

Tackling the gender gap in mathematics with active learning methodologies

Maria Laura Di Tommaso^{a,b,c,*}, Dalit Contini^a
Dalila De Rosa^{a,d}, Francesca Ferrara^e, Daniela Piazzalunga^{f,g,h,i}, Ornella Robutti^e

^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, Department of Finance, Italy

^e Department of Mathematics “Giuseppe Peano”, University of Torino, Italy

^f Department of Economics and Management, University of Trento, Italy

^g IRVAPP, Fondazione Bruno Kessler, Trento, Italy

^h CHILD – Collegio Carlo Alberto, Torino, Italy

ⁱ IZA, Bonn, Germany

19 February 2021

Abstract

We implement a teaching methodology that aims at improving children’s mathematical skills in primary school and we evaluate its impact on the gender gap in mathematics in Italy. The methodology focuses on peer interaction, sharing of ideas, students’ engagement, and problem-solving. The causal effect is evaluated using a randomized controlled trial, conducted in the province of Torino, involving 1,044 students. The treatment significantly improves math performance for girls (0.14 s.d.), with no impact on boys, contributing to reduce the gender gap by more than 40%. The results indicate that properly designed methodologies could help reduce the gender gap in mathematics.

JEL codes: I21, I24, J16, C93

Keywords: Gender gap in mathematics, School achievement, Primary school, Active learning, Teaching methodologies

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

We gratefully acknowledge the financial support from the University of Torino and the Fondazione Compagnia di San Paolo (Progetto di Ateneo 2016 “Tackling the gender gap in mathematics in Piedmont”, MATHGAP). Web site of the project: <https://sites.google.com/view/mathgendergap/>

The project, conducted by the Department of Economics and Statistics “Cognetti de Martiis” and the Department of Mathematics “Giuseppe Peano”, was realized in close cooperation with the Fondazione Agnelli, the Regional Board of Education in Piedmont (Ufficio Scolastico Regionale), and the Centro Servizi Didattici of Torino Città Metropolitana. In particular, we thank Andrea Gavosto and Martino Bernardi (Fondazione Agnelli), Giulia Ferrari (University of Torino), and Laura Tomatis (USR) for their very valuable contribution and support throughout the project. The realization of the project would not have been possible without the fundamental work of the tutors Isabella Boasso, Laura De Conti, Serena Gallipoli, Federica Lucco-Castello, the administrative support of Silvia D’Incau, the help of the external consultant Ketty Savioli. We have benefitted from discussion with Davide Azzolini, Simone Balestra, Nicola Bazoli, Giorgio Bolondi, Camilla Borgna, Ylenia Brilli, Pietro Di Martino, Chiara Giberti, Stefania Marcassa, Ignacio Monzon, Pauline Morault, Simone Moriconi, Chiara Pronzato, Enrico Rettore, Claudia Senik, Loris Vergolini, Rosetta Zan, and several seminar and conference participants. A particular thanks goes to the principals and the teachers involved in the project, and to pupils, who have actively participated in the program.

The trial has been registered at the AEA Registry: [AEARCTR-0003651](https://doi.org/10.1257/rct.3651-1.0) (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

Over the past decades, the traditional female disadvantage in education has disappeared and turned into an advantage in most subjects. But there is still one area in which girls are lagging behind boys, and this is mathematics. International learning assessments reveal that girls underachieve boys in mathematics in most countries (OECD 2014).

A wide range of factors have been proposed that can explain the existence of a gender gap in mathematics, from gender inequality to parental and teacher attitudes and stereotypes. A largely unexplored factor is the way mathematics is taught to children. This paper is an experimental evaluation of a program implementing teaching practices based on active and cooperative learning, aimed at improving children's mathematical skills in primary school. We evaluate if this program is effective in reducing the gender gap in math in a large Italian province.

According to the latest PISA survey to have a specific focus on mathematics (PISA-2012), the average OECD gender difference in math competency at age 15 amounted to 0.11 standard deviations in favor of boys, with considerable country variation, from -0.07 in Iceland to 0.24 in Austria. In Italy, the gap was 0.18 standard deviations.¹ As shown in several studies, the gender gap in mathematics (GGM) already exists at an early age and increases as children grow older (Fryer and Levitt 2010, Ellison and Swanson 2010, Contini et al. 2017, Meinck and Brese 2019). Italy is of particular interest because it has a very large gender gap in mathematics. Italy has the highest gap among the 57 countries participating in TIMSS 4th grade (Mullis et al. 2016), and the largest gap among OECD countries in the PISA test administered to 15-year-old students for the year 2018 (OECD 2019).

The presence of a substantial female disadvantage in math is of particular importance because it is likely to be a cause of the low share of women choosing STEM (Science Technology Engineering and Mathematics) disciplines at university (Turner and Bowen 1999, Card and Payne 2021). Furthermore, there is compelling evidence that this gender imbalance in choosing academic disciplines critically affects gender occupational choices and differences in wages (Paglin and Rufolo 1990, Machin and Puhani 2003, Black et al. 2008, Piazzalunga 2018). Women are still highly underrepresented in the most productive sectors of the economy and in high-paying occupations, often in STEM fields, with long-run

¹ Instead, results from the TIMSS-2015 study point to no GGM on average. This result is not surprising, because TIMSS focuses on curricular knowledge whereas PISA assesses competencies (the ability to use math in daily life) and it is widely recognized that girls are more engaged in school duties than boys are. Nonetheless, some countries including Italy still display a large average GGM in favour of boys in both 4th and 8th grade.

effects on gender differences in wages and wealth (Sierminska et al. 2019). Moreover, recent research underlines the importance of mathematical skills also in non-STEM degrees and occupations (Grinis 2019, Delaney and Devereux 2020).

Our paper examines the role of teaching methodologies in influencing gender differences in math achievement. Teaching methodologies and the school environment are part of the cultural and societal factors determining academic achievement. The high variability of the size of the gender gap in math across countries suggests that such cultural and societal factors play the major role. It has been shown that the female's disadvantage in mathematics narrows in countries with higher gender equality (Guiso et al. 2008, Pope and Syndor 2010, Nollenberger et al. 2016, Lippman and Senik 2018, Gevrek et al. 2020), and that parental attitudes towards gender equality are positively correlated with the proficiency in mathematics of girls (Dossi et al. 2019). Gender stereotypes also affect teachers' beliefs. Stereotypes often lead parents and teachers to attribute girls' achievements to diligence instead of talent (Ertl et al. 2017). Furthermore, teachers' implicit gender bias has a sizable influence on the gender gap in math (Carlana 2019). Role models are also important. For example, the assignment to a same-gender teacher improves the achievement of both girls and boys as well perceptions of student performance and student engagement (Dee 2007). These mechanisms may also be responsible for the girls' lower self-confidence, lower self-efficacy, and higher level of anxiety in doing math (Ho et al. 2000, Gneezy et al. 2003, Niederle and Vesterlund 2010, OECD 2015, Di Tommaso et al. 2018). Some scholars have considered biological differences in brain functioning to be important (e.g. Baron-Cohen 2003), but 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). Also, in some countries there is no difference between genders in mathematics achievement, indicating that gender differences in brain functioning do not play a role.

A critical divide in educational research is between teacher- and student-centered instruction. The first conceives teaching as a top-down activity and focuses on direct transmission. In this view, the teachers' role is to "communicate knowledge in a clear and structured way, to explain correct solutions, to give students clear and resolvable problems, and to ensure calm and concentration in the classroom" (pg. 92, OECD 2009). The second, based on the constructivist approach, views students as active participants in the process of learning. More value is attached to the development of thinking and reasoning processes rather than the acquisition of specific knowledge (Staub and Stern 2002). Students should become capable of developing solutions to problems on their own (Gutierrez and Boero

2006). Qualitative research suggests that when mathematics' teaching is centered upon problem-solving, involving students in discussions and investigative work, the gender gap narrows and can even disappear (Boaler and Greeno 2000, Boaler 2002a, Boaler 2002b, Zohar and Sela 2003, Boaler 2009, OECD 2016). Empowering children with a “growth mindset” seems to be particularly beneficial to girls (Boaler 2013).² Yet, we are not aware of any quantitative analysis of the effectiveness of this approach on the gender gap in math.

Our research contributes to the literature by implementing a mathematics teaching practice based on active and cooperative learning and assessing its impact with a randomized controlled trial (RCT). To the best of our knowledge, this is the first attempt to evaluate the causal impact of a teaching methodology on the gender gap in mathematics.

We implemented a set of activities based on the “Mathematics Laboratory”, an approach to teaching mathematics developed in Italy in the early 2000s (Anichini et al. 2004). In this approach, teaching practices are based on the active involvement of children, engaged in individual and peer work in a collaborative and non-competitive environment. Children are encouraged to frame problems and attempt to solve them by sharing and comparing ideas within small groups and in-class discussions. Mistakes are welcome, being a crucial means to understanding. The central idea is that learning involves active participation on the part of the learner (Lave and Wenger 1991). In the following, we denote this type of intervention as the “Math Active Learning” (MATL) program. It is worth noticing that according to the OECD teaching and learning international survey, TALIS-2008, teachers in Italy show the strongest preference for a teacher-centered approach over a student-centered approach.

The MATL program consisted of 15 hours of laboratory activities delivered to third-grade pupils over five consecutive weeks during spring 2019. We chose to focus on the third grade because at the end of the second grade pupils are assessed for the first time at the national level. The gender gap in math appears already then and increases throughout primary and secondary school (Contini et al. 2017). We wish to intervene as soon as the gender gap appears and before it becomes too big. All schools in the Torino province were invited to participate with two classes. Among the schools wishing to participate, we randomly selected 25 and randomly assigned one class to the treated group and one class to the control group. The final sample consisted of 1,044 children, with 519 children in the treatment group

² According to the growth mindset approach, as opposed to a fixed mindset, ability is malleable and intelligence can be learned. It gives importance to mistakes: 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).

and 525 children in the control group. The treatment was delivered at the class level during regular math hours by tutors specially trained in the new teaching methodology. Thus, the intervention did not provide additional math instruction, but it substituted teachers' lessons with the MATL activities. During these activities, math teachers remained in the classroom with the role of observers. Children in the control classes followed the usual curriculum with their own teachers. To assess the impact of MATL on children's skills, we administered math tests one month before the intervention (pre-test) and one month after the intervention (post-test). The tests were developed under the external supervision of scholars involved in the design of the national assessment test (INVALSI) and marked blindly. This ensured that these tests had a conceptual framework and structure in line with the national assessment tests.

The findings from the impact evaluation of the MATL program are encouraging regarding the gender gap in math. The MATL program increased girls' math achievement by 0.14 standard deviations, without hampering boys' performance. Given that the MATL activities were limited in time, this effect should be considered quite large in magnitude and to be policy-relevant.³ Overall, the intervention contributed to a reduction in the gender gap in math by over 40%.

In the paper, we evaluate how the impact of the MATL program varies with prior ability, as measured by the pre-test. We find that the treatment has no effect on boys, irrespective of their starting level, but that girls with above-average pre-test scores benefit the most from the treatment. We also find heterogeneous effects by migratory background and parental education. Given prior ability, the treatment has a larger impact on migrant girls and girls with low educated parents compared to girls from more advantaged family backgrounds.

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

³ As a comparison, based on a review of previous research, Bloom (2008) reports that one full year of attendance improves primary school pupils' achievement on average by 0.25 standard deviations, for both math and reading comprehension, and that decreasing class size by 10 children from 22-26 students improves performance by 0.10-0.20 standard deviations.

2. Institutional context and design of the program

2.1. 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 of public institutions, less than 7% of the children attend private primary school. Families can choose between two time schedules: a 40-hour school week, where children spend the whole day at school, and a more concentrated 27/30-hour week.⁴ 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 teaching methods. In each class, two-three teachers are covering the entire set of disciplines (sometimes except for foreign languages, gymnastics, and music). Didactic continuity is highly valued in the Italian school system. Children are grouped into classes that remain the same throughout primary schools and are normally taught by the same teachers for the entire 5-year cycle. Primary school teachers receive training enabling them to teach all subjects,⁵ although they often specialize in specific disciplines. They do not change the subjects they teach a group of children within a cycle. The school year starts in the first half of September and finishes in the mid of June.

In primary school, math instruction covers the domains of numbers, relations, data and predictions, space, and figures. National curricular guidelines recommend providing instruction on the different domains throughout the entire school year. In grade 3, when the MATL intervention was delivered, math instruction is usually provided for 6 to 8 hours weekly

2.2. The MATL intervention

Features of MATL program

The intervention consists in classroom-based activities aimed at improving children mathematical understanding. The teaching practices adopted are based on the theoretical framework of social constructivism, according to which: i) learning is inherently a social process because it is embedded within a social context as students and teachers work together to build knowledge; ii) knowledge cannot be directly imparted to students, so the goal of teaching is to provide experiences that facilitate the construction of knowledge. Rather than

⁴ The share of schools delivering the 40-hours schedule is much higher in the Northern regions.

⁵ 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*).

just passively take in information, as children make experiences and reflect upon them, they build their own representations and incorporate new information into their pre-existing knowledge (Thompson 2014). Another conceptual pillar of the approach is the “growth mindset” paradigm, according to which ability is malleable, intelligence can be learned, and the brain can grow from exercise (Dweck 2006a, Boaler 2013), as 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).

More specifically, the MATL intervention builds on the “*Laboratorio Matematico*”, a math education methodology developed in Italy in the early 2000s and widely acknowledged in the international mathematics education community (Anichini et al. 2004, Arzarello and Robutti 2008, 2010, Arzarello, Ferrara and Robutti, 2012, Ferrara and Ferrari 2020).

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

- (i) *Doing instead of Listening*. Focusing on problem framing and problem-solving as opposed to procedural work, the approach reverses the traditional teacher-centered instruction by putting children at the center of the learning process.
- (ii) *Cooperative learning*. Students are engaged with individual and peer-group work, and are encouraged to enter into dialogue with the teacher, both individually and collectively.
- (iii) *No pressure*. There is no demand for immediate answers or solutions at the individual level. Students are given suitable time to analyze the problem, explore different solutions, share and compare ideas, avoiding pressure and competition.
- (iv) *Learning from mistakes*. Mistakes are conceived as crucial means to understanding. By giving positive attention to their own and others’ mistakes, children explore their learning processes and develop a deeper understanding of the discipline.
- (v) *Manipulative activities*. Children get engaged with materials (caps, straws, buttons of different size, boxes, cards...) that they manipulate with their hands and move physically around, as perceptual-motor learning has been proven to be effective in mathematics understanding (Antinucci 2001, Nemirovsky et al. 2004).

All these elements aim at activating children’s thinking, helping them construct mathematical meanings, through self-reflection and interaction with the teacher and peers. The different activities occur within a collaborative and non-competitive environment, where the teacher – the tutor, in our case – has the role of “orchestrating” class activities.

MATL focuses on the subject area of Numbers, recognized as the most fundamental domain in the math field at this age and because we found that the GGM is highest in this domain.^{6,7} In our experiment, the MATL program was implemented using two activities. In the first, named *Thousandville*, children must enlarge a city without changing the proportions of the different components. The learning processes involved are counting, performing operations, estimating the order of magnitude, dealing with large numbers. The second activity, named *Forest Elves*, concerns a family of elves who must go to different places, at different speeds, and arriving at different times. The issues at stake are “who will arrive first in a given place?” and “when/where will they meet?”. The learning processes involved are measuring quantities, comparing quantities, discover relations between quantities in terms of multiples and submultiples.⁸

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

Laboratory teaching practices are devised to help to develop a growth mindset. As shown by Dweck (2006a, 2006b) fixed mindset messages prevail among students across the entire achievement distribution, but high-achieving girls are especially damaged by fixed ability beliefs. Girls suffer most by the fixed ability conception that implies giving labels, like being or not being smart, or being good or not being good at math (Dweck 2006b).

The teaching practices embodied in the MATL intervention have the potential to reduce the gender gap in math for different reasons. Firstly, the activities are meant to reduce pressure and competition. This should benefit girls, because girls are generally less competitive than boys, in competitive environments they tend to develop more anxiety, and anxiety is detrimental to learning (Bohnet 2016). Second, 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, girls could be more prone to learn from mistakes by better developing constructive reasoning on their own cognitive processes because they are more thoughtful (Boaler 2016). MATL could also improve girls’ test scores more than boys’ test scores because it was specifically devised to embody some mathematical activities into a narrative context and on average girls tend to be better than boys in reading comprehension and languages. A final element that might contribute to girls’

⁶ The other areas in the primary school curriculum are: relations, space and figures, data and predictions.

⁷ For a description of this point see the final report of the MATHGAP project (Di Tommaso et al. 2020).

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

activation and empowerment is the explicit support in the MATL guidelines for balanced participation in class discussions.

Delivery of the MATL intervention

The MATL program is delivered to children in grade 3, when they are about 8 years old. The reason behind this choice is to balance two different needs: (i) to tackle inequalities as early as possible and to contrast possible cumulative effects; (ii) to run the intervention at a point in time when the GGM already exists, to observe gender differences before the intervention and analyze their (short-term) development.⁹

MATL was delivered between February and April 2019. The intervention took place at the class level during school-time and during math hours, not to alter the total amount of time devoted to math instruction. It was organized over sessions of three hours, once per week for five consecutive weeks. Children were divided into small groups, heterogeneous with respect to gender and prior ability. All the pupils in the treated classes took part in the activities, including children with disabilities, special education needs, or learning difficulties. In the meantime, children in the control group followed the usual curriculum with their class teacher. The intervention was conducted by four tutors with a background in mathematics education at the Master or Ph.D. level. Class math teachers were present as observers.

A pilot study aimed at evaluating the intervention format was conducted a few months before the beginning of the RCT, in two schools not taking part in the experiment. The treatment was then revised to consider the comments and suggestions of the tutors and the class teachers. This pilot also gave 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.¹⁰

3. Design, Data, and Estimation

3.1. Research Design

We evaluate the effectiveness of the intervention by exploiting a randomized controlled

⁹ 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 first time children are assessed with a national test (INVALSI).

¹⁰ A full description of the pilot study and of the IRT analysis are available in the final report of the project (Di Tommaso et al. 2020).

trial research design. The intervention was planned to be delivered in public primary schools located in the province of Torino (Piedmont), in the north-west of Italy. There are 180 public primary schools in the province of Torino. We planned to enroll 25 schools and 50 classes, for a total of approximately 1000-1200 pupils.

The timeline of the implementation of the RCT is synthesized in Figure 1.

Fig.1 Timeline of the intervention

Enrollment in the project was on a voluntary basis. All principals of public primary schools in the province of Torino were informed about the project in March 2018 with an official letter signed by the Regional Board of Education¹¹ and were invited to a project presentation. The eligibility conditions were set as follows: (i) Schools had to enroll with at least two classes, one to be randomized to the treatment group and the other one to the control group. The reason was to control for potential self-selection issues: parents have substantial leeway in choosing the school for their children but cannot choose the specific class or teachers. Although random variability would ensure a fair allocation into the treated and control groups, due to the limited size of the sample of schools, some unbalance could occur. Including two classes per school eliminates school-specific effects related to school management, the socioeconomic composition of the student body, and school-level peer effects. In a broad sense, this procedure can be viewed as a matching method, set up to increase comparability of the treated and control group and improve the precision of the estimates. (ii) Classes in the same school had to have different mathematics teachers, to limit the risk of spillovers. (iii) Participating classes were not to be involved in other extra-curricular math projects in the same school year.

Thirty-one schools applied to the program. We excluded one school because it was already participating in another math-learning project and randomly selected 25 schools among the remaining ones. Since some schools applied with more than two classes, we also randomly selected the two participating classes (see Table A.1). In a second step, within each participating school we randomly assigned one class to the treatment group and the other to the control group.¹² The entire randomization process was public and took place at

¹¹ The Regional Board of Education is the highest authority for the organization of schools at the regional level.

¹² The sampling procedure was set before knowing how many schools and classes would have applied to the project, and different rules were defined in order to deal with different number of applications. Details can be found in the pre-analysis plan registered in the AEA RCT Registry (Contini et al. 2018).

the University of Torino in June 2018.

All children in the treatment and control classes attended the pre-test one month before the beginning of the MATL program (January 2019). The math laboratories were held between February and April 2019. The children attended the post-test approximately one month after the end of the intervention, between April and May 2019.

The trial was registered at the AEA Registry on December 10, 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.

3.2. Outcome measures and additional data

Outcome measures

The tests assessing children's math competencies before and after the treatment, designed by scholars of mathematics education, followed the same conceptual framework of the INVALSI national assessment for the numbers' domain.¹³ We could not use a pre-existing test because the INVALSI primary school assessments involve children in grades 2 and 5, and not children in grade 3. Each test consists of 20 items, to be completed in 40 minutes.¹⁴ The tests cover different topics, different mathematical dimensions (knowing, arguing, and problem-solving), and use both multiple choice-type answers and open answers.¹⁵

The tutors in charge of the laboratories administered the pre- and post-test inside the classrooms and later graded them blindly under the supervision of an external examiner.¹⁶ Correct answers are assigned 1 point each, incorrect and missing answers 0 points, thus raw scores range between 0 and 20 points. The individual raw score is then standardized as to have zero mean and standard deviation equal to 1.

The post-test is the main outcome variable 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, and it is included as a control variable to improve the precision of the estimates. Figure 2 shows the pre-test score distributions among girls and boys. On average, boys answered correctly to 11.23 items out of 20 and girls to 10.28; the difference is statistically significant and corresponds to 0.216 standard deviations (0.237 in

¹³ For an overview of the INVALSI test see:

https://invalsi-areaprove.cineca.it/docs/2018/INVALSI_tests_according_to_INVALSI.pdf

¹⁴ Both the results of the pre- and post- test have been analysed with an IRT model, available in the final report of the project (Di Tommaso et al. 2020).

¹⁵ The English translation of the tests is available in Appendix C (C.1 and C.2).

¹⁶ An expert in formulating and grading INVALSI tests.

the sample of children present both at the pre- and post-test). There is a gender gap in math across the entire distribution, confirming the findings from previous research (Contini et al. 2017). The gender gap measured by our test in grade 3 is close to the gap measured by INVALSI assessments in grade 2 in our experimental classes (0.241), but larger than the gap observed in the INVALSI tests in Piedmont (0.130) and Italy as a whole (0.099).¹⁷

Fig.2 Gender gap in the pre-test

We also collected information on children's attitudes towards math, as a second outcome variable, to explore possible mechanisms underlying the effect of the treatment on cognitive abilities. Attitudes are evaluated by means of a short questionnaire with five Likert-type questions, delivered immediately after the post-test. Details are provided in Section 5.2.

Additional data

A definition of all the variables used in the paper is available in the Appendix (Table A.2).

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, attention disorders (ADHD).¹⁸ The schools' administrative office gave us information on parental education and migratory background. The tutors recorded absenteeism during the math labs for the children in the treated classes.

Data on the math teachers were collected via a brief questionnaire recording gender, age, degree, experience overall and in the class, tenure, and type of contract. The tutors collected pieces of information at the class level, such as class size and the time schedule (full time i.e. 40 hours per week, or normal time i.e. 27-30 hours per week).

INVALSI provided class-level data on math and language scores as well as socio-economic background at the national assessment in grade 2. These data were used for evaluating external validity, comparing average ability and social composition in the experimental classes with the corresponding statistics at the regional and national levels.

¹⁷ See also Section 6.2 on external validity.

¹⁸ These data as all the other data collected in the project were treated with extreme confidentiality. They were collected following the code of ethics of the University of Torino and the Italian and European legislation for privacy.

3.3. Sample

Table 1 shows the sample selection and Table A.3 in the Appendix provides additional details.

No school or class dropped out of the project, so 25 primary schools participate in the project with two third-grade classes each, for a total of 50 classes, and 1,044 children. Of the 1,044 children in the full sample (sample a), 933 pupils were present at the pre-test (sample b), 983 were present at the post-test (sample c), and 888 were present at both (sample d).¹⁹ The sample used for the impact evaluation is sample d.

Tab.1 Sample selection

3.4. Balance, Attrition, and Compliance

Balance at baseline

Table 2 shows the balance between the treated and control groups at the baseline, i.e., before treatment, and descriptive statistics of the outcome variable (post-test). Panel A reports the mean values of the variables at the individual level, including pre-test scores of girls and boys, shares of girls and boys, native and migrant children, SEND and non-SEND children, parental education. Statistically significant differences exist in the maternal education variables. In the treatment group, we find a higher share of mothers with upper secondary education than in the control group, but the opposite occurs for tertiary education; considering the share with at least upper secondary education, the two groups appear perfectly balanced. The differences are slightly in favor of the control group, where mothers are more likely to have a tertiary degree. Panel B reports mean class size, time schedule, mean class composition, and teachers' characteristics. All variables are balanced. The only exception is the number of years the math teacher has been teaching in the class,²⁰ in favor of the control group (2.79 years in control classes, 2.40 in treatment classes). It is worth mentioning that the number of statistically significant differences is similar to the figure expected due to random variability (i.e., close to 3, the expected number of times we would reject a correct null hypothesis using a level of significance of 0.10 in 30 independent tests).

The treated and control groups are well balanced on most characteristics, both at the

¹⁹ 4 children are excluded from the analysis because they were present at the post-test, but did not answer to any item (probably very serious disability).

²⁰ As explained in the institutional context section, the Italian system values continuity and the teacher stays in the same class from grade 1 till grade 5 of primary schools.

overall level and by gender, indicating that the randomization was successful. In addition, we find that the two groups are very similar in terms of math performance not only at the mean but also across the entire distribution, as shown in Figure 3.

Tab.2 Baseline characteristics of treated and control children, full sample
Fig.3 Pre-test score distribution by treatment status

Attrition

Randomization normally yields similar groups at baseline, but the two groups also need to be equivalent at follow-ups. Attrition occurs when the outcome is not measured for all individuals in the original sample. The overall attrition rate is the share of units that are lost in the entire sample; the differential attrition rate is the difference in the attrition rates between the treated and control groups. Both overall and differential attrition create a potential for bias when the characteristics of sample members in one group differ systematically from those in the other (WWC-What Works Clearinghouse 2013).

In this study, there are two relevant levels of attrition: absences at the post-test and absences at either the pre- or the post-test (the latter matters because we control for pre-test scores in the preferred specification). We measure overall and differential attrition for all children, and separately for boys and girls. Attrition rates are reported in Table 3. The upper panel reports the attrition rates in the post-test relative to the full sample (1,044 children). Overall, 5.4% were absent at the post-test, with small differences between treated and control children and between girls and boys. The lower panel of Table 3 reports the share of children absent at either the pre- or the post-test (14.9%). More absences occurred at the pre-test, presumably because the test was administered in winter 2019, during the flu peak. This attrition rate is significantly higher among treated than among control children (16.7% vs. 12.4%), with a larger gap among girls than among boys. The overall and the differential attrition rates are small enough not to raise concern over the validity of the estimates of the intervention effect.²¹

Tab.3 Attrition pre-test and post-test

²¹ See the guidelines in WWC-What Works Clearinghouse (2013) that are based on an extensive simulation study.

We rerun balance checks for the sample of children who attended the post-test but not the pre-test (sample b - Tab A.4) and for the sample of children who were present at both tests (sample d - Tab A.5). Treatment and control groups appear well balanced also after attrition, and no substantial difference is found between the original and the analytical samples.

The comparison between the treated and control group in sample (d) has been further analyzed in a multivariate regression, by estimating a logit with the treatment status as the dependent variable and individual, teacher, and class characteristics as independent variables. Results are presented in Table A.6 and confirm the groups' comparability.

In the main empirical analyses, our preferred specification includes individual and class characteristics at the baseline as control variables, to account for the minor observed differences between treated and controls (despite the favorable results of the attrition analysis).

Compliance and spillover effects

Imperfect compliance occurs when some members of the comparison group manage to participate in the program or when some members of the treatment group do not receive the treatment. In this experiment, the first situation was not feasible and never occurred. Instead, children assigned to the treated classes were left untreated if they were absent in the days of treatment delivery. Noncompliance dilutes the treatment causing it to understate the average treatment effect (Bloom 2008).

In Table 4 we report statistics on MATL participation. No children missed all the lab sessions, 99.3% attended at least half of the hours and 73.8% attended all sessions, with a small difference in favor of boys (4 p.p. in the full participation). This may reduce the estimated impact on the GGM, yielding to conservative estimates of the actual treatment effect. Given that full participation in the program was not reached, the impact evaluation estimates represent estimates of the Intention-to-Treat (ITT) effect.

Spillover effects are also not a matter of concern. First, it is highly unlikely that interactions between eight years old children in different classes would involve mathematics. Second, it is also unlikely that teachers in the control group learned sufficient details about MATL to alter their teaching practices in a few-month time. Math teachers were different in the treated and control classes and the intervention was delivered by external tutors while treatment class teachers were present as observers. Moreover, the methodological materials were released to teachers only 1 year after the end of the project. If spillover occurred somehow, the treatment effect would be underestimated.

We cannot exclude that teachers of the treatment classes learned from observing the intervention. This is not problematic because we aim to assess the total effect of the program, consisting of the direct effect of MATL on children’s math achievement and the (potential) synergic indirect effect generated by the action of class teachers. Both channels are intended effects of the intervention.

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’ maths competences, and more specifically on boys’ and girls’ outcomes. The successful randomization into the treated and control groups ensures that the two groups can be safely compared, without incurring in selection bias. Nevertheless, to control for possible differences between the two groups generated by random variability, we do not simply compare the post-test scores of treated and control children but analyze these differences within a regression framework where we control for individual characteristics and pre-test scores. We estimate the effect of MATL using the following OLS specification, overall and separately for boys and girls:²²

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

where Y_{1iks} is the post-test score of individual i in class k of school s . T_{ks} is the binary treatment indicator, equal to 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 observable individual and class characteristics potentially predictive of the outcome (gender, special education needs or disability, migratory background, parental education, class size, and time schedule). θ_s is a vector of school fixed effects (our randomization strata), and ϵ_{iks} are random errors normally distributed and clustered at the class level k . β is the coefficient of interest, capturing the intention-to-treat (ITT) effect of being offered the MATL program. β cannot be interpreted as the average treatment effect (ATE), because some pupils did not attend all the lab sessions. However, since most of the

²² See the pre-analysis plan (Contini et al. 2018). Our empirical analysis is as close as possible to the pre-analysis plan. The analyses and outcomes investigated were pre-specified, unless otherwise indicated.

students did, we can expect ATE to be similar to the ITT in this case. We assess whether the treatment has a different impact on the two genders estimating equation (1) separately for boys and girls.

We then include an interaction effect between the pre-test score and the treatment dummy, for estimating heterogeneous effects by prior ability.

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

The coefficient λ captures the differential impact of the treatment according to the level of the pre-test.

We cannot simply compare gender gaps in the pre- and in the post-test scores to evaluate the effect of the treatment on the GGM because the two tests are not equated. Although they were designed within the same conceptual framework, they do not have the same level of difficulty and are not measured on the same scale.²³ A better strategy consists in comparing the raw GGM in treated and control groups after treatment. Due to the successful randomization, we consider the post-test in the control group as a valid estimate of what would have happened to the children in the treated classes had they not been exposed to MATL (and vice versa). To account for the small differences in the pre-test, we estimate the counterfactual as the outcome of control group children had they been treated, using the coefficients estimates from (2) and setting value 1 to the treatment indicator. Similarly, we obtain a counterfactual outcome for treated children. Since there are two possible comparisons, we will obtain two distinct estimates of the magnitude of the change of the GGM due to treatment.

Explanatory variables

In addition to pre-test scores, we control for gender, special education needs or disability (dummy) (SEND), migratory background, parental education, class size, and time schedule, as well as school dummies, to account for school fixed effects. We also estimate simpler specifications where not all the control variables are included in the estimation.

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

²³ See Di Tommaso et al. 2020.

Family background variables included in models (1) and (2) above are defined in Table A.2. Parental education is denoted “high education” if at least one parent has a tertiary degree, and 0 otherwise. The child’s migratory background is coded as 3 dummies variables: native if the child and at least one parent were born in Italy, first-generation migrant if the child and both parents were born abroad, and second-generation migrant if the child was born in Italy and both parents were born abroad. In order not to lose many observations and avoid self-selection issues, we include a dummy variable for each characteristic that is equal 1 if the characteristics is missing.²⁴

We use pre- and post-test scores in standardized version, thus the effect of the treatment reported in the results represents by how many standard deviations the test scores of the treated pupils differ on average from those of the control group.

Robustness checks

The main analytical sample includes only children who sat both the pre- and the post-test. In a robustness check, we also include the children who were absent at the pre-test, identifying them with a dummy variable and assigning a zero value for the pre-test score. As for children absent to the post-test, we had scheduled a deferred session on a different date, as close as possible to the original one, and we use such data in a second robustness check.²⁵

In additional robustness checks, we exclude children with special education needs or disabilities. 15% of the pupils were reported by the teachers to have learning problems, with a slightly higher share among boys.²⁶ 8.1% are certified as children with special needs or disabilities. Often children with mild problems have not yet obtained a certification by grade 3. The tests were designed for typically developing children, in line with the national assessments administered periodically at the national level by INVALSI. They may be not appropriate for children with severe learning problems. For this reason, in the pre-analysis plan we stated that we would exclude SEND children’s results from the analysis. Due to the

²⁴ We have been able to collect information on teachers’ characteristics in 49 out of 50 classes (one teacher refused the consent to data processing). To avoid losing an entire (control) class, we do not include teachers’ characteristics in the estimations at class level. Teachers’ characteristics are used in the balance tests.

²⁵ In the regular session, the tutors administered the post-test within the classroom. In the deferred session, the post-test was administered by the class teacher while the other children were involved in normal class activities. These tests were then sent by mail to the research team. Of the 57 children absent at the post-test, 35 children sat the deferred session. Due to the impossibility to have full control of this process, we preferred not to include these children in the main analyses.

²⁶ 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 can be partly ascribed to an existing gender bias against boys in the referrals for special education (Anderson 1997, Wehmeyer and Schwartz 2001). This finding supports the choice of including also SEND children in the analysis.

difficulty to identify children with severe problems that we were not aware of before going on the field, we decided to deviate from the previous plan. We include all SEND children in the main specification, leaving the estimation without them as robustness checks.

4. 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. In section 4.1, we estimate the average impact on the entire children population, on girls and on boys. In section 4.2 we analyze whether the treatment has heterogeneous effects according to prior achievement, parental education, and migratory background. In section 4.3, we describe the results of robustness checks.

4.1. Core results

Table 5 presents the main results. Considering all the children who sat the post-test (columns 1-3), we find that the intervention has significantly improved math performance (effect size 0.116 s.d.). The analyses by gender reveal that girls drive this effect (effect size 0.154 s.d.). Instead, the treatment did not influence boys' achievement. We then focus on our preferred sample, including the children who took both the pre- and the post-test. Columns 4-6 refer to the baseline specification with only treatment as an explanatory variable; columns 7-9 refer to the model that also includes pre-test scores; columns 10-12 to the model with school fixed effects and additional control variables. The ITT estimates in the baseline specification (columns 4-6) are always positive and not significantly different from zero (for all children, and separately for girls and boys). When accounting for pre-test scores, the impact of the treatment becomes larger and statistically significant for girls, remaining zero for boys. Our preferred estimates eventually include also school fixed effects, individual and family background characteristics, class size, and time schedule. The average effect of the treatment is positive and significant, at 0.083 s.d., driven by a large and positive effect for girls.^{27,28}

Overall, results on the treatment effect are quite stable across specifications: MATL increases girls' test scores by 0.142 standard deviations and has no effect on boys'

²⁷ Full results are presented in Table A.7 in the Appendix.

²⁸ In Appendix B, we present the main and the heterogeneous results using the latent ability estimated with IRT models as a dependent variable, instead of the standardized test-score. Results are confirmed and similar in magnitude.

performance. For educational interventions, this effect is quite large in magnitude. As a term of comparisons, Bloom (2008) reports that in primary school one full year of attendance improves pupils' achievement by 0.25 standard deviations on average for 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' practices lasting at least 12 weeks have a median effect size of 0.33 and Pellegrini et al. (2018) find a median effect size of 0.25 for similar programs.

A core question is how this impact translates into a raw reduction of the GGM. In the control group, the gender gap in math is 0.324, while in the treated group is 0.221, implying a reduction of 31.7% in the treated group as compared to the control group.

To account for differences in the pre-test, we compute the GGM reduction as follows. Firstly, we estimate counterfactual outcomes (of the control group children had they been treated, and of the treatment group had they not been treated) using the coefficients estimates from (2) and applying value 0 to the treatment indicator of the treated group children and value 1 to the treatment indicator of the control group children. Secondly, we compare each counterfactual GGM with the corresponding observed value. The actual GGM for the control group is 0.324, and the counterfactual GGM for this group had they been treated is 0.170, implying a reduction of 47.5%. The actual GGM for the treated group is 0.221, and the counterfactual GGM for this group had they not been treated is 0.369, implying a reduction of 40.1%.

Tab.5 Main results: effect of the treatment

4.2. Heterogeneity in treatment effects

Table 6 summarizes the heterogeneous effects, according to pre-test scores in a standardized form. Results confirm that MATL intervention has no effect for boys, irrespective of their ability level, and a positive effect for girls, increasing with pre-test scores. An increase in 1 standard deviation in girls' pre-test scores increases the effect of the treatment by 0.127 standard deviations in girls' post-test scores. Figure 4 shows the treatment effect by pre-test scores and corresponding confidence intervals. The effect of the treatment is significantly larger than zero for girls with standardized pre-test scores higher than -0.3 (the girls' average pre-test score is -0.09) and becomes as large as 0.4 for very well-

performing girls at the pre-test.²⁹

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

Fig.4 Treatment effects by prior achievement's levels

We also explore heterogeneity in the treatment effect by parental education and migratory background, by estimating the following equation twice.

$$Y_{1iks} = \alpha + \beta T_{ks} + \gamma Y_{0iks} + \delta X_{iks} + \lambda T_{ks} * Z_{iks} + \theta_s + \epsilon_{iks} \quad (3)$$

First, we include an interaction term between the treatment dummy and the vector of dummies for parental education Z_{iks} .³⁰ Second, we include an interaction term between the treatment dummy and the dummies for the migratory background M_{iks} instead of Z_{iks} .

Table 7 presents the estimated treatment effects on each of the subgroups, controlling for pre-test scores and the other additional explanatory variables.³¹ Girls with low-educated parents benefit most from the treatment, their test scores increase by 0.182 s.d. Native girls benefit from attending MATL (0.104 s.d.), but the average effect, given pre-test, is much larger for migrant girls (0.399 s.d.). The intervention has no effect on native boys, but it seems to decrease male migrants' performance (-0.285 s.d.).

We also estimate equation (2) separately for native and migrant girls. We find that the effect of treatment for migrant girls steeply increases with prior performance. Fig. 5 reports the treatment effect by prior performance for native and migrant girls. Among migrant girls, even mid-low performers benefit from the intervention, whereas only higher than average native girls do so.

MATL labs improve math skills for migrant girls and girls with low educated parents, but they seem to worsen math skills for migrant boys. This finding is not fully consistent with previous research. Two best-evidence syntheses by Slavin and coauthors (Slavin and Lake 2008, Pellegrini et al. 2018) indicate that students coming from different backgrounds benefit in a similar way and that low achievers benefit most by attendance to long active

²⁹ As a robustness check, we replicated the analysis by interacting the treatment variable with pre-test quintiles instead of a continuous variable, allowing the treatment to be non-linearly related to pre-test score. The results are consistent with the described findings and indicate that the effect is approximatively linear.

³⁰ Low education when both parents do not have tertiary education qualifications. High education if at least one parent has a tertiary degree.

³¹ Full estimates are available from the authors upon request.

learning math programs. We may speculate that MATL is a short-term program and that the skills of boys from disadvantaged backgrounds might improve if the intervention was implemented over a longer period. Further investigation is needed to shed light on this point.

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

Fig. 5 Treatment effect by prior achievement's levels, migrant and native girls

4.3. Robustness checks

We replicate the main analyses on different samples. The results are reported in Table 8. First, we exclude from the analysis the children with a certified special education need or disability (SEND, narrow definition). Second, we exclude children reporting special educational needs and disabilities even if not formally certified (SEND, broad definition). Third, we use the entire sample of children present at the post-test and we include a dummy variable for children absent at the pre-test. Fourth, we include the children who were absent at the post-test but were given a post-test on a deferred date.³² In all models, we include pre-test scores, school fixed effects, and the usual additional controls.

The robustness checks largely confirm the results. The treatment has an impact on girls (effect size 0.12-0.17) but not on boys. The impact of the treatment is larger if we exclude children with any type of special educational needs and if we include all children. It is the smallest if we include children who took the test in the deferred session. Absences at the pre-test do not affect the performance at the post-test, confirming our hypothesis that absences occurred randomly and that the peak observed in the pre-test was probably due to the flu season.

Tab.8 Robustness checks

5. Mechanisms

The MATL intervention has proven to be effective on girls. We now explore the role played by potential channels through which the program might have improved girls' math skills. The program could improve abilities by increasing problem-solving competences,

³² In the pre-analysis plan (PAP), we had decided to: exclude SEND children; include post-test taken in the deferred session; include children absent at the pre-test by labeling them with a missing dummy. Afterwards we decided to operate differently in the core analysis, but the choices specified in the PAP are presented here as Robustness checks.

engagement and fun, reducing competitiveness, motivating discussion, and valuing the role of mistakes. MATL might act directly on children's competencies or/and indirectly via an effect on self-confidence and more generally on attitudes towards math.

Firstly, we investigate whether the intervention improves mathematical skills overall or only in some dimensions. The question is whether MATL works by enhancing the competencies on some dimensions but not others, or by improving children's skills in facing specific item formats. Secondly, we assess the role of attitudes towards math. We measure attitudes directly via a short questionnaire administered to children after the post-test and evaluate whether these measures vary between children exposed and not exposed to the treatment. We also analyze if treated children are more likely than controls not to leave some items blank. Except the role of attitudes, these analyses were not specified in the pre-analysis plan, and should be considered exploratory.

We can anticipate that we find no evidence of the importance of these channels. The success of the intervention does not seem to be driven by improvement in specific cognitive dimensions or by raising the ability to answer specific types of questions, nor by improving attitudes towards math, or by reducing the chances to leave questions unanswered. At the moment, this drives to the conclusion that MATL has worked by directly improving girls' general math skills.

5.1. Type of question: item format, cognitive dimension, level of difficulty

We analyze whether the treatment has a differential impact by item format, cognitive dimension, or level of difficulty of the single items. We classified the 20 items of the post-test by format, dimension, and difficulty. The item format can be open-response or multiple choice. The level of difficulty has been established with a one-parameter IRT analysis on the control group: we consider *easy* the items with difficulty below -0.5 (corresponding to 5 items), *difficult* those above or equal to 0.5 (5 items), and *medium* those in between (10 items). The cognitive dimension of the items – arguing, knowing, problem-solving – was assigned by experts in the field. The classification is shown in Table A.8 in the Appendix.

We calculate a new set of outcome scores, one for each category of items, by computing the share of correct answers within each category and standardizing the score. We have one post-test score constructed using only multiple-choice items, one constructed using only open-response items, one using only easy items, etc. We estimate the impact of the treatment on each one of the “new” outcome scores, applying a model similar to equation (1), but

allowing for correlation among the error terms of the different equations for each group of outcomes (difficulty, format, dimension), by implementing a SUR (Seemingly Unrelated Regression) model.

Results are reported in Table 9. These models were estimated separately for boys and girls, controlling for pre-test scores and school fixed effects.³³ For each group of items, we tested the equality of the treatment coefficients across item categories.³⁴

We find no significant effects for boys, so we concentrate on girls. The point estimate of the treatment effect on the multiple-choice score (0.163) is larger than the corresponding effect on the open-answer score (0.125), and both are significant at least at the 10% level. However, the difference between the effects is not significant. We find that the treatment effect is larger on the knowing dimension than on the other two scores (arguing and problem-solving), although the direction is the same and the magnitude is not very different. The treatment has no effect on the easy-items score, a substantial (but not highly significant) effect on the medium-items score, and a very large effect on the difficult-items score. This result is not surprising if we recall that high achieving girls are those who benefit the most.

These results suggest that the treatment enhances girls' math skills and is not driven by improvements in specific cognitive dimensions or in items with a specific format.

Tab.9 Treatment effect by type of item

5.2. Children's attitudes towards math

Girls generally display less positive attitudes towards math than boys and in particular lower interest and enjoyment, lower self-confidence in solving problems, lower beliefs in own abilities, and higher anxiety and stress (Di Tommaso et al. 2018; Else-Quest et al. 2010; Hill et al. 2016; Mullis et al. 2008; OECD 2016). Attitudes are a key factor to understand performance in math: although the direction of causality is difficult to assess, there is empirical evidence of a strong relationship between attitudes and math achievement.

³³ Since the test-scores in this section are based on the answers to few items, they are subject to larger measurement error (in the dependent variable). To simplify the model and avoid introducing many irrelevant variables, in these specifications we do not include all the controls included in the main specification. This should not be a problem, because all control variables are well balanced between treated and control groups (results with all control variables are similar and available from the authors upon request). To allow appropriate comparisons, the estimate of the treatment effect from the comparable all-items model is reported in the first panel of Table 9.

³⁴ As reported in Table 9, the Breuch-Pagan test always rejects the null hypothesis of independent equations. As a comparison, we have also estimated single equation OLS models, with standard errors clustered at the class level. Results are very similar and available upon request.

To explore whether MATL enhances children’s attitudes towards math, we administered a short questionnaire on math self-beliefs and emotional response, right after the conclusion of the post-test.³⁵ The questionnaire consisted of 5 items with four-level Likert scale answers, ranging from 1 (more negative attitude) to 4 (more positive attitude). Our measure of attitudes is the raw sum of scores.

Consistently with the existing literature, we observe a sizable gender gap in attitudes in favor of boys (Table A.9 in the Appendix). We find a small negative effect of the treatment on the attitudes of both boys and girls, although the estimates are either weakly or not statistically significant, depending on the specification (Table A.10 in the Appendix).³⁶

We may conclude that the success of MATL on girls’ math skills was not mediated by a positive change in their attitudes towards math. The absence of an effect on attitudes has come as a surprise to us. Yet, if the concept of what mathematics is, is grounded on traditional teaching practices and already heavily rooted in children’s minds, it might be difficult to change it. In particular, with a short intervention delivered not by the teacher, but by personnel external to the school. This does not rule out that longer programs could have an impact on pupils’ attitudes.

5.3. Item non-response

The reduction of the gender gap in math observed for children exposed to treatment could be due to the propensity to leave questions unanswered. If girls in the treatment group experienced a strong reduction of non-response whereas boys did not, we could speculate that the effect of MATL on the gender gap in math test scores might be driven by a change in the propensity to give answers (even in the absence of a real improvement in math skills).

To estimate the effect of MATL on the propensity to leave items blank we use two models. An OLS linear model for the number of non-response items in the post-test, and a logit model for the probability to leave at least two items blank (see Table A.11). In addition to the treatment variable, we include usual controls, school fixed effects, and the corresponding missing indicator in the pre-test. We find a negative and significant effect of the treatment on the number of non-response items. On average, the difference in the number of blank items in the post-test between treated and control children is approximately 0.14 and statistically significant. In terms of the probability to leave at least 2 items blank, the

³⁵ The English translation of the full questionnaire is available in Appendix C (C.3).

³⁶ We also perform the analysis using as a dependent variable the first component delivered by principal component analysis, obtaining very similar results.

average marginal effect of the treatment is -0.082 . Hence there is evidence that MATL is effective in reducing non-response, but the effect is small.

When analyzing the probability to leave items blank separately by gender, we find similar results for girls and boys. We may conclude that there is no evidence that the decline in the GGM is related to differential changes in the propensity to leave items blank.

Finally, we may ask whether the observed improvement in test scores for girls could be largely driven by a decline in non-response. Back-of-the-envelope calculations show that this is not the case, because the change on item non-response is much too small to drive a substantial improvement in test scores.³⁷ Overall, these results do not support the hypothesis MATL improves girls' performance by reducing the propensity to leave questions unanswered and suggests that the observed change is due to a real improvement in girls' math skills.

6. Limitations of the study

6.1. Threats to internal validity

We envisage two potential threats to internal validity. The first is related to the test design, the second to the awareness of the gender perspective of the mathematics laboratory.

Pre- and post-tests were designed by members of the research team, under the supervision of a member of the advisory board of the National Institute of Evaluation (INVALSI). There is some concern over the appropriateness of assessments made by developers of the program, as such measures have been found to overstate program impacts (Pellegrini et al. 2018). This feature could represent a weakness of the study. We believe that our results are still valid. First, the tests were standardized and scored blindly by the tutors (leaving no leeway to conscious or unconscious bias in grading, both versus one gender and versus the treatment classes). Second, they were conceived as comprehensive measures of Numbers abilities. Moreover, if a bias still existed, we would expect it to influence the results of both boys and

³⁷ If this were the case, the estimated improvement in test scores would have to be roughly the same as the number of questions that were previously left blank multiplied by the probability to get the right answer by chance. This probability is difficult to establish, because some questions are open-answer, and the multiple-choice ones have a variable number of options. If the effect of treatment on the number of missing items for girls is -0.11 (meaning that treatment makes the number of blank items decrease by 0.11), even if the probability to give the correct answer by chance was equal to 1 (obviously far from truth) we would end up with an increase of 0.11 correct answers (in a 20-items test). This value, still an upper bound of the true impact of treatment on the number of correct answers, is much smaller than the estimated impact of MATL for girls, amounting to 0.14 standard deviations in the post-test score variable and approximately equivalent to 0.6 questions. Employing a more reasonable figure for the probability to give the correct answer (say, $0.2-0.5$), the distance would become even larger.

girls; yet, this is not the case, as in our experiment the results of treated and controls differ only for girls.

A second issue of possible concern is related to the awareness of the ultimate goal of the intervention – reducing the gender gap in math – by the actors involved: the tutors conducting the laboratories and the school teachers. The crucial point is what exactly constitutes the intervention. Is it the teaching methodology or is there also a “gender awareness”?

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 because of transparency requirements set by the regional authorities. Teachers of both treated and control classes were aware of the gender perspective, and there are no major reasons to expect a difference between the two groups. Recall also that teachers were not actively involved in the conduction of the program, but had the role of observer. The teachers were also asked not to reveal the goal of the project to the children.

Also the tutors were aware of the aim of their work: this was inevitable, as the inclusive participation of all children is a distinctive element of the program. The tutors were sensitized to conduct the activities in order to promote the active participation of the entire classroom. The tutors’ awareness of the gender perspective of the program might have contributed to some extent to raise the performance of girls more than that of boys. In this light, we must recognize that the program has two elements that cannot be disentangled. In future work aimed at evaluating a scale-up of the intervention, we should consider the implementation of two alternative programs: one similar to the current one, another one with only the teaching component. This is difficult to realize, however, because it would require to deliberately give incomplete information about the program to the school boards and regional authorities endorsing the project.

6.2. External validity

The study did not involve a representative sample of schools. Participation in the RCT was voluntary, so principals and teaching staff of experimental units are likely to be positively selected in terms of interest in gender issues or in experimenting new teaching methods.

To examine whether and how participating units differ from the regional and national level, we exploit data from the second grade INVALSI national assessment of students’

competencies held in the previous scholastic year 2017-18, and compare individual and family characteristics of the children in the experimental classes (treated and controls) with the children population at large.³⁸ The results show that the children in experimental classes are substantially better performing in both math and Italian INVALSI tests than children at the regional and national level (Table 10). It may be noticed that the gender gap in math is much larger in the participating classes: this is consistent with the common finding that girls lag behind boys in math test scores in particular among well performers. Also the educational level of the parents and the share of children who attended kindergarten are higher in the experimental group.

Altogether, these results indicate that our study has limited external validity. Hence, further research is needed to evaluate ex-ante the potential effects of a scale-up of the intervention introducing the proposed teaching methodology in different contexts.

Tab.10 Comparison of experimental classes with Piedmont and Italy

7. Discussion and conclusions

We implement a teaching methodology grounded on active and cooperative learning practices, which aims to improve children's mathematical skills in primary school, and evaluate its impact on students in grade 3 in Italy. The teaching methodology, applied during 15 hours of math laboratories, focuses on peer interaction, the sharing of ideas, students' engagement, problem posing, and problem-solving. The methodology is evaluated using a randomized controlled trial conducted in the province of Torino, involving 50 third grade classes in 25 schools, and 1,044 students.

The key finding of the paper is that active learning methodologies for teaching mathematics have the potential to reduce the gender gap in math. In our implementation of these methodologies, the treatment had a positive and statistically significant effect on girls' achievement (on average 0.14 standard deviations) without hampering boys' performance. In educational studies, an effect of this magnitude can be considered large and policy-relevant. As a consequence, the intervention reduced the gender gap in mathematics by

³⁸ Upon schools' authorization, for the experimental classes we obtained from the INVALSI institute the class averages of test scores in math and Italian, oral marks in math and Italian, shares of pupils' childcare attendance, mother and father education levels. To analyze regional and national test scores, we analyzed the representative sample of classes where the test was administered under external supervision (to reduce cheating).

somewhere in the range of 40.1% to 47.5%. In addition, we found that both girls with high pre-test scores and girls with low educated parents benefit the most. In terms of migratory background, the average effect of the treatment is four times larger for migrant girls than for native girls.

There are many studies on the gender gap in mathematics, but there are few or no rigorous evaluations of the impact of different teaching methodologies. This paper fills this gap and provides a very important contribution to research on the causes of the gender gap in mathematics.

Given the concern and effort that many countries and the international community have shown on the issue of the gender gap in math and the career of women in STEM subjects, it is rather surprising that not much has been done until now about the role that teaching methodologies could play to tackle these issues.

Our experiment could be expanded in many respects. The intervention could be scaled up. The class-based intervention could be extended to a longer period (in this experiment it was only 15 hours) and delivered over more years. The sample could be increased and it could involve different Italian regions and/or different countries. It would also be of interest to look at whether the intervention has a longer-term effect. In addition, the teachers themselves could implement the new teaching methodology and it could be included in a teachers' professional development program. If teachers instead of tutors delivered the intervention, its effect would be more permanent. Nevertheless, our results are encouraging and suggest that properly designed teaching methodologies may improve math performance among girls.

References

- 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 curriculum di matematica (ciclo secondario)*. Lucca: Matteoni Stampatore.
- Arzarello, F., Ferrara, F. & Robutti, O. (2012). Mathematical modelling with technology: the role of dynamic representations. *Teaching Mathematics and its Applications*, 31(1), 20-30.
- Arzarello, F., Robutti, O. (2008). Framing the embodied mind approach within a multimodal paradigm in L. D. English, Lyn D., M. B. Bussi, G. A. Jones, R. A. Lesh, B. Sriraman and D. Tirosh (eds.) *Handbook of International Research in Mathematics Education*, Abingdon: Routledge.
- Arzarello, F., & Robutti, O. (2010). Multimodality in multi-representational environments. *ZDM*, 42(7), 715-731.
- Baron-Cohen, S. (2003). *The essential difference: The truth about the male and female brain*. New York: Basic Books.
- 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.
- Boaler, J. (2013). Ability and mathematics: The mindset revolution that is reshaping education. *Forum* 55, 1, 143-152.
- Boaler, J. (2016). *Mathematical mindsets: Unleashing students' potential through creative math, inspiring messages and innovative teaching*. Jossey-Bass.
- Bohnet, I. (2016). *What works Gender Equality by design*. Harvard: Harvard University Press.
- Card, D., & Payne, A. A. (2021). High school choices and the gender gap in STEM. *Economic Inquiry*, 59(1), 9-28.
- Carlana, M. (2019). Implicit Stereotypes: Evidence from Teachers' Gender Bias. *The Quarterly Journal of Economics*, 134(3), 1163–1224.
- 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>.
- Dee, T. S. (2007). Teachers and the Gender Gaps in Student Achievement. *Journal of Human Resources*, 42(3), 528-554

- Delaney, J. M., & Devereux, P. J. (2020). Math matters! The importance of mathematical and verbal skills for degree performance. *Economics Letters*, 186, 108850.
- Dweck, C.S. (2006a) *Mindset: the new psychology of success*. New York: Ballantine Books.
- Dweck, C.S. (2006b) Is Math a Gift? Beliefs that Put Females at Risk, in S.J. Ceci & W. Williams (Eds) *Why Aren't More Women in Science? Top Researchers Debate the Evidence*. Washington DC: American Psychological Association
- Di Tommaso, M.L., Maccagnan, A., & Mandolia S. (2018). The Gender Gap in Attitudes and Test Scores: A New Construct of the Mathematical Capability. *IZA Discussion Paper* 11843.
- Di Tommaso, M.L., Bernardi, M., Contini, D., De Rosa, D., Ferrara, F., Ferrari, F., Piazzalunga, D., & Robutti, O. (2020). *Tackling the gender gap in math with active learning teaching practices*. Final Report of the project Tackling the gender gap in mathematics in Piedmont. Version 1.
- Dossi, G., Figlio, D., Giuliano, P. & Sapienza, P. (2019) Born in the Family: Preferences for Boys and the Gender Gap in Math. *IZA Discussion Paper* 12156.
- Ellison, G., & 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.
- 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.
- Ferrara, F. & Ferrari, G. (2020). Reanimating tools in mathematical activity. *International Journal of Mathematical Education in Science and Technology*, 51(2), 307–323.
- 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., Gevrek, D., & Neumeier, C. (2020). Explaining the gender gaps in mathematics achievement and attitudes: The role of societal gender equality. *Economics of Education Review*, 76, 101978.
- Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences, *The Quarterly Journal of Economics*, 118(3), 1049-1074.
- Grinis, I. (2019). The STEM requirements of “Non-STEM” jobs: Evidence from UK online vacancy postings. *Economics of Education Review*, 70, 144-158.
- 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).
- 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.
- 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.
- Lave, J. & Wenger, E. (1991). *Situated Learning. Legitimate peripheral participation*. Cambridge: University of Cambridge Press.
- Lippmann, Q., & Senik, C. (2018). Math, girls and socialism. *Journal of Comparative Economics*, 46(3), 874-888.
- 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.
- Meinck S., & Brese F. (2019). Trends in gender gaps: using 20 years of evidence from TIMSS, *Large Scale Assessments in Education*, 7, 8.
- Mullis, I.V.S., Martin, M.O., & Foy, P. (with Olson, J.F., Preuschoff, C., Erberber, E., Arora, A., & Galia, J.) (2008). *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- 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.
- Niederle, M., & Vesterlund, L. (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 (2009). *Creating Effective Teaching and Learning Environments First Results from TALIS*, OECD Publishing, Paris.
- OECD (2014). *PISA 2012 results: What students know and can do - Student performance in reading, mathematics and science* Vol. I. OECD Publishing, Paris.
- 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.
- OECD (2019). *PISA 2018 Results (Volume II): Where All Students Can Succeed*, PISA, 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). 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, D. G., & Sydnor, J. R. (2010). Geographic variation in the gender differences in test scores. *Journal of Economic Perspectives*, 24(2), 95-108.
- 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.
- Staub, F. C., & Stern, E. (2002). The nature of teachers' pedagogical content beliefs matters for students' achievement gains: Quasi-experimental evidence from elementary mathematics. *Journal of Educational Psychology*, 94(2), 344-355.
- 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.
- Wehmeyer, M. L., & Schwartz, M. (2001). Disproportionate representation of males in special education services: Biology, behavior, or bias?. *Education and treatment of children*, 24(1), 28-45.
- What Works Clearinghouse (2013). *Standard Handbook. Version 4.0*. Institute of Education Sciences, Washington, DC.
- Zohar, A., & Sela, D. (2003). Her physics, his physics: gender issues in Israeli advanced placement physics classes. *International Journal of Science Education*, 25(2), 245–26.

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, and post-test, full sample

<i>Panel A – Individual level</i>	Control group	Treated group	P-value of the difference
Girl	0.500	0.514	0.663
SEND – broad definition	0.149	0.156	0.736
SEND – broad definition (F)	0.106	0.139	0.260
SEND – broad definition (M)	0.191	0.175	0.635
SEND – narrow definition	0.086	0.083	0.868
SEND – narrow definition (F)	0.046	0.064	0.362
SEND – narrow definition (M)	0.126	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
Parents low educated	0.670	0.724	0.057
Parents high educated	0.329	0.275	0.057
Parents education missing	0.160	0.131	0.184
Observations	525	519	1,044
Raw pre-test score	10.786	10.703	0.774
Raw pre-test score (F)	10.394	10.152	0.540
Raw pre-test score (M)	11.179	11.274	0.816
Observations	481	452	933
Raw post-test score	9.842	10.335	0.067
Raw post-test score (F)	9.133	9.817	0.067
Raw post -test score (M)	10.566	10.924	0.385
Observations	493	490	983
<i>Panel B – Class level</i>			
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
Full time	0.800	0.720	0.517
Observations	25	25	50
Permanent contract teachers	1.000	0.920	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 with a university degree	0.375	0.400	0.861
Teacher's age (years)	48.33	50.00	0.501
Observations	24	25	49

Notes: SEND stands for “special educational needs and disability”. “SEND - broad definition” includes children with any form of special education needs or disability, “SEND - narrow definition” includes only children with a certified form of special education need or disability. Parents low educated: no parent has a tertiary degree; parents high educated: at least one parent has a tertiary degree. Summary statistics refer to full sample (a). Summary statistics of pre-test refers to 933 observations (sample b), those of post-test refers to 983 observations (sample c). 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 (0.141)	0.006 (0.194)	-0.009 (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** (0.021)	0.053* (0.031)	0.037 (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

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

Tab. 5 Main results: effects of the treatment

Variable	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 variables		
	Overall (1)	Girls (2)	Boys (3)	Overall (4)	Girls (5)	Boys (6)	Overall (7)	Girls (8)	Boys (9)	Overall (10)	Girls (11)	Boys (12)
Treatment	0.116*	0.154*	0.081	0.091	0.143	0.041	0.077	0.164**	-0.015	0.083**	0.142**	-0.009
	(0.065)	(0.086)	(0.086)	(0.068)	(0.090)	(0.092)	(0.048)	(0.069)	(0.068)	(0.033)	(0.055)	(0.046)
Pre-test score							0.760***	0.733***	0.788***	0.739***	0.737***	0.748***
							(0.023)	(0.037)	(0.024)	(0.025)	(0.035)	(0.033)
Constant	-0.048	-0.208***	0.115*	-0.030	-0.191***	0.133**	0.007	-0.132**	0.048	0.163	-0.194	0.290
	(0.045)	(0.053)	(0.058)	(0.046)	(0.051)	(0.063)	(0.040)	(0.058)	(0.045)	(0.157)	(0.225)	(0.249)
Observations	983	501	482	888	448	440	888	448	440	888	448	440
R-squared	0.003	0.007	0.002	0.002	0.006	0.000	0.592	0.572	0.601	0.616	0.603	0.641
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 (children with any form of special education needs or disability), parental education (parents high educated: at least one parent has a tertiary degree; parents’ education missing), migratory background (migrant I generation, II generation, information missing), class size, and time schedule. Full results (columns 10-12) are available in Table A.7.

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

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

Variable	Overall (1)	Girls (2)	Boys (3)
Treatment	0.081** (0.033)	0.155*** (0.053)	-0.013 (0.048)
Pre-test score	0.719*** (0.038)	0.679*** (0.050)	0.735*** (0.041)
Treatment* Pre-test score	0.062 (0.048)	0.127* (0.064)	0.028 (0.058)
Constant	0.139 (0.159)	-0.159 (0.224)	0.292 (0.251)
Observations	888	448	440
R-squared	0.614	0.607	0.641
School FE	YES	YES	YES
Additional controls	YES	YES	YES

Notes: Standardized test scores. Standard errors clustered at the class level in parenthesis. Sample (d). Additional controls include girl (in the Overall specification), SEND (special education needs and disability) dummy broad definition (children with any form of special education needs or disability), parental education (parents high educated: at least one parent has a tertiary degree; parents' education missing), migratory background (migrant I generation, II generation, information missing), class size, and time schedule. Full results are available upon request.

*** 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)
	Treatment on children with low educ. parents	0.060 (0.051)	0.182** (0.072)	-0.075 (0.068)
Effect of treatment by parents' level of education	Treatment on children with high educ. parents	0.086 (0.070)	0.082 (0.096)	0.044 (0.107)
	Treatment on children with educ parents missing	0.163** (0.076)	0.113 (0.186)	0.179 (0.151)
	<i>Observations</i>	888	448	440
	<i>R-squared</i>	0.616	0.604	0.643
Effect of treatment by migrant status	Treatment on natives	0.092** (0.041)	0.104* (0.062)	0.032 (0.062)
	Treatment on migrants	0.020 (0.094)	0.399*** (0.131)	-0.285* (0.164)
	<i>Observations</i>	888	448	440
	<i>R-squared</i>	0.615	0.605	0.643
	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. Sample (d). Additional controls include girl (in the Overall specification), SEND (special education needs and disability) dummy broad definition (children with any form of special education needs or disability), class size and time schedule; migratory background (migrant I generation, II generation, information missing) in the first panel and parental education (parents high educated: at least one parent has a tertiary degree; parents' education missing) in the second panel. Natives include children born in Italy with at least one parents born in Italy, migrants include first- and second- generation migrants (i.e., those with both parents born abroad) and children with migratory background information missing. Children with low educated parents have no parents with a tertiary degree, children with high educated parents have at least one parent with a tertiary degree. Full results are available upon request.

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

Tab. 8 Robustness checks

Variables	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 (1)	Girls (2)	Boys (3)	Overall (4)	Girls (5)	Boys (6)	Overall (7)	Girls (8)	Boys (9)	Overall (10)	Girls (11)	Boys (12)
Treatment	0.093** (0.035)	0.144*** (0.053)	0.008 (0.051)	0.111*** (0.037)	0.159*** (0.053)	0.017 (0.054)	0.110*** (0.037)	0.165*** (0.056)	0.035 (0.047)	0.074** (0.032)	0.118** (0.050)	-0.002 (0.046)
Pre-test scores	0.764*** (0.027)	0.740*** (0.036)	0.771*** (0.033)	0.769*** (0.026)	0.734*** (0.034)	0.786*** (0.034)	0.733*** (0.029)	0.716*** (0.037)	0.731*** (0.034)	0.744*** (0.026)	0.737*** (0.035)	0.739*** (0.033)
Pre-test sc. missing							-0.069 (0.097)	-0.195 (0.128)	0.078 (0.151)			
Constant	0.032 (0.174)	-0.228 (0.213)	0.092 (0.309)	0.090 (0.159)	-0.034 (0.194)	0.152 (0.338)	-0.012 (0.185)	-0.419 (0.261)	0.262 (0.234)	0.153 (0.152)	-0.055 (0.204)	0.242 (0.271)
Observations	818	425	393	757	396	361	983	501	482	916	462	454
R-squared	0.608	0.606	0.623	0.595	0.588	0.616	0.557	0.550	0.583	0.608	0.594	0.637
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
SEND def.	Narrow version	Narrow version	Narrow version	Broad version	Broad version	Broad version						

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 (children with any form of special education needs or disability) when appropriate (i.e., excluding models 4 to 6), parental education (parents high educated: at least one parent has a tertiary degree; parents' education missing), migratory background (migrant I generation, II generation, information missing), class size, and time schedule. Full results are available upon request.

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

Tab. 9 Treatment effect by type of item

		Girls		Boys	
	Outcome	Treatm. Effect	S.E.	Treatm. Effect	S.E.
All items	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: Standardized test scores. Sample (d). 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, school fixed effects and the pre-test score are included as controls. Below the SUR results, the results of the Breusch-Pagan test for independent equations and the tests of equivalence among the treatment coefficients of interest are reported, 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.8.

*** 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 difference experimental vs. Piedmont classes (3)	Italian Classes (4)	P-value of the difference experimental vs. Italian classes (5)
Invalsi score in Italian	0.393	0.067	0.000	0.000	0.000
Invalsi score in Math	0.559	0.023	0.000	0.000	0.000
Invalsi score Italian Female	0.389	0.113	0.000	0.017	0.000
Invalsi score Italian Male	0.407	0.021	0.000	-0.044	0.000
Invalsi score Math Female	0.439	-0.052	0.000	-0.070	0.000
Invalsi score Math Male	0.681	0.086	0.000	0.029	0.000
Gender Gap Math	-0.241	-0.139	0.000	-0.099	0.000
School grade Italian	8.140	8.105	0.354	8.058	0.011
School grade Math	8.224	8.230	0.863	8.143	0.014
Kindergarten attendance	0.420	0.326	0.000	0.381	0.000
Girl	0.510	0.504	0.007	0.489	0.000
Mother's education					
Lower secondary	0.258	0.339	0.000	0.331	0.000
Upper secondary	0.405	0.405	0.869	0.409	0.000
Tertiary	0.337	0.257	0.000	0.261	0.000
Father's education					
Lower secondary	0.360	0.469	0.000	0.427	0.000
Upper secondary	0.405	0.353	0.000	0.391	0.000
Tertiary	0.235	0.178	0.000	0.183	0.000
Parents low educated	0.697	0.743	0.012	0.754	0.000
Parents high educated	0.302	0.256	0.012	0.245	0.000
Parents' educ. level missing	0.145	0.097	0.000	0.154	0.409
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 1,020 (min) – 1,044 (max); Piedmont classes 689 (min) – 1,391 (max); Italian classes 12,766 (min) – 26,142 (max). Invalsi scores are standardized (at the Italian level). The Gender Gap in Math is defined as the Invalsi score for Female minus the Invalsi score for boys. School grade refer to the teacher grades given by teacher, which varies between 1 and 10. Parents low educated: no parent has a tertiary degree; parents high educated: at least one parent has a tertiary degree.

FIGURES

Fig. 1 Timeline of the intervention

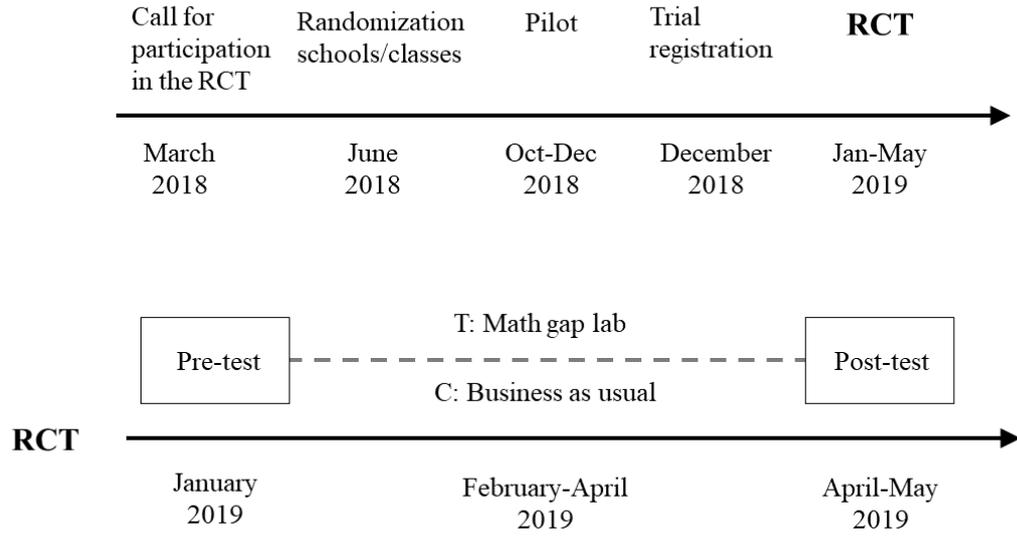
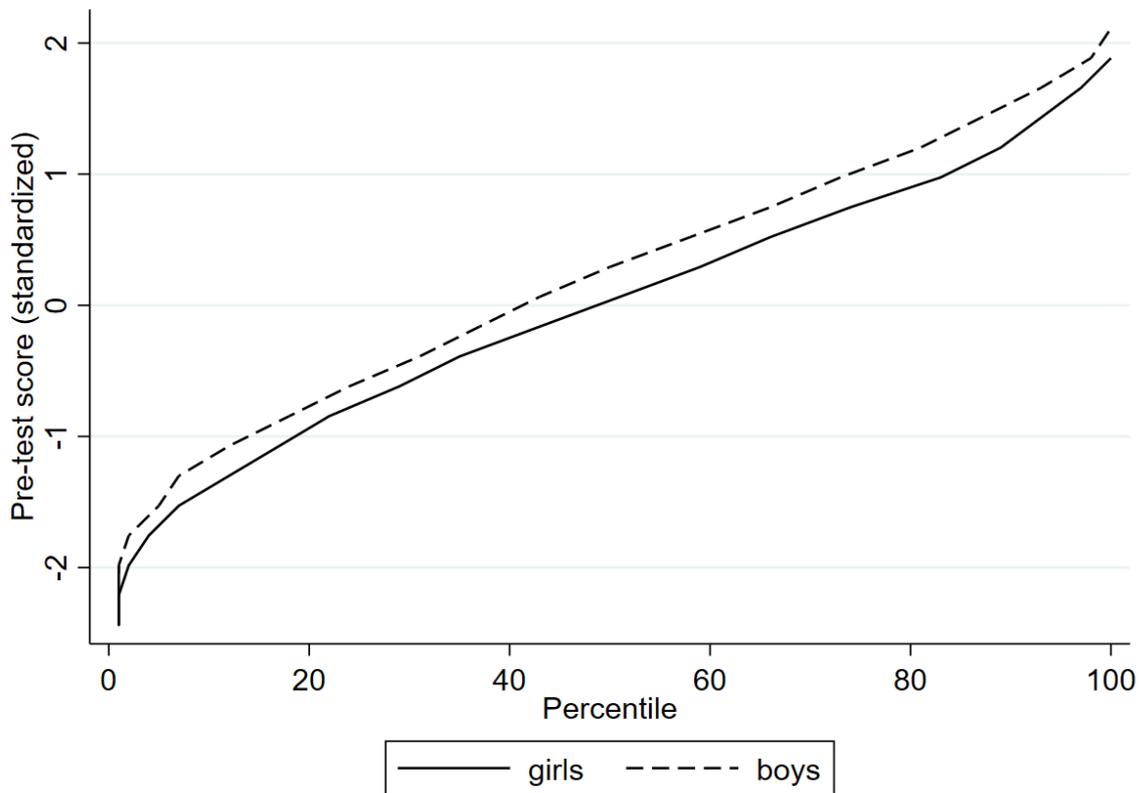
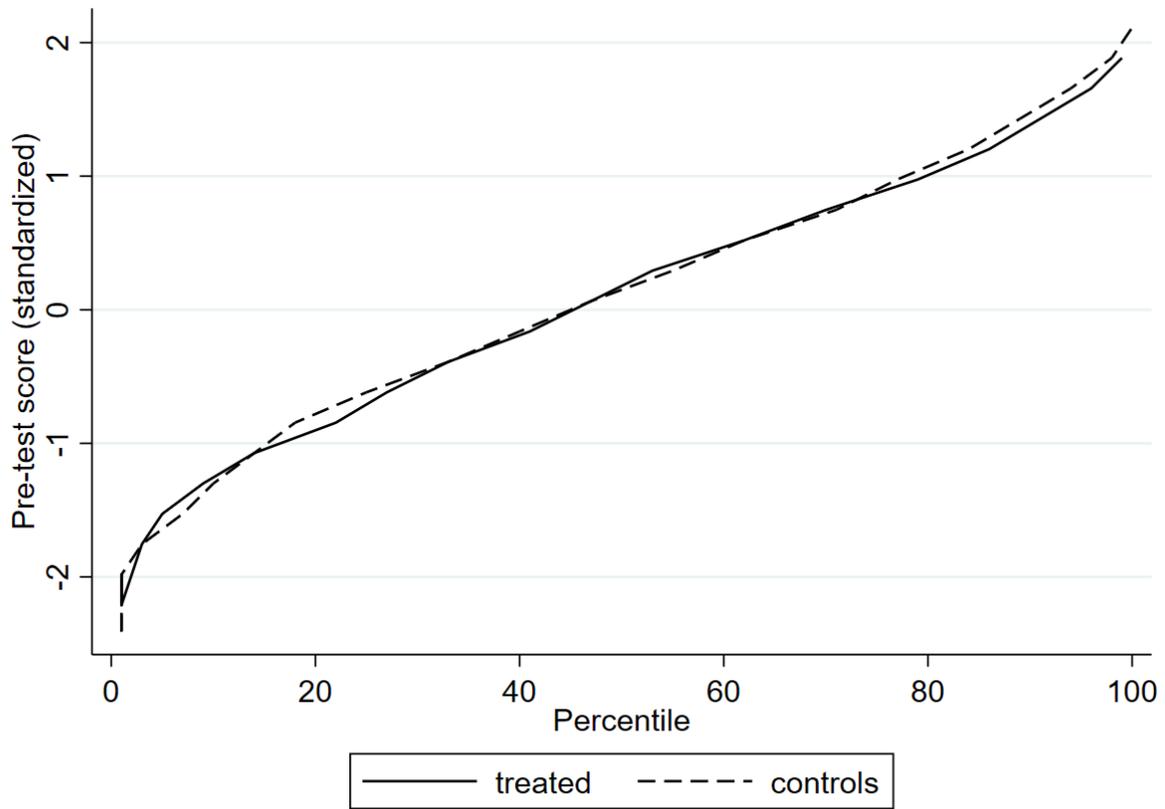


Fig. 2 Gender gap in the pre-test



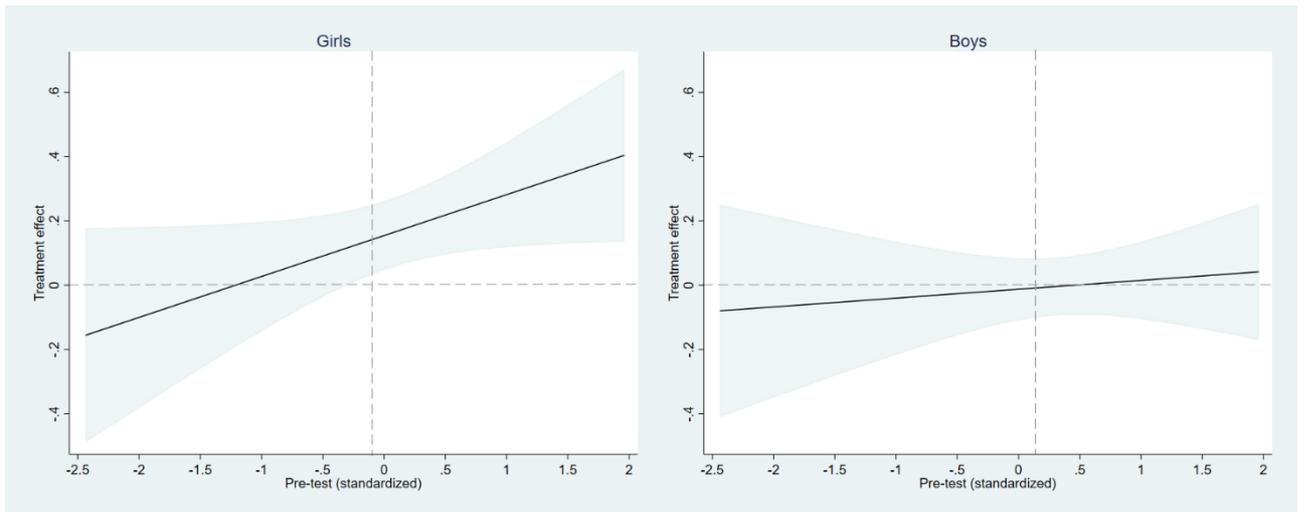
Note: Children present at the pre-test (sample b), 933 observations.

Fig. 3 Pre-test score distribution by treatment status



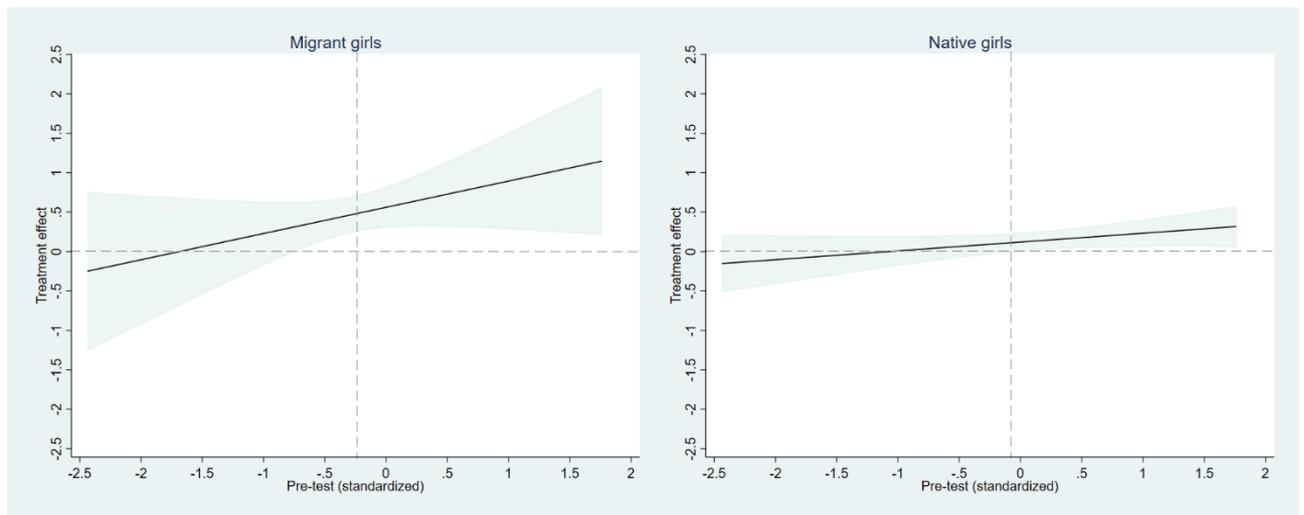
Note: Children present at the pre-test (sample b), 933 observations.

Fig. 4 Treatment effect by prior achievement's levels



Notes: Effect of the treatment by pre-test scores for boys and girls (estimates from regression in Table 6). Sample (d), 888 observations. 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.

Fig. 5 Treatment effect by prior achievement's levels, migrant and native girls



Note: Effect of the treatment by pre-test scores for migrant and native girls (estimates available upon request). Sample (d), 888 observations. The dashed horizontal line represents a zero-treatment effect, whereas the dashed vertical line represents the pre-test score mean for each of the two groups respectively. Native girls include children born in Italy with at least one parents born in Italy, migrant girls include first- and second- generation migrants (i.e., those with both parents born abroad) and children with migratory background information missing.

APPENDIX

Appendix A – Additional tables

Tab. A.1 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.2 Variables definition

Variable	Definition
Individual level	
Pre-test score	Pre-test score
Post-test score	Post-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
Parents low educated	1= no parent has tertiary degree; 0 = otherwise
Parents high educated	1= at least one parent has tertiary degree; 0 = otherwise
Parents education missing	1= at least one parent has missing education; 0 = otherwise
Class level	
Class size	Number of children in each class
Full time	1= class with full time schedule (40 hours per week); 0 = otherwise (27/30)
Pre-test score (mean)	Mean of pre-test score at class level
Pre-test score (s.d.)	Standard deviation of pre-test 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	1= Teacher with a permanent contract; 0 = otherwise
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 with university degree	1= Teacher with a permanent contract; 0 = otherwise
Teacher's age (years)	Age of teacher

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 items missing (post-test)	4	1	3
Number of SEND narrow def. in the full sample	88	43	45
Number of SEND broad def. 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”. “SEND - narrow definition” includes only children with a certified form of special education need or disability, “SEND - broad definition” includes children with any form of special education needs or disability.

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

	Control group	Treated group	P-value of the difference
Girl	0.505	0.514	0.772
SEND – broad definition	0.139	0.148	0.687
SEND – broad definition (F)	0.100	0.126	0.349
SEND – broad definition (M)	0.180	0.172	0.816
SEND – narrow definition	0.079	0.077	0.927
SEND – narrow definition (F)	0.040	0.059	0.320
SEND – narrow definition (M)	0.118	0.094	0.432
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
Parents low educated	0.677	0.716	0.185
Parents high educated	0.322	0.283	0.185
Parents education missing	0.154	0.136	0.439
Observations	493	490	983
Pre-test score	10.772	10.856	0.774
Pre-test score (F)	10.358	10.232	0.756
Pre-test score (M)	11.188	11.500	0.455
Observations	457	431	888

Notes: SEND stands for “special educational needs and disability”. “SEND - broad definition” includes children with any form of special education needs or disability, “SEND - narrow definition” includes only children with a certified form of special education need or disability. Parents low educated: no parent has a tertiary degree; parents high educated: at least one parent has a tertiary degree. Summary statistics refer to children present at the post-test (sample c). Summary statistics of pre-test refers to 888 observations (sample d).

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
Pre-test score (F)	10.358	10.232	0.756
Pre-test score (M)	11.188	11.500	0.455
Girl	0.501	0.508	0.834
SEND – broad definition	0.144	0.150	0.788
SEND – broad definition (F)	0.104	0.127	0.447
SEND – broad definition (M)	0.184	0.174	0.792
SEND – narrow definition	0.080	0.076	0.808
SEND – narrow definition (F)	0.043	0.059	0.453
SEND – narrow definition (M)	0.118	0.094	0.415
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
Parents low educated	0.671	0.740	0.025
Parents high educated	0.328	0.259	0.025
Parents education missing	0.153	0.141	0.625
Observations	457	431	888

Notes: SEND stands for “special educational needs and disability”. “SEND - broad definition” includes children with any form of special education needs or disability, “SEND - narrow definition” includes only children with a certified form of special education need or disability. Parents low educated: no parent has a tertiary degree; parents high educated: at least one parent has a tertiary degree. 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	0.100 (0.080)	0.098 (0.074)
Girl	0.008 (0.076)	0.049 (0.081)
SEND – broad definition	0.198 (0.196)	0.109 (0.177)
Migrant I generation	1.074* (0.550)	0.727 (0.545)
Migrant II generation	-0.366** (0.164)	-0.379** (0.162)
Migrant missing	-0.771 (0.790)	-1.039 (0.850)
Parents high educated	-0.628*** (0.164)	-0.549*** (0.139)
Parents education missing	-0.522*** (0.163)	-0.516*** (0.165)
Class size	0.020 (0.316)	-0.062 (0.208)
Full time	-3.898** (1.902)	-1.343 (1.357)
Teaching experience	-0.035 (0.059)	
Teacher’s university degree	1.187 (1.153)	
Teacher’s age	0.090 (0.087)	
Constant	-1.767 (6.847)	2.172 (3.858)
Observations	845	888
Wald test of joint significance	93.97 (0.000)	65.70 (0.000)
School FE	YES	YES

Notes: Standardized pre-test scores. Standard errors clustered at the class level in parenthesis. Sample (d). Results of a logit model. “SEND - broad definition” includes children with any form of special education need or disability. Parents high educated: at least one parent with a tertiary degree. Reference categories are: boy, typically developed child, native child, parent’s low educated.

*** 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	0.083** (0.033)	0.142** (0.055)	-0.009 (0.046)
Pre-test score	0.739*** (0.025)	0.737*** (0.035)	0.748*** (0.033)
Girl	-0.097** (0.047)		
SEND broad definition	-0.106 (0.067)	0.034 (0.129)	-0.184* (0.101)
Migrant I generation	-0.061 (0.156)	-0.061 (0.237)	-0.059 (0.146)
Migrant II generation	0.047 (0.073)	0.004 (0.099)	0.126 (0.129)
Migrant missing	-0.152 (0.122)	-0.063 (0.244)	-0.484 (0.351)
Parents high educated	0.121** (0.055)	0.083 (0.081)	0.158* (0.085)
Parents education missing	-0.043 (0.095)	-0.159 (0.121)	0.145 (0.110)
Class size	-0.012 (0.008)	0.006 (0.013)	-0.023* (0.012)
Full time	0.008 (0.051)	-0.076 (0.065)	0.057 (0.074)
Constant	0.163 (0.157)	-0.194 (0.225)	0.290 (0.249)
R-squared	0.616	0.603	0.641
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. SEND - broad definition" includes children with any form of special education need or disability. Parents high educated: at least one parent with a tertiary degree. Reference categories are: boy, typically developed child, native child, parents' low educated.

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

Tab. A.8 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.9 Attitudes, summary statistics

	Obs.	Mean	S.D.	Min	Max
Overall	882	15.147	3.351	5	20
Boys	438	15.554	3.299	5	20
Girls	444	14.745	3.358	5	20
	Obs.	Diff	S.E.	P-value of the diff	
Mean diff. Boys vs. Girls	882	0.809	0.224	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.

Tab. A.10 Effect of the treatment on attitudes towards mathematics

Variable	Attitudes (1)	Attitudes (2)
Girls	-0.819** (0.393)	-0.841** (0.376)
Treatment effect on boys	-0.508 (0.371)	-0.538* (0.301)
Treatment effect on girls	-0.485 (0.381)	-0.544* (0.325)
Constant	15.801 (0.272)	15.899*** (0.629)
Observations	882	882
R-squared	0.020	0.075
School FE		YES
Additional controls		YES

Notes: Standard errors clustered at the class level in parenthesis. Sample (d). 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 is an index build as a sum of such points, Additional controls include SEND (special education needs and disability) dummy broad definition (children with any form of special education needs or disability), parental education (parents high educated: at least one parent has a tertiary degree; parents' education missing), migratory background (migrant I generation, II generation, information missing), class size, and time schedule. Full results available upon request.

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

Tab. A.11 Treatment effect on blank items

Variables	OLS			LOGISTIC		
	Overall (1)	Boys (2)	Girls (3)	Overall (4)	Boys (5)	Girls (6)
Treatment	-0.146** (0.061)	-0.142* (0.077)	-0.137* (0.072)	0.284*** (0.101)	0.298*** (0.113)	0.223** (0.161)
Gender	0.008 (0.054)			0.799 (0.173)		
N. of blank items at pre-test	0.138*** (0.041)	0.146** (0.057)	0.115*** (0.039)			
Pre-test score std.	-0.037 (0.038)	-0.028 (0.055)	-0.056 (0.042)	1.009 (0.167)	0.916 (0.188)	1.183 (0.357)
At least 2 blank items pre-test				5.579*** (1.650)	3.955*** (1.741)	7.307*** (4.749)
Constant	0.070 (0.243)	-0.260 (0.282)	0.441 (0.369)	0.043 (0.114)	0.257 (0.636)	0.000*** (0.000)
Observations	888	448	440	888	440	448
R-squared	0.159	0.191	0.212			
School FE	YES	YES	YES	YES	YES	YES
Additional Controls	YES	YES	YES	YES	YES	YES
Dependent Variable	Num. of blank items at post-test	Num. of blank items at post-test	Num. of blank items at post-test	Dummy (at least 2 blank items at post-test)	Dummy (at least 2 blank items at post-test)	Dummy (at least 2 blank items at post-test)

Notes: Standardized test scores. Standard errors clustered at the class level in parenthesis. In columns (1), (2), and (3) the dependent variable is the number of blank items at the post-test; in columns (4), (5) and (6) the dependent variable is a dummy variable equal to 1 if at least 2 items are left blank at the post-test, and a logistic model is estimated (coefficients reported in terms of Odd Ratio). Additional controls include SEND (special education needs and disability) dummy broad definition (children with any form of special education needs or disability), parental education (parents high educated: at least one parent has a tertiary degree; parents' education missing), migratory background (migrant I generation, II generation, information missing), class size, and time schedule. Full results available upon request. *** p<0.01, ** p<0.05, * p<0.1.

Appendix B – IRT analysis

In this Appendix, we present results from our preferred specification using as outcome variable Y_1 and as baseline control Y_0 the latent abilities estimated with IRT (Item Response Theory) models instead of pre- and post-test standardized results (Table B.1), and the heterogeneous results by prior achievement (Table B.1). The first two columns present our main results to ease the comparison. 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.

All the results are confirmed and similar in magnitude to results using, 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.

Tab. B.1 Main results with IRT scores as dependent variable

Dependent var.	Post-test std.		Ability from					
	Girls	Boys	IRT 1 p.	IRT 1 p.	IRT 2 p.	IRT 2 p.	IRT 2 p.	IRT 2 p.
Variable	(1)	(2)	Girls	Boys	Girls	Boys	Girls	Boys
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.142** (0.055)	-0.009 (0.046)	0.138*** (0.048)	-0.009 (0.043)	0.117*** (0.043)	-0.011 (0.043)	0.121*** (0.044)	-0.019 (0.041)
Pre-test score std.	0.737*** (0.035)	0.748*** (0.033)						
Pre-test ability IRT 1p.			0.743*** (0.038)	0.732*** (0.035)				
Pre-test ability IRT 2p.					0.748*** (0.038)	0.759*** (0.034)	0.766*** (0.039)	0.778*** (0.033)
Constant	-0.194 (0.225)	0.290 (0.249)	-0.150 (0.200)	0.230 (0.224)	-0.204 (0.179)	0.257 (0.211)	-0.084 (0.177)	0.442** (0.196)
Observations	448	440	448	440	448	440	448	440
R-squared	0.603	0.641	0.601	0.625	0.607	0.641	0.605	0.635
School FE	YES							
Additional controls	YES							

Notes: Standard errors clustered at the class level in parenthesis. Sample (d). 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 (children with any form of special education needs or disability), parental education (parents high educated: at least one parent has a tertiary degree; parents' education missing), migratory background (migrant I generation, II generation, information missing), class size, and time schedule. Full results available upon request.

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

Tab. B.2 Heterogeneous results by prior achievements with IRT scores as dependent variable

Dependent var.	Post-test	Post-test	Ability from	Ability from IRT				
	std.	std.	IRT 1 p.	IRT 1 p.	IRT 2 p.	IRT 2 p.	IRT 2 p. (controls)	2 p. (controls)
Variable	Girls (1)	Boys (2)	Girls (3)	Boys (4)	Girls (5)	Boys (6)	Girls (7)	Boys (8)
Treatment	0.155*** (0.053)	-0.013 (0.048)	0.153*** (0.046)	-0.012 (0.045)	0.131*** (0.043)	-0.016 (0.045)	0.137*** (0.044)	-0.008 (0.046)
Pre-test score	0.679*** (0.050)	0.735*** (0.041)	0.683*** (0.055)	0.722*** (0.044)	0.696*** (0.055)	0.741*** (0.042)	0.705*** (0.055)	0.753*** (0.041)
Treatment* Pre-test score	0.127* (0.064)	0.028 (0.058)	0.128* (0.069)	0.021 (0.063)	0.114 (0.068)	0.039 (0.061)	0.115 (0.069)	0.031 (0.062)
Constant	-0.159 (0.224)	0.292 (0.251)	-0.120 (0.195)	0.231 (0.225)	-0.170 (0.178)	0.261 (0.214)	-0.114 (0.189)	0.327 (0.223)
Observations	448	440	440	440	448	440	448	440
R-squared	0.607	0.641	0.604	0.625	0.610	0.642	0.611	0.638
School FE	YES							
Additional controls	YES							

Notes: Standard errors clustered at the class level in parenthesis. Sample (d). 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 (children with any form of special education needs or disability), parental education (parents high educated: at least one parent has a tertiary degree; parents' education missing), migratory background (migrant I generation, II generation, information missing), class size, and time schedule. Pre-test scores are always the appropriate ones (e.g. standardized, IRT 1p., or IRT 2p.) depending on the outcome used. Full results available upon request.

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

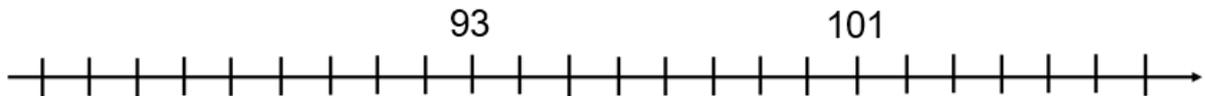
Appendix C
Test: Pre- and post-test, and non-cognitive questionnaire

C1. Pre-test on math competences

NAME
☺

GOOD LUCK!

1) Look at the number line.



Add these numbers to the line: 89 and 97 and 105.

2) a. Which number has 3 units and 2 tens?

- A. 23
- B. 203
- C. 302

b. Complete the sentence:

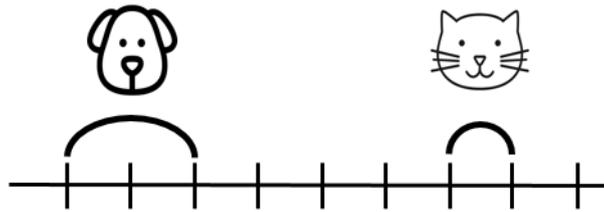
There are tens in the number 703.

3) Martha has to organize the bookshelves in her room:

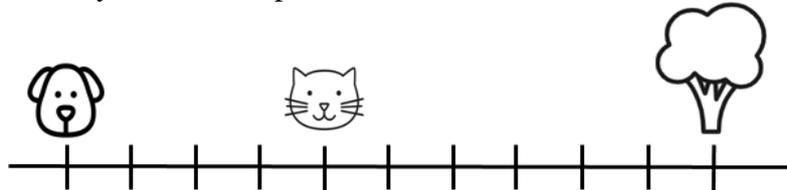


Martha wants to have the same number of books on each shelf.
How many books does she have to move from shelf A to shelf B?
Answer: books.

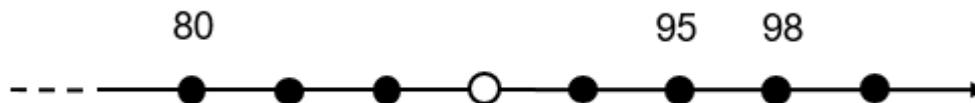
- 4) A dog and a cat are playing at chasing each other.
This is what the dog's step and the cat's step are like:



At a certain point they are in these positions:



- a. How many steps does the cat have to take to reach the tree?
Answer: steps.
- b. How many steps does the dog have to take to reach the tree?
Answer: steps.
- 5) Add 7 units and 3 tens to the number two hundred and ten: what number do you get?
A. 283
B. 247
C. 220
- 6) Look at this figure:



What number can you put over the white circle?
Answer:

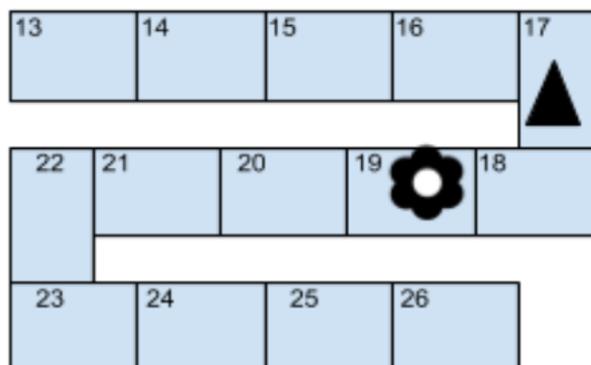
- 7) Today is Peter's birthday.
Peter has brought candy and cakes to celebrate with his friends.
This is how he distributes them:

	<p>1 cake for every 8 children</p>
	<p>3 pieces of candy for each child</p>

There are 48 children at the birthday party.

- a. How many cakes did Peter bring?
 A. 8
 B. 6
 C. 11
- b. How many pieces of candy did Peter have in all?
 Answer: pieces

- 8) Martina and Christian are playing Snakes and Ladders.
Martina's piece is shaped like a flower: she rolled a 6 and moved to the space shown in the figure.



- a. What space was Martina's piece on before she rolled the dice?
 Answer: On space
- b. Christian's piece is shaped like a triangle and before he moved was on space 15.
 What number did Christian roll last?
 Answer: He rolled

9) Amanda is preparing a box of beads for a friend.
She bought 4 hundreds, 2 tens and 23 units.

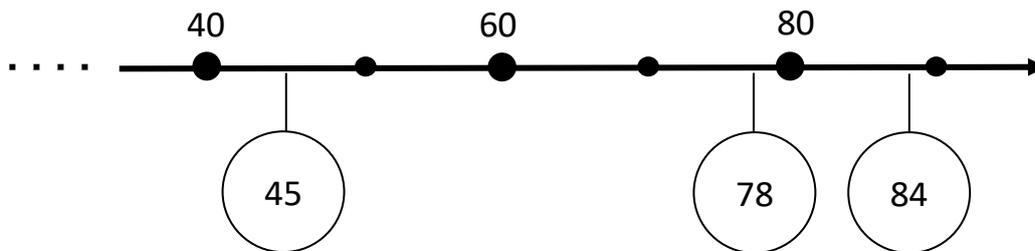
a. Choose the operation to be used to count how many beads Amanda bought:

- A. $4 + 23 + 2$
- B. $400 + 23 + 2$
- C. $400 + 20 + 23$

b. How many beads does Amanda have in all?

Answer: beads

10) Look at the number line.

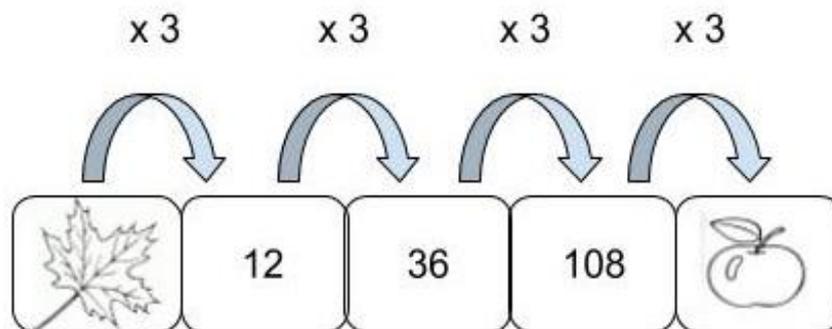


The number in one of the circles is wrong.

The wrong number is:

- A. 45
- B. 78
- C. 84

11) Look at the sequence in the boxes and the operation indicated by the arrows.



The leaf and the apple cover two numbers.

a. What number is hidden behind the leaf?

Answer:

b. What number is hidden behind the apple?

Answer:

12) Frank's birthday is February 22 and his brother Luke's is 3 weeks earlier.

FEBBRAIO 2017						
Lunedì	Martedì	Mercoledì	Giovedì	Venerdì	Sabato	Domenica
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	1	2	3	4	5
6	7	8	9	10	11	12

a. When is Luke's birthday?

- A. February 1
- B. February 19
- C. January 31

b. Luke and Frank's father celebrates his birthday on March 8.

Complete the sentence by writing a number on the dotted line:
The father's birthday is exactly weeks after Frank's.

13) A t-shirt costs 8 euros and 70 cents.
 Three friends have this much money:

Matt	
Mark	
Burt	

Who can't buy the t-shirt?

- A. Matt
- B. Burt
- C. Mark

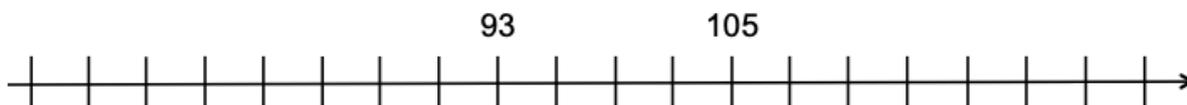
C2. Post-test on math competences

NAME



GOOD LUCK!

- 1) Look at the number line.

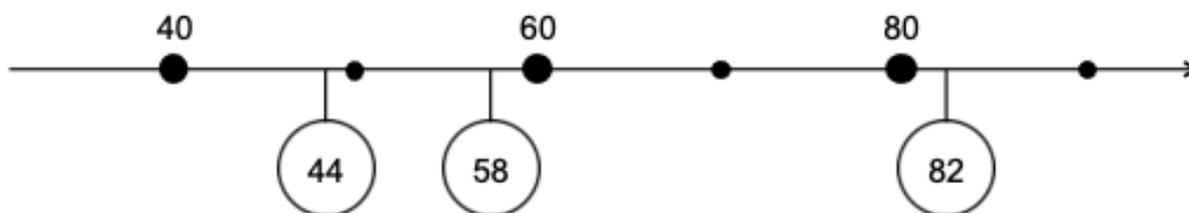


Write these numbers on the line: 90 and 99 and 114.

- 2) Think about the number 940.
a. What digit is in the tens place?
Answer:

b. How many tens make up the number 940?
Answer: tens

- 3) Look at the number line:



The number in one of the circles is wrong.

The wrong number is:

- A. 44
- B. 58
- C. 82

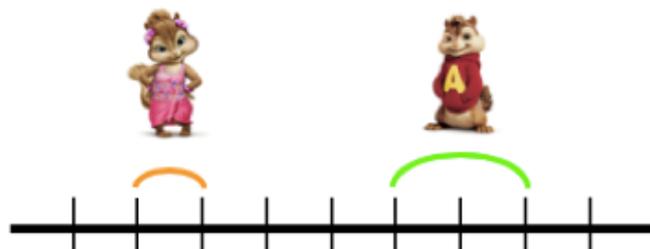
- 4) Look at this figure:



What number can you put over the white circle?

Answer:

- 5) Chippie and Chip are racing to get an acorn.
Here is Chippie's step and Chip's step:

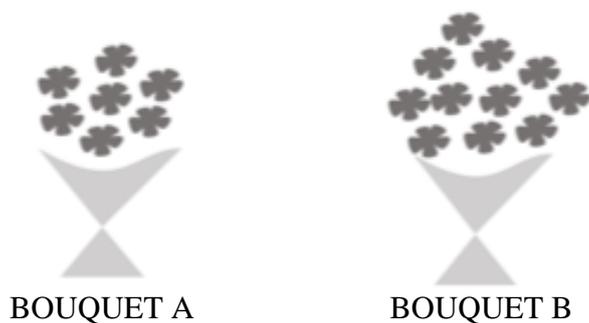


These are their positions at the beginning:



- a. How many steps does Chippie have to take to arrive exactly at the acorn?
Answer: steps
- b. How many steps does Chip have to take?
Answer: steps

- 6) Eliza has two bouquets:



Eliza wants both bouquets to have the same number of flowers.
What does she have to do?

Complete the sentence:

Eliza moves flowers from bouquet to bouquet

7) Mr. Andrew, the teacher, prepares colored pencils for the class. He has 5 hundreds, 68 units and 3 tens.

a. What operation does Mr. Andrew use to count how many pencils he has?

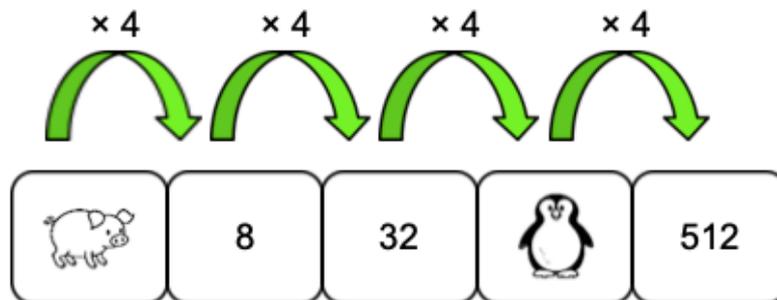
- A. $50 + 3 + 68$
- B. $500 + 30 + 68$
- C. $68 + 3 + 500$

b. Mr. Andrew takes only the red, blue and green pencils: he counts 120. He has 25 students in his class.

Can Mr. Andrew give 5 pencils of these colors to each student?

- A. Yes, with 5 pencils left over.
- B. No, he doesn't have enough pencils.
- C. Yes, and he has no red, blue or green pencils left over.

8) Look at this picture:



a. What number is hidden behind the piglet?

Answer:

b. What number is hidden behind the penguin?

Answer:

9) A doll costs 7 euros and 90 cents.

Three friends have this much money:

Vittoria	
Lucrezia	
Sara	

Complete the sentence:

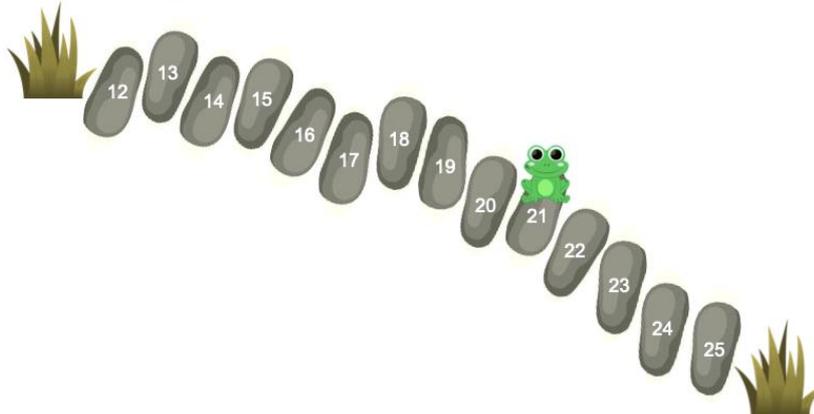
One of the three friends can't buy the doll: it's

- 10) Today the school cafeteria is serving pizza and French fries for lunch.
The cook made:

	1 pizza for every 7 children
	5 French fries for every child

There are 35 children in the cafeteria.

- a. How many pizzas did the cook make?
A. 12
B. 5
C. 7
- b. How many French fries did the cook have to make?
A. more than 170
B. fewer than 150
C. 165
- 11) A frog is hopping from stone to stone along a path.
Each stone is numbered as shown in the picture.
Look where the frog is now.



- a. The frog hopped 7 times to get there.
What stone was she on before hopping 7 times?
Answer: she was on stone No.
- b. Complete the sentence:
If the frog had been on stone No. 25, she would have had to hop
times to return to stone No. 13.

- 12) If you add 4 units and 2 tens to the number four hundred and thirty, you get:
A. 454
B. 472
C. 436

- 13) Julia's birthday is January 29 and her friend Alexandra's is exactly 1 week later.

Gennaio 2019						
Lunedì	Martedì	Mercoledì	Giovedì	Venerdì	Sabato	Domenica
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

- a. When is Alexandra's birthday?
A. January 22
B. February 2
C. February 5
- b. Alexandra's sister celebrates her birthday exactly three weeks before Julia.
What day of the week did Alexandra's sister's birthday fall on in 2019?
A. Monday
B. Tuesday
C. Wednesday

C.3 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

Appendix D

Methodological guidelines of the activities for the teacher

D.1 Activity 1 - Thousandville: The City Gets Bigger

Lesson Plan (methodological guidelines for the teacher)



Thematic unit: Numbers

Level: Primary school (3rd Grade)

Average time: 7 hours

Concepts

- Base-ten natural number system
- Writing natural numbers
- Place value in centesimal notation
- Comparing and ordering natural numbers
- Estimates and quantities

The lesson plan provides methodological guidelines for each stage of work.

The description of each stage is followed by the worksheet with the activities covered in it.

STAGE 0: Preliminaries and treasure hunt

Method: Group work

Time: A few minutes (around 10 minutes in all)

Materials needed:

- 120 bottle caps
- 500 drinking straws
- 100 small buttons, 50 medium buttons, 20 large buttons
- 1 container for each group of children

Classroom preparation

- Before starting the activity, the teacher hides piles of bottle caps, straws and buttons around the classroom. The teacher then divides the children into groups mixed by gender and aptitude level. The desks are arranged so that each group has a station, with a container for collecting the objects.

Description of activity

- The activity starts by reading the first part of the story:

Reesykle, the mayor of Thousandville, wants to make his city bigger. To do this, he has to make a model showing the plan for the new part of Thousandville. The model will be very large and will be made out of bottle caps, straws and buttons. Reesykle needs many helpers to make the model.

- Then proceed to the “treasure hunt”: in 3 minutes, the children go around the classroom, collecting the required objects and putting them in the container assigned to their group. *Do not let the children count the collected objects* (if any of them try to count, the teacher should tell them not to).

STAGE 1: Narration, estimation and counting

Method: Group work, class discussion

Time: Around one hour (do not take too long over this stage)

Materials used:

- Objects collected in the container
- 2 additional containers for each group of children
- 3 colored cards (in three different colors) for each group

DESCRIPTION OF ACTIVITY

- The activity resumes with a reading of the second part of the **story** (in which Reesykle gives additional information to the children):

Reesykle has something else to tell us: “I’m such an airhead! I forgot to tell you how many objects we have to use! We need 100 bottle caps, 500 straws and 1000 buttons. Do you have everything we need for the model?”

Are we sure we managed to collect all the material we need to make the model of Thousandville with Mayor Reesykle?

- If necessary, the children should be told again *not to count the collected objects yet*. Before asking the children to count the objects on the desk, the teacher should ask the children a number of stimulus **questions** (it is not necessary to ask all of the questions listed below, which are provided only as suggestions):
 1. ***How can we figure out how many objects of each type we’ve collected?***
OBJECTIVE: Give the groups two more containers and see how they divide the objects
Some children will propose dividing the collected objects by type. At this point, the teacher will give them two more containers so that they can collect objects by type. The teacher should not suggest that the three materials be divided into the three boxes, but should wait for the children to do so themselves. If they do not, it will be necessary to lead them to this solution via class discussion.
 2. ***Which container do you think has the most objects in it?***
OBJECTIVE: Quantity-dimension of the objects in the box [...]
 3. ***Without counting them, how many straws do you think you’ve collected? How many bottle caps? And how many buttons?***
OBJECTIVE: Rough estimates and concept of estimation [...]
 4. ***What methods would you use to count the objects in a container quickly?***
OBJECTIVE: Different counting strategies [...]
 5. ***Now use your methods to count exactly how many objects of each kind you’ve collected.***
OBJECTIVE: Counting [...]

This stage involves working on the concepts of estimation and quantity. The first three questions are intended to stimulate the children’s capacity to estimate/picture quantities. The

last question provides the lead-in to the next stage, where the class identifies the best strategies for counting large quantities.

- After a suitable length of time, the students are asked to give the results of their count. Each group will be given **3 cards** in different colors, one for each type of object. The children will write the number of objects they counted on each card (for example, as shown in the figure, the number of straws on the yellow card, the number of bottle caps on the red card, and the number of buttons on the green card).

Group A # straws	Group A # buttons	Group A # bottle caps
---------------------	----------------------	--------------------------

It is recommended that each group's cards be kept, so that the same situation as regards the materials can be replicated in the next session.

Guidelines for the class discussion:

The teacher:

- Starts the discussion by asking questions to draw attention to the differences between the students' choices, fostering balanced participation between girls and boys and children of different backgrounds;
- Asks more questions in order to discuss the choices;
- Reinforces good contributions by the children by means of approving looks, gestures, words, tone of voice and facial expression;
- Writes all of the answers on the blackboard or a poster;
- Encourages peer interaction and exchanges of views about different approaches, paying attention to the sensitive and multimodal aspects of understanding to promote the construction of mathematical meaning (e.g., by making use of sketches, turns of phrase introduced by the children, as well as their errors, silences, facial expressions and so forth);
- Shows willingness to listen to everyone in the class, aware of their diversity, without expressing judgements such as right/wrong, correct/incorrect;
- Stimulates the discussion to reach a consensus about the reasoning and strategies that can help answer the questions, paying particular attention to the strategies that are most effective in dealing with multiple-choice questions (for example, pointing out that not answering can be even more counterproductive than just guessing at an answer).

STAGE 2: Counting strategies

Method: Class discussion

Time: Around half an hour

Materials used:

- Objects collected in the containers
- Cards

The class discussion now focuses the children's attention on the number found by counting the objects in each container (written on the card) and on their initial estimates (written on the blackboard). [...]

	TAPPI	CANNUCCIE	BUTTONI
A	18 13	60 40	40 26
B	25 36	45 42	36 20
C	44 30	16 20	66 40
D	58 30	163 60	51 28
E	30 26	110 89	9 10
F	28 20	70 32	15 10

Figure 1. The teacher can write the estimates (shown in green in the photo) made by each group on the blackboard, and then the number which was counted (in white) in order to address the idea of estimation in the class discussion.

STAGE 3: Place value

Method: Group work, individual work

Time: Around 2 hours (approximately 1 hour for each worksheet, on separate occasions if necessary)

Materials needed:

- Worksheet 1A (group work)
- Worksheet 1B (individual work)
- Colored cards used in the previous stage

[...]

STAGE 4: Ordering on the number line

Method: Class discussion, group work

Time: Under two hours

Materials needed:

- String (3 pieces approximately 3 meters long each)
- Masking tape
- Colored cards marked with the numbers counted in the previous stage
- Flags to be placed on the target numbers
- Sufficient space (in the classroom, hall or gym)
- Worksheet 2

In this stage, the children must place their cards on lines marked on the floor. It is thus advisable to find a place that offers sufficient space for this activity. If the classroom is large enough, the desks can be moved to the sides and the three lines placed at a certain distance from them in the center of the room. Otherwise, the activity can be performed in the hall or gym. [...]

STAGE 5: The value of 1000

Method: Class discussion, group work, individual work

Time: Around one hour

Materials:

- Worksheet 3 (group work)
- Worksheet 4 (individual work)
- Collected buttons
- New card or the new value of the buttons to be placed on the number line
- *Guidelines:* This stage is designed to make the children think about the difference between counting the number of objects and calculating their actual value (as is the case with coins, for example).

[...]

D.2 Activity 2 - Forest Elves

Lesson Plan (methodological guidelines for the teacher)



Thematic unit: Numbers

Level: Primary school (3rd Grade)

Average time: 8 hours

Concepts:

- Number as measure
- Multiplicative reasoning
- Use of tables and the number line

The lesson plan provides methodological guidelines for each stage of work.
The description of each stage is followed by the worksheet with the activities covered in it.

STAGE 1: Narration and drawing

Method: Individual work, class discussion

Time: 1 ½ hours

Materials:

- Worksheet 1

DESCRIPTION OF ACTIVITY

- Hand out **worksheet 1** (individual).
- **Read** the worksheet (the worksheet should be read out loud, either by the teacher or a student)

Once upon a time, a family of forest elves lived in a house in the woods. The family was made up of Mummy Elf, Daddy Elf and their two children.

It was autumn, time to start gathering provisions for the long cold winter ahead.

The first to go out was Elf Girl. She left the house with her basket and went down the path. She took twenty steps towards the mountain and reached an apple tree. She filled her basket with apples and went back home.

Then Elf Boy left the house, with his basket. He went down the path towards the mountains, took twenty steps and reached a chestnut tree. He gathered chestnuts until his basket was full and went home.

A bit later, Mummy Elf came out of the house carrying an empty bucket. She went down the path towards the lake, took twenty steps and reached the pump. She filled

- Before proceeding to individual work with the worksheet, it is advisable to ask the children to **repeat the content of the story** to make sure they have a firm grasp of its basic narrative (the elves walk along the path, and each takes 20 steps) and can thus picture it clearly to themselves. This will prevent them from representing steps that do not follow the path, which would prevent the exercise from reaching its goal.
[...]

Guidelines for the class discussion:

The teacher monitors the children's work, moving around the classroom to see what kinds of representation are being used and organize the class discussion. When all the students have finished (those who finish very early can color their drawings), the class discussion begins, directed by the teacher.

The teacher:

- Starts the discussion by asking questions to draw attention to the different possible choices (for example: *How did you draw the elves' routes? Where did you put the apple tree? And the chestnut tree?*), fostering balanced participation between girls and boys and children of different backgrounds;
- Asks more questions in order to discuss the choices;
- The end goal of the discussion is to reach a consensus about a representation that effectively captures the routes taken by the characters in the story and the points they reach. It can be helpful to draw this consensus representation on the blackboard.

Figure 1. On the blackboard, the teacher or one of the students can draw the different strategies used by the children to represent the routes, and the consensus representation chosen by the class at the end of the discussion.



STAGE 2: The length of the steps

Method: group work (groups mixed by gender and aptitude level), class discussion

Time: Around two hours

Materials:

- Reference worksheet
- Worksheet 2a
- Worksheet 2b

[...]

STAGE 3: New relationships and representations

Method: Group work (groups mixed by gender and aptitude level: the same as in the previous stage), class discussion, use of teaching aids

Time: Around two and a half hours

Materials:

- **Drinking straws** of different colors, cut into pieces whose length is proportional to that of the elves' steps, e.g.:
 - o Four 12 cm pieces (Daddy Elf)
 - o Six 8 cm pieces (Mummy Elf)
 - o Eight 6 cm pieces (Elf Boy)
 - o Twelve 4 cm pieces (Elf Girl)
- Worksheet 3a
- Worksheet 3b

DESCRIPTION OF ACTIVITY

- Hand out **worksheet 3a** (individual work).
The worksheet requires the children to recognize a link between the steps taken by Mummy Elf and Daddy Elf, comparing different representations shown on the worksheet, which starts with a written description of this link. The teacher can direct a short class discussion of the children's answers, talking about the relationships shown by the different representations and then converging on the correct one.
- Afterwards, the teacher can draw a table on the blackboard showing how the number of steps taken by each character relates to the number taken by the others. In particular, the teacher can start by asking: "*If Daddy Elf reaches a place in 2 steps, how many steps will Mummy Elf have to take to reach the same place? And Elf Boy? Elf Girl?*"
- Hand out **worksheet 3b** (group work).
For this activity, which is the most complex in the entire sequence:
 - Drinking straws cut into different lengths can be used to help represent the elves' steps during the group work.
 - It is also useful to employ concrete perceptual experiences (for example, reproducing the elves' steps by having two children and/or the teacher walk) to represent the paired relationships between the elves' steps that are to be compared on the worksheet.

Stage 4: Let's all go to Uncle and Aunt Elf's house!

Method: Class discussion

Time: Around 2 hours

Materials:

- Worksheet 4
- Large roll of graph paper
- Pictures of the characters and the points they reach

DESCRIPTION OF ACTIVITY

- Hand out **worksheet 4** "Let's all go to Uncle and Aunt Elf's house!", which contains a new piece of the story (may be read together)
- The teacher can give the children time to work individually on the worksheet, or proceed directly to producing a consensus representation using the roll of graph paper to make a poster.

Guidelines:

The poster shows only the elves' house, the path, the lake and the mountains (as on worksheet 1). The objective is to add all the other places reached by the elves: the apple tree, the chestnut tree, the pump, the market and Uncle and Aunt Elf's house. To do so, it is necessary to establish the length of the four elves' steps, knowing that (for example) Elf Girl's step is 1 square (1 cm) long. In this case, then, Elf Boy's step will be one and a half squares long, Mummy Elf's 2 squares long, and Daddy Elf's 3 squares long. Once the units have been found, the positions of the apple tree, the chestnut tree, the pump, the market and Uncle and Aunt Elf's house can be found. The poster can also be used to illustrate and understand the relationships between the different characters' steps, and the number of steps each character has to take to reach a given point on the map, as covered in the previous worksheets.

In this final stage, the class shares the discoveries made in following the story, and the children are normally highly involved in making the poster: they can color and paste on the characters and places mentioned in the story in the appropriate points.



Figure 2. An example of a poster made by a class, with the routes and the places mentioned in the story