Tackling the gender gap in math with active learning teaching practices

Maria Laura Di Tommaso [a,b,c,*], Dalit Contini [a]

Dalila De Rosa [a,d], Francesca Ferrara[h] Daniela Piazzalunga [e,f,g] Ornella Robutti[h]

[a] Department of Economics and Statistics "Cognetti de Martiis", University of Torino, Italy
[b] Collegio Carlo Alberto, Torino, Italy
[c] Frisch Center for Economic Research, Oslo, Norway
[d] Ministry of Economy and Finance
[e] IRVAPP, Fondazione Bruno Kessler, Trento, Italy
[f] CHILD – Collegio Carlo Alberto, Torino, Italy
[g] IZA, Bonn, Germany
[h]Department of Mathematics Giuseppe Peano, University of Torino, Italy

Preliminary version, 30 June 2020

Updated version available online at bit.ly/3dLIUk0

Abstract

We design an innovative teaching method that aims to narrow the Gender Gap in Mathematics (GGM) in primary school and we evaluate its impact in grade 3 in Italy. The teaching methodology consists of 15 hours of math laboratories, which focus on peer interaction, sharing of ideas, students' engagement, problem solving, and problem posing. The causal effect is evaluated using a randomized controlled trial, conducted in the province of Torino, involving 50 third grade classes in 25 schools, and 1044 students. The treatment significantly improves math performance for girls (0.15 s.d.), with no impact on boys, contributing to reduce the gender gap in math by 39.5-46.2%. The results indicate that properly designed innovative methodologies have the potential to reduce the gender gap in math and call for further research on the role of teaching methodologies on math learning.

JEL codes: I21, I24, J16, C93

Keywords: Gender gap in mathematics, School achievement, Inequalities, Primary school, Mathematics laboratory

* Corresponding author. E-mail: marialaura.ditommaso@unito.it.

We gratefully acknowledge the financial support from the University of Torino and the Compagnia di San Paolo (Progetto di Ateneo 2016 "Tackling the gender gap in mathematics in Piedmont"). The project was realized in close cooperation with the Fondazione Agnelli and with the Regional Board of Education in Piedmont (Ufficio Scolastico Regionale). In particular we thank the Director of Fondazione Agnelli, Andrea Gavosto and Martino Bernardi for their very valuable contribution throughout the project.

We thank Davide Azzolini, Simone Balestra, Nicola Bazoli, Camilla Borgna, Ylenia Brilli, Stefania Marcassa, Ignacio Monzon, Pauline Morault, Simone Moriconi, Chiara Pronzato, Enrico Rettore, Claudia Senik, Loris Vergolini as well as seminar participants at the Frisch Center, FBK-IRVAPP, the University of Torino, Webinar in Gender and Family Economics, EALE/SOLE/AASLE conference for helpful discussion and comments.

The trial has been registered at the AEA Registry: AEARCTR-0003651 (Contini, D., Di Tommaso, M.L., Piazzalunga, D. (2018). "Tackling the Gender Gap in Mathematics in Italy", AEA RCT Registry. December 10. https://doi.org/10.1257/rct.3651-1.0).

1. Introduction

International assessments of children skills in mathematics show that the gender gap in mathematics varies among countries. Among 15 years old children, PISA data for OECD countries (OECD 2018) show that the average gender gap in mathematics (girls minus boys) varies between minus 16 for Italy and plus 10 for Iceland with an average equal to minus 5 (the average mean score for girls in OECD countries is equal to 487). The gender gap in math increases with age, and varies along the performance distribution, being generally negligible among poor performers and largest at the top of the achievement distribution (Fryer and Levitt 2010, Ellison and Swanson, 2010, Contini et al. 2017). Given that girls do better than boys in most academic outcomes, we may question whether the girls' disadvantage in math should be viewed as a policy relevant issue. Although differences in preferences and labor market expectations are important determinants of the choice of the field of study in college, the females' relative weakness in math is also one of the reasons of the critically low share of women in STEM disciplines at university (Turner and Bowen 1999, Card and Payne 2017). In addition, recent research underlines the importance of mathematical skills also in non-STEM occupations. The unbalance in academic choices critically affects gender occupational choices and differences in wages (Paglin and Rufolo 1990, Machin and Puhani 2003, Black et al. 2008); women are highly underrepresented in the most productive sectors of the economy and in high-paying occupations (European Commission 2006; European Commission 2012; European Commission 2015; National Academy of Science 2007; Piazzalunga 2018; Sierminska et al. 2019).

A wide range of theories have been proposed for the existence of the gender gap in mathematics. Some scholars refer to biological differences in brain functioning (e.g. Baron-Cohen 2003), although recent research on the neural processes in young children finds that boys and girls engage the same neural system during mathematics development (Kersey et al. 2019). The high variability of the gender gap in math across countries points to cultural and societal factors. In countries with higher gender equality, girls perform better than boys in mathematics (Guiso et al. 2008; Pope and Syndor 2010; Gevrek et al. 2018, Nollenberger et al 2016). Parental attitudes towards gender equality are positively correlated with girls' test scores in mathematics (Dossi et al 2019). Gender stereotypes may contribute to shape gender differences in math achievement by affecting not only parental behavior towards girls and boys, but also teachers' beliefs. Stereotypes often lead to attribute girls' achievements to diligence instead of talent (Ertl et al. 2017) and the existing teachers' implicit gender biases (measured with the Implicit Association Test[2]) has a sizable influence on the gender gap in

[2] IAT; Greenwald, McGhee, and Schwartz (1998)

math (Carlana 2019). Another related explanation involves the interactions between students and teachers associated to gender role-models (Dee 2007). These mechanisms may also be responsible of the girls' lower self-confidence in math, higher level of anxiety, lower competitiveness (Ho et al. 2000, Gneezy, Niederle and Rustichini 2003, Niederle and Vesterlund 2010; OECD 2015, Di Tommaso et al. 2018).

A less explored potentially relevant factor refers to educational methods and practices. Some studies suggest that when mathematics' teaching is centered upon problem-solving, involving students in discussions and investigative work, the gender gap decreases and can even disappear (Boaler 2002a; Boaler 2002b; Boaler 2009; Boaler and Greeno 2000; OECD 2016; Zohar and Sela 2003). These scholars frame the problem of the gender gap in math within the consolidated stream of 'constructivist and social' methods in mathematical teaching/learning (Gutierrez and Boero 2006). In a nutshell, these methods are based on the idea that mathematical learning involves proactivity on part of the learner, leading to the idea that learners 'make things' together and 'communities of practice' are created (Lave and Wenger 1991). More specifically, in this project, we design a teaching method grounded on the notion of "mathematics laboratory" (Anichini et al., 2003; Robutti et al., 2004), based on group and peer work, sharing and comparison of ideas, class discussions led by the teacher, "doing" instead of 'listening' through problem posing and problem solving. The hypothesis is that by empowering children with a "growth mindset" (Boaler 2013)[3], this methodology could contribute to reduce the gender gap in mathematics. We name this methodology "Math Active Learning Teaching practices" (MATL) program.

Our main research question is to assess if the MATL program can help to reduce the gender gap in mathematics. We evaluate the impact of the intervention with a Randomized Control Trial. The trial takes place in the province of Torino, a large city in the north-west of Italy. Italy is a case-study of particular interest because it displays a very large gender gap in mathematics in all international assessments: it has the highest gap among the 57 countries participating in TIMMS 4th grade (Mullis et al. 2016), and the largest gap (among OECD countries) in the PISA test administered to 15-year-old students for year 2018 (OECD 2018)[4].

[3] The growth mindset, as opposed to a fixed mindset, gives importance to mistakes; in particular it is important how teachers treat mistakes in the classroom. Mistakes should be valued for the opportunities they provide for brain development and learning. Fixed mindset beliefs contribute to inequalities in education as they particularly harm minority students and girls (Boaler 2013).

[4] The reason why the gender gap in math is so large in Italy is difficult to establish and out of the scope of the present work. However, according to the TIMSS teachers' questionnaire, math teaching in Italy is still largely traditional, based on frontal lessons, large home workload, frequent school assessments, and only a minority of teachers rely on frequent group-work and lab sessions within the classroom.

The intervention has been developed by a team of educational mathematicians, and consists of 15 hours of MATL program delivered to third grade pupils. The intervention takes place in third grade classes because previous studies (Contini et al. 2017) show that the gender gap in Italy starts in second grade and it increases until 10th grade. All the schools in Torino province were invited to participate into the program with at least two classes.[5] Among participant schools, we first randomly selected 25 schools and then within each school we randomly assigned one class to the treated and one class to the control group. This procedure was adopted to ensure balance between control and treated group. The final sample consists of 1,044 children of which 519 are assigned to the treatment group and 525 to the control group. The treatment is delivered at the class level and it is carried out by tutors trained in mathematics education as a substitute to the traditional math hours. Teachers were present in the classroom with the role of observers. Children in control classes follow the usual curriculum with their own teachers. The laboratories are organized over five sessions of three hours each, once per week for five consecutive weeks between February and April 2019. In order to assess the impact of the MATL program on children's performances, we also deliver a test in mathematics one month before the intervention (pre-test) and a month after the intervention (post-test). The test is developed with the external supervision of scholars involved in the design of the national assessment test (INVALSI). Its structure is similar to the national assessment test and it consists of 20 items.

The findings from the impact evaluation of the MATL program are encouraging. The treatment significantly improves math performance for all children but the effect is driven by girls' improvement whereas the treatment does not have a significant effect on boys. In particular, we can conclude that MATL methodology increases math achievement for girls of 0.15 standard deviations, without damaging boys' performance. For educational interventions, this effect is large in magnitude and policy-relevant. As a term of comparisons, for primary school Bloom (2008) reports that one full year of attendance improves pupils' achievement by 0.25 standard deviations in both math and reading, while decreasing class-size by 10 children from 22-26 students improves performance by 0.10-0.20 standard deviations.

As far as the main research question is concerned, i.e. the impact of the educational program on the gender gap in mathematics, we find that the MATL program reduces the math gender gap in between 39 % and 46 %.

We also assess whether the MATL has different impacts for pupils with different levels of prior math ability, measured with the pre-test. We find that the treatment has no effect for boys, irrespective

[5] The Regional Board of Education, the main regional authority for education at the regional level, has been involved in the project and directly invited the schools to participate into the program.

of their starting level, and a positive effect for girls, increasing with pre-test scores. Girls with above average pre-test scores benefit the most from the treatment. Heterogeneous effects by migrant status vs. native and by mother and father educational achievements are also found. Migrant girls and girls with low educated mothers improve their math achievements more than native girls and girls with highly educated mothers. For boys the heterogeneous effect of the treatment goes in the opposite direction. Boys with highly educated father and native boys benefit more than the other boys.

Our paper contributes to the existing literature showing that teaching methodologies can influence the gender gap in mathematics. In particular, that Mathematics Active Learning Teaching Practices (MATL) can contribute to the reduction of the gender gap in mathematics. This is the first paper that evaluate the causal impact of a teaching methodology on the gender gap in mathematics. Literature on the causes of the gender gap in mathematics have until now pointed to biological causes or parents and teachers' beliefs, expectations and biases but there has been a lack of evaluation of the impact of teaching methodology on the gender gap in math. Our paper fills this gap.

The rest of the paper is organized as follows. In section 2, we provide an overview of the Italian institutional context. In Section 3, we describe the treatment. Section 4 is devoted to the research design of the RCT, as well as to data and estimation strategy. Results are presented in Section 5, while we explore potential mechanisms in Section 6. We discuss critical issues and problems in section 7 and we conclude in Section 8.

2. Institutional context and design of the program

Institutional context

In the Italian educational system, children enter formal schooling at age 6. Primary education lasts for five years, until age 11. The system is largely composed by public institutions: only 6% of the children attends private schools in primary school. In principle, families can choose between two time schedules: a 40-hour school week, where children spend the whole day at school and a more concentrated 28/30-hour week. However, the first option is not highly available in all areas of the country[6]. Curricula and learning targets are defined at the national level and do not vary across time regimes, but teachers have full leeway in the choice of the teaching methods. In each class, there are two-three teachers covering the entire set of disciplines (sometimes with the exception of foreign language, gymnastics and music). Didactic continuity is extremely valued in the Italian school

[6] The share of schools delivering the 40-hours schedule is much higher in the Northern regions.

system: children are grouped into classes that remain the same and are normally taught by the same teachers for the entire 5-year cycle. At the primary school level, teachers receive a training enabling them to teach all subjects[7], although they often accumulate experience in specific disciplines. In any case, they do not change subjects within each cycle with the same group of pupils.

The school year starts in the first half of September and finishes in the mid of June. In grade 3, when the MATL intervention has been delivered, math instruction is provided for 6 to 8 hours weekly, amounting to 198-264 hours per year. The different domains covered in primary school are numbers, relations, data and predictions, space and figures. National curricular guidelines recommend providing instruction on the different domains in alternation throughout the entire school year, and there is anecdotal evidence that this is the actual practice in all schools.

The MATL intervention

The math intervention consists in classroom-based activities aimed at improving children mathematical understanding and at reducing the gender gap. The teaching practices adopted are based on the theoretical framework of social constructivism, embodying theories of active learning that emphasize the need for students to "construct" their own understanding. More specifically, the math intervention MATL (Math Active Learning Teaching practice) builds on the "*laboratorio matematico*" (mathematics laboratory), an Italian math didactics approach developed at the beginning of the twenty-first century and widely acknowledged in the math education international community (Anichini et al., 2003; Robutti et al., 2004). Although not stated explicitly, the general approach follows the "growth mindset" paradigm, according to which ability is malleable, intelligence can be learned, and the brain can grow from exercise (Dweck 2006, Boaler 2013). There is evidence that students who acquire a growth mindset learn more effectively "displaying a desire for challenge and resilience in the face of failure" (Boaler 2013, pg…).

The fundamental elements of the MATL program can be summarized as follows:

(i) *Doing instead of Listening*. Focusing on problem-solving, it reverses the traditional teacher-centered instruction by putting the child at the center of the learning process. Students are engaged with small peer-group work, and dialogue with the teacher both individually and with collective discussions.

(ii) *No pressure*. There is no request for immediate answers or solutions at the individual level; instead, students are given suitable time to analyze the problem, explore different solutions,

[7] The required qualification to become a primary school teacher is now a university degree in primary school teaching education. Before 2001 the required qualification was a specific high school diploma (*Istituto magistrale*).

share and compare ideas, avoiding pressure and competitive tasks.

(iii) *Learning from mistakes*. Mistakes are conceived as crucial means to understand. By giving positive attention to theirs and others' mistakes, children explore their learning processes and develop a deeper understanding of the discipline.

These elements aim at activating children's thinking, constructing mathematical meanings and solving problems, through interaction and communication. The teacher – or the tutor, in our case – has the role of "orchestrating" class activities. Another peculiar feature of the lab is the use of different objects and poor materials like caps, straws, buttons of different size, boxes, cards.

MATL is based on two activities, named *Forest Elves* and *Thousandville*. The first moves around a family of elves who have to go in different places at different paces and in different moments, and the issues at stakes are " who will get first in given place?" "when/where will they meet"? The second involves the task of enlarging a city while keeping the same proportions of the different elements that constitute it. The curricular contents covered are: writing numbers in the base-ten natural number system, using the decimal notation, comparing and ordering natural numbers, estimating quantities, using numbers as measures, multiplicative reasoning, using tables and the number line[8].

Why should MATL contribute to reduce the gender gap in math?

Laboratory teaching practices are devised to help developing a growth mindset. As shown by Dweck (2006 a,b) "fixed mindset messages prevail among students across the achievement range and some of the students who are most damaged by fixed ability beliefs are high-achieving girls". These studies show that girls suffer most by the fixed ability conception that favors the attachment of labels like being or not being smart, or being good or not being good at math (Dweck, 2006b).

In this perspective, the teaching practices embodied in the MATL intervention have the potential to reduce the gender gap in math for different reasons. Firstly, the laboratory is meant to reduces pressure and competition. This should benefit girls in particular, because girls are generally less competitive than boys, in competitive environments they tend to develop more anxiety, and anxiety is detrimental to learning . One of the channels through which MATL works in this direction, is giving positive value to mistakes. Transforming mistakes from a failure into an opportunity to learn is even more important for girls, because girls have been shown to be more risk averse and have more fear of giving the wrong answer (Bohnet 2016). Moreover, being in general more reflective, girls could

[8] Extracts from the methodological guidelines (English translation) are available in Appendix C. The full methodological guidelines are available in English (translation) or in Italian (original) upon requests.

be more prone to learn from mistakes by better developing constructive reasoning on their own cognitive processes (Boaler, 2016). This could also be enhanced by the fact that girls tend to be better achieving than boys in reading comprehension and the mastering of language. In this vein, the MATL intervention has been specifically devised to embody some mathematical activities into a narrative context. A final element that might contribute to girls' activation and empowerment is the explicit support in the MATL guidelines for gendered balanced participation to class discussions.

Implementation of the MATL intervention

According to the literature, the GGM is often observed at very young age and increase as children grow older; in Italy it is already in place in grade 2 (Contini et al. 2017). The MATL program is implemented in grade 3, when children are about 8 years old. The reason behind this choice is to balance two different needs: (i) to tackle inequalities as early as possible, to hinder their onset and contrast possible cumulative effects; (ii) to run the intervention at a point in time when the GGM already exists, so to actually observe gender differences before the intervention and analyze their (short-term) development.

MATL has been delivered between February and April 2019. The intervention was run during school-time, at the class level; more specifically, it was delivered during math hours, in order not to alter the total amount of time devoted to math instruction. It took place over five sessions of three hours, once per week for five consecutive weeks.

Children were divided into small heterogeneous groups (in terms of gender and level of ability) and asked to do both group and individual work. All the students in the treated classes took part in the activities, including student with disabilities, special education needs, or learning difficulties. Instead, children in the control group followed the usual curriculum.

MATL focuses on the subject area of numbers, recognized as the most fundamental domain in the math field at this age (the other areas in the primary school curriculum are: relations, space and figures, data and predictions). Incidentally, there is evidence that the GGM is highest in the domain of numbers (Contini, Di Tommaso, Ferrara et al. 2018). The intervention was conducted by four tutors with a background in mathematics education at the Master or Ph.D. level, specifically trained by the field scholars in our team. Class teachers were present as observers.

A pilot study to evaluate the intervention format was conducted a few months before in two different schools that were not participating to the RCT. The treatment was then reviewed to take into account the comments and suggestions of the tutors and the class teachers. This pilot gave also the opportunity to assess the length, difficulty, and discriminatory power of the items included in earlier versions of

the pre- and post-tests. These tests were analyzed with item-response-theory (IRT) models and modified accordingly.

3. Design, Data, and Estimation

3.1. *Research Design*

We evaluate the intervention using a randomized control trial, which took place in primary schools of the province of Torino (Piedmont), located in the North West of Italy, where there are about 180 primary public schools altogether. Due to budget constraints, we planned to enroll 50 third-grade classes in 25 schools, implying approximately 1000-1200 pupils.

Enrollment to the project was on a voluntary basis, under the eligibility rules described below, and it was open between April and May 2018. All principals of public primary schools in the province of Torino were informed about the project in March 2018, with an official letter by the Regional Board of Education, and invited also to an open meeting along with math teachers, during which the project has been presented in detail. Due to transparency requirements set by the regional authorities, the schools were informed that the aim of the project was an evaluation of the effects of the intervention on the gender gap in math. Anecdotical evidence also suggests that the focus on GGM raised the interest in the project, compared to other several projects proposed to schools every year. As long as this knowledge does not affect teachers in treated and control classes differently, it does not hamper the validity of the trial, and there is no reason to expect that teachers in the two groups reacted differently to the information. In addition, teachers were not actively involved in the conduction of the program, and all participants were aware that what was under evaluation was not their own work or their approach to girls and boys, but the proposed intervention itself.

Eligibility conditions were set as follows: (i) Each school had to participate with at least two classes, one of which to be randomized to the treatment group and the other one to the control group. The scope is to eliminate potentially large school-specific effects on girls' and boys' math achievement, related to school management, socioeconomic composition of the student body and school-level peer effects, considering that classes are more similar within schools than across schools. It can be seen as a matching procedure, set up to improve the precision of the estimates and to increase the similarities between the treated and control group. (ii) The two participant classes had to have different mathematics teachers, to limit the risk of spillover and contamination. (iii) Participating classes were not to be involved in other extra-curricular math projects in the same school year.

Thirty-one schools applied to participate to the program, but one was excluded because

participating to another math project. Among the remaining schools, we randomly selected 25 of them, and since some schools applied with more than two classes, we also randomly selected the two participating classes (see Table A.2).

In a second step, within each participating school we randomly assigned one class to the treatment group and the other to the control one.[9] The entire randomization process was public and took place at University of Torino on June 2018. No school or class dropped out of the project, thus altogether, twenty-five primary schools participate in the project with two third-grade classes each, one assigned to the treated and one to the control group, for a total of 50 classes, and 1,044 children (classes have on average 21 students).

The timeline of the implementation of the RCT is as follows (Figure 1). We conducted the pre-test on math skills in control and treated classes in January 2019. The math laboratories took place in the treated classes once per week over five consecutive weeks between February and April 2019, while control classes follow the usual curriculum. We conducted the post-test between April and May 2019.

The trial was registered at the AEA Registry on December 6, 2018, along with a pre-analysis plan (PAP), before the start of the intervention. The paper presents analyses on pre-specified outcomes, unless differently specified (e.g. in the mechanism section).

Fig.1 Timeline of the intervention

3.2 Outcome measures and additional data

Outcome measures

All children in the treatment and control classes sat a pre-test on math numeracy one month before the start of the treatment; the post-test, following the same framework, was administered one month after the end of the intervention. The tests were designed by the scholars of mathematics education members of the research team and followed the same conceptual framework of the INVALSI assessment. The reasons of using tests designed ad hoc is two-fold: first, INVALSI assessments cover also other domains (relations; data and previsions; and space and figures) and has a lower number of items on numeracy only (around 15) than our test; most importantly, the INVALSI assessment is only available for other grades, with a level of difficulty not appropriate to grade 3.

The tests consist in 13 questions and 20 items each, to be completed in 40 minutes. The pre-test

[9] The sampling procedure plan was set before knowing how many schools and classes would have applied to participate to the project, and different rules were defined depending on the number of applications. Details can be found in the pre-analysis plan (Contini, Di Tommaso, and Piazzalunga 2018).

has been analyzed with an IRT model, to give further insights to the design of the post-test, and also the post-test, to have also an ex-post analysis of its level of difficulty and discriminatory power. The tests cover topics such as the number line; tens and hundreds; additions and subtractions; times tables; easy questions with money; calendar time. As the characteristics of the item can influence the gap (see Contini, Di Tommaso, Ferrara et al. 2018), the tests are designed to include different types of items: more specifically, they cover different mathematical dimensions (knowing, arguing, and problem solving) and use both multiple choice-type answers and open answers; moreover, there are items without figures and with different types of figures (e.g. complementary or necessary to solve the problem).

The tests were administered in class by the tutors in charge of the laboratories and graded blindly by them under the supervision of an external examiner.[10] Correct answers are given 1 point each, incorrect and missing answers 0 points, for a maximum score of 20. In the main analysis, the individual total score is then standardized to have 0 mean and a standard deviation of 1.

While the post-test is the main outcome variable employed to assess the effectiveness of the intervention, the pre-test is used to evaluate the gender gap before the intervention, to assess the balance between treated and control classes in terms of baseline achievements, and included as a control variable to improve the precision of the estimates. Figure 2 shows the pre-test score distributions among girls and boys who sit the pre-test (sample (b)): a gender gap in math is evident before the intervention, across the entire distribution, confirming findings from other papers, which show that a GGM is present already in grade 2 (Contini et al. 2017). On average, boys answered correctly to 11.23 items out of 20 and girls 10.28, with a significant difference of almost one correct answer, which corresponds to 0.216 standard deviations (0.237 in our preferred sample of children present both at the pre- and post-test). As we will discuss later on (see Section 7.2 on external validity), this difference is larger than the one measured with INVALSI data at the end of grade 2 in Piedmont and Italian classes (respectively, 0.13 and 0.11), but close to the INVALSI one measured in our experimental classes.

Fig.2 Gender gap in the pre-test

In addition to math achievement, we consider as an additional outcome/possible mechanism children's attitudes towards math, evaluated by means of a short questionnaire with five Likert-type

[10] An expert in formulating and grading INVALSI tests.

questions, delivered immediately after the post-test. More details are provided in Section 6.2.

Additional data

In addition to math numeracy tests and to children's attitudes towards math, we collected additional data at the individual and at the class level. For treated children, tutors monitored absenteeism during the math labs. The school teachers provided information on children's special educational needs and disability (SEND), including any form of learning difficulty, such as physical or mental disability, learning disorders, hyperactivity (ADHD).[11]

Background information on parental education and migratory background was recovered from the school administrative office, which had collected such information in previous years, either for administrative purposes or for INVALSI, through parents' questionnaire; thus, missing data can be due to a new child in the class or, more often, to missing answers. For each parent, the level of education is recoded into lower secondary (if s/he has at most a professional qualification), upper secondary, and tertiary or above. According to the country of birth of the child and the parents, the child's migratory background is recoded into native, first generation migrant, second generation migrant.

We gathered data on the math teachers of classes involved in the project through a brief questionnaire asking gender, age, degree, experience in the class, tenure, and type of contract. Class level information was registered by the tutors for the scope of the project (e.g. class size).

A description of the variables can be found in the Appendix (Table A.1)

In addition, to position the schools taking part to the RCT within the general population in terms of achievement and socio-economic background and evaluate the external validity of the experiment, additional average information at the class level were obtained from INVALSI for the classes participating to the project (scores in math and Italian at the assessment in grade 2, socio-economic background), to be compared with the corresponding statistics at the regional and national level.

3.3 Sample

Of the initial sample composed by 1,044 children, some children were absent at the pre-test, other at the post-test, and for some of them background information is missing. Children absent either at the pre-test or post-test are excluded from the main sample in the core analysis. For children absent at

[11] Two different versions of the variable are codified as dummy variables: a restricted version of the variable takes value 1 only for children with certified educational needs, whereas the broad version of the variable takes value 1 for all children reporting any kind of learning disorder/special needs, either certified or only displayed.

the pre-test, in a robustness check, we assign a zero value and include a dummy variable for missing pre-test scores. For children absent to the post-test, we scheduled a deferred session in a different date, as close as possible to the original one, and we use such data in a second robustness check.[12] In order not to lose too many observations, in case of missing values for each of the individual characteristics, we assign a zero value and include a dummy variable accounting for missing. For this reason, also estimates without control variables are presented. Finally, we have information on teachers' characteristics in 49 out of 50 classes, because one teacher refused the consent to the use of her data; thus, teachers' characteristics are used for the balance tests, but only class size is included as a control variable.

To have a clearer picture of the sample(s) we work on, Table 1 summarizes the number of pupils in the initial and in the preferred samples, and Table A.3 in the Appendix provides additional details. Of the 1,044 children in the full sample (sample (a)), we exclude 4 children who have all item missing in the post-test even if they were present; 933 pupils were present at the pre-test and 983 were present at the post-test, for a total of 888 present both at the pre- and post-test, which constitute our main sample.

Tab.1 Sample selection

3.4 Balance, Attrition and Compliance

Balance at baseline

To validate the randomization process, Table 2 presents the balance of baseline variables across treatment status at the individual and class level, respectively. Panel A reports individual characteristics of treated and control children – math competences before the intervention (pre-test), gender and special educational needs and disabilities – and their family background, namely parental education and migratory background. Small differences emerge only in terms of maternal education: however, it is not an overall tendency of being more educated in one group than in the other (consider the variable "at least upper secondary"), but mothers' of control and treated children are switched in the probability of having a tertiary or an upper secondary degree. Moreover, if anything the differences favor the control group, where mothers are more likely to have a tertiary degree. Panel B

[12] In the normal session the test is administered by the external tutor and all the class sits the test at the same time. On the contrary, in the deferred session, the post-test id administered by the class teacher, while the rest of the class is having lesson, and the completed test was posted to the project team. For these differences, we preferred not to include the core analysis. Of the 57 children absent at the post-test, we received 35 tests taken in the deferred session, the remaining 22 either being absent also during the deferred session or their tests have not been mailed.

presents teachers' and class characteristics, which are also well balanced, except for the number of years a teacher is teaching math in the experimental class, which is higher in control classes (2.79 years versus 2.40). It is also worth mentioning that the number of significant differences is similar to the one expected due to chance variation (around 3).

Altogether, the baseline sample is well balanced on individual and class-level characteristics, overall and by gender, indicating that the randomization was successfully implemented. In addition, the two groups are comparable in terms of math performance not only at the mean but also across the entire distribution, as can be seen from Figure 3.

Tab.2 Baseline characteristics of treated and control children, full sample

Fig.3 Pre-test score distribution by treatment status

Note that 15 percent of the children display special educational needs or disabilities in a broader sense, while 8.1 percent of the children were already certified as children with learning problems or disabilities.[13] This information is important because our tests are designed for typically developing children, and they may be not appropriate for children with some learning problems. For this reason, in the PAP we pre-specified that SEND children's results would have been excluded from the analysis. However, often SEND children have not yet been certified in grade 3, and the probability of being certified is not necessarily correlated with the seriousness of the condition (unless in the worst cases), thus differences between certified and not certified SEND children are vague. Thus, in the core analysis, we include SEND children, whereas we exclude them in two robustness checks.

Attrition

Attrition occurs when individuals who were included in the initial sample, either assigned to the treatment or the control group, have missing values in the outcome variable and are not part of the final analysis. The overall attrition rate is the rate for the entire sample, measured as the percentage of the initial sample that has been lost; the differential attrition rate is the percentage point difference in the rates of attrition for the treated and control groups. Both overall and differential attrition can create potential biases in the estimates, by influencing the baseline equivalence of the two groups.

[13] Differences in the percentage of SEND between boys and girls are well known and documented in the literature (e.g. Vogel 1990; Nass 1993) and part of it can be ascribed to gender bias against boys in the referrals for special education (Anderson 1997; Wehmeyer and Schwartz 2001), a finding which is out-of-the scope of our paper, but supports including also SEND children in the analysis.

We have two levels of attrition: the first one is due to absences at the post-test among the full sample of children, whereas the second one is due to absences at the pre- *and* post-test, which is important as we include pre-test scores in our core analysis. For these two levels, we measure overall and differential attrition for all children and separately for boys and girls.

Starting from the full sample of 1,044 children, the first part of Table 3 reports the attrition rates to get sample (b): 5.4 percent of children were absent only at the post-test, with small differences between treated and control children (-0.1 percent); the same is true across the two genders. The second part of Table 3 presents the attrition rates considering children absent both at the pre- and post-test, i.e. to get our main sample (d) starting from the full sample of 1,044 children: altogether, 14.9 percent of children did not sit either the pre- or post-test, due to more absences at the pre-test, carried out in January and February, during the flu peak. The percentage of absences is significantly higher among treated children than among control ones (16.7 vs. 12.4 percent respectively), with a more pronounced gap among girls than among boys. Overall, our core analysis is conducted on a sample of children composed of 85 percent of the initial one.

Tab.3 Attrition pre-test and post-test

Attrition can potentially undermine the initial equivalence between the control and treated group, thus being a threat to validity to our estimates. Hence, Table A.4 in the Appendix presents the complete set of balance checks for the sample of children present at the post-test only (sample (c)), and more importantly, Table A.5 for the sample of children present both at the pre- and post-test, our main sample. On most dimensions, treatment and control groups are well balanced also after attrition has occurred, the only difference with the full sample being in the percentage of tertiary educated fathers, which marginally favors the control group. Moreover, there are no differences between children's characteristics in the full sample and in the preferred one. Therefore, attrition has not worsened the balance between the two groups. Nevertheless, to account for the differences mentioned, we present the main results with and without controlling for baseline variables.

The comparison between the treated and control group in the final sample has also been replicated with a regression including all control variables: we estimate a logit regression with treatment status as the dependent variable and individual and class characteristics as independent variables and conduct a Wald test of the joint significance of the variables. Results are presented in Table A.6, including teachers' variables (fewer observations) and excluding them, thus using the entire sample of children present both at the pre- and post-test. Results confirm that the two groups are comparable,

with the small differences discussed earlier and a lower probability of second-generation migrants of being treated. Fewer variables are significant in our preferred sample, one needs to remember that in the sample including teachers' characteristics one entire control class is missing.

Compliance

Another threat to the validity of an RCT is the level of attendance of treated individuals to the treatment, which in our case means the participation of treated children to the math laboratories, information collected by the tutors. While it is not possible for children assigned to the control group to participate in the labs (crossovers), absences to part or all the laboratory sessions are possible, and children assigned to the treated classes may remain untreated, leading to one-sided non-compliance (no-shows).[14] Nevertheless, participation has been very high (Table 4): no children missed all the lab sessions; 99.3 percent of children attended at least half of the hours of laboratories and 73.8 percent of them attended all sessions, with a small difference in favor of boys (4 p.p. in the full participation), which if anything may reduce the estimated impact on the GGM.

Instead, we do not expect contamination for several reasons, namely teachers in the control group cannot propose the math lab: (i) the laboratory nature of the intervention, performed by external tutors, (ii) math teachers are different in the two classes; (iii) the intervention and the evaluation phase were carried out within a few-months span; (iv) the lab materials were not released to the teachers until the end of the project.[15]

Tab.4 Attendance to the laboratory sessions

3.5 Empirical strategy

We aim to assess the impact of participating in the math laboratories on pupils' math competences, and more specifically on boys' and girls' outcomes. For our core analysis, the effects are estimated using the following OLS specification, overall and separately for males and females:

$$Y_{1iks} = \alpha + \beta T_{ks} + \gamma Y_{0iks} + \delta X_{iks} + \theta_s + \epsilon_{iks} \qquad (1)$$

[14] "Crossovers" are control group members who receive the treatment; "no-shows" are treatment group members who do not receive treatment. Noncompliance dilutes the experimental treatment contrast, causing it to understate the average treatment effect (Bloom 2008).

[15] If this contact occurred somehow, our estimates would represent a lower bound of the actual treatment effect.

where Y_{1iks} is the post-test score of individual *i* in class *k* of school *s*. T_{ks} is the binary indicator, which equals one if the pupil is in a class randomly assigned to the treatment group and zero otherwise. Y_{0iks} is the outcome variable at baseline (pre-test score). X_{iks} is a vector of observables potentially predictive of the outcome (gender, special education needs or disability dummy, migratory background, parental education, and class size). θ_s is a vector of school fixed effects (i.e. our randomization strata), and ϵ_{iks} are random errors normally distributed clustered at the class level *j*. β is the coefficient of interest, which captures the intention-to-treat (ITT) impact of participating to the math lab, as full participation to the math lab has not been reached (absences due to sickness; more details below). The pre- and post-test scores are standardized, thus the effect of the treatment β represents by how many standard deviations test scores of the treated pupils differ on average from those of the control ones.

In more basic specifications, pre-test scores, control variables, and school fixed effects are not included in the estimation; they are progressively added until the full specification of our core analysis.

Estimating equation (1) separately for boys and girls allow us to assess the possible different treatment impacts on the two genders. The GGM in the pre-test is measured as the difference between boys' and girls' scores, but to evaluate by how much the GGM changes as a consequence of the treatment, one cannot compare the treated group before and after, for the same reason that the before-after comparison does not identify a causal effect (other changes in the same period; different tests, which cannot be anchored). One possibility is comparing the raw GGM of the control and treated group after the treatment; to better take into account for differences in the pre-test, albeit small, we use coefficients estimated from Equation 1 and estimate a counterfactual GGM among control children would they receive the treatment and a counterfactual GGM among treated children would they not receive the treatment.

It is important to note that in carrying out our empirical analysis we remain as close as possible to the pre-analysis plan. Unless otherwise indicated, the analyses and outcomes investigated were pre-specified. In particular, in addition to the core analysis, we evaluate if the impacts differ by children's prior math skills and the impact on children's attitudes, as pre-specified in the PAP. Moreover, we present additional heterogeneous effects and additional possible mechanisms, both shaped mainly by questions and feedback received in seminars; we consider these analyses exploratory.

4. Results

4.1 Core results

To evaluate the ITT impact of the intervention on math performance, we compare post-test results between the treated and control group, overall and by gender, as described in the previous section. The analysis separated by gender is of most interest for the scope of the research. In this section, we focus on the main results, namely the effect of the intervention on math performance by gender, whereas in the next subsection we investigate if the impact heterogeneity, namely if the treatment has had differential impacts according to prior achievement and if it affected education inequalities by parental education and migratory background; finally, we assess the robustness of the results.

Table 5 first presents base results, namely the raw differences in math performance between treated and control children, for all children who sat the post-test (983 children, sample (c)). The treatment significantly improves math performance for all children (effect size 0.116); however, the effect is driven by girls' improvement (column 2, effect size 0.154), whereas the treatment did not have a significant effect on boys. We then focus on the sample of the children who took both the pre- and the post-test, our preferred sample (sample (d)): columns 4 to 6 summarize the ITT impacts without baseline controls, and progressively including pre-test scores (columns 7 to 9), school fixed effects and additional control variables (columns 10 to 12). While not statistically significant, the point estimates of the basic specification are positive and similar to the one above for girls, smaller and even closer to zero for boys. Accounting for pre-test scores, the impact of the treatment has a similar effect for girls, more precisely estimated and statistically significant, and is virtually zero for boys. As expected, pre-test scores are highly correlated with post-test results: one standard deviation increase in the pre-test implies higher achievement at the post-test by more than 0.7 standard deviations. Our core and preferred estimates include school fixed effects, taking into account the possible correlation of treated and control classes in the same school, additional individual and family background characteristics, as well as for class size. The effect of the treatment is robust and confirms an impact on girls of 0.150 standard deviations and a zero effect on boys, with an overall effect of 0.091 s.d. (full results are presented in Table A.7 in the Appendix).[16]

Overall, results are very stable across specifications, and we can conclude that the teaching

[16] In the Appendix, we present also results from our preferred specification using as outcome variable Y_1 and as baseline control Y_0 the latent ability estimated with IRT (Item Response Theory) models instead of pre- and post-test standardized results (Table A.8), and the heterogenous results by prior achievement (Table A.9). The first two columns present our main results to ease the comparison. Let us anticipate that all the results are confirmed and similar in magnitude, and thus we decided to keep the standardized test-scores in the main analysis first to adhere as much as possible to the pre-analysis plan, and second because the treatment itself could partially affect the estimated latent ability. More specifically, we have estimated three IRT models: (i) one-parameter IRT logistic model, which accounts for the level of difficulty of the items; (ii) two-parameters IRT logistic model, which accounts for the level of difficulty and the discriminatory power of the items; (iii) two-parameters IRT logistic model estimated only on the control group, and predicted latent ability for both treated and control children, to reduce the risk that the treatment impacts on the estimated latent ability.

methodology introduced enhances math achievement for girls of 0.15 standard deviations, without hampering boys' performance. For educational interventions, this effect is not only statistically significant but also large in magnitude and policy-relevant. As a term of comparisons, for primary school Bloom (2008) reports that one full year of attendance improves pupils' achievement by 0.25 standard deviations in both math and reading, while decreasing class-size by 10 children from 22-26 students improves performance by 0.10-0.20 standard deviations. Slavin and Lake (2008) find that programs targeting teachers' instructional practices – as the implemented math labs do – lasting at least 12 weeks have a median effect size of 0.33 and Pellegrini et al. (2018) for the same find a median effect of 0.25; the magnitude of our results, due to a five weeks intervention, is thus in line.

A core question is how this impact translate into a reduction of the GGM. In the control group. the gender gap in math is 0.324, whereas in the treated group the GGM is 0.221 standard deviations, i.e. the GGM is 31.7% lower in the treated group compared to the control group.[17] To assess more precisely the percentage GGM reduction due to the treatment, taking into account differences in pre-test scores and in individual variables, albeit small, we would like to compare the GGM among exactly the same pupils receiving and not receiving the intervention, which is impossible by definition. Hence, we first estimate the counterfactual GGM among control pupils would they receive the treatment (0.174) and compare it with their actual GGM (0.324); a second comparison is made between the real GGM among treated children (0.221) and the counterfactual one, should they not receive the treatment (0.365).[18] The reduction in the math gender gap is thus 39.5 - 46.21%.

Tab.5 Main results: effect of the treatment

4.2 Heterogeneity in treatment effects

We now assess whether the treatment has different impacts for pupils with different levels of prior math ability, measured with the pre-test: as recalled in the literature, the math gender gap is generally larger at the top of the performance distribution (Contini et al. 2017, Fryer and Levitt 2010), and we aim to investigate who benefit the most from the math program. This analysis was pre-specified in the pre-analysis plan.

[17] The GGM is measured as the difference in the standardized post-test between boys and girls (in the treated and control group, respectively). The percentage reduction is calculated as the difference between the GGM in the treated group and the GGM in the control group, divided by the GGM in the control group.

[18] The counterfactuals are estimated from a regression including also the interaction between treatment and pre-test scores, as specified in the next section. Nevertheless, counterfactual estimates of the GGM reduction without the interaction term are very similar.

To do this, we estimate the following regression, which includes an interaction term between the treatment dummy and the pre-test score ($T_{ks} * Y_{0iks}$):

$$Y_{1iks} = \alpha + \beta T_{ks} + \gamma Y_{0iks} + \delta X_{iks} + \lambda T_{ks} * Y_{0iks} + \theta_s + \epsilon_{iks} \qquad (2)$$

In addition to β, we are now interested in the coefficient λ, which captures the differential impact of the treatment according to the level of the pre-test.

Table 6 summarizes the heterogeneous effects, according to pre-test scores, which being standardized varies between -2.4 and 2.1. Results confirm no effect for boys, irrespective of their starting level, and a positive effect for girls, increasing with pre-test scores. More specifically, better performing girls benefit the most from the intervention –with an increase in 1 s.d. in the pre-test score increasing the effect of the treatment by 0.128 s.d; the effect of the treatment is significantly larger than zero for girls with at least -0.3 standardized pretest scores (girls' average pre-test score: -0.09), as shown in Figure 4, which presents the treatment effects according to the different levels of pre-test scores.[19]

Tab.6 Heterogeneous effects of the treatment by prior achievement's levels

Fig.4 Treatment effects by prior achievement's levels

Having assessed that the treatment has a differential impact depending on the level of math skills before the intervention, we now explore impact heterogeneity, overall and by gender, along the additional domains of migratory background and parental education. As above (Equation 2), the analysis is performed by including the appropriate interaction terms between the treatment dummy and the categorical dummies in three different equations which account for i) the migratory background (natives versus first and second generation migrants); ii) maternal level of education (lower secondary or below, upper secondary, and tertiary education) and iii) paternal education. Table 7 presents the estimated treatment effects on each of the subgroups, controlling for pre-test scores, as well as for other additional variables, i.e. the effects should be interpreted being equal the pre-test score.

[19] To assess if the treatment has a non-linear impact depending on different levels of the pre-test, we replicated the analysis interacting the treatment variable with quintiles of the pre-test instead than with a continuous variable (overall and by gender). The results indicate that the effect is approximatively linear.

Overall, opposite effects arise for boys and girls. In terms of migratory background, native girls benefit from attending the math lab, but the effect is three times as large for migrants (0.374); instead, for native boys the treatment has no effect, but it hampered male migrants' performance. Also, girls with low educated parents benefit the most (the treatment has a positive impact also on the other two groups, but not significant); on the contrary, math skills of boys with a low educated father worsen as a consequence of the treatment (maternal education has an effect in the same direction but smaller), while those with a high educated father largely benefit. Consequently, the math labs improve math skills for disadvantaged girls, once controlling for pre-test scores, but they worsen them for disadvantaged boys. Why is this the case? Two best-evidence syntheses by Slavin and coauthors (Slavin and Lake 2008; Pellegrini et al. 2018) indicate that for math programs lasting at least 12 weeks with similar teaching practices as the one implemented during "our" math labs, either students coming from different backgrounds benefit in a similar was or low achievers benefit most. Hence, it is possible that our math labs would improve also skills for boys from disadvantaged background if implemented over a longer period.

Tab.7 Heterogeneous effects of the treatment by migrant status and parents' education

4.3 Robustness checks

We test the robustness of our results to the choices done in the core analysis, mainly in terms of sample selection: results are reported in Table 8. More specifically, we assess the robustness of the main results in four ways: (i) excluding from the analysis children with a certified special education need or disability (SEND, narrow definition) (columns 1-3); (ii) excluding those reporting any special educational need and disability (SEND, broad definition) (columns 4-6); (iii) assigning a zero value to missing values in the pre-test and including a dummy variable for children not sitting the pre-test, and using the entire sample of children present at the post-test (columns 7-9); (iv) including post-test scores of children not present the scheduled day of the post-test, who sat it in the deferred session (columns 10-12), increasing the number of observations by 35 children.[20] In all the robustness check specifications, pre-test scores, school fixed effects, and additional controls are included.

The results of the main analysis are confirmed, with the treatment having an impact on all children, driven by the effect on girls (effect size 0.12-0.18), while the effect on boys is virtually zero. In

[20] In the pre-analysis plan, we had decided to: exclude SEND children; include post-test taken in the deferred session; include children absence at the pre-test with a missing dummy. Afterwards, we have decided to operate differently in the core analysis, but these choices – as specified in the PAP – are those presented here as Robustness checks.

particular, if we exclude children with any type of special educational need/learning disorder, the impact of the treatment is larger (0.17), while it is smaller if we include children who took the test in the deferred session. Also, being absent at the pre-test does not affect the performance at the post-test, confirming our idea that absences were random and probably due to the flu season.

Tab.8 Robustness checks

5. Further analysis and mechanisms

5.1 Type of question: effect of the treatment by difficulty, dimension, and item's format

As a further analysis, we investigate the possible differential impact of the treatment by type of question. As discussed in the Introduction, existing literature shows that the GGM tends to be larger in multiple-choice items than in open-response ones (e.g. Bolger and Kellaghan 1990; Wilder and Powell 1989), confirmed for Italy by our work on INVALSI data (Contini, Di Tommaso, Ferrara et al. 2018). Besides, our paper shows that other characteristics related to the type of question are correlated with different levels of the gender gap, in particular the so-called Dimension, which classifies the main elements of mathematical thinking behind a specific question: Knowing, Arguing, and Problem Solving. In particular, Contini, Di Tommaso, Ferrara et al. (2018) found a larger gap in favor of boys for Problem-solving, followed by Knowing, while there was no gap in the Arguing dimension. Keeping in mind these results, we investigate if the treatment has different impacts depending on the following characteristics of the question: the difficulty of the question, the format, and the dimension. Results should be considered exploratory and interpreted with caution: it is not possible to claim a causal effect in terms of such differences and the different characteristics cannot be accounted for at the same time; also, the score may be prone to measuring error, due to the fact that it is calculated on a reduced number of items.

To perform the analyses, we have classified the 20 items of the post-test according to the above categories,[21] recalculated the post-test score as the sum of the corrected item in each group, and then standardized the score. To estimate the effect of the treatment, for each group of outcomes (difficulty, format, and dimension) we performed a SUR (Seemingly Unrelated Regression) model, which

[21] To classify the level of difficulty, we have performed a one-parameter IRT analysis on the control group, sorted the items according to the resulting difficulty, and considered *easy* the items with a level below -0.5, *difficult* those with a level above or equal to 0.5, and *medium* those in between. The professors in Mathematics Education involved in the project have classified the items by Dimension. Within each characteristic (difficulty, format, dimension) an item may be classified in one category only (i.e. the categories are mutually exclusive). The classification is shown in Table A.9 in the Appendix.

assumes the error terms are assumed to be correlated across equations. The equations are estimated separately for boys and girls, controlling for pre-test scores and school fixed effects.[22] Results with the complete post-test score are also presented as a reference point, because with respect to our core results, additional controls are not included and the standard errors here are not clustered at the class level and are slightly larger. For each sex and groups of outcomes, we test the equivalence of the treatment coefficient among the different item characteristics.

Results are reported in Table 9. First, the Breuch-Pagan test always rejects the null hypothesis of independent equations. [23] It can be noticed that the treatment has no effect on boys, with one interesting exception, a positive effect on the five most difficult items. For girls, the treatment has a larger effect on items of medium difficulty and especially on very difficult ones, confirming the heterogeneous results by pre-test scores (high achieving girls benefit most from the treatment). On the other hand, there is no relevant difference in terms of format (type of answer) or dimension. These results suggest that the treatment has an overall positive effect, improving general girls' math skills irrespective from the type of question.

Tab.9 Treatment effect by type of item

5.2 Mechanisms

We now turn to investigate (and exclude) possible mechanisms through which the intervention generated positive results. As far as possible, we support our claims with data, nevertheless the analyses that follow should be considered as suggestive and exploratory.

Attitudes

Among the explanations proposed for the existence of a gender gap in mathematics, the presence of different attitudes towards math between boys and girls is a relevant one, even if the direction of causality is difficult to assess. The concept of attitude is multidimensional and according to one of the most widespread definition, three components can be identified: emotional response, beliefs, and behavior (Hart 1989). Empirical evidences confirm that girls display lower math self-efficacy, lower math self-concept and higher anxiety in doing math related activities (Else-Quest, Hyde, & Linn,

[22] To avoid capturing part of the effects with the controls, we do not include other control variables - as they are well balanced and the main results show that there is little difference when we include them. However, results with the control variables are similar to the one presented here, and available from the authors upon request.

[23] Despite the results of the Breusch-Pagan test, we have also estimated results without the SUR model, i.e. without accounting for correlated error terms, but with standard errors clustered at the class level, and results are very similar.

2010). According to a study conducted in Italian primary school, girls have higher levels of math anxiety than boys, despite having no difference in math performance (Mammarella et al. 2016). However, attitudes towards math were found strongly correlated to test scores (Di Tommaso et al. 2020).

Within the project, we collected students' attitudes towards math through a short non-cognitive questionnaire submitted immediately after the post-test, on the same day. The questionnaire is composed by five questions such as "Do you like math?" with four-level Likert scale answers (the English translation of test is available in Appendix B). It was developed based on the INVALSI non-cognitive questionnaire for grade 5 with the collaboration of a specialist in the field of math education and non-cognitive learning. Thus, we explore if attitudes towards math are one of the possible channels of changes in math performance. However, it is prone to some limitations. First, as mentioned above, we did not conduct a questionnaire on attitudes before the treatment, in order not to influence children behavior and "standard" attitudes toward math, so we cannot control for pre-treatment values. Second, attitudes tend to change less than achievements over time, in particular over a short span. Third, usually, attitudes are assessed for older children (e.g. in INVALSI starting from grade 5), and it may be questionable if the questionnaire is a good instrument to assess attitudes of younger children.

Each answer to one of the five question was given a point from 1 (Not at all) to 4 (A lot), then two indexes were recovered: one as the sum of all answers and the other one from a principal component analysis (PCA). Even though the design of the five questions aimed at capturing the different dimensions of attitudes towards math, the factor analysis suggests the presence of one main latent construct: PCA results converge in one component explaining the 48 percent of total variance with an overall KMO value of 0.766.

From a descriptive point of view, the gender gap in attitudes towards math is confirmed, and this is true both if attitude is considered as the sum of items' score or as the standardized latent component form PCA. Boys displays a sum-score of 15.5 against a 14.7 score for girls, with a significant difference of almost one point (Table A.11 in the Appendix).

Table A.12 highlights the effects of the treatment on boys' and girls' attitude towards math, estimated from a regression which includes the treatment dummy and an interaction between being a girl and being treated, as well as school fixed effects and the additional controls, consistently with the main analysis. As dependent variable, we use both the sum of single items' score (column (1)) and the standardized factor scores of the first latent component resulted from PCA (column (2)).

Results confirm the presence of a gender gap in math's attitude with girls experiencing lower attitude, which is quite striking considering the age of the children (around 8 years old). However, the treatment has no effect: even if the sign is negative, the size is small and it is not statistically significant. Hence, we can confidently conclude that the success of the treatment did not pass through an improvement of attitudes, which were not influenced either for their stable nature or for the shortness of the intervention.

6. External validity

The classes participating to the RCT are particular under at least two aspects: first, they are primary schools in the province of Torino, one of the biggest Italian cities,[24] located in the North of the country; thus, the schools are likely to be different from other schools in different geographical area. Second, they decided to participate to the study voluntary, showing a high level of interest in testing innovative teaching practices.

We compare children's individual and family characteristics to understand the similarity of the participating classes with other classes both in Piedmont and in Italy. This information may prove useful also to evaluate the possible scale up of the intervention.

All data used in this section come from the INVALSI. For the experimental classes, upon schools' authorization we asked to INVALSI to provide the class average of the following variables, recorded in grade 2 in the previous scholastic year (2017-2018): INVALSI test scores in Math and Italian, oral marks in Math and Italian, pupils' childcare attendance, mother and father education. Similarly, we collected the national INVALSI data of same information for grade 2 in the year 2017-2018, using only information for classes with the external examiner.[25]

Table 10 reports results of the comparison between our experimental classes, the Piedmont average, and the national average: most variables are statistically different. While even small differences are likely to be significant, given the large number of observations, the numerical differences also point to large differences between the experimental classes and the Piedmont and national average. Our classes report higher performance at the INVALSI test score, both for Math and for Italian tests. This is not the case for school marks, which however have less variability. Also the educational levels of the parents and the probability of attendance to kindergarden are higher in the experimental group than in the rest of Italy. These differences are probably due to the geographical location of the classes. Moreover, the gender gap in Math is larger in our classes. These results

[24] The population of Torino is around 900,000 individuals.
[25] The presence of the external inspector reduces the cheating possibility.

indicate that further research is needed to understand the possible effects of the innovative teaching methodology in different settings.

Tab.10 Comparison of experimental classes with Piedmont and Italy

7. Conclusions

We design an innovative teaching practice that aims to narrow the Gender Gap in Mathematics in primary school and evaluate its impact in grade 3 in Italy. The teaching practice consists of 15 hours of math laboratories, which focus on peer interaction, sharing of ideas, students' engagement, problem solving, and problem posing. The causal effect is evaluated using a randomized controlled trial conducted in the province of Torino, involving 50 third grade classes in 25 schools, and 1044 students.

The key message of the paper is that innovative methodologies for teaching mathematics have the potential to reduce the gender gap in math. In particular, the treatment has a positive and statistically significant effect on girls' achievement (on average: 0.15 standard deviations). In educational studies, this effect can be considered large in magnitude and policy-relevant. In addition, we find that girls with high pre-test scores benefit the most while there are no benefits for low performing girls and for the boys.

While there are many studies on the gender gap in mathematics and its possible causes, this project is the first attempt to find a causal link between teaching methodologies and the gender gap in mathematics. So, the paper provides a very important contribution to research on the causes of the gender gap in mathematics.

Further investigation is needed. In particular, we intend to explore if changes in girls' attitudes towards math are causing their improvement in achievement (very preliminary estimates seem to exclude this channel). Two main limitations remain, due to the design of the intervention: (i) small scale; (ii) short-term results. The first issue could be addressed by testing the same methodology on a larger sample; to tackle the second one, the class-based intervention should be extended over longer periods (in this experiment was only 15 hours) and delivered for more years. Nevertheless, results are encouraging and suggest that properly designed teaching methodologies may improve math performance among girls.

References

Ajello, A.M., Caponera, E. & Palmerio, L. (2018). Italian students' results in the PISA mathematics test: does reading competence matter? *European Journal of Psychology of Education, 33*(3), 505–520.

Anderson, K. (1997). Gender Bias and Special Education Referrals. Annals of Dyslexia, 47, 151-162.

Anichini, G., Arzarello, F., Ciarrapico, L. & Robutti, O. (Eds.). (2004). *Matematica 2003. Attività didattiche e prove di verifica per un nuovo curricolo di matematica (ciclo secondario)*. Lucca: Matteoni Stampatore.

Ashcraft, M.H. (2002). Math anxiety: personal, educational, and cognitive consequences. *Curr. Dir. Psychol. Sci., 11*, 181–185.

Black, D. A., Haviland, A. M., Sanders, S. G., & Taylor, L. J. (2008). Gender wage disparities among the highly educated. *Journal of human resources*, *43*(3), 630-659.

Bloom, H.S. (2008). Chapter 9. The core analytics of randomized experiments for social research, in Alasuutari, P., Bickman, L. & Brannen, J. (eds.) *The SAGE Handbook of Social Research Methods*. London: SAGE Publications Ltd.

Boaler, J. & Greeno, J. (2000). Identity, Agency and Knowing in Mathematics Worlds. In J. Boaler (Ed.), *Multiple Perspectives on Mathematics Teaching and Learning* (pp. 171–200). Westport, CT: Ablex Publishing.

Boaler, J. (2002a). The development of disciplinary relationships: Knowledge, practice and identity in mathematics classrooms. *For the learning of mathematics*, *22* (1), 42–47.

Boaler, J. (2002b). *Experiencing School Mathematics: Traditional and Reform Approaches to Teaching and Their Impact on Student Learning*. Mahwah, NJ: Lawrence Erlbaum Association.

Boaler, J. (2009). *The Elephant in the Classroom: Helping Children Learn and Love Maths*. London: Souvenir Press.

Bolger, N. & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement, 27*(2), 165–174.

Caponera, E., Sestito, P. & Russo, P.M. (2016). The influence of reading literacy on mathematics and science achievement. *The Journal of Educational Research, 109*(2), 197–204.

Card, D., & Payne, A. A. (2017). *High school choices and the gender gap in STEM* (No. w23769). National Bureau of Economic Research.

Carlana, M. (2019) Implicit Stereotypes: Evidence from Teachers' Gender Bias, The Quarterly Journal of Economics, 134(3), 1163–1224. doi:10.1093/qje/qjz008

Contini, D., Di Tommaso, M.L. & Mendolia, S. (2017) The gender gap in mathematics achievement: Evidence from Italian data. *Economics of Education Review, 58*, 32–42.

Contini, D., Di Tommaso, M.L. & Piazzalunga, D. (2018). Tackling the Gender Gap in Mathematics in Italy. *AEA RCT Registry*. December 10. https://doi.org/10.1257/rct.3651-1.0.

Contini, D., Di Tommaso, M.L., Ferrara, F., Piazzalunga, D. & Robutti, O. (2018) The gender gap in mathematics test: Exploring variation across items. Mimeo.

De Simone, G. (2013). Render unto primary the things which are primary's: Inherited and fresh learning divides in Italian lower secondary education. *Economics of Education Review, 35*(C), 12–23.

Dee Thomas S. (2007) Teachers and the Gender Gaps in Student Achievement, *Journal of Human Resources*, 42(3), 528-554

Devine, A., Fawcett, K., Szucs, D. & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and Brain Functions, 8*:33.

Di Tommaso, M.L., Maccagnan, A., and Mandolia S. (2018) The Gender Gap in Attitudes and Test Scores: A New Construct of the Mathematical Capability, IZA Discussion Paper 11843.

Dossi, G., Figlio, D., Giuliano, P. and Sapienza, P. (2019) Born in the Family: Preferences for Boys and the Gender Gap in Math, IZA Discussion Paper 12156.

Ellison, G., and Swanson, A. (2010) The Gender Gap in Secondary School Mathematics at High Achievement Levels: Evidence from the American Mathematics Competitions, *Journal of Economic Perspectives*, 24 (2): 109-28. DOI: 10.1257/jep.24.2.109

Else-Quest, N.M., Hyde, J.S. & Linn, M.C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*(1), 101–127.

Ertl, B., Luttenberger, S., & Paechter, M. (2017). The impact of gender stereotypes on the self-concept of female students in STEM subjects with an under-representation of females. *Frontiers in psychology*, *8*, 703.

European Commission (2006). Women in Science and Technology - the Business Perspective. Luxembourg: Office for Official Publication of the European Communities.

European Commission (2012) Enhancing excellence, gender equality and efficiency in research and innovation. Luxembourg: Office for Official Publication of the European Community.

European Commission (2015). Science is a girls' thing, Newsletter Nov. 2015.

Fryer, R. G., & Levitt, S. D. (2010). An Empirical Analysis of the Gender Gap in Mathematics. *American Economic Journal: Applied Economics*, 2(2): 210–40.

Gevrek, Z. E., Neumeier, C., & Gevrek, D. (2018). Explaining the Gender Test Score Gap in Mathematics: The Role of Gender Inequality. IZA DP11260

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The quarterly journal of economics*, *118*(3), 1049-1074.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6), 1464.

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. Science, 320(5880), 1164-1165.

Gutierrez, A., & Boero, P. (2006). *Handbook of Research on the Psychology of Mathematics Education Past, Present and Future*. Rotterdam: Sense publ. (pp. 305–428).

Heniz, M., Normann, H.T. & Rau, H.A. (2016). How competitiveness may cause a gender wage gap: Experimental evidence. *European Economic Review, 90*, 336–349.

Hill, F., Mammarella, I. C., Devine, A., Caviola, S., Passolunghi, M. C. & Szucs, D. (2016). Maths anxiety in primary and secondary school students: Gender differences, developmental changes and anxiety specificity. *Learning and Individual Differences, 48*, 45–53.

Ho, H. Z., Senturk, D., Lam, A. G., Zimmer, J. M., Hong, S., Okamoto, Y., ... & Wang, C. P. (2000). The affective and cognitive dimensions of math anxiety: A cross-national study. *Journal for research in mathematics education*, 362-379.

Husain, M. & Millimet, D. (2009). The mythical ''boy crisis''? *Economics of Education Review, 28*, 38–48.

Indire (2014) The Italian Education System. *I quaderni di Euridice*, 30, available at http://www.indire.it/lucabas/lkmw_img/eurydice/quaderno_eurydice_30_per_web.pdf

Jiban, C.L. & Deno, S.L. (2007). Using math and reading curriculum-based measurements to predict state mathematics test performance: are simple one-minute measures technically adequate? *Assessment for Effective Intervention, 32*(2), 78–89.

Kersey, A. J., Csumitta, K. D., & Cantlon, J. F. (2019). Gender similarities in the brain during mathematics development. *npj Science of Learning*, *4*(1), 1-7.

Lafontaine, D. & Monseur, C. (2009). Gender gap in comparative studies of reading comprehension: to what extent do the test characteristics make a difference? *European Educational Research Journal, 8*(1), 69–79.

Lave, J. & Wenger, E. (1991). *Situated Learning. Legitimate peripheral participation*. Cambridge:

University of Cambridge Press.
Li, Q. (1999). Teachers' beliefs and gender differences in mathematics: A review. *Educational Research, 41*(1), 63–76. Qing
LoGerfo, L., Nichols, A. & Chaplin, D. (2006). *Gender gaps in math and reading gains during elementary and high school by race and ethnicity*. Washington, DC: Urban Institute. Retrieved from http://webarchive.urban.org/UploadedPDF/411428_Gender_Gaps.pdf.
Machin, S., & Puhani, P. A. (2003). Subject of degree and the gender wage differential: evidence from the UK and Germany. *Economics Letters*, *79*(3), 393-400.
Marks, G.N. (2008). Accounting for the gender gaps in student performance in reading and mathematics: evidence from 31 countries. *Oxford Review of Education*, *34*(1), 89–109.
Mullis, I.V.S., Martin, M.O. & Foy, P. (2013). The impact of reading ability on TIMSS mathematics and science achievement at the fourth grade: an analysis by item reading demands. In Martin, M.O. & Mullis, I.V.S. (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics and science achievement at the fourth grade—implications for early learning* (pp. 67–108). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College; International Association for the Evaluation of Educational Achievement (IEA).
Mullis, I.V.S., Martin, M.O., Foy, P. & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College; International Association for the Evaluation of Educational Achievement (IEA).
Nass, R. D. (1993). Sex differences in learning abilities and disabilities. Annals of Dyslexia, 43(1), 61-77.
National Academy of Science (2007). Beyond Bias and Barriers: Fulfilling the Potential of Women in Academic Science and Engineering.
Niederle, Muriel, and Lise Vesterlund. 2010. "Explaining the Gender Gap in Math Test Scores: The Role of Competition." *Journal of Economic Perspectives*, 24 (2): 129-44.
Nollenberger, N., Rodríguez-Planas, N., & Sevilla, A. (2016). The math gender gap: The role of culture. *American Economic Review*, *106*(5), 257-61.
OECD (2010). PISA 2009 results:What students know and can do—Student performance in reading,mathematics and science Vol. I. http://dx.doi.org/10.1787/9789264091450-en.
OECD (2015) The ABC of Gender Equality in Education. Aptitude, Behaviour, Confidence, OECD Publishing, Paris.
OECD (2016), PISA 2015 Results (Volume I): Excellence and Equity in Education, OECD Publishing, Paris.
Paglin, M., & Rufolo, A. M. (1990). Heterogeneous human capital, occupational choice, and male-female earnings differences. Journal of Labor Economics, 8(1, Part 1), 123-144.
Pellegrini, M., Lake, C., Inns, A., & Slavin, R. E. (2018, October). Effective programs in elementary mathematics: A best-evidence synthesis. In Annual meeting of the Society for Research on Educational Effectiveness, Washington, DC, Available online at: http://www.bestevidence.org/word/elem_math_Oct_8_2018.pdf.
Piazzalunga, D. (2018) The Gender Wage Gap among College Graduates in Italy. *Italian Economic Journal*, 4(1), 33–90.
Pope, Devin G., and Justin R. Sydnor. 2010. Geographic Variation in the Gender Differences in Test Scores. Journal of Economic Perspectives, 24 (2): 95-108.
Rathbun, A.H., West, J. & Germino-Hausken, E. (2004). From kindergarten through third grade: Children's beginning school experiences (NCES 2004-007). Washington, DC: National Center for Education Statistics.
Robinson, J.P. & Lubiensky, S.A. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school examining direct cognitive

assessments and teacher ratings. *American Educational Resources Journal*, 48, 2268–2302.
Sierminska, E., Piazzalunga, D., & Grabka, M.M. (2019) Transitioning towards more equality? Wealth gender differences and the changing role of explanatory factors over time, IZA DP 12404.
Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence *synthesis. Review of Educational Research, 78(3), 427-515.*
Stoet, G. & Geary, D.C. (2013). Sex differences in mathematics and reading achievement are inversely related: within-and across-nation assessment of 10 years of PISA data. *PLoS One, 8*(3), e57988.
Stylianides, A.J. & Stylianides, G.J. (2013). Seeking research-grounded solutions to problems of practice: classroom-based interventions in mathematics education. *ZDM Mathematics Education, 45*(3), 333–341.
Turner, S. E., & Bowen, W. G. (1999). Choice of major: The changing (unchanging) gender gap. *ILR Review*, *52*(2), 289-313.
Vogel, S. A. (1990). Gender differences in intelligence, language, visual-motor abilities, and academic achievement in students with learning disabilities: A review of the literature. Journal of Learning Disabilities, 23(1), 44-52.
Walshaw, M., Chronaki, A., Leyva, L., Stinson, D. W., Nolan, K., & Mendick, H. (2017). Beyond the box: Rethinking gender in mathematics education research. In A. Chronaki (Ed.), *Proceedings of the 9th International Mathematics Education and Society Conference* (MES9, Vol. 1, 184–188). Volos, Greece: MES9.
Wehmeyer, M. L., & Schwartz, M. (2001). Disproportionate representation of males in special education services: Biology, behavior, or bias?. Education and treatment of children, 28-45.
Wilder, G.Z. & Powell, K. (1989). Sex differences in test performance: a survey of the literature. *ETS Research Report Series, 1989*(1), i–50.
Zohar, A., & Sela, D. (2003). Her physics, his physics: gender issues in Israeli advanced placement physics classes. *International Journal of Science Education*, 25(2), 245–268.

TABLES

Tab.1 Sample selection

Sample	Children	Treated	Controls
Full sample (a)	1,044	519	525
Present at the pre-test (b)	933	452	481
Present at the post-test (c)	983	490	493
Present at the pre-test and post-test (d)	888	431	457

Tab.2 Baseline characteristics of treated and control children, full sample

Panel A – Individual level	Control group	Treated group	P-value of the difference
Girl	0.500	0.514	0.663
SEND – broad definition	0.148	0.156	0.736
SEND – broad def. (F)	0.106	0.138	0.260
SEND – broad def. (M)	0.190	0.174	0.634
SEND – narrow definition	0.085	0.082	0.868
SEND – narrow definition (F)	0.045	0.063	0.362
SEND – narrow definition (M)	0.125	0.103	0.419
Native Child	0.847	0.876	0.176
Migrant I generation	0.011	0.021	0.212
Migrant II generation	0.127	0.096	0.109
Migrant missing	0.013	0.005	0.210
Mother educ (lower secondary)	0.219	0.229	0.691
Mother educ (upper secondary)	0.280	0.354	0.096
Mother educ (tertiary)	0.299	0.236	0.023
Mother educ (missing)	0.201	0.179	0.350
Mother at least upper secondary	0.579	0.591	0.682
Father educ (lower secondary)	0.224	0.254	0.263
Father educ (upper secondary)	0.417	0.443	0.396
Father educ (tertiary)	0.163	0.142	0.341
Father educ (missing)	0.194	0.159	0.146
Father at least upper secondary	0.580	0.585	0.875
Observations	525	519	1,044
Pre-test score	10.394	10.152	0.816
Pre-test score (F)	10.351	10.152	0.615
Pre-test score (M)	11.179	11.274	0.671
Observations	481	452	933
Panel B – Class level			
Class size	21.000	20.760	0.818
Pre-test score (mean)	10.783	10.646	0.728
Pre-test score (s.d.)	4.310	4.219	0.621
Percent of female students	0.500	0.512	0.630
Percent of I gen migrant students	0.011	0.018	0.422
Percent of II gen migrant students	0.136	0.098	0.254
Percent of SEND (broad)	0.146	0.155	0.718
Percent of SEND (narrow)	0.083	0.082	0.954
Observations	25	25	50
Permanent contract teachers %	100.00	92.00	0.164
Teaching experience (years)	21.375	22.560	0.720
Teaching exp in math (years)	13.695	14.200	0.867
Teaching math in the class (years)	2.791	2.400	0.093
Teacher's university degree %	0.375	0.400	0.861
Teachers age (years)	48.33	50.00	0.501
Observations	24	25	49

SEND stands for "special educational needs and disability". Summary statistics refer to full sample (a). Summary statistics of pre-test refers to 933 observations (sample (b)). Teaching experience includes the year of the intervention, but some teachers started teaching in the second semester, thus, they reply that they have been teaching for less than one year, i.e. 0 years.

Tab.3 Attrition at pre-test and post-test

		Overall	Girls	Boys
Post-test [a]	Overall attrition	0.054	0.052	0.056
	Control	0.055	0.049	0.061
	Treated	0.054	0.056	0.051
	Difference (T-C)	-0.001	0.006	-0.009
		(0.141)	(0.194)	(0.020)
Pre and post-test [b]	Overall attrition	0.149	0.153	0.138
	Control	0.124	0.125	0.123
	Treated	0.167	0.179	0.155
	Difference (T-C)	0.043**	0.053*	0.037
		(0.021)	(0.031)	(0.303)

Notes: Standard errors of the difference in parenthesis. [a] Sample (c); [b] Sample (d).
*** p<0.01, ** p<0.05, * p<0.1

Tab.4 Attendance to the laboratory sessions

Share of labs attended	% children	% boys	% girls
0%	0.00%	0.00%	0.00%
≥ 50%	99.30%	100%	98.63%
≥ 70%	95.82%	97.16%	94.52%
≥ 80%	94.19%	95.75%	92.69%
100%	73.78%	75.94%	71.68%
Observations	431	212	219

Note: 100% of laboratories corresponds to 15 hours. Sample (d) (children present at pre- and post-test).

Tab.5 Main results: effects of the treatment

	Post-test scores			Post-test scores			Post-test scores controlling for pre-test scores			Post-test scores controlling for pre-test, school FE, family background and class size		
Variable	Overall	Girls	Boys	Overall	Girls	Boys	Overall	Girls	Boys	Overall	Girls	Boys
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treatment	0.116*	0.154*	0.081	0.091	0.143	0.041	0.077	0.164**	-0.015	0.091***	0.150***	0.004
	(0.065)	(0.086)	(0.086)	(0.068)	(0.090)	(0.092)	(0.048)	(0.069)	(0.068)	(0.032)	(0.056)	(0.046)
Pre-test score							0.760***	0.733***	0.788***	0.736***	0.735***	0.742***
							(0.023)	(0.037)	(0.024)	(0.026)	(0.034)	(0.032)
Constant	-0.048	-0.208***	0.115*	-0.030	-0.191***	0.133**	0.007	-0.132**	0.048	-0.072	-0.286	-0.069
	(0.045)	(0.053)	(0.058)	(0.046)	(0.051)	(0.063)	(0.040)	(0.058)	(0.045)	(0.210)	(0.252)	(0.318)
R-squared	983	501	482	0.002	0.006	0.000	0.592	0.572	0.601	0.622	0.607	0.656
Obs.	0.003	0.007	0.002	888	448	440	888	448	440	888	448	440
School FE										YES	YES	YES
Addit. controls										YES	YES	YES

Notes: Standardized test scores. Standard errors clustered at the class level in parenthesis. Columns 1 to 3 use sample (c) (children present at the post-test); columns 4 to 12 use sample (d) (children present at the pre- and post-test). In columns 7, and 10 the control variable "Girl" is also included. Additional controls include SEND (special education needs and disability) dummy, broad definition; mother's level of education; father's level of education; migratory background; and class size.

*** p<0.01, ** p<0.05, * p<0.1

Tab.6 Heterogeneous effects of the treatment by prior achievement's levels

Variable	(1) Overall	(2) Girls	(3) Boys
Treatment	0.089***	0.164***	0.001
	(0.032)	(0.054)	(0.048)
Pre-test score	0.714***	0.676***	0.731***
	(0.039)	(0.049)	(0.041)
Treatment* Pre-test score	0.064	0.128**	0.024
	(0.047)	(0.062)	(0.055)
Constant	-0.097	-0.254	-0.068
	(0.213)	(0.250)	(0.322)
Observations	888	448	440
R-squared	0.621	0.611	0.656
School FE	YES	YES	YES
Additional controls	YES	YES	YES

Notes: Standardized test scores. Standard errors clustered at the class level in parenthesis. Additional controls include girl (in the Overall specification), SEND (special education needs and disability) dummy, broad definition; mother's level of education; father's level of education; migratory background; and class size. Sample (d)
*** p<0.01, ** p<0.05, * p<0.1

Tab.7 Heterogeneous effects of the treatment by migrant status and parents' education

		Overall (1)	Girls (2)	Boys (3)
Effect of treatment by migrant status	Treatment on natives	0.104**	0.118*	0.044
		(0.041)	(0.065)	(0.062)
	Treatment on migrants	0.007	0.374***	-0.264*
		(0.089)	(0.134)	(0.150)
	Observations	888	448	440
	R-squared	0.622	0.608	0.656
Effect of treatment by mother's level of education	Treatment on children with mother lower sec. educ	0.064	0.299***	-0.108
		(0.078)	(0.111)	(0.119)
	Treatment on children with mother upper sec. educ	0.063	0.055	0.035
		(0.064)	(0.104)	(0.090)
	Treatment on children with mother tertiary educ	0.062	0.136	-0.088
		(0.077)	(0.110)	(0.103)
	Observations	888	448	440
	R-squared	0.617	0.607	0.646
Effect of treatment by father's level of education	Treatment on children with father lower sec. educ	-0.020	0.259*	-0.323**
		(0.094)	(0.134)	(0.122)
	Treatment on children with father upper sec. educ	0.096**	0.124	0.027
		(0.051)	(0.084)	(0.073)
	Treatment on children with father tertiary educ	0.249**	0.143	0.355***
		(0.099)	(0.152)	(0.160)
	Observations	888	448	440
	R-squared	0.622	0.605	0.664
	Pre-test scores	YES	YES	YES
	School FE	YES	YES	YES
	Additional controls	YES	YES	YES

Notes: Standardized test scores. Standard errors clustered at the class level in parenthesis. Additional controls include girl (only in the Overall specification), SEND (special education needs and disability) dummy, broad definition; mother's level of education; father's level of education; migratory background; and class size.
*** p<0.01, ** p<0.05, * p<0.1

Tab.8 Robustness checks

	Post-test scores excluding children with certified special educational needs or disabilities			Post-test scores excluding children with any special educational needs or disabilities			Post-test scores including pre-test score missing dummy			Post-test score including children sitting the post-test deferred session		
	Overall	Girls	Boys	Overall	Girls	Boys	Overall	Girls	Boys	Overall	Girls	Boys
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treatment	0.101***	0.151***	0.019	0.118***	0.171***	0.015	0.116***	0.180***	0.046	0.082**	0.124**	0.011
	(0.035)	(0.056)	(0.050)	(0.036)	(0.057)	(0.051)	(0.037)	(0.055)	(0.049)	(0.031)	(0.051)	(0.047)
Pre-test scores	0.760***	0.739***	0.766***	0.764***	0.732***	0.786***	0.728***	0.710***	0.726***	0.741***	0.738***	0.733***
	(0.028)	(0.036)	(0.032)	(0.027)	(0.034)	(0.033)	(0.029)	(0.037)	(0.035)	(0.026)	(0.034)	(0.033)
Pre-test sc. missing							-0.051	-0.177	0.101			
							(0.098)	(0.127)	(0.150)			
Constant	-0.197	-0.281	-0.272	-0.113	-0.083	-0.148	-0.244	-0.606**	-0.011	-0.056	-0.116	-0.080
	(0.234)	(0.244)	(0.381)	(0.212)	(0.232)	(0.393)	(0.238)	(0.292)	(0.310)	(0.196)	(0.223)	(0.321)
Observations	818	425	393	757	396	361	983	501	482	916	462	454
R-squared	0.615	0.608	0.640	0.599	0.590	0.630	0.567	0.557	0.597	0.614	0.598	0.650
School FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Additional controls	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Class Size	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
SEND	Restricted Version	Restricted Version	Restricted Version	Broader Version	Broader Version	Broader Version						

Notes: Standardized test scores. Standard errors clustered at the class level in parenthesis. Additional controls include girl (only in the Overall specification), SEND (special education needs and disability) dummy when appropriate, broad definition; mother's level of education; father's level of education; migratory background; and class size.
*** p<0.01, ** p<0.05, * p<0.1

Tab.9 Treatment effect by type of item

		Girls		Boys	
All items	**Outcome**	**Treatm. Effect**	**S.E.**	**Treatm. Effect**	**S.E.**
	Post-test score	0.152**	0.059	-0.028	0.061
DIFFICULTY	**Outcome**	**Treatm. Effect**	**S.E.**	**Treatm. Effect**	**S.E.**
	Easy items score	0.014	0.077	0.032	0.073
	Medium items score	0.123*	0.067	-0.100	0.064
	Difficult items score	0.258***	0.071	0.080	0.078
		Chi2	**P-value**	**Chi2**	**P-value**
	Breusch-Pagan test	48.46	0.000	86.99	0.000
	Easy = Medium	1.392	0.238	2.445	0.118
	Easy = Difficult	5.586	0.018	0.238	0.626
	Medium = Difficult	2.627	0.105	4.660	0.031
FORMAT	**Outcome**	**Treatm. Effect**	**S.E.**	**Treatm. Effect**	**S.E.**
	Open Answers score	0.125*	0.065	-0.052	0.066
	Multiple Choice score	0.163**	0.067	0.013	0.066
		Chi2	**P-value**	**Chi2**	**P-value**
	Breusch-Pagan test	37.37	0.000	59.19	0.000
	Open Ans. = Multiple Choice	0.241	0.624	0.773	0.379
DIMENSION	**Outcome**	**Treatm. Effect**	**S.E.**	**Treatm. Effect**	**S.E.**
	Knowing score	0.162***	0.063	0.002	0.067
	Arguing score	0.108	0.080	-0.118	0.089
	Problem Solving score	0.101	0.069	-0.008	0.066
		Chi2	**P-value**	**Chi2**	**P-value**
	Breusch-Pagan test	75.53	0.000	79.62	0.000
	Knowing = Arguing	0.341	0.559	1.338	0.247
	Knowing = Problem Solving	0.615	0.433	0.018	0.893
	Arguing = Problem Solving	0.006	0.937	1.321	0.250
	Observations	448		440	
	School FE	YES		YES	
	Pre-test score	YES		YES	
	Additional controls	NO		NO	

Notes: Test scores standardized. The treatment effect is estimated with an OLS regression in the "All item" case. For each group of outcomes (difficulty, format, dimension) the treatment effects are estimated with a SUR (seemingly unrelated regression) model, in which the error terms are assumed to be correlated across equations. In all equations, we control for school fixed effects and the pre-test score. Below the SUR results, we report the results of the Breusch-Pagan test for independent equations and the tests of equivalence among the treatment coefficients of interest, together with the corresponding p-values. Difficulty classifies the item's difficulty into three categories (easy, medium, high), using a one-parameter IRT model and (+/-) 0.5 as a threshold. Format classifies items by the type of answer (open answer vs. multiple choice). Dimension classifies the item according to the mathematical thinking behind a specific question (Knowing, Arguing, Problem-solving). The classification of single items can be seen in Table A.10.
*** p<0.01, ** p<0.05, * p<0.1

Tab.10 Comparison of experimental classes with Piedmont and Italy

Variable	Experimental Classes (1)	Piedmont Classes (2)	P-value of the differences (3)	Italian Classes (4)	P-value of the differences (5)
Invalsi score in Italian	215.727	203.296	0.000	200.634	0.000
Invalsi score in Math	222.451	201.860	0.000	200.929	0.000
Invalsi score Italian Female	215.532	205.313	0.000	201.674	0.000
Invalsi score Math Female	217.684	199.308	0.000	198.643	0.000
Invalsi score Italian Male	216.264	201.237	0.000	199.695	0.000
Invalsi score Math Male	227.333	204.484	0.000	203.129	0.000
Gender Gap Math	-9.649	-5.175	0.000	-4.485	0.000
Oral marks Italian	8.140	8.105	0.354	8.058	0.011
Oral marks Math	8.224	8.230	0.863	8.143	0.014
Kindergarten attendance	41.995	32.724	0.000	38.086	0.000
Girl	51.005	50.467	0.000	50.467	0.021
Mother's education					
Primary school	0.790	1.450	0.176	1.910	0.000
Lower secondary	17.610	22.440	0.000	23.570	0.000
Lower secondary-professional qualification	7.376	9.440	0.005	6.740	0.002
Upper secondary	40.525	40.760	0.675	41.310	0.000
Upper secondary –technical diploma	2.084	3.630	0.004	2.210	0.445
Tertiary	31.613	22.280	0.000	24.230	0.000
Father's education					
Primary school	1.309	1.310	0.999	2.340	0.000
Lower secondary	25.374	31.830	0.000	31.180	0.000
Lower secondary-professional qualification	9.286	13.580	0.000	8.490	0.000
Upper secondary	40.532	35.270	0.000	39.950	0.000
Upper secondary –technical diploma	1.491	1.800	0.556	1.640	0.260
Tertiary	22.005	16.20	0.000	16.390	0.000
Max n. of obs.	1,044	1,391		26,142	

Notes: Maximum number observation reported. The number of observations varies depending on the variable and the missing values. Range of variation: Experimental classes 863 (min)–1,044 (max); Piedmont classes: 16 (min)-1,391 (max); Italian classes 347 (min) – 26,142 (max).

FIGURES

Fig.1 Timeline of the intervention

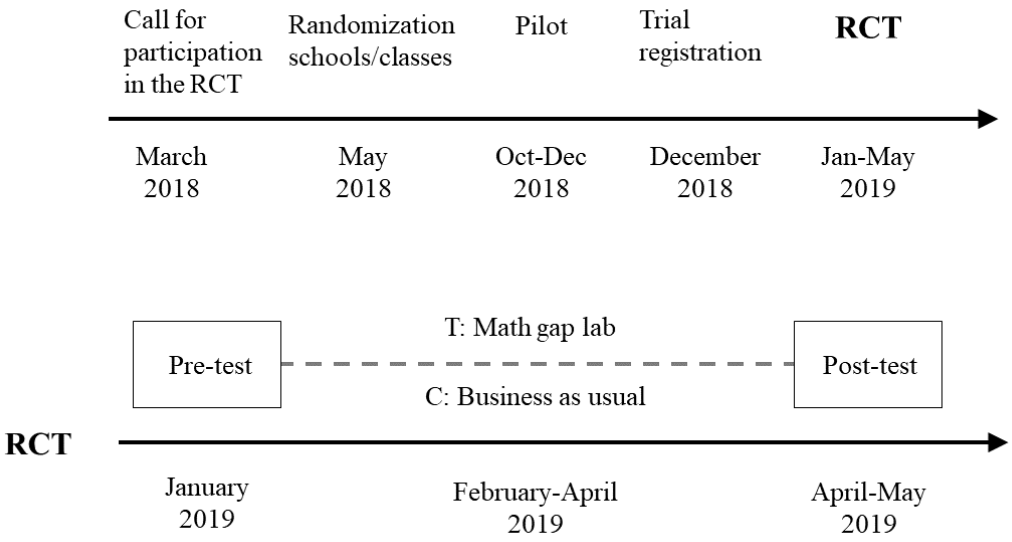
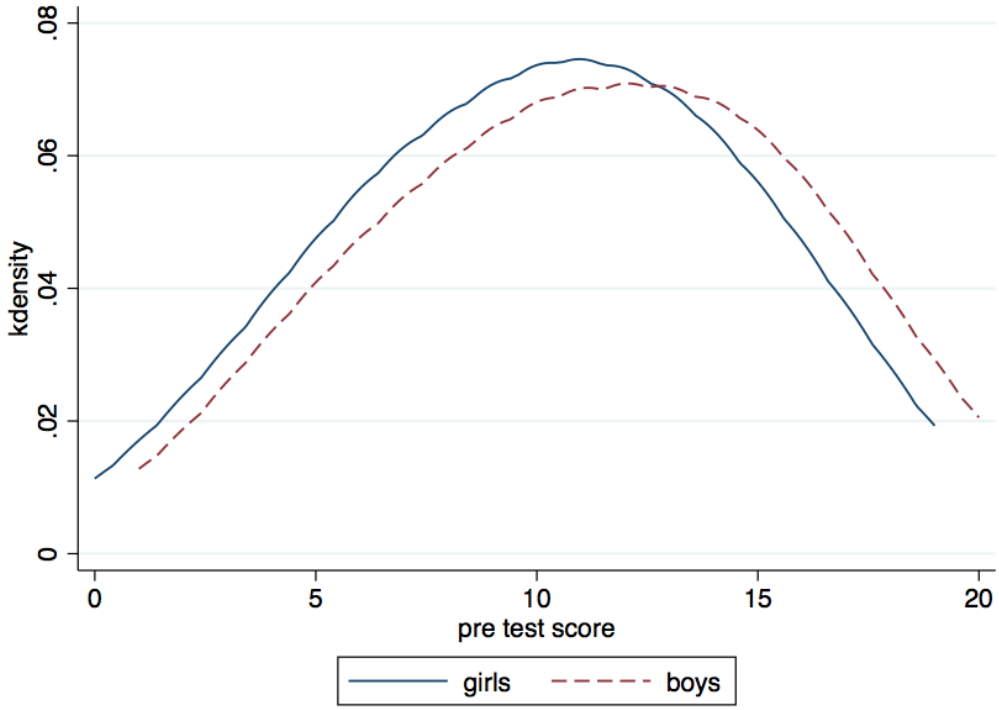
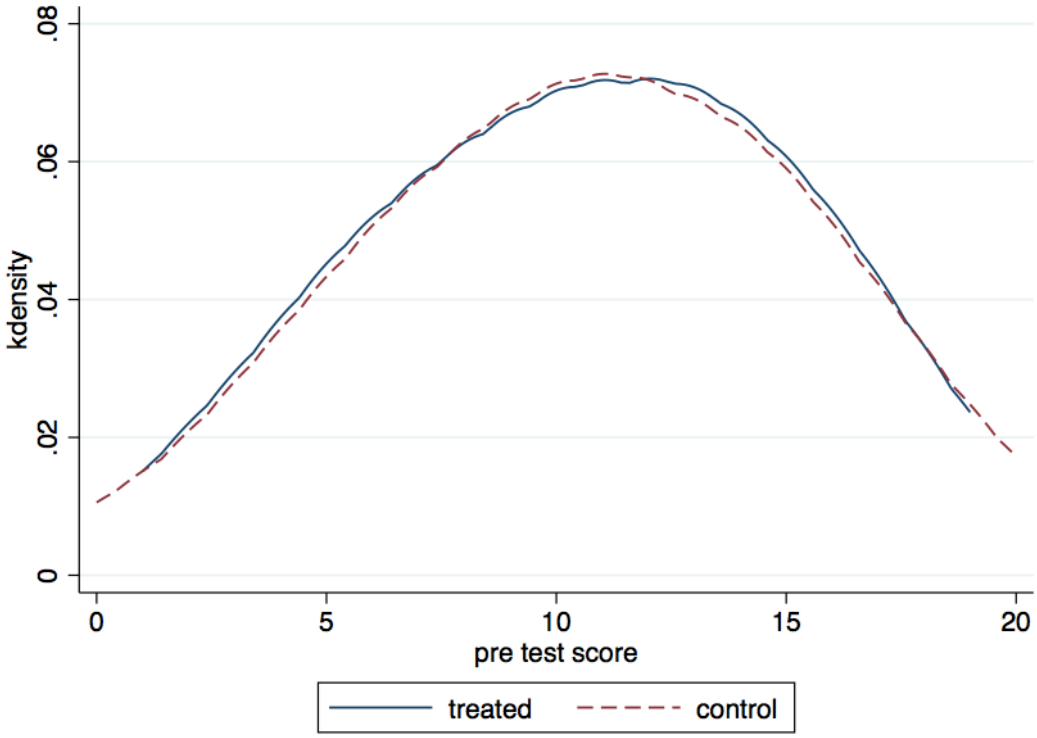


Fig.2 Gender gap in the pre-test



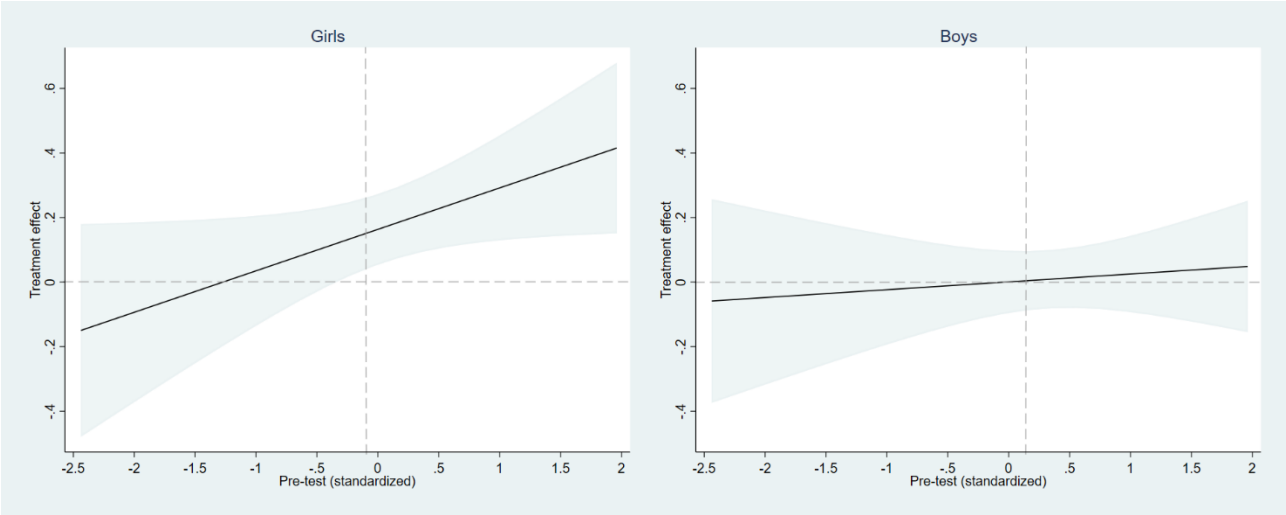
Note: Children present at the pre-test (sample (b)), 933 observations.

Fig.3 Pre-test score distribution by treatment status



Note: Children present at the pre-test (sample (b)), 933 observations.

Fig.4 Treatment effect by prior achievement's levels



Note: Effect of the treatment by pre-test scores for boys and girls (estimates from regression in Tab.8). Sample (d), 888 observations 888. The dashed horizontal line represents a zero treatment effect, whereas the dashed vertical line represents the pre-test score mean for girls and boys respectively.

APPENDIX

Appendix A

Tab.A.1 Variables definition

Variable	Description
Individual level	
Pre-test score	Pre-test score
Girl	1= girl; 0 = boy
SEND – broad definition	1= child with any form of special education needs or disability; 0 = otherwise
SEND – narrow definition	1= child with only certified special educ. needs or disability; 0 = otherwise
Native Child	1= child born in Italy with at least one parent born in Italy; 0 = otherwise
Migrant I generation	1= child born abroad with both parents born abroad; 0 = otherwise
Migrant II generation	1= child born in Italy with both parents born abroad; 0 = otherwise
Migrant missing	1= missing info on child and parents' birthplace; 0 = otherwise
Mother educ (lower secondary)	1= mother level of education is lower secondary or less (including 3 years of professional education at high school); 0 = otherwise
Mother educ (upper secondary)	1= mother level of education is upper secondary; 0 = otherwise
Mother educ (tertiary)	1= mother level of education is tertiary or above; 0 = otherwise
Mother educ (missing)	1= mother level of education is missing; 0 = otherwise
Mother at least upper secondary	1= mother level of education is at least upper secondary; 0 = otherwise
Father educ (lower secondary)	1= father level of education is lower secondary or less (including 3 years of professional education at high school); 0 = otherwise
Father educ (upper secondary)	1= father level of education is upper secondary; 0 = otherwise
Father educ (tertiary)	1= father level of education is tertiary; 0 = otherwise
Father educ (missing)	1= father level of education is missing; 0 = otherwise
Father at least upper secondary	1= father level of education is at least upper secondary; 0 = otherwise
Class level	
Class size	Number of children in each class
Pre-test score (mean)	Mean of pretest score at class level
Pre-test score (s.d.)	Standard deviation of pretest score at class level
Percent of female students	Percent of female students in the class
Percent of I gen migrant students	Percent of I generation migrants in the class
Percent of II gen migrant students	Percent of II generation migrants in the class
Percent of SEND (broad)	Percent of children with any form of special educ. needs or disability in the class
Percent of SEND (narrow)	Percent of children with only certified special educ. needs or disability in the class
Permanent contract teachers %	Percentage of teachers with a permanent contract
Teaching experience (years)	Number of years teacher has been teaching
Teaching exp in math (years)	Number of years teacher has been teaching math
Teaching math in the class (years)	Number of years teacher has been teaching math in the class
Teacher's university degree %	Percentage of teachers with university degree
Teacher's age (years)	Age of teacher

Tab.A.2 Primary schools in the province of Torino, application and participation into the program

	Schools	Classes
Population	180	-
Applicants	31	100
Eligible	30	82
Sampled	25	50

Tab.A.3 Sample selection, details

Sample	Children	Treated	Controls
Full sample (a)	1,044	519	525
Present at the pre-test (b)	933	452	481
Present at the post-test (c)	983	490	493
Present at the pre-test and post-test (d)	888	431	457
Provide background information (e)	759	385	374
Present at the pre-test and post-test and provide background information (f)	659	327	334
Number of pupils with all item missing	4	1	3
Number of SEND narrow in the full sample	88	43	45
Number of SEND broad in the full sample	159	81	78
Post-test in the deferred session	35	20	15

Note: SEND stands for “special educational needs and disability”.

Tab.A.4 Baseline characteristics of treated and control children, sample (c)

	Control group	Treated group	P-value of the difference
Pre-test score [a]	10.772	10.856	0.774
Girl	0.505	0.514	0.772
SEND – broad definition	0.139	0.148	0.687
SEND – narrow definition	0.079	0.077	0.927
Native Child	0.849	0.885	0.097
Migrant I generation	0.012	0.020	0.308
Migrant II generation	0.123	0.089	0.085
Migrant missing	0.014	0.004	0.096
Mother educ (lower secondary)	0.223	0.224	0.959
Mother educ (upper secondary)	0.290	0.348	0.047
Mother educ (tertiary)	0.290	0.246	0.127
Mother educ (missing)	0.196	0.179	0.491
Mother at least upper secondary	0.580	0.595	0.615
Father educ (lower secondary)	0.227	0.251	0.381
Father educ (upper secondary)	0.419	0.438	0.550
Father educ (tertiary)	0.164	0.144	0.400
Father educ (missing)	0.188	0.165	0.338
Father at least upper secondary	0.584	0.583	0.987
By gender			
Pre-test score (F) [a]	10.358	10.232	0.756
Pre-test score (M) [a]	11.188	11.500	0.455
SEND – broad def. (F)	0.100	0.126	0.349
SEND – broad def. (M)	0.180	0.172	0.816
SEND – narrow definition (F)	0.040	0.059	0.320
SEND – narrow definition (M)	0.118	0.094	0.432
Observations	493	490	983
[a] Observations	457	431	888

Notes: SEND stands for "special educational needs and disability". Summary statistics refer to children present at the post-test (sample (c)).

Tab.A.5 Baseline characteristics of treated and control children, sample (d)

	Control group	Treatment group	P-value of the difference
Pre-test score	10.772	10.856	0.774
Girl	0.501	0.508	0.834
SEND – broad definition	0.144	0.150	0.788
SEND – narrow definition	0.080	0.076	0.808
Native Child	0.879	0.851	0.221
Migrant I generation	0.002	0.008	0.133
Migrant II generation	0.095	0.126	0.133
Migrant missing	0.004	0.013	0.181
Mother educ (lower secondary)	0.218	0.234	0.581
Mother educ (upper secondary)	0.293	0.364	0.024
Mother educ (tertiary)	0.295	0.225	0.017
Mother educ (missing)	0.192	0.176	0.534
Mother at least upper secondary	0.588	0.589	0.983
Father educ (lower secondary)	0.216	0.262	0.111
Father educ (upper secondary)	0.424	0.438	0.674
Father educ (tertiary)	0.168	0.127	0.087
Father educ (missing)	0.190	0.171	0.470
Father at least upper secondary	0.592	0.566	0.418
By gender			
Pre-test score (F)	10.358	10.232	0.756
Pre-test score (M)	11.188	11.500	0.455
SEND – broad def. (F)	0.104	0.127	0.447
SEND – broad def. (M)	0.184	0.174	0.792
SEND – narrow definition (F)	0.043	0.059	0.453
SEND – narrow definition (M)	0.118	0.094	0.415
Observations	457	431	888

Notes: SEND stands for “special educational needs and disability”. Summary statistics refers to children present at pre and post-test (sample (d)).

Tab. A.6 Effect of baseline characteristics on the probability of being treated

Variables	Treatment	Treatment
Pre-test score	0.092	0.087
	(0.086)	(0.079)
Girl	0.023	0.048
	(0.078)	(0.081)
SEND – broad definition	0.190	0.111
	(0.190)	(0.179)
Migrant I generation	1.256**	0.861
	(0.572)	(0.547)
Migrant II generation	-0.421***	-0.408**
	(0.152)	(0.162)
Migrant missing	-0.417	-0.912
	(0.905)	(0.935)
Mother educ (upper secondary)	0.173	0.150
	(0.162)	(0.158)
Mother educ (tertiary)	-0.502**	-0.354
	(0.218)	(0.224)
Mother educ (missing)	-0.591**	-0.193
	(0.283)	(0.306)
Father educ (upper secondary)	-0.118	-0.174
	(0.194)	(0.192)
Father educ (tertiary)	-0.402*	-0.561**
	(0.233)	(0.252)
Father educ (missing)	-0.287	-0.460*
	(0.289)	(0.274)
Class size	-0.138	-0.119
	(0.195)	(0.176)
Teaching experience	-0.045	--
	(0.054)	
Teacher's university degree	0.432	--
	(1.111)	
Teachers age	0.103	--
	(0.077)	
Constant	-1.265	2.667
	(5.170)	(3.479)
Observations	845	888
	188.69	103.60
Wald test of joint significance	(0.000)	(0.000)
School FE	YES	YES

Notes: Standardized pre-test scores. Standard errors clustered at the class level in parenthesis. Sample (d). Reference categories are: boy, typically developed child, native child, mother's lower education, fathers' lower education.

*** p<0.01, ** p<0.05, * p<0.1

Tab.A.7 Effect of the treatment controlling for individual and family background characteristics – full results

Variables	Overall (1)	Girls (2)	Boys (3)
Treatment	0.091***	0.150***	0.004
	(0.032)	(0.056)	(0.046)
Pre-test score	0.736***	0.735***	0.742***
	(0.026)	(0.034)	(0.032)
Girl	-0.090*	--	--
	(0.046)		
SEND broad def.	-0.098	0.046	-0.194*
	(0.065)	(0.127)	(0.098)
Migrant I generation	-0.022	0.012	-0.090
	(0.140)	(0.223)	(0.187)
Migrant II generation	0.068	0.016	0.152
	(0.075)	(0.099)	(0.125)
Migrant missing	-0.246**	-0.051	-0.672*
	(0.110)	(0.224)	(0.368)
Mother educ (upper secondary)	0.055	-0.008	0.116
	(0.060)	(0.085)	(0.092)
Mother educ (tertiary)	0.060	0.043	0.090
	(0.072)	(0.119)	(0.111)
Mother educ (missing)	-0.120	-0.235*	0.106
	(0.080)	(0.117)	(0.139)
Father educ (upper secondary)	0.109	0.098	0.127
	(0.080)	(0.106)	(0.119)
Father educ (tertiary)	0.289***	0.165	0.386**
	(0.099)	(0.138)	(0.166)
Father educ (missing)	0.271**	0.120	0.381***
	(0.115)	(0.175)	(0.129)
Class Size	-0.009	0.006	-0.016
	(0.009)	(0.013)	(0.014)
Constant	-0.072	-0.286	-0.069
	(0.210)	(0.252)	(0.318)
R-squared	0.622	0.607	0.656
Observations	888	448	440
School FE	YES	YES	YES

Notes: Standardized test scores. Standard errors clustered at the class level in parenthesis. The Table corresponds to columns 10, 11, 12 of Table 5. Reference categories are: boy, typically developed child, native child, mother's lower education, fathers' lower education.

*** p<0.01, ** p<0.05, * p<0.1

Tab.A.8 Main results with IRT scores

Dependent var.	Post-test std.	Post-test std.	Ability from IRT 1 p.	Ability from IRT 1 p.	Ability from IRT 2 p.	Ability from IRT 2 p.	Ability from IRT 2 p. (controls)	Ability from IRT 2 p. (controls)
	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.150***	0.004	0.145***	0.004	0.128***	-0.000	0.132***	0.005
	(0.056)	(0.046)	(0.049)	(0.043)	(0.044)	(0.043)	(0.045)	(0.043)
Pre-test score std.	0.735***	0.742***						
	(0.034)	(0.032)						
Pre-test ability IRT 1p.			0.739***	0.726***				
			(0.038)	(0.034)				
Pre-test ability IRT 2p.					0.744***	0.751***	0.754***	0.758***
					(0.037)	(0.034)	(0.038)	(0.034)
Constant	-0.286	-0.069	-0.232	-0.132	-0.315	-0.081	-0.249	-0.013
	(0.252)	(0.318)	(0.228)	(0.278)	(0.204)	(0.253)	(0.214)	(0.269)
Observations	448	440	448	440	448	440	448	440
R-squared	0.607	0.656	0.604	0.641	0.611	0.655	0.611	0.651
School FE	YES	YES	YES	YES	YES	YES	YES	YES
Additional controls	YES	YES	YES	YES	YES	YES	YES	YES

Notes: Columns (1) and (2) report the results of our preferred specification and use standardized pre- and post-test scores (they correspond to columns (11) and (12) of Table 5). Columns (3) and (4) use as outcome and pre-test the latent abilities predicted with a one-parameter IRT (Item Response Theory) logistic model; columns (5) and (6) the latent abilities predicted with a two-parameters IRT model; columns (7) and (8) use as outcome the latent abilities predicted with a two-parameters IRT model estimated on the control group only (predicted abilities for both control and treated pupils). Additional controls include SEND (special education needs and disability) dummy, broad definition; mother's level of education; father's level of education; migratory background; and class size.

Standard errors clustered at the class level in parenthesis. Sample (d). *** p<0.01, ** p<0.05, * p<0.1

Tab.A.9 Heterogeneous results by prior achievements with IRT scores

Dependent var.	Post-test std.	Post-test std.	Ability from IRT 1 p.	Ability from IRT 1 p.	Ability from IRT 2 p.	Ability from IRT 2 p.	Ability from IRT 2 p. (controls)	Ability from IRT 2 p. (controls)
	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.164***	0.001	0.161***	0.002	0.141***	-0.005	0.146***	0.002
	(0.054)	(0.048)	(0.047)	(0.044)	(0.043)	(0.045)	(0.044)	(0.045)
Pre-test score	0.676***	0.731***	0.679***	0.717***	0.691***	0.733***	0.701***	0.744***
	(0.049)	(0.041)	(0.054)	(0.045)	(0.054)	(0.043)	(0.054)	(0.042)
Treatment* Pre-test score	0.128**	0.024	0.131*	0.019	0.116*	0.039	0.116*	0.031
	(0.062)	(0.055)	(0.068)	(0.060)	(0.067)	(0.059)	(0.068)	(0.060)
Constant	-0.254	-0.068	0.161***	-0.132	-0.286	-0.079	-0.220	-0.011
	(0.250)	(0.322)	(0.047)	(0.280)	(0.202)	(0.258)	(0.211)	(0.273)
Observations	448	440	440	440	448	440	448	440
R-squared	0.611	0.656	0.656	0.641	0.614	0.655	0.614	0.651
School FE	YES	YES	YES	YES	YES	YES	YES	YES
Additonal controls	YES	YES	YES	YES	YES	YES	YES	YES

Notes: Columns (1) and (2) report the heterogeneous results of our preferred specification and use standardized pre- and post-test scores (they correspond to columns (2) and (3) of Tab.6). Columns (3) and (4) use as outcome and pre-test the latent abilities predicted with a one-parameter IRT (Item Response Theory) logistic model; columns (5) and (6) the latent abilities predicted with a two-parameters IRT model; columns (7) and (8) use as outcome the latent abilities predicted with a two-parameters IRT model estimated on the control group only (predicted abilities for both control and treated pupils). Additional controls include SEND (special education needs and disability) dummy, broad definition; mother's level of education; father's level of education; migratory background; and class size. Pre-test scores are always the appropriate ones (e.g. standardized, IRT 1p., or IRT 2p. depending on the outcome used.

Standard errors clustered at the class level in parenthesis. Sample (d). *** p<0.01, ** p<0.05, * p<0.1

Tab.A.10 Item classification, post-test

Question	Item	Difficulty score	Difficulty level	Format	Dimension
D1	1	1.244	Difficult	Open	Knowing
D2_a	2	-1.357	Easy	Open	Knowing
D2_b	3	1.323	Difficult	Open	Knowing
D3	4	-0.252	Medium	Multiple	Knowing
D4	5	0.207	Medium	Open	Knowing
D5_a	6	-0.991	Easy	Open	Problem solving
D5_b	7	2.897	Difficult	Open	Problem solving
D6	8	-0.272	Medium	Open	Problem solving
D7_a	9	-1.466	Easy	Multiple	Knowing
D7_b	10	1.270	Difficult	Multiple	Arguing
D8_a	11	-0.242	Medium	Open	Knowing
D8_b	12	0.246	Medium	Open	Knowing
D9	13	-0.410	Medium	Open	Problem solving
D10_a	14	-0.086	Medium	Multiple	Problem solving
D10_b	15	0.838	Difficult	Multiple	Problem solving
D11_a	16	0.276	Medium	Open	Arguing
D11_b	17	-0.164	Medium	Open	Arguing
D12	18	-0.802	Easy	Multiple	Knowing
D13_a	19	-0.696	Easy	Multiple	Problem solving
D13_b	20	-0.500	Medium	Multiple	Problem solving

Tab.A.11 Attitudes, summary statistics

	Variable	Obs.	Mean	Std. Dev.	Min	Max
Overall	Attitudes (sum)	882	15.147	3.351	5	20
	Attitudes (PCA)	882	0.685	0.218	0	1
Boys	Attitudes (sum)	438	15.554	3.299	5	20
	Attitudes (PCA)	438	0.713	0.214	0	1
Girls	Attitudes (sum)	444	14.745	3.358	5	20
	Attitudes (PCA)	444	0.658	0.219	0	1
	Variable	Obs.	Diff	Std. Err	P-value of the diff	
Mean diff. Boys vs. Girls	Attitudes (sum)	882	0.809	0.224	0.000	
	Attitudes (PCA)	882	0.054	0.014	0.000	

Notes: The indexes for attitudes are constructed from five questions, with four possible Likert-type answers, coded from 1 (not at all) to 4 (a lot). Attitudes (sum) is an index build as a sum of such points, whereas attitudes (PCA) is an index extracted with a Principal Component Analysis.

Tab.A.12 Effect of the treatment on attitudes towards mathematics

Variable	Attitudes (Sum) (1)	Attitudes (PCA) (2)
Girls	-0.849**	-0.056**
	(0.380)	(0.025)
Treatment effect on boys	-0.487	-0.031
	(0.297)	(0.019)
Treatment effect on girls	-0.469	-0.031
	(0.359)	(0.024)
Constant	16.093***	0.752***
	(0.669)	(0.044)
Observations	882	882
R-squared	0.075	0.079
School FE	YES	YES
Additional controls	YES	YES

Notes: The indexes for attitudes are constructed from five questions, with four possible Likert-type answers, coded from 1 (not at all) to 4 (a lot). Attitudes (sum) is an index build as a sum of such points, whereas attitudes (PCA) is an index extracted with a Principal Component Analysis. Additional controls include SEND (special education needs and disability) dummy, broad definition; mother's level of education; father's level of education; migratory background; and class size.

Standard errors clustered at the class level in parenthesis. Sample (d). *** p<0.01, ** p<0.05, * p<0.1

Appendix B

Non-cognitive questionnaire

Name_____ Surname _____

1. Do you like math?
☐ not at all
☐ a little
☐ to some extent
☐ a lot

2. Are you good at math?
☐ not at all
☐ a little
☐ to some extent
☐ a lot

3. Are you worried to make a mistake when you do math?
☐ not at all
☐ a little
☐ to some extent
☐ a lot

4. Do you feel relaxed when doing math?
☐ not at all
☐ a little
☐ to some extent
☐ a lot

5. Are you worried not to finish the required tasks when you do math exercises in class?
☐ not at all
☐ a little
☐ to some extent
☐ a lot