Assessing the classroom wide impacts of a mental health program for disruptive students.

Pre-analysis plan.

Clement de Chaisemartin[1] and Nicolas Navarrete[2]

28/04/2017

Abstract

There has been a lot of research studying whether students can benefit from being exposed to cognitively more able peers. On the other hand, there has been much less research studying whether students can be hampered by disruptive peers. In this study we measure the effects of "Habilidades para la vida", a program aiming to improve the behaviour of the most disruptive students in 2nd grade in Chile. We will study the effects of this program on disruptive students, on their teachers, and on their non-disruptive classmates.

1. Introduction

This plan outlines the hypotheses to be tested and specifications to be used in the analysis of the impact of "Habilidades para la vida" (HPV) a program intended to improve the behaviour of disruptive 2nd grade students in Chile, and implemented by "Junta Nacional de Auxilio Escolar y Becas" (JUNAEB).

The authors completed the plan after baseline and follow-up data have been collected, but before JUNAEB provided the authors with the data that will allow them to identify disruptive students benefiting from the program in the classes participating in the experiment. JUNAEB officials will provide the authors with a signed certificate showing the date when they will have provided the authors with these data, as well as all the variables they will have provided to them. As most of the analysis presented below amounts to looking at the effects of this program separately for disruptive students and for their non-disruptive classmates, at the time we write this document we cannot conduct the analysis presented below, so the plan can provide a useful reference in evaluating the final results of the study.

The plan is outlined as follows: Section 2 reviews the motivation for the study and describes the HPV program. Section 3 describes the randomization, the data we will use in our analysis, and the study population. Section 4 describes the internal validity checks we will perform. Section 5 describes how we will measure treatment implementation and compliance with randomization. Section 6 enumerates the hypotheses we will test regarding the effect of the treatment in the whole sample. Section 7 describes the subgroup analysis we will perform.

2. Motivation and the HPV program

There has been a lot of research studying whether students can benefit from being exposed to cognitively more able peers. On the other hand, there has been much less research studying whether students can be hampered by disruptive peers. This study is the first evaluation of the

[1] UC Santa Barbara, clementdechaisemartin@ucsb.edu
[2] Warwick University, n.a.navarrete-hernandez@warwick.ac.uk

classroom-wide effects of a program aimed at improving the behaviour of the most disruptive students in a classroom. Some papers have already looked at the effects of such programs on treated students, but no paper has ever looked at their effects on non-disruptive classmates and teachers.

100 000 students in the most vulnerable schools of Chile benefit from HPV each year, thus making it the largest school behavioural program in the world. In the end of 1st grade, the HPV municipal teams have 1st grade teachers fill the Teacher Observation of Classroom Adaptation questionnaire (see Kellam et al., 1977, and Werthamer-Larsson et al., 1989) for each of their student. Based on this questionnaire, students receive scores on the following 6 scales: authority acceptance (AA), social contact (SC), motivation for schooling (MS), emotional maturity (EM), attention and focus (AF), and activity levels (AL). Students scoring above some threshold in the AA, AF, and AL scales and below some threshold in the MS scale are categorized in the "blue" profile (aggressive, disobedient, and hyperactive). Students scoring below some threshold in the SC scale and above some threshold either in the AA or AL scale are categorized in the "green" profile (shy, disobedient, and hyperactive or aggressive). Students scoring below some threshold in the CS, MS, and EM, and either in the AA or AL scales are categorized in the "yellow" profile (obedient, shy, and immature). "Blue", "green", and "yellow" students are all assigned to the HPV treatment in 2nd grade. In practice, "yellow" students account for 7% of eligible students, while "blue" and "green" students respectively account for 40% and 53%. Overall, the HPV program mostly targets disruptive students.

In 2nd grade, students assigned to the program and enrolled in schools with 6 or more disruptive students follow 10 weekly group sessions. These 10 sessions each last 2 hours and they take place over the course of one school semester. During these sessions, 2 psychologists try to help students achieve the following things: respect themselves and respect others; recognize their feelings and share them with others; manage anger and find non-violent solutions to conflicts. Sessions consist of activities (games, role play, drawing, or singing). Groups cannot bear strictly more than 10 students: if a school has strictly more than 10 eligible students, 2 groups of eligible students are formed.

3. Randomization, data, study population, and definition of the treatment dummy.

Our sample consists of 172 classes. All municipal teams of the HPV program in the Santiago and Valparaiso regions 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 these municipalities, and collected data on the number of students eligible for the program enrolled in each 2nd grade class. 172 classes with 4 or more eligible students and in schools with 6 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 spill-over effects.

Randomization took place both within schools and within municipalities. There were 29 schools with two classes included in our population and where it was possible to form 2 groups of 6 students or more without grouping students of the two classes together. In such instances, we conducted a lottery within the school, to assign one of the two classes to receive the treatment in the first semester of 2015, and the second class to receive it in the second semester. For the remaining 114 classes in our study population, 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, and 83 to receive it in the second semester.

In our analysis, we will use: baseline data we collected in March 2015, before the beginning of the first semester group sessions; endline data we collected in August-September 2015, after the end of the first semester and before the beginning of the second semester group sessions; data we will collect on treatment implementation; data produced by JUNAEB.

For each class, baseline and endline data consist of:

- A list of students enrolled in the class (Class list baseline: CLB, Class list endline: CLE)
- A student questionnaire (SQB, SQE)
- Students' standardized tests in Spanish and mathematics (STSB, STSE, STMB, STME).
- A student sociogram whereby students identify students they usually study with, play with, and students who they think give good opinions in class (SSB, SSE).
- Two class maps filled by the two enumerators we sent to the class to collect data, and including 6 observations of the behaviour of each student during one lecture (CMB, CME).
- Two questionnaires filled by our enumerators (EQB, EQE).
- A teacher questionnaire (TQB, TQE).
- Two 50 minutes recordings of the decibels levels in the class made by our enumerators (DLB, DLE).

Our questionnaires are available upon request.

The variables we will collect on treatment implementation is merely the number of group sessions that had taken place in each class before endline data collection.

The variables that JUNAEB has agreed to provide to us are:

- The TOCA scores collected for 1st and 2nd graders in 2013, 2014, and 2015.
- The Pediatric symptoms checklist scores (see Jellinek et al., 1988) collected for 1st and 2nd graders in 2013, 2014, and 2015.
- Some socio-demographic variables about students and their family: gender of the student, age of the mother at birth (in brackets), whether the student lives with her biological father, whether her family benefits from the "Chile Solidario" program, the vulnerability index of her family, her parents' income, the copayment rate her family has to pay when purchasing drugs or health services in the public health system, her mother's and father's education, whether she benefits from the Chilean free lunch program (Programa de Alimentacion escolar).
- Some students' schooling outcomes: whether the student repeated a grade in 2014 and 2015, and students' average monthly school attendance in 2015 school year.
- Some variables about treatment implementation: number of group sessions attended by each student during the first and the second semester.

In case JUNAEB can finally not provide us some of those variables, we will still follow this pre-analysis plan except that these variables will be dropped from all the estimations described below. In case one of those variables is missing for more than 50% of students, it will also be dropped.

In our analysis, we will use class-level variables, student-level variables, and teacher-level variables. Our population of classes are the 172 classes included in the randomization. Our population of students are all students appearing either on the CLB or CLE of one of these 172 classes. Similarly, our population of teachers will be teachers teaching in one of these 172 classes, either in baseline or in endline.

In all of what follows, we will distinguish between two groups of students. Students eligible for the treatment will be referred to as "disruptive students", while other students will be referred to as "classmates". Coming up with a definition of disruptive students comprehensive and comparable in the treatment and in the control group raises some issues. The HPV municipal teams do not have access to a centralized data base with the TOCA scores of all students in schools covered by the program in 2014. Moreover, students in schools not covered by the program do not have a TOCA score in 2014. Therefore, the teams need to ask teachers to fill the TOCA questionnaire in 2015 for students who are new in the school and who do not come from a school in the same town covered by the program. In classes where the group sessions took place in the first semester, this process took place in March 2015. In classes where the group sessions took place in the second semester, this process sometimes took place at a later point in the semester because there was no rush to make these measures. Therefore, it might be the case that incoming students assigned to the program based on their 2015 TOCA score are not comparable in the treatment and in the control group. For instance, teachers in the control group might have known better these incoming students at the time they filled their TOCA questionnaire.

Therefore, before defining the group of disruptive students we will conduct the two following tests. First, we will compare the share of incoming students assigned to the program in the treatment and in the control group, as well as the baseline characteristics (the same as in H4 below) of incoming students assigned to the program in the two groups.

If these two tests do not show systematic differences between the two groups, we will define disruptive students as students belonging to either of the two following subgroups of students:

Subgroup a)
Students:
i) appearing on the CLB[3] of one of the 172 classes;
ii) who either did not change school between 2014 and 2015 or changed school but were in a school covered by the program and in the same municipality in 2014;
iii) who were declared eligible to the program in 2014 based on their TOCA score.

Subgroup b)
Students:
i) appearing on the CLB of one of the 172 classes;
ii) who changed school between 2014 and 2015 and were not in a school covered by the program and in the same municipality in 2014;

[3] HPV teams report that they very rarely assign to first semester groups sessions a student joining the school after March, even if that student comes from a school they cover and is at risk.

iii) who were declared eligible to the program in 2015 based on their TOCA score (variable AT1 in the 2015 TOCA dataset for 2nd graders);

On the other hand, if these two tests show systematic differences between the two groups, throughout the paper we will define disruptive students as subgroup a) only. Even if these two tests are satisfied, as a robustness check we will also estimate our main specifications considering subgroup b) as classmates instead of disruptive students.

In all the class- or teacher-level regressions described below, we will define the treatment group dummy as a variable equal to 1 for the 89 classes assigned to receive the treatment in the first semester. In all the student-level regressions, the treatment group dummy will be a variable equal to 1 for students appearing in the CLB of one of the 89 classes assigned to receive the treatment in the first semester, and for students not appearing in any CLB and appearing in the CLE of one of the 89 classes assigned to receive the treatment in the first semester.

4. Internal validity checks

4.1. Questions with Limited Variation and standardization

In order to limit noise caused by variables with minimal variation, variables listed in the remainder of this plan and for which 95 percent or more of observations have the same value within the relevant sample will be omitted from the analysis and will not be included in any indicators or hypothesis tests. In the event that omission decisions result in the exclusion of all constituent variables for an indicator, the indicator will be not be calculated.

For all the standardized scores, standardization will be done using the mean and standard deviations of the variables in the entire population.

4.2. Statistical methods

a) For variables measured at the student level

<u>Estimation method</u>

For each of the student-level variables listed below, we will run an OLS regression of that variable on:

-a dummy for students in the treatment group

-56 dummies for each randomization group

<u>Standard errors</u>

Cluster-robust standard errors, clustered at the class level.

<u>Adjustment for multiple testing</u>

For each variable, we will report both the unadjusted p-value of the coefficient of the treatment variable, and the p-value adjusted for control of the False Discovery Rate (see Benjamini, Y. and Y. Hochberg, 1995) within each hypothesis.[4]

b) For variables measured at the class level

Estimation method

For each of the class- or teacher-level variables listed below, we will run an OLS regression of that variable on a dummy for classes in the treatment group.

Let D denote the treatment group dummy, and let S denote the lottery within which a class was included. In these regressions, each treated class will be weighted by the square-root of P(D=1)/P(D=1|S). Each control class will be weighted by the square-root of P(D=0)/P(D=0|S). This propensity score reweighting will ensure that the coefficient of the dummy for treatment is identified out of comparisons of treated and control classes within the same lottery, without having to include the dummies for the 56 lotteries we conducted in the regression.

Standard errors

Heteroskedasticity-robust standard errors.

Adjustment for multiple testing

For each individual outcome, we will report both the unadjusted p-value of the coefficient of the treatment variable, and the p-value adjusted for control of the False Discovery Rate (see Benjamini, Y. and Y. Hochberg, 1995) within each hypothesis.

4.3. *Hypothesis tested*

H1: Among disruptive students, attrition is balanced between the treatment and control groups

Dummy for classes observed at endline.

Number of disruptive students per class at endline.

Dummy for students not appearing on CLB.[5]

Dummy for students appearing on CLE, within the sample of students appearing on CLB.

Dummy for whether a student took the SQE, the STSE & STME, and the SSE, and has a sixth observation in the CME by at least one enumerator.

Dummy for whether a student's teacher answered question 0 for that student in the TQE.

If worrying levels of attrition are found in these last two tests, we will adjust for the potential effect of such attrition using Lee bounds.

[4] All the hypothesis we test are denoted by H1, H2, etc. in what follows.

[5] This will measure the effect of the program on the probability that a student joins the school between baseline and endline. If we were to find a significant effect here, we will drop students appearing only on the CLE from our analysis.

H2: Among classmates, attrition is balanced between the treatment and control groups

The analysis will be the same as in H1, within the sample of classmates.

H3: Among teachers, attrition is balanced between the treatment and control groups

Dummy for whether a teacher is teaching in endline the same class as in baseline.

Dummy for whether a teacher filled the TQE.

If worrying levels of attrition are found in these two tests, we will adjust for the potential effect of such attrition using Lee bounds.

H4: Baseline characteristics of disruptive students are balanced between the treatment and the control group.

Happiness in school: question 12 in SQB standardized.

Self-control: standardized score constructed from questions I6-I10 in SQB.

Self-esteem: standardized score constructed from questions I1-I5 in SQB.

Students' disruptiveness as measured by their teacher: question 0 of TQB, standardized.

Pollsters' assessment of students' disruptiveness: average of observations 6 by enumerators 1 and 2 in the CMB, standardized.

Spanish score: percentage of correct answers across all questions in STSB, standardized.

Maths score: Percentage of correct answers across all questions in STMB, standardized.

Normalized degree centrality in the friendship network: percentage of her classmates who filled the SSB and who named a student in the second column of the SSB.

Average of Spanish and math test scores at baseline of endline friends.

Average disruptiveness at baseline of endline friends, where the disruptiveness score is constructed as explained in S13 below.

Dummy for students who took the SSB and who were nominated as friend by no other student.

Distance between student and teacher in CMB: square-root of (row student – row teacher)^2 + (column student – column teacher)^2.

2014 AA TOCA score, standardized (2013 AA TOCA score if 2014 TOCA is not available).

2014 SC TOCA score, standardized (2013 SC TOCA score if 2014 TOCA is not available).

2014 MS TOCA score, standardized (2013 MS TOCA score if 2014 TOCA is not available).

2014 EM TOCA score, standardized (2013 EM TOCA score if 2014 TOCA is not available).

2014 AF TOCA score, standardized (2013 AF TOCA score if 2014 TOCA is not available).

2014 AL TOCA score, standardized (2013 AL TOCA score if 2014 TOCA is not available).

Summary question A, standardized ("Puntaje Global A") in 2014 TOCA questionnaire (Summary question A in 2013 TOCA questionnaire if 2014 TOCA is not available).

Summary question B, standardized ("Puntaje Global B") in 2014 TOCA questionnaire (Summary question B in 2013 TOCA questionnaire if 2014 TOCA is not available).

2014 PSC score, standardized (2013 PSC score if 2014 PSC is not available).

Dummy for whether student is a boy.

Dummy for whether the student's mother was below 18 when the student was born.

Dummy for whether the student's mother was above 36 when the student was born.

Dummy for whether the student's lives with his biological father.

Dummy for whether the student's family is in the "Chile solidario" program.

Vulnerability index of the student's family.

Copayment rate the student's family has to pay when purchasing drugs or health services in the public health system.

Mother's education.

Father's education.

Percentage of school days missed in March 2015.

Dummy for students who repeated a grade.

H5: Baseline characteristics of disruptive students are balanced between the treatment and the control group when we restrict the sample to students for which SQE, SSE, STSE&STME, and a sixth observation in CME by at least one enumerator are available.

Same variables as in H4.

H6: Baseline characteristics of disruptive students are balanced between the treatment and the control group when we restrict the sample to students for which question 0 of TQE is available.

Same variables as in H4.

H7: Baseline characteristics of classmates are balanced between the treatment and the control group.

Same variables as in H4, within the sample of classmates.

H8: Baseline characteristics of classmates are balanced between the treatment and the control group when we restrict the sample to students for which SQE, SSE, STSE&STME, and a sixth observation in CME by at least one enumerator are available.

Same variables as in H7.

H9: Baseline characteristics of classmates are balanced between the treatment and the control group when we restrict the sample to students for which question 0 of TQE is available.

Same variables as in H7.

H10: Baseline characteristics of teachers are balanced between the treatment and the control group

Teacher gender: question 1 in TQB

Teacher age: question 2 in TQB

Teacher qualifications: dummy for whether the teacher has a university degree or is currently being trained to receive a university degree (question 3 in TQB)

Teacher's experience: question 8 in TQB

Teacher's experience in this school: question 9 in TQB

Teacher's absenteeism: question 11 in TQB (categorical variable, we will take the lowest value of each category to transform it into a discrete variable)

Teachers' taste for their job: standardized score formed from questions 21.1 and 21.3 in TQB

Teacher's confidence to make a difference in students' life: standardized score formed from questions 20 and 21.2 in TQB

Teacher's level of stress: standardized score formed from questions 28-29 in TQB

Teacher's level of happiness: standardized score formed from questions 30-31 in TQB

Teacher's effort to prepare lectures: question (15+16)/14 in TQB

Teacher's effort to implement a variety of pedagogical methods in the classroom: standardized score formed from questions 17, 18, and 19 in TQB

Teacher's amount of control on his/her life: standardized score formed from questions 32-33 in TQB.

H11: Baseline characteristics of teachers are balanced between the treatment and the control group when we restrict the sample to teachers who filled the TQE.

Same variables as in H10.

H12: Baseline characteristics of classes are balanced between the treatment and the control group

Academic level of the class assessed by teacher: standardized score formed from questions 12 and 13 in TQB.

Students' disruptiveness assessed by teacher: standardized score formed from questions 22, 23.3 to 23.7 in TQB.

Prevalence of bullying in class assessed by teacher: standardized score formed from question 27 in TQB.

Students' disruptiveness assessed by enumerators: standardized score from questions 1 to 9 in EQB.

Classroom average decibel levels: variable constructed from DLB.

Number of minutes between the moment the class is supposed to start and the moment it actually starts: variable constructed from CMB.

5. Treatment implementation

5.1. The strike and the measurement of treatment implementation.

In some classes, the first-semester group sessions were delayed due to a teachers' strike, and ended in the beginning of the second semester. However, in these municipalities the HPV teams accepted to delay the start of the second semester group sessions. Therefore, in almost all the municipalities we could make our measures after the end of the group sessions in the treatment group and before the start of the group sessions in the control group. Hence, by comparing outcomes at endline in the two groups, we will measure the full effect of the program.

Still, the strike makes it more complicated for us to measure perfectly students' actual exposure to the treatment at endline. JUNAEB produces comprehensive data sets where they record the attendance of each student to their group sessions. Unfortunately, these data sets go by semester, and they do not include the date at which each session took place. Because of the strike, comparing the number of sessions followed by disruptive students in the treatment and in the control group in the end of the first semester will underestimate the true differential exposure to treatment in the two groups at endline, because in some municipalities disruptive students in the treatment group followed some sessions between the end of the first semester and endline.

To address this issue, for all the treatment group classes where the group sessions could not be terminated before the end of the first semester, we will collect from the municipal teams the date at which they conducted the remaining sessions during the second semester. By matching this information with the date at which endline took place in each class, and with JUNAEB data on students' attendance to 2nd semester group sessions, we will be able to measure perfectly the number of sessions each student had attended before endline.

5.2. Statistical methods

a) For variables measured at the student level

Same as in 4.2.

b) For variables measured at the class level

Same as in 4.2.

5.3. *Hypothesis tested*

H13: At endline, disruptive students have attended more group sessions in the treatment than in the control group, and classmates have barely attended any session in any group.

Dummy for whether at least one group session was conducted in each class before endline.

Number of group sessions conducted in each class before endline.

Dummy for whether disruptive students attended at least one group session before endline.

Number of group sessions attended by disruptive students before endline.

Dummy for whether classmates attended at least one group session before endline.

Number of group sessions attended by classmates before endline.

6. Effects of the treatment

6.1. Statistical methods

a) For outcomes measured at the students' level, and regressions estimated in the subsample of disruptive students.

Estimation method

For each of the students-level outcomes listed below, we will run an OLS regression of that variable on:

-a dummy for students in the treatment group

-56 dummies for each randomization group

-some student-level controls that will get selected through the following procedure. We will estimate a Lasso regression of the outcome on the 32 baseline student-level variables listed in H4, and on dummies for each randomization group, within the sample of disruptive students. We will use the subset of the 32 baseline student-level variables selected by the Lasso regression as our student-level controls. In the context of a randomized experiment, this procedure to select controls corresponds to the post-double selection procedure proposed in Belloni et al. (2014). Missing values of the selected controls will be replaced by the mean of these controls, and dummies for students for which these controls are missing will also be included in the regression.

-some class-level controls that will get selected through the following procedure. We will estimate a Lasso regression of the class-average (among disruptive students) of the outcome on the class-average (among disruptive students) of the 32 baseline student-level variables listed in H4 and on the 19 class-level variables listed in H10 and H12. We will also use propensity score reweighting to account for the fact that treatment probability varies across treatment groups (see part c) below for more details on the weights). We will use the subset of these 55 baseline class-level variables selected by the Lasso regression as our class-level controls. Missing values of the selected controls will be replaced by the mean of these controls, and dummies for students for which these controls are missing will also be included in the regression.

Standard errors

Cluster-robust standard errors, clustered at the class level.

Adjustment for multiple testing

For each individual outcome, we will report both the unadjusted p-value of the coefficient of the treatment variable, and the p-value adjusted for control of the False Discovery Rate (see Benjamini, Y. and Y. Hochberg, 1995) within each hypothesis. We will also report the unadjusted p-value of the standardized treatment effect within each hypothesis constructed following the same method as in Anderson (2008) and Haushofer and Shapiro (2013).

Robustness checks (to be presented in an appendix)

We will estimate the baseline specification dropping all the control variables except the 56 dummies for each randomization group.

We will compute the unadjusted p-value of the coefficient of the treatment variable in the baseline specification with control variables by using randomization inference.

b) For outcomes measured at the students' level, and regressions estimated in the subsample of classmates.

Estimation method

For each of the students-level variables listed below, we will run an OLS regression of that variable on:

-a dummy for students in the treatment group

-56 dummies for each randomization group

- some student-level controls that will get selected through the following procedure. We will estimate a Lasso regression of the outcome on the 32 baseline student-level variables listed in H4, and on dummies for each randomization group, within the sample of classmates. We will use the subset of the 32 baseline student-level variables selected by the Lasso regression as our student-level controls. Missing values of the selected controls will be replaced by the mean of these controls, and dummies for students for which these controls are missing will also be included in the regression.

- some class-level controls that will get selected through the following procedure. We will estimate a Lasso regression of the class-average (among classmates) of the outcome on the class-average (among classmates) of the 32 baseline student-level variables listed in H4 and on the 19 class-level variables listed in H10 and H12. We will also use propensity score reweighting to account for the fact that treatment probability varies across treatment groups (see part c) below for more details on the weights). We will use the subset of these 55 baseline class-level variables selected by the Lasso regression as our class-level controls. Missing values of the selected controls will be replaced by the mean of these controls, and dummies for students for which these controls are missing will also be included in the regression.

Standard errors

Same as in a)

Adjustment for multiple testing

Same as in a)

Robustness checks (to be presented in an appendix)

Same as in a)

c) For outcomes measured at the class level:

Estimation method

For each of the class-level variables listed below, we will run an OLS regression of that variable on:

-a dummy for classes in the treatment group

- some class-level controls that will get selected through the following procedure. We will estimate a Lasso regression of the class-average of the outcome on the class-average of the 32 baseline student-level variables listed in H4 and on the 19 class-level variables listed in H10 and H12. We will also use propensity score reweighting to account for the fact that treatment probability varies across treatment groups (see the next paragraph for more details on the weights). We will use the subset of these 55 baseline class-level variables selected by the Lasso regression as our controls. Missing values of the selected controls will be replaced by the mean of these controls, and dummies for students for which these controls are missing will also be included in the regression.

Let D denote the treatment group dummy, and let S denote the randomization group a class belongs to. In these regressions, each treated class will be weighted by the square-root of P(D=1)/P(D=1|S). Each control class will be weighted by the square-root of P(D=0)/P(D=0|S). This propensity score reweighting will ensure that the coefficient of the dummy for treatment is identified out of comparisons of treated and control classes within the same randomization group, without having to include the dummies for the 56 randomization groups in the regression.

Standard errors

Heteroskedasticity-robust standard errors.

Adjustment for multiple testing

Same as in a)

Robustness checks (to be presented in an appendix)

Same as in a)

6.2. *Hypothesis tested*

a) Effects on disruptive students

H14: The treatment has an effect on the emotional stability of disruptive students

Happiness in school: I6 in SQE, standardized

Self-control: standardized score constructed from II7-II9 in SQE

Self-esteem: standardized score constructed from II1 to II6 in SQE

Standardized treatment effect constructed following the same method as in Anderson (2008) and Haushofer and Shapiro (2013).

H15: The treatment has an effect on the disruptiveness of disruptive students

Teacher assessment of a student's disruptiveness: 0 in TQE, standardized

Pollsters' assessment of students' disruptiveness: average of observations 6 by enumerators 1 and 2 in the CME, standardized.

Standardized treatment effect constructed following the same method as in Anderson (2008) and Haushofer and Shapiro (2013).

H16: The treatment has an effect on the academic outcomes of disruptive students

Percentage of school days missed in the first semester of 2015 (April to July).

Spanish score: percentage of correct answers in STSE, standardized.

Maths score: percentage of correct answers in STME, standardized.

Standardized treatment effect constructed following the same method as in Anderson (2008) and Haushofer and Shapiro (2013).

H17: The treatment has an effect on the integration of disruptive students in the class network

Dummy for students who took the SSE and who were nominated as friend by no other student.

Normalized degree centrality in the friendship network: percentage of her classmates who filled the SSE and who named a student in the second column of the SSE.

Average of Spanish and math test scores at baseline of endline friends. Contingency plan: if the treatment has an effect on the proportion of students nominated by at least one friend, we can no longer study this outcome because the populations for which it is defined (students with at least one friend) are no longer comparable in the treatment and in the control group. Instead, we will look at the following measure. Normalized degree centrality in the friendship network, weighted according to students' baseline ability: percentage of her classmates who filled the SSE and who named a student in the second column of the SSE, where all classmates receive a weight proportional to the average of their score in the STSB and STMB.

Average disruptiveness at baseline of endline friends, where the disruptiveness score is constructed as explained in S13 below. Contingency plan: if the treatment has an effect on the proportion of students nominated by at least one friend, we can no longer study this outcome because the

populations for which it is defined (students with at least one friend) are no longer comparable in the treatment and in the control group. Instead, we will look at the following measure. Normalized degree centrality in the friendship network, weighted according to students' baseline disruptiveness: percentage of her classmates who filled the SSE and who named a student in the second column of the SSE, where all classmates receive a weight proportional to their disruptiveness score in baseline (the score will be constructed as explained in S13 below).

Standardized treatment effect constructed following the same method as in Anderson (2008) and Haushofer and Shapiro (2013).

b) Effects on teachers

H18: The treatment has an effect on teachers' job satisfaction and mental health

Teacher's taste for her job: standardized score formed from questions 25.1 and 25.3 in TQE

Teacher's confidence to make a difference in students' life: standardized score formed from questions 24 and 25.2 in TQE

Teacher's stress levels: standardized score constructed from question 33 and 34 in TQE.

Teacher's happiness levels: question 35 in TQE, standardized.

Standardized treatment effect constructed following the same method as in Anderson (2008) and Haushofer and Shapiro (2013).

H19: The treatment has an effect on teachers' effort

Teacher's effort to prepare lectures: question (19+20)/17 in TQE

Teacher's effort to implement a variety of pedagogical methods in the classroom: standardized score formed from questions 21, 22, and 23 in TQE

Standardized treatment effect constructed following the same method as in Anderson (2008) and Haushofer and Shapiro (2013).

H20: The treatment has an effect on which students are targeted by teachers

Teacher's target level of instruction: question 32 in TQE.

Distance between student and teacher in CME: square-root of (row student – row teacher)^2 + (column student – column teacher)^2, or the minimum of square-root of (row student – rowteacher)^2 + (column student – column teacher)^2 and square-root of (row student – row assistant)^2 + (column student – column assistant)^2 for classes with an assistant (question 12 in TQE). We will estimate this regression in the sample of disruptive students.

Standardized treatment effect constructed following the same method as in Anderson (2008) and Haushofer and Shapiro (2013).

c) Effects on classmates

H21: The treatment has an effect on the emotional stability of classmates

Same as H14, within the sample of classmates.

H22: The treatment has an effect on the disruptiveness of classmates

Same as H15, within the sample of classmates.

H23: The treatment has an effect on the academic outcomes of classmates

Same as H16, within the sample of classmates.

H24: The treatment has an effect on the integration of classmates in the class network

Same as H17, within the sample of classmates.

d) Effects on the classroom environment

H25: The treatment has an effect on the classroom environment

Students' disruptiveness assessed by teacher: standardized score formed from questions 26 in TQE.

Prevalence of bullying assessed by teacher: standardized score formed from questions 28 in TQE.

Students' disruptiveness assessed by enumerators: standardized score formed from questions 1 to 9 in the 2 EQE.

Classroom average decibel levels: variable constructed from DLE.

Number of minutes between the moment the class is supposed to start and the moment it actually starts: variable constructed from CME.

Standardized treatment effect constructed following the same method as in Anderson (2008) and Haushofer and Shapiro (2013).

7. Subgroup analysis

7.1. Effects on disruptive students

S1: Classes with the lowest number of disruptive students in the group sessions

We will repeat the analysis in H14 to H17 in classes meeting either of the two conditions:

i) At least 2 treatment and 2 control classes belong to their lottery group, and their number of disruptive students is below or at the median of their lottery * treatment assignment group.

ii) Less than 2 treatment or less than 2 control classes belong to their lottery group, and the average number of disruptive students in classes belonging to their lottery group is below or at the median of this average across all lottery groups with less than 2 treatment or less than 2 control classes.

S2 and S3: Students with mild / severe psychological disorder.

For disruptive students in each profile ("blue", "green", and "yellow"), we will compute the average of the students' TOCA sub-scores relevant for her profile (e.g.: AA, AF, AL, and MS for "blue" students). We will then compute the median of this quantity for each profile within each class, and finally we will repeat the analysis in H14 to H17 among students below or at the median, and among those strictly above.

S4 and S5: Students below/above the median of the vulnerability index.

We will compute the median of the vulnerability index among disruptive students of each class, and we will repeat the analysis in H14 to H17 among students below or at the median, and among those strictly above.

S6: Among classes with teachers with a low self-confidence as to their professional abilities.

We will form an index of teachers' self-confidence as to their professional abilities. Our index will be the average of the following standardized variables at the teacher level:

- Teachers' confidence that they can make students improve: standardized score formed from question 20 in TQB.
- Teachers' confidence that they can make a difference in their students' life: question 21.2 in TQB, standardized.
- Teachers' levels of stress: standardized score formed from questions 28 and 29 in TQB.
- Teachers' feeling of control: standardized score formed from questions 32 and 33 in TQB.

We will then repeat the analysis in H14 to H17 in classes meeting either of the two conditions:

i) At least 2 treatment and 2 control classes belong to their lottery group, and the score of their teacher is below or at the median of their lottery * treatment assignment group.

ii) Less than 2 treatment or less than 2 control classes belong to their lottery group, and the average of the score of the teachers belonging to their lottery group is below or at the median of the average of this score across all lottery groups with less than 2 treatment or less than 2 control classes.

Adjusting for multiple testing

For each hypothesis tested (H14 to H17), we will report the unadjusted p-values of the coefficients of treatment estimated in each subgroup, and the p-values adjusted for control of the False Discovery Rate (see Benjamini, Y. and Y. Hochberg, 1995) accounting for all the coefficients estimated in S1 to S6 within that hypothesis.

7.2. Effects on teachers

S7: Unexperienced teachers.

We will repeat the analysis in H18 to H20 in classes meeting either of the two conditions:

i) At least 2 treatment and 2 control classes belong to their lottery group, and the experience of their teacher is below or at the median experience of their lottery * treatment assignment group.

ii) Less than 2 treatment or less than 2 control classes belong to their lottery group, and the average experience of the teachers belonging to their lottery group is below or at the median of the average experience across all lottery groups with less than 2 treatment or less than 2 control classes.

S8: Teachers in disruptive classes.

We will form an index of a class's disruptiveness at baseline. Our index will be the average of the following standardized variables at the class level:

- Students' disruptiveness assessed by teacher at baseline: standardized score formed from questions 22, 23.3 to 23.7 in TQB.
- Average disruptiveness of students as explained in S13.

We will repeat the analysis in H18 to H20 in classes meeting either of the two conditions:

i) At least 2 treatment and 2 control classes belong to their lottery group, and their score is above or at the median score of their lottery * treatment assignment group.
ii) Less than 2 treatment or less than 2 control classes belong to their lottery group, and the average score of classes belonging to their lottery group is above or at the median of the average score across all lottery groups with less than 2 treatment or less than 2 control classes.

S9: Teachers with a low self-confidence as to their professional abilities.

We will repeat the analysis in H18 to H20 in the same subgroup of classes as in S6.

Adjusting for multiple testing

For each hypothesis tested (H18 to H20), we will report the unadjusted p-values of the coefficients of treatment estimated in each subgroup, and the p-values adjusted for control of the False Discovery Rate (see Benjamini, Y. and Y. Hochberg, 1995) accounting for all the coefficients estimated in S7 to S9 within that hypothesis.

7.3. Effects on classmates

S10: In classes with the highest number of disruptive classmates.

We will repeat the analysis in H21 to H24 in classes meeting either of the two conditions:

i) At least 2 treatment and 2 control classes belong to their lottery group, and their number of disruptive classmates is above or at the median of their lottery * treatment assignment group.
ii) Less than 2 treatment or less than 2 control classes belong to their lottery group, and the average number of disruptive classmates in classes belonging to their lottery group is above or at the median of this average across all lottery groups with less than 2 treatment or less than 2 control classes.

S11: Among classmates who are reciprocal friends of at least one disruptive student at baseline.

Based on the SSB, we will identify classmates who are reciprocal friends of at least one disruptive student at baseline (students reciprocally name each other in the 2nd column of the SSB). We will repeat the analysis in H21 to H24 among those students.

S12: Among classmates who are seating next to at least one disruptive student at baseline.

Based on the CMB, we will identify classmates who are seating side by side with at least one disruptive student at baseline. We will repeat the analysis in H21 to H24 among those students.

S13: Among the most disruptive classmates.

We will form an index of a classmate's disruptiveness at baseline. Our index will be the average of the following standardized variables: students' AA, AF, AL scores, and Summary question B in 2014 TOCA (the same variables in TOCA 2013 if TOCA 2014 is not available).

We will then repeat the analysis in H21 to H24 among classmates above or at the median of this score within their class.

S14: Among classes with teachers with a low self-confidence as to their professional abilities.

We will repeat the analysis in H21 to H24 in the same subgroup of classes as in S6. We will conduct this analysis only if more than 10% of our FDR-adjusted p-values are below 0.1 either in S6 or in S9.

Adjusting for multiple testing

For each hypothesis tested (H21 to H24), we will report the unadjusted p-values of the coefficients of treatment estimated in each subgroup, and the p-values adjusted for control of the False Discovery Rate (see Benjamini, Y. and Y. Hochberg, 1995) accounting for all the coefficients estimated in S10 to S14 within that hypothesis.

7.4. *Effects on classes*

S15: Classes with the highest number of disruptive classmates.

We will repeat the analysis in H25 in the same subgroup of classes as in S10.

S16: Classes with an unexperienced teacher.

We will repeat the analysis in H25 in the same subgroup of classes as in S7.

S17: Disruptive classes.

We will repeat the analysis in H25 in the same subgroup of classes as in S8.

S18: Classes with teachers with a low self-confidence as to their professional abilities.

We will repeat the analysis in H25 in the same subgroup of classes as in S6.

Adjusting for multiple testing

For each hypothesis tested (H25), we will report the unadjusted p-values of the coefficients of treatment estimated in each subgroup, and the p-values adjusted for control of the False Discovery

Rate (see Benjamini, Y. and Y. Hochberg, 1995) accounting for all the coefficients estimated in S15 to S18.

Bibliography

Anderson, Michael L. "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 (2008).

Amy Finkelstein, Sarah Taubman, Heidi Allen, Jonathan Gruber, Joseph P. Newhouse, Bill Wright, Kate Baicker, and the Oregon Health Study Group. "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)

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. "Inference on treatment effects after selection among high-dimensional controls." *The Review of Economic Studies* 81.2 (2014): 608-650

Benjamini, Y. and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society B* 57.1 (1995): 289-300

Haushofer, Johannes, and Jeremy Shapiro. "Welfare Effects of Unconditonal Cash Transfers: Pre-Analysis Plan." (2013).

Kellam, Sheppard G., 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 (1977): 1012-1022.

Werthamer-Larsson, L., S. Kellam, and K. E. Ovesen-McGregor. "Teacher observation of classroom adaptation-revised (TOCA-R)." Johns Hopkins Prevention Center Training Manual. Baltimore, MD: Johns Hopkins University (1989).

JUNAEB, Gobierno de Chile. Manual de Apoyo Tecnico/Metodologico Talleres Preventivos Habilidades Para la Vida. (2008)

Jellinek, Michael S., et al. "Pediatric Symptom Checklist: screening school-age children for psychosocial dysfunction." The Journal of pediatrics 112.2 (1988): 201-209.