the friendship nominations received by very disruptive eligible students. Unadjusted p-values and p-values controlling the False Discovery Rate (FDR) (see [Benjamini and Hochberg, 1995]) across all the tests in the table are presented.

First, Panel A shows that the program does not have any statistically significant effect on the disruptiveness and test scores of very disruptive eligible students, but increases by 50% the percentage of their classmates who nominate them as friends (unadjusted p-value=0.042, adjusted p-value=0.102). Second, in Panel B we estimate SFL's effects among the other eligible students. The program increases their teachers' disruptiveness ratings by 0.496σ (unadjusted p-value=0.0006, adjusted p-value=0.0102), may reduce their Spanish scores by 0.201σ (unadjusted p-value=0.033, adjusted p-value=0.112), does not have a significant effect on their enumerators' disruptiveness ratings and maths scores, and doubles the proportion that nominate at least one very disruptive

[19] A growing literature in economics has also found that programs for disruptive youth targeting non-cognitive skills can have large positive short- and long-run effects.

student as a friend (unadjusted p-value=0.012, adjusted p-value=0.051). Very disruptive eligible students may then have a negative influence on other eligible students, which could explain the negative effects the program has on them. Third, in Panel C we estimate that the program increases teachers' and enumerators' disruptiveness ratings of ineligible students, respectively by 0.477σ (unadjusted p-value=0.008, adjusted p-value=0.045) and 0.137σ (unadjusted p-value=0.083, adjusted p-value=0.176). On the other hand, the program does not have a significant effect on the test scores of those students and on the proportion of them who nominate a very disruptive student as a friend. The mechanism whereby the program makes ineligible students more disruptive may be a contagion effect: eligible students become more disruptive, and ineligible students imitate them. Finally, in Panel D we estimate that the program increases teachers' and enumerators' overall disruptiveness ratings of the classes, respectively by 0.669σ (unadjusted p-value=0.005, adjusted p-value=0.043) and 0.516σ (unadjusted p-value=0.035, adjusted p-value=0.099). The regressions in the table are estimated with the controls selected by the Lasso. Treatment effects are similar when those controls are dropped (see Appendix Table C1), and when the few covariates that are imbalanced at baseline in the relevant subsample are added as controls (see Appendix Table C2).

Classes with at least one very disruptive eligible student have slightly more eligible students than classes that do not have any (8.3 versus 6.2). This difference is not very large, but we still checked if we also find negative effects of the program in the subsample of classes that have more eligible students than the median. The answer is negative, so it does not seem that the negative effects we find in classes with at least one very disruptive eligible student are mediated by the slightly higher number of eligible students in those classes.

Table 11: Treatment effect in classes with at least one very disruptive student

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: Very disruptive eligible students						
Disruptiveness, teacher	0.985	-0.175	0.330	0.600	0.850	86
Disruptiveness, enumerator	0.286	0.613	0.439	0.162	0.306	85
Spanish test score	-0.460	0.047	0.304	0.878	1	88
Math test score	-0.230	0.063	0.512	0.902	1	88
% class friends with student	0.051	0.025	0.012	0.042	0.102	109
Panel B: Not very disruptive eligible students						
Disruptiveness, teacher	0.294	0.496	0.145	0.001	0.010	391
Disruptiveness, enumerator	0.162	0.118	0.124	0.339	0.576	393
Spanish test score	-0.349	-0.201	0.097	0.033	0.112	397
Math test score	-0.349	-0.008	0.181	0.965	0.965	397
Friends with ≥ 1 very dis.	0.065	0.075	0.030	0.012	0.051	397
Panel C: Ineligible students						
Disruptiveness, teacher	-0.205	0.477	0.181	0.008	0.045	1517
Disruptiveness, enumerator	-0.093	0.137	0.079	0.083	0.176	1576
Spanish test score	0.035	0.012	0.122	0.924	0.982	1579
Math test score	0.115	0.053	0.151	0.725	0.948	1579
Friends with ≥ 1 very dis.	0.067	0.015	0.028	0.584	0.903	1577
Panel D: Class-level outcomes						
Disruptiveness, teacher	-0.250	0.669	0.236	0.005	0.043	72
Disruptiveness, enumerator	-0.250	0.516	0.245	0.035	0.099	76

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 potential controls, following Belloni et al. [2014]. To account for the fact the randomization is stratified, the regressions in Panels A, B, and C have lottery fixed effects, while in the regressions in Panel D we use propensity score reweighting. 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.

Overall, we find suggestive evidence that SFL's effectiveness is hampered by the presence of very disruptive students, who may be less present in the other contexts where SEL programs have been shown to work. In fact, a substantial fraction of researchers designing SEL programs seem to consider they may be less effective with very disruptive students: Panel B of Table D1 shows that unlike SFL, at least 30% of the studies in the meta-analysis of Payton et al. [2008] excluded those students.

We still do not find statistically significant effects of SFL in the 93 classes that do not have any

very disruptive student. This may be because the effects we can reject in this subsample are too large, though we can for instance still reject at the 5% level an effect larger than 0.13σ on Spanish test scores. Another potential explanation is that even those classes may still have some students that are more disruptive than the typical students benefiting from selected SEL programs in the US.

6.2 SFL may not be implemented with high-enough fidelity

The interventions studied in the psychology literature are typically small-scale programs mounted by researchers, that are either delivered by researchers, or by personnel very closely monitored by researchers.[20] Similarly, most of the interventions studied in the economics literature are implemented either by researchers, or by personnel trained and supervised by researchers, or by personnel employed by the NGO that created the program. On the other hand, SFL is delivered by government employees. SFL implementers have a high turnover rate and are barely monitored by the government. As shown in Section 2, SFL's curriculum closely resembles that of other successful interventions, but implementation fidelity may be lower than in those successful interventions, as delivery is not supervised or monitored.

Intervention fidelity (IF) is the extent to which an intervention is implemented as conceived and planned [Schulte et al., 2009]. IF encompasses implementers' expertise (whether they feel they can deliver the intervention as planned) and adherence (whether they do deliver it as planned), as well as participants' responsiveness (their engagement with the intervention). Rojas-Andrade [2018] measured IF for 73 SFL teams, many of which are also part of our sample, in 2017. Using a questionnaire constructed following the implementation fidelity literature [see Hulleman et al., 2013, Abry et al., 2015], and administered to implementers, he constructed IF scales ranging from 0 to 100%. In this literature, values greater than 80% are deemed as high IF [Gresham, 2009]. Rojas-Andrade [2018] finds that across the 73 SFL teams, the median IF was 70% for expertise, 79% for adherence, and 81% for responsiveness: on the first two scales, less than half of the 73 teams can be deemed as having high IF. As is commonly the case in this literature, the IF measures in Rojas-Andrade [2018] are based on self-reports rather than external observations, so those measures could be upward biased. Overall, this suggests that IF may be low in a number of teams, which

[20] Researchers' involvement is very high in all the studies reviewed by Payton et al. [2008], but another meta-analysis of universal SEL interventions suggests that they can produce large effects without researchers' involvement. Wigelsworth et al. [2016] review 25 interventions implemented without researchers' involvement and find large effects: -0.15σ for conduct problems, +0.47σ for emotional stability, and +0.22σ for academic performance. However, looking at a random sample of 10 of those 25 studies, it appears that in 6 of the 8 studies where monitoring was discussed, monitoring was frequent and intensive, and was often conducted by an NGO promoting the program.

could explain SFL's lack of effect. To test this hypothesis directly, one could estimate SFL's effects separately in teams with high and low IF. Unfortunately, the data protection agreement under which Rojas-Andrade [2018] collected his data does not allow him to share it with other researchers. All of this suggests that SFL's IF may be lower than the IF in programs studied in the economics and psychological literature, where the close involvement of program creators in implementation presumably leads to a very high IF.

7 Conclusion

We explore the effects of "Skills for life" (SFL), a nationwide school-based SEL program for disruptive second graders in Chile. Eligibility to the program is based on first-grade teachers' ratings of students' disruptiveness, and SFL workshops consist in 10 two-hours sessions during which psychologists help students recognize and express their emotions, and teach them techniques to improve their behavior. We randomly assigned 172 classes to either receive SFL 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 emotional stability, disruptiveness, and test scores. This is at odds with an extensive literature that has found large effects of SEL programs (see [Payton et al., 2008], [Durlak et al., 2011], [Dymnicki et al., 2012], [Sklad et al., 2012], [Wigelsworth et al., 2016], [Taylor et al., 2017], and [Corcoran et al., 2018] for recent meta-analyses). We even find some negative effects of the program on teachers' and enumerators' ratings of the overall disruptiveness of treated classes.

To understand SFL's lack of effect, we investigate the differences between SFL and the programs studied in the literature. SFL is not less intensive than those other programs. But the literature has only considered small-scale programs mounted by researchers or NGOs, and either delivered by the researchers or NGO personnel, or by personnel closely monitored by them. On the other hand, SFL is a nationwide governmental program, implemented by government employees with a high turnover rate, and whose quality is not monitored by the government. While SFL's curriculum is closely aligned with that of successful programs, Rojas-Andrade [2018] find that implementation fidelity is low in some SFL teams, which could explain SFL's lack of effect. To remediate this, SFL could attempt to reduce implementers' turnover and monitor teams' delivery more systematically and frequently.

However, low implementation fidelity cannot explain the negative effects we find on our classroom-

level measures of disruption. Another difference between SFL and the interventions studied in the literature is that its population may be harder to treat. All the programs studied in the literature take place in high-income countries, where the prevalence of ADHD, a disorder correlated with conduct problems, is much lower than in Chile. Moreover, many of those programs exclude very disruptive students, unlike SFL. We find some evidence that very disruptive students may hamper the program's effectiveness: in classes with at least one eligible student in the top decile of eligible students' disruptiveness, the program actually has negative effects. The mechanism seems to be that SFL increases the friendships between very disruptive and other eligible students. Then, very disruptive students may have a negative influence on those other eligible students. To remediate this, SFL could exclude very disruptive students from its workshops, and offer them another type of treatment.

References

Tashia Abry, Chris S Hulleman, and Sara E Rimm-Kaufman. Using indices of fidelity to intervention core components to identify program active ingredients. *American Journal of Evaluation*, 36(3):320–338, 2015.

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.

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.

Rio Bianchini, Valentina Postorino, Rita Grasso, Bartolo Santoro, Salvatore Migliore, Corrado Burlò, Carmela Tata, and Luigi Mazzone. Prevalence of adhd in a sample of italian students: a population-based study. *Research in developmental disabilities*, 34(9):2543–2550, 2013.

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.

Roisin P Corcoran, Alan CK Cheung, Elizabeth Kim, and Chen Xie. Effective universal school-based social and emotional learning programs for improving academic achievement: A systematic review and meta-analysis of 50 years of research. *Educational Research Review*, 25:56–72, 2018.

JW Cornejo, O Osío, Y Sánchez, J Carrizosa, G Sánchez, H Grisales, H Castillo-Parra, and J Holguín. Prevalencia del trastorno por déficit de atención-hiperactividad en niños y adolescentes colombianos. *Rev neurol*, 40(12):716–722, 2005.

Clément de Chaisemartin and Jaime Ramirez-Cuellar. At what level should one cluster standard errors in paired experiments? *arXiv preprint arXiv:1906.00288*, 2019.

Flora De La Barra, Virginia Toledo, and Jorge Rodríguez. Prediction of behavioral problems in chilean schoolchildren. *Child Psychiatry and Human Development*, 35(3):227–243, 2005.

Flora Eloísa de la Barra, Benjamin Vicente, Sandra Saldivia, and Roberto Melipillan. Epidemiology of adhd in chilean children and adolescents. *ADHD Attention Deficit and Hyperactivity Disorders*, 5(1):1–8, 2013.

Joseph A Durlak, Roger P Weissberg, Allison B Dymnicki, Rebecca D Taylor, and Kriston B Schellinger. The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child development*, 82(1):405–432, 2011.

A Dymnicki, K Kendziora, and D Osher. Adolescent development for students with learning disabilities and behavioral disorders: The promise of social emotional learning. *Classroom behavior, contexts, and interventions*, 25:131–166, 2012.

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.

Tanya E Froehlich, Bruce P Lanphear, Jeffery N Epstein, William J Barbaresi, Slavica K Katusic, and Robert S Kahn. Prevalence, recognition, and treatment of attention-deficit/hyperactivity disorder in a national sample of us children. *Archives of pediatrics & adolescent medicine*, 161 (9):857–864, 2007.

Myriam George, Ximena Siraqyan, Randa Morales, Flora Barra De La, Jorge Rodríguez, Carmen López, and Virginia Toledo. Adaptación y validación de dos instrumentos de pesquisa de problemas de salud mental en escolares de 1 básico. *Revista de Psicología*, 5:ág–26, 1994.

Frank M Gresham. Evolution of the treatment integrity concept: Current status and future directions. *School Psychology Review*, 38(4):533, 2009.

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.

Chris S Hulleman, Sara E Rimm-Kaufman, and Tashia Abry. Innovative methodologies to explore implementation: Whole-part-whole—construct validity, measurement, and analytical issues for intervention fidelity assessment in education research. 2013.

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.

Sheppard G Kellam, Amelia CL Mackenzie, C Hendricks Brown, Jeanne M Poduska, Wei Wang, Hanno Petras, and Holly C Wilcox. The good behavior game and the future of prevention and treatment. *Addiction science & clinical practice*, 6(1):73, 2011.

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

Michel Lecendreux, Eric Konofal, and Stephen V Faraone. Prevalence of attention deficit hyperactivity disorder and associated features among children in france. *Journal of Attention Disorders*, 15(6):516–524, 2011.

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.

J Michael Murphy, Madelaine R Abel, Sharon Hoover, Michael Jellinek, and Mina Fazel. Scope, scale, and dose of the world's largest school-based mental health programs. *Harvard review of psychiatry*, 25(5):218–228, 2017.

John Payton, Roger P Weissberg, Joseph A Durlak, Allison B Dymnicki, Rebecca D Taylor, Kriston B Schellinger, and Molly Pachan. The positive impact of social and emotional learning for kindergarten to eighth-grade students: Findings from three scientific reviews. technical report. *Collaborative for Academic, Social, and Emotional Learning (NJ1)*, 2008.

Kathryn E Perry and Rhona S Weinstein. The social context of early schooling and children's school adjustment. *Educational Psychologist*, 33(4):177–194, 1998.

Rodrigo Miguel Rojas-Andrade. Efectos de la fidelidad de la implementación sobre los resultados de un programa chileno de salud mental escolar. *PhD thesis*, 2018.

Rodrigo Miguel Rojas-Andrade and Loreto Leiva. La salud mental escolar desde la perspectiva de profesionales chilenos. *Psicoperspectivas*, 17(2):151–162, 2018.

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.

Parvin Safavi, Forouzan Ganji, and Atenasadat Bidad. Prevalence of attention-deficit hyperactivity disorder in students and needs modification of mental health services in shahrekord, iran in 2013. *Journal of clinical and diagnostic research: JCDR*, 10(4):LC25, 2016.

Ann C Schulte, Julia E Easton, and Justin Parker. Advances in treatment integrity research: Multidisciplinary perspectives on the conceptualization, measurement, and enhancement of treatment integrity. *School Psychology Review*, 38(4), 2009.

Marcin Sklad, Rene Diekstra, Monique de Ritter, Jehonathan Ben, and Carolien Gravesteijn. Effectiveness of school-based universal social, emotional, and behavioral programs: Do they enhance students' development in the area of skill, behavior, and adjustment? *Psychology in the Schools*, 49(9):892–909, 2012.

Giuseppe Sorrenti, Ulf Zölitz, Denis Ribeaud, and Manuel Eisner. The causal impact of socio-emotional skills training on educational success. *University of Zurich, Department of Economics, Working Paper*, (343), 2020.

Rebecca D Taylor, Eva Oberle, Joseph A Durlak, and Roger P Weissberg. Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. *Child development*, 88(4):1156–1171, 2017.

Belén Vargas and Felipe Peña. *Orientaciones técnico metodologícas. Talleres Preventivos Habilidades para la Vida I 2016*. Santiago de Chile: JUNAEB, 2016.

Cristóbal Villalobos Dintrans. Los conflictos sociales en el campo educativo en el Chile Post-Dictadura (1990-2014). Análisis de su evolución, principales características y factores relacionados. *PhD thesis*, 2019.

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.

M Wigelsworth, A Lendrum, J Oldfield, A Scott, I ten Bokkel, K Tate, and C Emery. The impact of trial stage, developer involvement and international transferability on universal social and emotional learning programme outcomes: a meta-analysis. *Cambridge Journal of Education*, 46 (3):347–376, 2016.

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 Supplementary tables

Table A1: 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	0.7	0.397	0.051	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 A2: 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 A3: 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	0.389	1178
With all enumerators' measures	0.748	-0.035	0.03	0.247	0.308	1238
With teacher's disruption measure	0.768	-0.084	0.071	0.235	0.392	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 A4: 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	0.691	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	0.371	4466
With teacher's disruption measure	0.753	-0.059	0.067	0.383	0.638	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 A5: Balancing tests of eligible students' baseline characteristics

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	N (5)
Panel A: demographic characteristics					
Male	0.581	-0.004	0.046	0.937	1238
Teen mother	0.343	0.018	0.031	0.549	991
Student lives with father	0.563	-0.012	0.034	0.726	899
Social security score	5564.943	137.239	173.203	0.428	1124
Payment rate in health services	2.879	0.327	0.361	0.365	1122
Mother's education	8.813	-0.292	0.32	0.362	1080
Father's education	8.743	-0.565	0.38	0.137	913
Panel B: TOCA scores, PSC scores, and baseline measures					
Authority Acceptance TOCA	1.027	-0.084	0.063	0.181	1223
Social Contact TOCA	0.842	-0.025	0.072	0.723	1223
Motiv. for Schooling TOCA	0.842	-0.036	0.06	0.543	1223
Emotional Maturity TOCA	0.563	-0.12	0.076	0.117	1223
Attention and Focus TOCA	0.834	-0.054	0.063	0.391	1223
Activity Level TOCA	0.831	-0.054	0.064	0.404	1223
Academic ability TOCA	0.667	-0.016	0.071	0.82	1222
Overall disruptiveness TOCA	0.891	-0.046	0.076	0.548	1220
PSC	0.477	-0.011	0.08	0.889	903
Distance to teacher's desk	4.361	-0.079	0.18	0.66	863

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. Finally, Column (5) reports the number of observations used in the regression.

Table A6: 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	0.632	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
Attention 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	0.201	0.203	0.095	0.033	0.957	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	0.657	765
Distance to teacher's desk	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 A7: 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	0.952	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
Attention 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	0.999	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	0.912	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	0.193	0.107	0.105	0.31	1	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's desk	4.441	0.061	0.178	0.732	1	643
% school days missed, March	37.204	-2.795	4.143	0.5	0.966	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 A8: Balancing tests of ineligible students' baseline characteristics.

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	N (5)
Panel A: demographic characteristics					
Male	0.486	0.026	0.027	0.327	4466
Teen mother	0.328	0.016	0.02	0.434	3449
Student lives with father	0.639	-0.012	0.017	0.501	2866
Social security score	5965.036	-108.938	107.006	0.309	3944
Payment rate in health services	4.132	-0.019	0.313	0.951	3927
Mother's education	9.239	-0.19	0.2	0.341	3647
Father's education	9.181	-0.017	0.177	0.925	3204
Panel B: TOCA scores, PSC scores, and baseline measures					
Authority Acceptance TOCA	-0.356	0.059	0.054	0.278	3654
Social Contact TOCA	-0.346	0.14	0.055	0.01	3654
Motiv. for Schooling TOCA	-0.312	0.071	0.047	0.132	3654
Emotional Maturity TOCA	-0.171	0.024	0.092	0.795	3654
Attention and Focus TOCA	-0.32	0.092	0.053	0.086	3654
Activity Level TOCA	-0.33	0.124	0.066	0.059	3645
Academic ability TOCA	-0.244	0.043	0.041	0.292	3633
Overall disruptiveness TOCA	-0.335	0.075	0.041	0.068	3630
PSC	-0.171	0.043	0.044	0.333	2882
Distance to teacher's desk	4.519	0.168	0.158	0.286	3129

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. Finally, Column (5) reports the number of observations used in the regression.

Table A9: 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	0.903	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	0.847	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
Attention 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	0.807	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	-0.06	0.048	0.061	0.434	0.787	2805
Spanish test score	0.171	-0.067	0.071	0.347	0.838	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's desk	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 A10: 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	0.506	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
Attention 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	0.56	2637
Academic ability TOCA	-0.245	0.06	0.049	0.222	0.536	2632
Overall disruptiveness TOCA	-0.328	0.104	0.048	0.032	0.229	2630
PSC	-0.172	0.075	0.06	0.212	0.558	2084
Panel C: baseline measures						
School happiness score	0.047	-0.061	0.05	0.227	0.506	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	-0.095	0.083	0.065	0.201	0.582	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	0.523	2259
No friends in the class	0.088	0.023	0.021	0.277	0.536	2659
Distance to teacher's desk	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 A11: 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	0.958	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, taste for their job, and assessment of the class level						
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
Academic level of the class, teacher	0.059	–0.086	0.14	0.538	1	162
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	0.444	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 A12: 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						
Disruptiveness, teacher	-0.145	0.326	0.17	0.055	0.276	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	0.171	155
Average decibels during class	-0.108	1.391	0.815	0.088	0.146	153
Delay in class's start (minutes)	8.885	1.424	1.412	0.313	0.391	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.

Table A13: Treatment effect on teacher's and enumerator's disruptiveness ratings, using different sets of controls

Variables	Specification (1)	Control (2)	T-C (3)	S.E. (4)	Unadj. P (5)	N (6)
Disruptiveness, teacher	No controls	-0.187	0.39	0.131	0.003	160
Disruptiveness, teacher	Imb. charact.	-0.187	0.247	0.143	0.084	160
Disruptiveness, teacher	Lasso controls	-0.187	0.232	0.137	0.091	160
Disruptiveness, enumerator	No controls	-0.186	0.389	0.148	0.009	167
Disruptiveness, enumerator	Imb. charact.	-0.186	0.282	0.154	0.066	167
Disruptiveness, enumerator	Lasso controls	-0.186	0.389	0.148	0.009	167

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator, computed with propensity score weights. Column (1) gives the set of control variables used in the specification. *No controls* does not control for any variable. *Imb. charact.* controls for teacher's and enumerator's baseline disruptiveness ratings. *Lasso controls* controls for the variables selected by a Lasso regression of the dependent variable on all potential controls, following Belloni et al. [2014]. Column (2) reports the mean of the outcome variable for the control group. Column (3) reports the coefficient of the treatment indicator. Column (4) reports the standard error of this coefficient, clustered at the lottery level. Column (5) reports the unadjusted p-value of this coefficient. Finally, Column (6) reports the number of observations used in the regression. All the dependent variables were collected by the authors at endline.

Table A14: Differences between classes in the top and bottom quartiles of the predicted effect.

	Bottom quartile of predicted effect (1)	Δ top and bottom quartile (2)	P-value (3)
Panel A: teachers' disruptiveness ratings of the classes			
Gender	0.522	0.021	0.962
Household social security score	5781.638	-140.857	0.950
Class with a very disruptive eligible student	0.435	0.092	0.824
Average Spanish and mathematics score	-0.023	-0.072	0.983
Class size	33.801	0.171	0.964
TOCA disruptiveness measures	0.026	0.083	0.759
Panel B: enumerators' disruptiveness ratings of the classes			
Gender	0.523	0.019	0.948
Household social security score	5802.868	-151.243	0.908
Class with a very disruptive eligible student	0.465	0.069	1.000
Average Spanish and mathematics score	-0.030	-0.044	1.000
Class size	33.929	-0.933	1.000
TOCA disruptiveness measures	0.027	0.051	1.000

Notes: This table shows differences between classes in the top and bottom quartiles of the predicted effect, for two of the outcomes in Table 9. Column (1) shows the median characteristics of students predicted to be in the bottom quartile, across 100 split-sample replications. Column (2) shows the median difference between the characteristics of students predicted to be in the bottom and top quartiles. Column (3) show the p-value of this difference.

Table A15: Treatment effect on medium-term outcomes, in municipalities with first stage ≥15%

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	N (5)
Panel A: Workshop attendance in 2015, in the first or second semester					
Eligible students attended ≥1 session, 2015	0.534	0.365	0.055	0.000	478
Panel B: Attrition in cognitive tests					
Took 2nd grade test, 2015	0.751	0.011	0.031	0.718	478
Took 4th grade test, 2017	0.671	0.027	0.037	0.467	478
Took 4th grade test, 2018	0.107	0.005	0.03	0.865	478
Panel C: 2SLS estimates					
Spanish test score, 2nd grade	–0.682	–0.097	0.275	0.723	364
Spanish test score, 4th grade	–0.439	–0.487	0.317	0.125	376
Math test score, 4th grade	–0.496	–0.334	0.301	0.267	376

Notes: Panels A and B of this table report results from OLS regressions of several dependent variables on an indicator for being assigned to the SFL workshop in the first semester of 2015, lottery fixed effects and control variables. Panel C reports results from 2SLS regressions, where participation in the SFL workshop in 2015 is instrumented by the assignment indicator. Individual-level control variables are 2014 school GPA and attendance, as well as an indicator variable for whether the student is female. We also control for class × gender average, among eligible students, of Spanish and math baseline test score, and for class × gender average, among eligible students, of teacher's and enumerator's baseline disruptiveness assessment. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment. 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.

Appendix B Robustness: results without controls, and randomization inference

Table B1: Treatment effect on eligible students, no controls

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	1	904
Disruptiveness, enumerator	0.157	0.017	0.083	0.842	0.842	948
Standardized Treatment Effect	-0.025	0.041	0.088	0.645		1110
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, no controls

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	-0.046	0.02	0.042	0.637	0.637	3518
Standardized Treatment Effect	-0.051	0.107	0.069	0.122		4033
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, no controls

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.

Table B4: Treatment effect on eligible student's medium-term outcomes, without controls

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	N (5)
Panel A: Workshop attendance in 2015, in the first or second semester					
Eligible students attended ≥1 session, 2015	0.000	0.161	0.039	0.000	1228
Panel B: Attrition in cognitive tests					
Took 2nd grade test, 2015	0.76	0.002	0.022	0.921	1228
Took 4th grade test, 2017	0.692	0.004	0.027	0.886	1228
Took 4th grade test, 2018	0.105	0.016	0.02	0.42	1228
Panel C: ITT estimates					
Student passed grade, 2015	0.876	0.002	0.023	0.92	1228
Student passed grade, 2016	0.934	-0.025	0.015	0.099	1228
Student passed grade, 2017	0.916	-0.002	0.019	0.899	1228
Attendance, 2015	88.046	-0.599	0.659	0.363	1212
Attendance, 2016	88.326	0.818	0.566	0.149	1207
Attendance, 2017	89.751	-0.192	0.554	0.729	1192
School dropout, 2015	0.01	0.003	0.006	0.677	1228
School dropout, 2016	0.012	0.011	0.008	0.176	1228
School dropout, 2017	0.028	0.001	0.01	0.926	1228
Spanish test score, 2nd grade	-0.755	0.041	0.064	0.517	930
Spanish test score, 4th grade	-0.523	-0.071	0.088	0.415	952
Math test score, 4th grade	-0.57	-0.067	0.078	0.395	957
Panel D: 2SLS estimates					
Student passed grade, 2015	0.876	0.014	0.138	0.917	1228
Student passed grade, 2016	0.934	-0.156	0.111	0.161	1228
Student passed grade, 2017	0.916	-0.015	0.117	0.896	1228
Attendance, 2015	88.046	-3.694	4.217	0.381	1212
Attendance, 2016	88.326	5.08	3.458	0.142	1207
Attendance, 2017	89.751	-1.194	3.379	0.724	1192
School dropout, 2015	0.01	0.016	0.038	0.677	1228
School dropout, 2016	0.012	0.068	0.056	0.228	1228
School dropout, 2017	0.028	0.006	0.063	0.924	1228
Spanish test score, 2nd grade	-0.755	0.247	0.375	0.51	930
Spanish test score, 4th grade	-0.523	-0.431	0.503	0.391	952
Math test score, 4th grade	-0.57	-0.403	0.453	0.373	957

Notes: Panels A, B, and C of this table report results from OLS regressions of several dependent variables on an indicator for being assigned to the SFL workshop in the first semester of 2015 and lottery fixed effects. Panel D reports results from 2SLS regressions, where participation in the SFL workshop in 2015 is instrumented by the assignment indicator. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment. 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.

Table B5: Treatment effect on medium-term outcomes for randomization group with a first stage above 15%, without controls

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	N (5)
Panel A: Workshop attendance in 2015, in the first or second semester					
Eligible students attended ≥1 session, 2015	0.000	0.353	0.069	0.000	478
Panel B: Attrition in cognitive tests					
Took 2nd grade test, 2015	0.751	0.031	0.029	0.285	478
Took 4th grade test, 2017	0.671	0.062	0.038	0.1	478
Took 4th grade test, 2018	0.107	-0.011	0.025	0.643	478
Panel C: 2SLS estimates					
Spanish test score, 2nd grade	-0.682	0.163	0.297	0.582	364
Spanish test score, 4th grade	-0.439	-0.349	0.361	0.333	376
Math test score, 4th grade	-0.496	-0.198	0.292	0.498	376

Notes: Panels A and B of this table report results from OLS regressions of several dependent variables on an indicator for being assigned to the SFL workshop in the first semester of 2015 and lottery fixed effects. Panel C reports results from 2SLS regressions, where participation in the SFL workshop in 2015 is instrumented by the assignment indicator. Column (1) reports the mean of the outcome variable for the control group. Column (2) reports the coefficient of the treatment. 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.

Table B6: Treatment effect on eligible students, randomization inference

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	0.137	0.411	876
Self-control score	-0.184	-0.04	0.087	0.68	0.68	880
Self-esteem score	-0.17	-0.106	0.079	0.172	0.258	903
Standardized Treatment Effect	0.015	-0.002	0.08	0.977		915
Panel B: disruptiveness						
Disruptiveness, teacher	0.353	0.1	0.102	0.362	0.724	904
Disruptiveness, enumerator	0.157	0.02	0.083	0.786	0.786	948
Standardized Treatment Effect	-0.025	0.062	0.089	0.489		1110
Panel C: academic outcomes						
% school days missed	12.82	1.055	1.016	0.283	0.849	1236
Spanish test score	-0.308	-0.049	0.069	0.486	0.729	956
Math test score	-0.274	-0.006	0.08	0.941	0.941	956
Standardized Treatment Effect	0.011	-0.035	0.071	0.622		1238
Panel D: integration in the class network						
No friends in the class	0.27	-0.028	0.027	0.348	0.464	1147
% class friends with student	0.07	0.007	0.005	0.142	0.568	1147
Friends' average ability	-0.061	-0.011	0.077	0.896	0.896	829
Friends' average disruptiveness	0.177	0.132	0.087	0.147	0.294	787
Standardized Treatment Effect	-0.008	0.038	0.063	0.54		1148

Notes: This table replicates results in Table 6 computing unadjusted p-values using randomization inference. 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, using randomization inference. 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 B7: Treatment effect on ineligible students, randomization inference

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	0.704	0.704	3360
Self-control score	0.097	-0.05	0.043	0.273	0.819	3404
Self-esteem score	0.084	-0.043	0.047	0.332	0.498	3446
Standardized Treatment Effect	0.027	-0.023	0.042	0.577		3476
Panel B: disruptiveness						
Disruptiveness, teacher	-0.212	0.208	0.106	0.057	0.114	3203
Disruptiveness, enumerator	-0.046	-0.003	0.046	0.952	0.952	3518
Standardized Treatment Effect	-0.051	0.063	0.072	0.384		4033
Panel C: academic outcomes						
% school days missed	13.089	0.382	0.634	0.566	0.849	4427
Spanish test score	0.128	-0.055	0.055	0.388	1	3517
Math test score	0.08	-0.013	0.056	0.837	0.837	3517
Standardized Treatment Effect	0.018	-0.019	0.044	0.66		4452
Panel D: integration in the class network						
No friends in the class	0.197	-0.035	0.013	0.009	0.036	4168
% class friends with student	0.087	0.004	0.003	0.195	0.39	4168
Friends' average ability	0.027	-0.011	0.077	0.844	0.844	3342
Friends' average disruptiveness	-0.11	0.051	0.053	0.349	0.465	3176
Standardized Treatment Effect	0.003	0.066	0.037	0.076		4171

Notes: This table replicates results in Table 7 computing unadjusted p-values using randomization inference. 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, using randomization inference. 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 B8: Treatment effect on classroom environment, randomization inference

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	0.078	0.195	160
Bullying in class, teacher	-0.038	0.105	0.153	0.451	0.451	160
Disruptiveness, enumerator	-0.186	0.389	0.148	0.011	0.055	167
Delay in class's start (minutes)	9.938	1.204	1.046	0.301	0.376	160
Average decibels during class	0.022	0.681	0.487	0.261	0.435	169
Standardized Treatment Effect	-0.1	0.308	0.095	0.001		169

Notes: This table replicates results in Table 8 computing unadjusted p-values using randomization inference. 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, using randomization inference. 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.

Appendix C Robustness checks for classes with at least one very disruptive student

Table C1: Treatment effect in classes with at least one very disruptive student, without controls

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: Very disruptive eligible students						
Disruptiveness, teacher	0.985	-0.325	0.355	0.359	0.611	86
Disruptiveness, enumerator	0.286	0.613	0.439	0.162	0.306	85
Spanish test score	-0.460	0.038	0.335	0.910	0.910	88
Math test score	-0.230	0.063	0.512	0.902	0.958	88
% class friends with student	0.051	0.025	0.012	0.042	0.103	109
Panel B: Not very disruptive eligible students						
Disruptiveness, teacher	0.294	0.451	0.128	0.000	0.007	391
Disruptiveness, enumerator	0.162	0.103	0.127	0.417	0.644	393
Spanish test score	-0.349	-0.176	0.106	0.095	0.202	397
Math test score	-0.349	-0.092	0.167	0.581	0.823	397
Friends with ≥ 1 very dis.	0.065	0.075	0.030	0.012	0.049	397
Panel C: Ineligible students						
Disruptiveness, teacher	-0.205	0.509	0.185	0.006	0.034	1517
Disruptiveness, enumerator	-0.093	0.172	0.077	0.025	0.086	1576
Spanish test score	0.035	-0.053	0.141	0.707	0.858	1579
Math test score	0.115	-0.031	0.177	0.862	0.977	1579
Friends with ≥ 1 very dis.	0.067	0.015	0.028	0.584	0.764	1577
Panel D: Class-level outcomes						
Disruptiveness, teacher	-0.250	0.669	0.236	0.005	0.039	72
Disruptiveness, enumerator	-0.250	0.516	0.245	0.035	0.100	76

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator. To account for the fact the randomization is stratified, the regressions in Panels A, B, and C have lottery fixed effects, while in the regressions in Panel D we use propensity score reweighting. 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.

Table C2: Treatment effect in classes with at least one very disruptive student, with extra controls

Variables	Control (1)	T-C (2)	S.E. (3)	Unadj. P (4)	Adj. P (5)	N (6)
Panel A: Very disruptive eligible students						
Disruptiveness, teacher	0.985	-0.097	0.403	0.811	0.984	86
Disruptiveness, enumerator	0.286	0.226	0.500	0.652	1	85
Spanish test score	-0.460	0.109	0.354	0.759	0.992	88
Math test score	-0.230	-0.076	0.576	0.895	1	88
% class friends with student	0.051	0.011	0.018	0.544	1	109
Panel B: Not very disruptive eligible students						
Disruptiveness, teacher	0.294	0.482	0.148	0.001	0.019	391
Disruptiveness, enumerator	0.162	0.074	0.130	0.567	0.963	393
Spanish test score	-0.349	-0.196	0.101	0.052	0.146	397
Math test score	-0.349	-0.002	0.176	0.991	0.991	397
Friends with ≥ 1 very dis.	0.065	0.075	0.032	0.019	0.108	397
Panel C: Ineligible students						
Disruptiveness, teacher	-0.205	0.476	0.183	0.009	0.078	1517
Disruptiveness, enumerator	-0.093	0.122	0.082	0.136	0.331	1576
Spanish test score	0.035	0.012	0.124	0.922	0.979	1579
Math test score	0.115	0.057	0.152	0.709	1	1579
Friends with ≥ 1 very dis.	0.067	0.017	0.023	0.448	0.952	1577
Panel D: Class-level outcomes						
Disruptiveness, teacher	-0.250	0.543	0.261	0.038	0.161	72
Disruptiveness, enumerator	-0.250	0.492	0.246	0.045	0.153	76

Notes: This table reports results from OLS regressions of several dependent variables on a treatment indicator and control variables. The control variables include those selected by a Lasso regression of the dependent variable on potential controls, following Belloni et al. [2014], the variables imbalanced at baseline in the relevant subsample, and the baseline value of the outcome variable. To account for the fact the randomization is stratified, the regressions in Panels A, B, and C have lottery fixed effects, while in the regressions in Panel D we use propensity score reweighting. 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.

Appendix D Comparing SFL to other socio-emotional and non-cognitive skills interventions

In this section, we conduct a thorough and systematic comparison of SFL and similar interventions studied in the psychology and economics literature.

D.1 Comparing SFL to SEL interventions studied in the psychology literature

In the psychology literature, Payton et al. [2008] conduct a meta-analysis of 80 selected SEL interventions. To our knowledge, this is the only meta-analysis reporting effects separately for selected SEL interventions comparable to SFL. They find that selected SEL interventions produce large effects: they reduce conduct problems by 0.47σ, and respectively improve emotional stability and academic performance by 0.50σ and 0.43σ.

Table D1 compares SFL to the selected SEL interventions reviewed in Payton et al. [2008]. Many features of the interventions reviewed in Payton et al. [2008] are readily available from Table 7 therein. We reviewed a random sample of 25 of the meta-analysis's papers and manually collected features that seemed important to us but were not reported in the paper. They appear in italic in Table D1.

Table D1: Comparing "Skills for Life" to the selected SEL interventions in Payton et al. [2008]

	Skills for Life	Payton et al. [2008]
Panel A: Intervention Intensity		
Number of sessions	10	12 (median)
Sessions' duration in minutes	120	50 (median)
Intervention duration in weeks	10	10 (median)
Number of students per workshop	7.2	6 (median)
Parental training	Yes	41%
Parental sessions	3	14 (median)
Parents' attendance	34%	49% (median)
Students pulled out of class	Yes	100%
Panel B: Study design		
Random assignment of treatment	Yes	80%
Journal's impact factor (for published studies)	NA	4.01 (median)
Outcomes based on teacher ratings per study	3	0.6
Outcomes based on enumerator ratings per study	2	0.3
Outcomes based on student ratings per study	5	1.2875
Outcomes based on parent ratings per study	0	0.175
Outcomes based on school records per study	1	0.35
Uses validated psychometric scale as outcome	Yes	69%
Weeks between end of intervention and endline	3	1 (median)
Panel C: Targeting of eligible students		
Primary school students	Yes	69%
Students with conduct problems	Yes	48%
Students with emotional problems	Yes	23%
Students with conduct and emotional problems	Yes	29%
Low SES students	Yes	73%
Exclusion of students with psychological disorder or very disruptive	No	30%
Exclusion of students with cognitive problems	No	4%
Panel D: Location		
United States	No	85%
High-income country	No	100%
Panel E: Delivery Personnel, Monitoring of Delivery, and Intervention Scale		
Intervention delivered by:		
Researchers (alone, or together with school staff)	No	43%
School staff trained and monitored by researchers	No	22%
Other personnel trained and monitored by researchers	No	35%
Frequency at which delivery is monitored:	Never	Weekly (median)
Number of treated students	8,570	36 (median)

Notes: This table compares the "Skills for Life" intervention to those in the meta-analysis of Payton et al. [2008]. For the meta-analysis's papers, the variables in italic were collected manually by the authors, by reviewing a random sample of 25 of the 80 articles reviewed by Payton et al. [2008]. The variables not in italic are directly available from Table 7 in Payton et al. [2008]. SFL's number of treated students is for 2013.

SFL's intensity is comparable to that of the meta-analysis's interventions

Panel A of Table D1 shows that SFL's intensity is similar to that of the meta-analysis's interventions. The median number of sessions across those interventions is slightly higher than SFL's number of sessions (12 versus 10), but their sessions are typically shorter (50 versus 120 minutes). The number of students per workshop is comparable (a median of 6 in the meta-analysis, versus 7.2 on average in our sample). Their median duration is the same as SFL's (10 weeks). 59% of those interventions only include sessions with students, while 41% also include a parental training, like SFL. Only seven of the papers we reviewed give the number of parental sessions, but among those the median number of sessions (14) is higher than in SFL (three parental sessions). Only three of the papers we reviewed mention parents' attendance, but among those the median attendance (49%) is comparable to that in SFL (34%, see Table 2). In all those interventions, selected students are pulled-out of their class during the class day, as in the SFL intervention.

Our study design is comparable to that of the metanalysis's studies

Our study design is also comparable to that of the meta-analysis's studies. Panel B of Table D1 shows that the treatment was randomly assigned in 80% of those studies. Many of the published studies appeared in high-impact-factor peer-reviewed journals (median impact factor=4.01).[1] Most of their outcome measures are teacher, enumerator, and student ratings, often made using validated psychometric scales, as in our study. We measured our outcomes three weeks after the end of the SFL intervention, while in the reviewed interventions, the median number of weeks between the end of the intervention and endline data collection is equal to one.

Students receiving SFL may be harder to treat than those in the meta-analysis's interventions.

Panel C of Table D1 shows that on some dimensions, SFL targets similar students as the programs reviewed by Payton et al. [2008]. Like SFL, 69% of those interventions target primary school students. 48% target students with conduct problems, 23% target students with emotional distress, and the remaining interventions target students with a combination of problems. 73% target low SES students, like SFL.

However, an important difference is that 30% of those programs mention that they exclude students with a psychological disorder or very disruptive students. For the most part, the remaining programs do not explicitly say that they do not exclude those students, so 30% is a lower bound on the proportion of programs that do so. SFL, on the other hand, does not exclude those students. Moreover, Panel D of Table D1 shows that 85% of the interventions in Payton et al. [2008] take place in the US, and all take place in high-income countries. Recent epidemiological studies show that the prevalence rate of ADHD, a disorder correlated with conduct problems, is equal to 15.5% among primary school children in Chile (see [de la Barra et al., 2013]), against 8.7% in the US (see [Froehlich et al., 2007]), 3.5 to 5.6% in France (see [Lecendreux et al., 2011]) or 3% in Italy (see [Bianchini et al., 2013]). Similarly, surveys indicate that domestic violence, a cause of conduct disorder problems in children (see [Carrell and Hoekstra, 2010]), is more prevalent in Chile than in the US. 4.3% of Chilean women report having been physically assaulted by their partner over the previous year (see [Ministerio de Interior y de Seguridad Pública, 2017]), against 1.3% in the US (see [Tjaden and Thoennes, 2000]). Overall, students receiving SFL may be harder to treat

[1] 85% of the 80 studies reviewed by Payton et al. [2008] were published in peer-reviewed journals.

than those in the meta-analysis's interventions, both because conduct problems are more prevalent among primary-school-age children in Chile than in high-income countries, but also because many of the meta-analysis's interventions exclude the hardest-to-treat children.

SFL's delivery is less monitored than the meta-analysis's interventions

Panel E of Table D1 shows that SFL strikingly differs from the meta-analysis's programs in terms of delivery. All of the meta-analysis's interventions are demonstration programs, mounted by researchers for research purposes. 43% of the interventions are entirely or partly delivered by the researchers, 22% are delivered by school staff trained and supervised by the researchers, and 35% are delivered by other personnel (most often psychologists) hired, trained, and supervised by the researchers. 69% of the studies where the intervention was not entirely delivered by the researchers mention the frequency at which the researchers monitored the delivery personnel, for instance by attending sessions, or by reviewing video- or audio-recorded sessions. The median is a weekly monitoring.

JUNAEB provides SFL implementers with a detailed manual describing the content of each of the workshop's session. SFL employees also attend "good practices" meetings every six months, during which they share with other teams what seems to work in their sessions. However, JUNAEB does not systematically and frequently monitor each team's delivery. Of the three teams we interviewed, only one had a workshop observed over the last two years.[2]

D.2 Comparing SFL to non-cognitive skills interventions studied in the economics literature

Table D2 below describes programs intended to improve non-cognitive skills that have been studied in the economics literature (hereafter, NCS-ECON). We focus on interventions targeting non-cognitive skills, because few SEL interventions have been studied in the economics literature. To select the papers included in our review, we looked at Figure 1 in Sorrenti et al. [2020] and at a J-PAL review of CBT interventions (`https://www.povertyactionlab.org/es/node/4521`), and selected all interventions therein primarily targeting non-cognitive skills. Overall, we included seven papers studying eight interventions. The unit of analysis in Table D2 is an intervention, and some variables in Table D2 are missing for some interventions. Many of those interventions produce large effects, such as increases of 18 to 35% in high school graduation rates, [Algan et al., 2014, Heller et al., 2017, Oreopoulos et al., 2017, Sorrenti et al., 2020] or a decrease of 0.2σ in classroom disruption Sorrenti et al. [2020].

NCS-ECON programs are more intensive than SFL (the median number of sessions is 33.5 and the median duration in weeks is 39.5, see Panel A) and treat an older population (the median age is 12.5 years, and only 50% of programs treat primary school students, see panel B). Again, a large fraction of the NCS-ECON programs are conducted in high-income countries (75%, see Panel D). Researchers have often looked at long-run outcomes (the median is 108 weeks after treatment, see Panel C). NCS-ECON programs also seem to be conducted under the close supervision of program designers. 14% of the interventions are delivered by the researchers alone or by the researchers and school staff, 29% are delivered by other personnel trained and supervised by the researchers, and 43% of interventions are delivered by personnel of the NGO promoting the program (see Panel E).

[2]SFL employees also do not have monetary or non-monetary incentives tied to the quality of their workshops.

Only one intervention was delivered by government employees, as SFL, in a very different setting (in prisons).

Table D2: Non-cognitive skills programs studied in the economics literature

	Non-cognitive skills programs
Panel A: Intervention Intensity	
Number of sessions	33.5 (median)
Intervention duration in weeks	39.5 (median)
Parental training	14%
Parental sessions	46 (median)
Parents' attendance	38%
Program cost per beneficiary (2015 USD)	1,332 (median)
Panel B: Target Population	
Primary school students	50%
Age (in years)	12.5 (median)
Low SES students	100%
Panel C: Study design	
Random assignment of treatment	86%
Weeks between end of intervention and endline	108 (median)
Panel D: Location	
United States	25%
High-income country	75%
Panel E: Delivery Personnel	
Intervention delivered by:	
Researchers (alone, or together with school staff)	14%
Other personnel trained and monitored by researchers	29%
NGO personnel	43%
Government employees (penitentiary personnel)	14%
Number of treated students	1,296 (median)

Notes: This table displays characteristics of programs aiming to improve non-cognitive skills and studied in the economics literature. The variables in the table were collected manually, by reviewing the following papers: Alan et al. [2019], Algan et al. [2014], Blattman et al. [2017], Heller et al. [2017], Kosse et al. [2020], Oreopoulos et al. [2017], Sorrenti et al. [2020].

Appendix E Examples of SFL activities

SFL teaches socio-emotional skills to students through different types of activities, such as games, story-telling, and cooperative activities. An example of a game is the "my feelings" game, that teaches eligible students to recognize their feelings. In this game, two students get together and receive one deck of cards with emotions written (e.g. the word "happy") and drawn (e.g. a happy face) on them. Students alternate on who picks a card. The one picking the card has to express "physically" the feeling shown in the card. The other one has to imitate the feeling and guess what it is.

Other activities involve story-telling. For instance, in the "Pito the mole" story, students learn to recognize fear and anxiety, and sympathize with others experiencing those emotions. Students first have to put in order the drawings that tell the "Pito the mole" story. Then, SFL implementers read the story about a mole (Pito) who goes out in a rainy night to look for a friend. Pito's friend confuses Pito with a monster and runs away from him. But in the end, Pito's friend realizes that the monster is indeed Pito and they both laugh. As they tell the story, SFL implementers make pauses and discuss with students how the characters feel, what are the consequences of being afraid, whether fear is justified, among others things.

Finally, cooperative activities are often used to reinforce the importance of following the rules when students interact among themselves. For instance, in the "we all play" activity, there is a path segmented in boxes. Each student has to color and paint one box in the path. The game consists in crossing the path, but before starting, each student has to set one rule (e.g. you cannot move more than one box per round, or each box needs to have an activity associated). Usually, conflicts between students' rules emerge, and SFL implementers help students solve those conflicts logically. In this activity, students learn the relevance of following the rules and understand why some rules are more appropriate than others.

Overall, SFL's manual covers 15 different activities in 40 pages. For each activity, the objective and the materials needed are described. The activity description also guides implementers on the discussion that they should have with students, encouraging them to pursue specific objectives in each part of the activity.

The activities described above present some remarkable similarities with other SEL programs, such as the Promoting Alternative Thinking Strategies (PATHS) program [Sorrenti et al., 2020]. For instance, students participating in PATHS learn to recognize and express their emotions by using cards that have emotions written and drawn on them, as in the SFL "my feelings" game. PATHS students also learn how to control themselves by listening and discussing a story about a girl who learned how to control herself by calming down and recognizing her emotions, as in the SFL "Pito the mole" story.

Appendix F SFL's cost

JUNAEB does not have an estimate of the total cost of the program per treated student. Here are the two indirect methods we used to estimated that cost.

First, the 2014 budget of one of the municipal teams in our sample shows that its program implementers earned on average 7.42 USD per hour in 2014. Then, based on interviews with two implementers, we estimated that it takes 149 hours of work to implement an SFL workshop. This includes the 52 hours that the two workshop implementers spend delivering 13 two-hours sessions to students and their parents, but also the time that they spend: preparing the sessions and buying the material they need; going to and returning from the school for each session; preparing the reporting documents JUNAEB asks them to send for each workshop; meeting with the school principal and 2nd grade teachers prior to the start of the workshop, to agree on the schedule and location of the workshop; and interview 1st grade teachers to fill the TOCA questionnaire for each of their students the year before the workshop. Then, the team's budget shows that digitizing the 2014 TOCAs of all the first grade students in the town costed 860 USD. Divided by the 10 workshops conducted that year, that leads to a cost of 86 USD per workshop. Implementers also received transportation vouchers worth 63 USD per workshop. Finally, the cost of the material needed for the workshop activities is estimated at 188 USD per workshop, based on a detailed list of all the items bought for a workshop provided by the implementers we interviewed. Overall, we estimate the total cost of a workshop at $7.42\times149+86+63+188=1{,}443$ USD. The team whose budget we used had 7.2 students per workshop in 2014, which finally yields our estimated cost of 200 USD per treated student. This estimate relies on one team's budget. Costs may vary between teams, but we do not have reasons to suspect that the program's average cost is orders of magnitude away from our estimate.

Second, the Chilean government reports that in 2015, SFL's budget was 5,687,985,000 Chilean pesos, while 252,695 individuals benefited from the program.[3] Per the government's definition, beneficiaries include all 2nd grade students in a school where an SFL workshop took place, their parents, and their teachers. Assuming that only one parent is counted in the beneficiary population, and that teachers account for a negligible fraction of beneficiaries, we obtain that 126,482 2nd graders (252,695/2) were in a school where an SFL workshop took place. The government also reports that around 19% of second graders are eligible for SFL's workshops, and that 85% of eligible students attend a workshop. This means that 20,426 2nd graders attended a workshop in 2015 ($126{,}482\times0.19\times0.85$). Then, SFL's cost was 278,445 Chilean pesos per treated student in 2015. Converting into dollars yields an estimated cost of 458 USD.

[3] https://programassociales.ministeriodesarrollosocial.gob.cl/pdf/2015/PRG2015_5_61477.pdf

Appendix G Measurements' quality, and departures from our pre-analysis plan

Some of the dimensions we are trying to measure are hard to observe. To get a sense of the reliability of our measures, Table G1 shows their baseline-endline correlation in the control group. Students' Spanish and mathematics test scores have high positive baseline-endline correlations, above 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 has a baseline-endline correlation of 0.32. Our school happiness, self-esteem, and self-control measures respectively have baseline-endline correlations of 0.22, 0.13, and 0.14.

Turning to disruptiveness measures, the rating of students' disruptiveness by teachers has a baseline-endline correlation of 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,[4] so our baseline and endline measures were not made by the same teacher. This suggests that students' disruptiveness is relatively stable, and that different teachers tend to agree in their ratings. Then, Table G2 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 G1 shows that teachers' rating of the disruptiveness of the class also has a high baseline-endline correlation, equal to 0.50.

In our PAP, we had planned to use the average of the two enumerators' ratings of a student's disruptiveness as our enumerator disruptiveness rating. However, this measure 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 different standards to assign a given grade on the disruptiveness scale. Therefore, we depart from our PAP, and slightly modify our measure. We start by regressing enumerators' ratings 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 use the average of those residuals, across the two enumerators that have rated a student, as our enumerators' rating. This modified measure is the difference between a student's average rating by the two enumerators and the average of the ratings made by the same enumerators in the control group. Panel A of Table G1 shows that it has a positive and significant baseline-endline correlation equal to 0.13, and Panel A of Table G2 shows that it correlates well with teachers' ratings, and reasonably well with students' academic ability. Overall, enumerators' ratings of students' disruptiveness seem noisier than teachers', but they are still meaningful. Then, Panel B of Table G1 shows that enumerators' ratings of classes' disruptiveness have a relatively high baseline-endline correlation, around 0.25, and Panel B of Table G2 shows that this measure correlates well with teachers' ratings. Contrary to teachers' ratings, enumerators' ratings are blinded: enumerators do not know if the class they observe has been treated or not.[5]

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 again from our PAP, and net out enumerators' fixed effects from

[4] 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.

[5] Previous literature on SEL interventions has also relied on non-blinded teacher ratings (see [Payton et al., 2008]).

decibel measures, exactly as we did for enumerators' disruptiveness ratings. This new measure has a higher baseline-endline correlation than the measure described in our PAP, though Table G1 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 G2.

Table G1: 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	1735
Self-control score	0.141	0.000	1816
Self-esteem score	0.134	0.000	1841
Disruptiveness, teacher	0.419	0.000	1782
Disruptiveness, enumerator	0.126	0.000	1871
% school days missed	0.033	0.084	2751
Spanish test score	0.525	0.000	1897
Math test score	0.508	0.000	1897
% class friends with student	0.324	0.000	2245
Friends' average ability	0.408	0.000	1644
Friends' average disruptiveness	0.349	0.000	1502
No friends in the class	0.099	0.000	2245
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 G2: Correlations between baseline disruptiveness measures

	Correlation (1)	P-value (2)	N (3)
Panel A: student-level measures			
Enumerator 1 - enumerator 2	0.504	0.000	4075
Teacher - enumerator	0.293	0.000	4035
Teacher dis. - avg. test score	-0.277	0.000	4139
Enumerator dis. - avg. test score	-0.17	0.000	4594
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.

Appendix references

Sule Alan, Teodora Boneva, and Seda Ertac. Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics*, 134(3): 1121–1162, 2019.

Yann Algan, Elizabeth Beasley, Frank Vitaro, Richard E Tremblay, et al. The impact of non-cognitive skills training on academic and non-academic trajectories: From childhood to early adulthood. *Sciences Po Working Paper*, 2014.

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.

Rio Bianchini, Valentina Postorino, Rita Grasso, Bartolo Santoro, Salvatore Migliore, Corrado Burlò, Carmela Tata, and Luigi Mazzone. Prevalence of adhd in a sample of italian students: a population-based study. *Research in developmental disabilities*, 34(9):2543–2550, 2013.

Christopher Blattman, Julian C Jamison, and Margaret Sheridan. Reducing crime and violence: Experimental evidence from cognitive behavioral therapy in liberia. *American Economic Review*, 107(4):1165–1206, 2017.

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.

Flora Eloísa de la Barra, Benjamin Vicente, Sandra Saldivia, and Roberto Melipillan. Epidemiology of adhd in chilean children and adolescents. *ADHD Attention Deficit and Hyperactivity Disorders*, 5(1):1–8, 2013.

Tanya E Froehlich, Bruce P Lanphear, Jeffery N Epstein, William J Barbaresi, Slavica K Katusic, and Robert S Kahn. Prevalence, recognition, and treatment of attention-deficit/hyperactivity disorder in a national sample of us children. *Archives of pediatrics & adolescent medicine*, 161 (9):857–864, 2007.

Sara B Heller, Anuj K Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold A Pollack. Thinking, fast and slow? some field experiments to reduce crime and dropout in chicago. *The Quarterly Journal of Economics*, 132(1):1–54, 2017.

Fabian Kosse, Thomas Deckers, Pia Pinger, Hannah Schildberg-Hörisch, and Armin Falk. The formation of prosociality: causal evidence on the role of social environment. *Journal of Political Economy*, 128(2):434–467, 2020.

Michel Lecendreux, Eric Konofal, and Stephen V Faraone. Prevalence of attention deficit hyperactivity disorder and associated features among children in france. *Journal of Attention Disorders*, 15(6):516–524, 2011.

Ministerio de Interior y de Seguridad Pública. *Tercera encuesta nacional de violencia intrafamiliar contra la mujer y delitos sexuales.* 2017.

Philip Oreopoulos, Robert S Brown, and Adam M Lavecchia. Pathways to education: An integrated approach to helping at-risk high school students. *Journal of Political Economy*, 125 (4):947–984, 2017.

John Payton, Roger P Weissberg, Joseph A Durlak, Allison B Dymnicki, Rebecca D Taylor, Kriston B Schellinger, and Molly Pachan. The positive impact of social and emotional learning for kindergarten to eighth-grade students: Findings from three scientific reviews. technical report. *Collaborative for Academic, Social, and Emotional Learning (NJ1)*, 2008.

Giuseppe Sorrenti, Ulf Zölitz, Denis Ribeaud, and Manuel Eisner. The causal impact of socio-emotional skills training on educational success. *University of Zurich, Department of Economics, Working Paper*, (343), 2020.

Patricia Godeke Tjaden and Nancy Thoennes. *Full report of the prevalence, incidence, and consequences of violence against women: Findings from the National Violence Against Women Survey.* US Department of Justice, Office of Justice Programs, National Institute of . . . , 2000.