Economics of Education Review 100 (2024) 102538

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Tackling the gender gap in mathematics with active learning methodologies☆

Maria Laura Di Tommaso [a], Dalit Contini [b,*], Dalila De Rosa [c], Francesca Ferrara [b], Daniela Piazzalunga [d], Ornella Robutti [b]

[a] University of Torino, Collegio Carlo Alberto, and Frisch Center for Economic Research, Italy
[b] University of Torino, Italy
[c] Italian Ministry of Economy and Finance, Department of Finance, Italy
[d] University of Trento, IZA, and FBK-IRVAPP, Italy

ARTICLE INFO

JEL codes:
I21
I24
J16
C93

Keywords:
Gender gap
Mathematics
Child development
Teaching methodologies
Randomised controlled trial

ABSTRACT

Gender gaps in mathematics are at the root of gender differences in human capital accumulation, but the role of teaching practices on such gaps has been underinvestigated. We implement a teaching methodology to improve children's mathematical skills and evaluate the causal effect of the intervention on the gender gap in mathematics in Italy with a randomised controlled trial. The methodology, grounded in active and cooperative learning, focuses on peer interaction, sharing of ideas, learning from mistakes, and problem-solving. The treatment significantly improves girls' math performance (0.14 standard deviations), with no impact on boys, and reduces the math gender gap by about 40 %. The effect is stronger for girls with high pre-test scores.

1. Introduction

Over the past decades, the traditional female disadvantage in education has disappeared and turned into an advantage in most subjects. International learning assessments nonetheless indicate that girls still lag behind boys in mathematics in most countries (OECD, 2019; Mullis et al., 2016). According to the latest PISA survey with a specific focus on mathematics (PISA-2022), the math competency at age 15 was on average 0.09 standard deviations greater for boys than for girls, albeit with considerable country variation (OECD, 2023).

The gender gap in mathematics is especially large at the top of the performance distribution (Fryer & Levitt, 2010; Ellison & Swanson, 2010; Ellison & Swanson, 2023). This difference, and in particular the girls' comparative disadvantage in mathematics with respect to verbal skills, seems to be one of the factors contributing to explain why women are less likely than men to choose STEM majors at university (OECD, 2019; Turner & Bowen, 1999; Delaney & Devereux, 2019). Gender imbalance in academic studies then translates into gender-based disparities in occupational choices and in human capital accumulation. Women are still underrepresented in the most productive sectors of the economy and in high-paying occupations, often in STEM fields, with long-term effects on gender differences in wages and wealth (Paglin & Rufolo, 1990; Machin & Puhani, 2003; Black et al., 2008; Piazzalunga, 2018; Francesconi & Parey, 2018; Sierminska et al., 2019; Card & Payne, 2021). Moreover, recent research underlines the importance of mathematical skills even for non-STEM degrees and occupations and suggests that the gender gap in numeracy among adults contributes to the gender wage gap (Grinis, 2019; Delaney & Devereux, 2020; Battisti et al., 2023).

A wide range of social and cultural factors contribute to the math gender gap, which is, in fact, narrower in countries with better gender equality (Guiso et al., 2008; Pope & Syndor, 2010; Nollenberger et al., 2016; Lippman & Senik, 2018; Gevrek et al., 2020). Several studies show that the gender gap in math is highly related with parents' and teachers' attitudes and stereotypes (Alan et al., 2018; Carlana, 2019; Dossi et al., 2021; Nicoletti et al., 2022), and is subject to the influence of role

☆ The views and opinions expressed in this article are those of the authors and do not necessarily reflect the positions of the institutions Dalila De Rosa represents.

* Corresponding author: University of Torino, Italy.
E-mail address: dalit.contini@unito.it (D. Contini).

https://doi.org/10.1016/j.econedurev.2024.102538
Received 13 July 2023; Received in revised form 17 February 2024; Accepted 2 April 2024
Available online 15 April 2024

models (Dee, 2007; Paredes, 2014; Coenen et al., 2018). Such forces can erode girls' sense of self-confidence and self-efficacy and increase their anxiety about doing math (Ho et al., 2000; OECD, 2015; Sansone, 2017; Di Tommaso et al., 2021). Other studies highlight the role of competition, showing that gender differences in math skills among high-achieving students can be explained by gender differences in self-confidence and in attitudes toward competition (Gneezy et al., 2003; Niederle & Vesterlund, 2010; Ellison & Swanson, 2010) and that the gender gap in the willingness to compete can be addressed by specific interventions (Alan & Ertac, 2019).

A largely unexplored factor in the math gender gap is the way mathematics is taught to children. Qualitative research suggests that when the teaching methodology is problem-solving oriented and the students are engaged in discussions and investigative learning activities in low-competition environments, the math gender gap narrows and can even disappear (Boaler & Greeno, 2000; Boaler, 2002a; Boaler, 2002b; Zohar & Sela, 2003; Boaler, 2009; OECD, 2016).

The literature provides many quantitative empirical studies of the effectiveness of programs aimed at improving children's mathematical skills. Specific attention has been paid to interventions designed to change daily teaching practices with active and cooperative learning approaches, classroom management, and motivation programs. Rigorous evaluations show that these programs generally improve student achievement overall (see the meta-analysis by Slavin & Lake, 2008). Yet, a small literature in economics finds sizable positive impacts of traditional teaching practices, such as lecturing and rote memorisation, on test scores (Lavy, 2016; Schwerdt & Wuppermann, 2011; Berlinski & Busso, 2017). Bietenbeck (2014) contends that traditional and modern teaching practices promote different cognitive skills in students. Traditional teaching practices increase students' factual knowledge and their competency in solving routine problems, while modern teaching practices foster reasoning skills. Positive evidence of the effectiveness of cooperative learning also on literacy outcomes is documented in Puzio and Colby (2013)).

Despite the interest in the role of teaching practices on children's learning, there appear to be no quantitative investigations to establish the effectiveness of active learning practices in mitigating the gender gap in mathematics. Our paper fills this gap. This study set out to implement and assess a mathematics teaching program based on active and cooperative learning aimed at improving children's mathematical skills in Italian primary school. We evaluate the program's impact with a randomised controlled trial (RCT). To the best of our knowledge, this is the first attempt to investigate the causal impact of a teaching methodology on the gender gap in mathematics.

Our approach to teaching mathematics is based on the "Mathematics Laboratory" ("*Laboratorio di matematica*"), a math teaching methodology developed by math education scholars in the early 2000s in Italy (Anichini et al., 2004). The basic building block of this approach is the active involvement of the children, who are engaged in individual and peer work in a collaborative and non-competitive environment. Children are encouraged to frame problems and to attempt to solve them by sharing and comparing ideas within small groups and in-class discussions. Mistakes are welcome and considered a crucial means to understanding. This approach can be classified within the broad family of active learning teaching styles, the central idea of which is that learning involves active participation on the part of the learner (Lave & Wenger, 1991).

The absence of pressure and competition and a positive attitude toward mistakes should especially benefit girls (Bohnet, 2016; Boaler, 2016; Sansone, 2017). In addition, many activities use a narrative context, which is generally attractive to girls (OECD, 2019), and inclusive and gender-balanced participation is supported. For all these reasons, we believe the methodology has the potential to improve girls' learning and reduce the gender gap in mathematics. In what follows, we refer to the intervention implemented for the purpose of this study as the "Math Active Learning" (MATL) program.

The MATL programme consisted of 15 h of laboratory activities delivered to grade 3 children over five consecutive weeks in the spring of 2019. We focus on third grade to intervene as early as possible, as there is strong evidence that the gap already exists in second grade and then steadily increases throughout primary and secondary school (Contini et al., 2017). This allows us to use as a baseline the national test administered at the end of second grade, when the gap is first detected.

Each school in the province of Torino was invited to choose at least two of its third-grade classes to apply for the program. We then randomly selected 25 of the schools that applied and randomly assigned one of each of those schools' classes to the treated group and the other to the control group. The final sample consisted of 1044 children, with 519 children in the treatment group and 525 children in the control group.

The laboratory activities were delivered by external instructors, postgraduate students in mathematics education, all of whom happened to be female. The intervention did not provide additional math lessons, but replaced the regular lessons with MATL activities, while children in the control classes followed the usual curriculum.

To assess the impact of MATL on the children's performance, we administered math tests one month before the intervention (pre-test) and one month after the intervention (post-test). External supervisors involved in the design of the national assessment test regularly administered to all children in school at given grades (INVALSI) helped developing the tests, which had a conceptual framework and structure in line with the national one.

Italy is of particular interest for two reasons. First, it had the highest gender gap among the 57 countries participating in TIMSS 4th grade test (Mullis et al., 2016) and the largest gender gap among OECD countries in the PISA test administered to 15-year-old students for the year 2022 (OECD, 2023). Second, Italian teachers show the strongest preference for a teacher-centred approach over a student-centred approach, as shown in the teaching and learning international survey TALIS-2008 (OECD, 2009).

The findings from the impact evaluation of the MATL program are encouraging. The MATL program increased girls' math achievement by 0.14 standard deviations, without hampering boys' performance. Since this is a short-lived intervention, this effect should be considered quite large in magnitude and thus it is highly policy relevant. We also evaluate how the impact of the MATL program varies with prior ability. We find that the treatment has no effect on boys irrespective of their starting level, and that the girls benefitting most from the treatment are those with above-average pre-test scores. Overall, the intervention led to an over 40 % reduction in the math gender gap.

We then explore the potential channels through which the program might have improved girls' math skills. We analyse whether girls improved particularly in specific cognitive dimensions or in given types of questions, and whether the program contributed to changing attitudes toward mathematics. We also analyse whether the increase in test scores is driven by a lower propensity to leave questions unanswered, perhaps due to increased self-confidence. However, the success of the intervention does not seem to be driven by any of these channels. This leads us to infer that, because of its specific features, MATL worked by directly improving girls' general math skills.

Finally, we discuss and rule out alternative mechanisms other than the methodology that might explain the positive estimates of the program effects and threaten the validity of our findings. Potential mechanisms relate to the characteristics of those who delivered the intervention, their awareness of the gender perspective, and the design of the test. We discuss each of them in details and conclude that they do not invalidate our results.

Our paper is the first to evaluate the causal effect of a teaching methodology on the gender gap in mathematics, when delivered by instructors who understand the rationale behind the teaching practices and how they should be implemented. The positive results obtained in reducing the gender gap open the field for scaling up the intervention through direct teachers training.

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 the data and estimation strategy. Results are presented in Section 4, while we explore some potential channels that might explain the results in Section 5. Alternative mechanisms are discussed in Section 6. Evidence on medium-term impacts of the intervention is presented in Section 7. We discuss external validity in Section 8 and conclude in Section 9.

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, with less than 7 % of children attending private primary school. Families can choose between two schedules: a 40-hour school week, where children spend the whole day at school, or a more concentrated 27/30-hour week.[1] Curricula and learning targets are set at the national level and are the same for both schedules, but teachers are completely free to choose the teaching methods they feel are best. Each class typically has two or three generalist teachers who cover all the subjects between them (with the occasional exception of specialist teachers for foreign languages, gymnastics, and music). Didactic continuity is highly prized in the Italian school system. Children are assigned to a class that then remains the same for all five years of primary school and are normally taught by the same teachers. Primary school teachers receive training enabling them to teach all subjects,[2] although they often specialize in specific disciplines. However, once they have started teaching certain subjects to a class, they continue to teach those subjects to those students for the entire five-year cycle. The school year starts in early September and finishes in mid-June.

In primary school, math instruction covers the domains of numeracy, relations, data and predictions, space and figures. National curricular guidelines recommend providing instruction in the different domains throughout the entire school year. In third grade, when the MATL intervention was delivered, math instruction is usually offered 6 to 8 h a week.

2.2. The MATL intervention

2.2.1. Features of the MATL program

Educational research generally identifies two broad models in the teaching and learning paradigm: teacher-centred and learner-centred. The first conceives of teaching as a top-down activity and focuses on direct transmission of knowledge. In this view, the teacher's 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 views students as active participants in the process of learning. More value is attached to the development of thinking and reasoning processes than to the acquisition of specific knowledge (Staub & Stern, 2002). Students should become capable of developing solutions to problems on their own (Gutierrez & Boero, 2006).

Our intervention consists in classroom-based activities aimed at improving children's mathematical understanding and is based on the learner-centred approach, according to which knowledge cannot be directly imparted to students, but rather students and teachers work together to build competences through active-learning. The goal of teaching is to provide experiences that facilitate the construction of knowledge (Thompson, 2014). Another pillar of the approach is adherence to the theory that skill is malleable, intelligence can be learned, and the brain can grow through exercise (Dweck, 2006; Boaler, 2013).

More specifically, the MATL intervention builds on the "*Laboratorio di matematica*", 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 & Robutti, 2008, 2010; Arzarello et al., 2012; Ferrara & Ferrari, 2020).

The basic components of the MATL program can be summarized as follows:

(i) *Active learning*. Focusing on problem framing and problem-solving as opposed to procedural work, the approach reverses the traditional teacher-centred instruction by putting children at the centre 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 analyse the problem, explore different solutions, share and compare ideas, avoiding pressure and competition.
(iv) *Learning from mistakes*. Mistakes are seen as a 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 are engaged with materials (caps, straws, buttons of different size, boxes, cards…) that they manipulate with their hands and move around physically, as perceptual-motor learning has been proven to be effective in improving mathematics understanding (Nemirovsky et al., 2004).

Each of these components aims at activating children's thinking and helping them construct mathematical meanings through self-reflection and interaction with the teacher and their peers. The different activities take place within a collaborative and non-competitive environment, where the teacher – the instructor, in our case – has the role of "orchestrating" the classroom activities.

MATL focuses on the subject area of numeracy, recognized as the most fundamental domain in the math field at this age and because we found that the math gender gap is highest in this domain.[3, 4]

2.2.2. 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 (2006, 2007) fixed mindset messages prevail among students across the entire achievement distribution,

[1] The share of schools delivering a 40-hour schedule is much higher in the northern regions.

[2] Qualifying as a primary school teacher now requires a university degree in primary school education. Before 2001, a specific high school diploma (*Istituto magistrale*) was required.

[3] For further details, see Ferrara et al. (2021).

[4] In our experiment, the MATL program was implemented using two activities. In the first, named *Thousandville*, children must increase the size of a city without changing the proportions of the different components. The learning processes involved are counting, performing arithmetic operations, estimating the order of magnitude, and 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, and discovering relations between quantities in terms of multiples and sub-multiples. Extracts from the methodological guidelines (English translation) are available as Online Appendix D. The full methodological guidelines are available in English (translation) or in Italian (original) upon request.

but high-achieving girls are especially damaged by fixed ability beliefs. Girls suffer most from the fixed ability concept that implies giving labels, like being or not being smart, or being good or not being good at math (Dweck, 2007). Sansone (2017) shows that girls in particular benefit from teachers who believe that all students can succeed (*growth mindset approach*).

The teaching practices embodied in the MATL intervention have the potential to reduce the gender gap in math for several reasons. First, the activities are meant to reduce pressure and competition. This should benefit girls, because girls are generally less competitive than boys (Niederle & Vesterlund, 2010, 2011); in competitive environments girls tend to develop more anxiety, and anxiety is detrimental to learning (OECD, 2015; Bohnet, 2016; Sansone, 2017). Second, the approach encourages a positive attitude to mistakes. Reframing mistakes as an opportunity to learn rather than as a sign of failure is particularly important for girls, because girls have been shown to be on average more risk-averse and afraid of giving the wrong answer (Bohnet, 2016). Moreover, girls might have a propensity for learning from mistakes through the development of constructive reasoning about 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 embed mathematical activities within a narrative context and girls are typically better than boys at reading comprehension and languages. Another factor that might contribute to girls' activation and empowerment is the explicit support in the MATL guidelines for balanced participation in class discussions.

2.2.3. Delivery of the MATL intervention

The MATL program is delivered to children in grade 3, when they are around 8 years old. The reasons for this choice were: (i) to tackle inequalities as early as possible and to contrast possible cumulative effects; (ii) to deliver the intervention at a time when the math gender gap already exists so we could observe gender differences before the intervention and analyse their (short-term) development;[5] (iii) to use the second grade assessment (the first to which children are exposed) as a baseline against which to compare our results.

The intervention was delivered by four young female external instructors with a background in mathematics education at Master level, who were specifically trained in the activities by the researchers in our team, while regular mathematics teachers (also female) remained in the classroom as observers.

Choosing who should deliver the educational interventions in experimental settings is never obvious. If the teachers implement the activities, it is difficult to disentangle the program's effectiveness from the adequacy of the teachers' training, especially in the case the program does not appear to have any effect (e.g., Berlinski & Busso, 2017). On the other hand, the involvement of external instructors ensures effective and homogeneous program implementation, although it increases the risk of not being able to identify a priori the effect of the program from instructors possibly being "better teachers". We chose to implement the intervention with external staff to assess whether the program has the potential to reduce the gender gap when it is well implemented, levelling the field to possible future interventions at scale and with teachers.[6]

MATL was delivered between February and April 2019. The intervention took place at the class level during school-time and during the usual math time and did not change the total amount of time devoted to math instruction. Each lab session lasted three hours and took place once a week for five consecutive weeks. The children were divided into small, heterogenous groups of mixed prior ability and gender. 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.

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 based on comments and suggestions from the instructors and the classroom teachers. This pilot also provided 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 analysed with item-response-theory (IRT) models and modified accordingly.[7]

3. Design, data, and estimation

3.1. Research design

We evaluate the effectiveness of the intervention by exploiting a randomised controlled trial research design. The intervention was designed for delivery 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 enrol 25 schools and 50 classes, for a total of approximately 1000–1200 pupils.

The timeline of the implementation of the RCT is outlined in Fig. A.1 in Appendix A. Enrolment in the project was on a voluntary basis. In March 2018, all of the public primary school principals in the province of Torino received an official letter by the Regional Board of Education[8] inviting them to a presentation about the project. To be eligible to participate in the project: (i) Schools had to apply with at least two classes, one to be randomised to the treatment group and the other to the control group.[9] (ii) Classes in the same school had to have different mathematics teachers, to limit the risk of spillover. (iii) Participating classes were not to be involved in other extra-curricular math projects in the same school year.

Thirty-one schools applied for the program, some with more than two classes. We excluded one school because of the eligibility criteria and randomly selected 25 schools among those remaining and the two participating classes (see Table A.1). We then randomly assigned one class from each school to the treatment group and the other to the control group.[10] The entire randomisation process was public and took place at the University of Torino in June 2018.

All the 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 and pre-analysis plan (PAP) were registered with the AEA RTC Registry on December 10, 2018, before the start of the intervention. This paper presents analyses on pre-specified outcomes, unless otherwise specified.

[5] According to the literature, the math gender gap is often observed at a very young age and increases as children grow older; in Italy, the gap is already apparent at the end of second grade (Contini et al. 2017), when children take their first standardized national achievement test (INVALSI).

[6] Section 6.1 discusses these points in context.

[7] A full description of the pilot study and of the IRT analysis are available from the authors upon request.

[8] The Regional Board of Education is the highest authority of scholastic management at the regional level.

[9] In Italy, parents have substantial leeway in choosing the children's school, but cannot choose the specific class or teachers.

[10] The sampling procedure was set before knowing how many schools and classes would apply for the project, and different rules were devised to deal with different numbers of applications. The details can be found in the pre-analysis plan registered with the AEA RCT Registry (Contini et al. 2018).

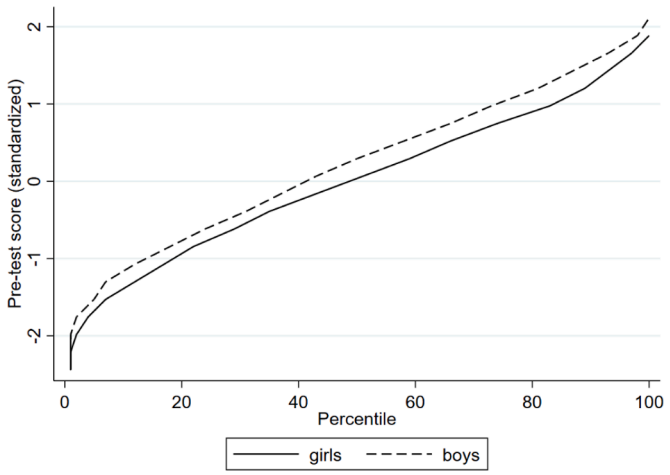

Fig. 1. Gender gap in the pre-test.
Notes: Children present at the pre-test (sample b), 933 observations.

3.2. Outcome measures and additional data

3.2.1. Outcome measures

The tests assessing children's math competencies before and after the treatment, designed by experts in mathematics education, followed the same conceptual framework as the INVALSI national assessment for the domain of numeracy.[11] 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.[12] Each test consists of 20 items, to be completed in 40 min.[13] The tests cover different topics and mathematical dimensions (knowing, arguing, and problem-solving), and use both multiple choice-type answers and open answers.[14]

The instructors in charge of the laboratories administered the pre- and post-tests in the classrooms and later graded them blindly under the supervision of an external examiner.[15] Correct answers are assigned 1 point each and incorrect and missing answers 0 points, for a total possible of raw scores between 0 and 20 points. The individual raw score is then standardized to have zero mean and standard deviation 1.

The post-test is the main outcome variable for assessing the effectiveness of the intervention. The pre-test is used to evaluate the gender gap before the intervention and to assess the balance between treated and control classes, and it is included as a control variable to improve the accuracy of the estimates. Fig. 1 shows the pre-test score distributions among girls and boys. On average, boys answered 11.23 items out of 20 correctly and girls 10.28; the difference is statistically significant and corresponds to 0.216 standard deviations (0.237 in 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 (0.171), but larger than the gap observed in the INVALSI tests in Piedmont (0.139) and Italy as a whole (0.100). Since children in the experimental classes perform substantially better on both the math and Italian INVALSI tests than children at the regional and national level, the larger gender gap in the former is consistent with the well-known fact that girls lag behind boys in math test scores particularly among high achievers. This suggests that our study has limited external validity, and this needs to be considered when thinking about scaling up.[16]

We also collected information about children's attitudes towards math, as a second outcome variable, to explore possible mechanisms underlying the effect of the treatment on cognitive abilities. Attitudes were 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.

3.2.2. Additional data

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

The schoolteachers provided information about children's special educational needs and disability (SEND), including any forms of learning difficulty, such as physical or mental disability, learning disorders, and attention disorders (ADHD).[17] The schools' administrative offices gave us information about parental education and migratory background. The instructors recorded absenteeism during the math labs for the children in the treated classes.

Data about the math teachers was collected via a brief questionnaire about gender, age, degree, experience overall and in the class, tenure, and type of contract. The instructors collected information about the class, including class size and the schedule (full time: 40 h per week, or normal: 27–30 h per week).

INVALSI provided data on math and language scores as well as class-level socio-economic background from the national assessment following grade 2. This data was 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.

3.3. Sample

No school or class dropped out of the project, so 25 primary schools participated in the project with two third-grade classes each, for a total of 50 classes, and 1044 children. Of the 1044 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 at both (sample d) (see Table A.3 in Appendix A).[18] The sample used for the impact evaluation is sample d.

3.4. Balance, attrition, and compliance

3.4.1. Balance at baseline

Tables 1 and 2 show the balance between the treated and control groups at the baseline, i.e., before treatment, and descriptive statistics of the outcome variable (post-test). Table 1 reports the mean values of the variables at the individual level and Table 2 reports class-level variables. The treated and control groups are well balanced for all characteristics, both at the overall level and by gender, indicating that the randomisation was successful. The only exception is for first-generation migrants among boys, with a small difference significant at 10 % (yet the proportion is very small: 0.004 in the control group and 0.020 in the treated group). 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 Fig. 2. From Table 2, it is also interesting

[11] For an overview of the INVALSI tests see INVALSI (2018).
[12] As mentioned above, the INVALSI grade 2 test was used as a reference to compare our results and to analyse external validity.
[13] The results of the pre- and post-tests were analysed with an IRT model and are available from the authors upon request.
[14] The English translation of the tests is available as Online Appendix C (C.1 and C.2).
[15] An expert in formulating and grading INVALSI tests.
[16] See also Section 8 on external validity.
[17] 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.
[18] 4 children are excluded from the analysis because they were present at the post-test, but did not answer any of the test items (probably due to very serious disability).

Table 1
Baseline characteristics of treated and control children, and post-test, full sample.

	All children			Girls			Boys		
Variable	Control	Treated	P-value of diff.	Control	Treated	P-value of diff.	Control	Treated	P-value of diff.
Girls	0.501	0.514	0.583						
SEND - broad definition	0.149	0.156	0.767	0.106	0.139	0.320	0.191	0.175	0.613
SEND - narrow definition	0.086	0.083	0.898	0.046	0.064	0.454	0.126	0.103	0.426
Native	0.848	0.877	0.429	0.848	0.884	0.395	0.847	0.869	0.591
Migrant I generation	0.011	0.021	0.253	0.019	0.022	0.778	0.004	0.020	0.085
Migrant II generation	0.128	0.096	0.339	0.114	0.086	0.472	0.141	0.107	0.345
Migrant status missing	0.013	0.006	0.465	0.019	0.007	0.395	0.008	0.004	0.672
Low-educated parents	0.670	0.724	0.382	0.669	0.723	0.447	0.672	0.726	0.394
High-educated parents	0.330	0.276	0.382	0.331	0.277	0.447	0.328	0.274	0.394
Parents' education missing	0.160	0.131	0.735	0.175	0.154	0.817	0.145	0.107	0.639
Observations	*525*	*519*	*1044*	*263*	267	*530*	*262*	*252*	*514*
Raw pre-test score	10.786	10.704	0.816	10.394	10.152	0.595	11.179	11.275	0.804
Observations	*481*	*452*	*933*	*241*	*230*	*471*	*240*	*222*	*462*
INVALSI math score [grade 2][a]	212.84	207.16	0.268	209.56	203.82	0.333	216.21	210.73	0.286
Observations	*474*	*465*	*939*	*245*	*236*	*481*	*229*	*229*	*458*
Teacher mark math [grade 2][a]	8.145	8.146	0.999	8.133	8.125	0.960	8.159	8.167	0.956
Observations	*461*	*474*	*935*	*234*	*240*	*474*	*227*	*234*	*461*
Raw post-test score (outcome)	9.842	10.355	0.083	9.1325	9.8175	0.080	10.566	10.924	0.353
Observations	*493*	*490*	*983*	*249*	*252*	*501*	*244*	*238*	*482*

Notes: Each row reports the mean of the control, the mean of the treated group, and the p-value of the difference, estimated from a regression of the variable shown in the first column on the treatment dummy, with standard errors clustered at class level.
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. 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).
[a] INVALSI data and teachers' mark in grade 2: INVALSI math test score refers to the test score on the national assessment in grade 2, and teachers marks at the end of the first semester of grade 2, released by INVALSI. The statistics refer to subsets of the full sample (a) for which data were available.

Table 2
Baseline characteristics of treated and control children, class-level variables.

Variable	Control	Treated	P-value of the difference
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*
INVALSI math score (mean) [grade 2][a]	212.80	208.20	0.426
INVALSI math score (s.d.) [grade 2][a]	33.50	31.91	0.512
Percent of children who attended childcare[a]	42.84	38.825	0.592
Percent of children who attended pre-primary educ.[a]	94.494	94.467	0.996
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
Female teacher	100.00	100.00	
Observations	*24*	*25*	*49*

Notes: Each row reports the mean of the control, the mean of the treated group, and the p-value of the difference, of class-level variables. "Childcare" refers to formal childcare for children aged 0–3 (observations: 20 control and 22 treated classes); "pre-primary education" refers to formal childcare for children aged 3–5 (observation: 20 control and 20 treated classes). 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.
[a] INVALSI data: INVALSI math test score refers to the test score on the national assessment in grade 2, released by INVALSI.

to note that all teachers are women.

3.4.2. Attrition

In this study, there are two relevant sources of attrition: absences at the post-test and absences at the pre-test, which matters because our identification strategy relies on controlling for pre-test scores. We measure both overall attrition and differential attrition for all children, and separately for boys and girls, and report attrition rates in Table A.4. 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 A.4 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 during the winter of 2019, during the peak flu season. 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

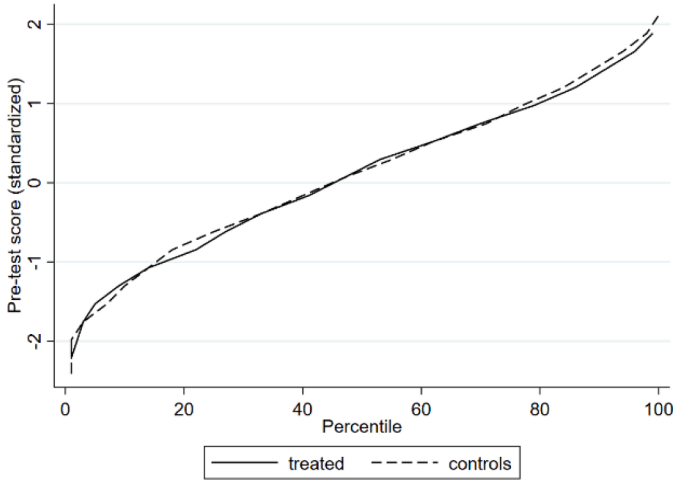

Fig. 2. Pre-test score distribution by treatment status.
Notes: Children present at the pre-test (sample b), 933 observations.

Table 3
Attendance of the laboratory sessions.

Percent of labs. attended	Percent of children	Percent of boys	Percent of girls
0 %	0.00	0.00	0.00
≥ 50 %	99.30	100.00	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 h. Sample (d) (children present at pre- and post-test).

attrition rates are small enough not to raise concern about the validity of the estimates of the intervention effect.[19]

We rerun balance checks for the sample of children who attended the post-test but not the pre-test (sample b)[20] and for the sample of children who were present at both tests (sample d - Table A.5). The treatment and control groups still appear to be well balanced after attrition, overall and by gender, and no substantial difference is found between the original and the analytical samples.

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 the treated and control groups (despite the favourable results of the attrition analysis).

3.4.3. Compliance and spillover effects

In this experiment, none of the children assigned to the control group took part in the program.[21] Children assigned to the treated classes, instead, were left untreated if they were absent on lab days. Noncompliance dilutes the treatment and yields underestimates of the average treatment effect (Bloom, 2008).

In Table 3, we report statistics on MATL participation. No children missed all the lab sessions, 99.3 % attended at least 50 % of the time, and 73.8 % attended all of the sessions, with a small difference in favour of boys. This may reduce the estimated impact on the math gender gap, yielding 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-year-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 modify their teaching practices in such a short space of time. The math teachers in the treated group were different from those in the control classes, and the intervention was delivered by external instructors with the treatment class teachers present as observers. While it is true that teachers may talk to each other, the methodological materials were released to teachers only a year after the project ended. If spillover did occur somehow, the treatment effect would be underestimated.

We cannot rule out the possibility that teachers in the treatment classes learned from observing the intervention and at least in part adopted the approach. However, this would not be problematic because our aim is to assess the total effect of the program, which consists of the direct effect of MATL on children's math achievement and the (potential) synergic indirect effect generated by the class teachers. Both channels are intended effects of the intervention.

3.5. Empirical strategy

3.5.1. Model

Our goal is to assess the impact of participation in the math laboratories on pupils' math skills, and more specifically on boys' and girls' outcomes. The successful randomisation into treated and control groups ensures that the two groups can be safely compared, without incurring 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 analyse 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 schedule). θ_s is a vector of school fixed effects (our randomisation 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 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 Eq. (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 post-test scores to evaluate the effect of the treatment on the math gender gap, 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. A better strategy consists in comparing the raw math gender gap in treated and control groups after treatment. Due to the successful randomisation, 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 math gender gap due to treatment.

3.5.2. Explanatory variables

In addition to pre-test scores, we control for gender, special education needs or disability (dummy variable) (SEND), migratory

[19] See the guidelines in WWC-What Works Clearinghouse (2013), which are based on an extensive simulation study.
[20] Available upon request from the author.
[21] No child switched class during the year.
[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.

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 assumes the value of 1 only for children with certified educational needs, and a broad version of the variable that assumes the value of 1 for all children reporting any kind of learning disorder/special needs, whether certified or merely demonstrated.

Family background variables included in models (1) and (2) above are defined in Table A.2. Parental education is denoted as "high education" if at least one parent has a tertiary degree, and 0 otherwise. The child's migratory background is coded as 3 dummy 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. To prevent the loss of numerous observations and to avoid self-selection issues, we include a dummy variable for each characteristic that is equal to 1 if the characteristic is missing.[23,24]

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.

3.5.3. Robustness checks

The main analytical sample includes only children who took both the pre- and the post-test. In a robustness check, we also include the children who were absent from the pre-test, identifying them with a dummy variable and assigning a zero value for the pre-test score. As for children absent from the post-test, we had scheduled a deferred session on a different date, as close as possible to the original one, and we use the resulting 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. It is not uncommon for children with mild problems not to have 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. Because of problems identifying children with severe problems that we were not aware of before going into the field, we decided to deviate from the original plan. We include all SEND children in the main specification, leaving the estimations without them as robustness checks.

4. Results

To evaluate the ITT impact of the intervention on math performance, we compare the post-test results of the treated and control groups, overall and by gender, as described in the previous section. In Section 4.1, we estimate the average impact on the entire group of participants, and on girls and on boys separately. In Section 4.2 we analyse whether the treatment has heterogeneous effects according to prior achievement and parental education. In Section 4.3, we describe the results of robustness checks.

4.1. Core results

Table 4 presents the main results.[27] We focus on our preferred sample, including the children who took both the pre- and the post-test, and control for school fixed effects and pre-test scores; the full model includes also controls for individual and family background characteristics, class size, and time schedule. Results indicate that MATL increases math test scores; as one may expect, the results on the treatment effect are quite stable across the two specifications. The overall effect (0.083 s.d.) is entirely attributable to the positive impact of the treatment on girls' skills (0.142 s.d. in the full model).[28,29,30] As mentioned above, these results can be interpreted as ITT. Considering as treated children who attended at least 80 % of the laboratories (94.19 % of children, see Table 3), the local average treatment effect for compliers (LATE) is very similar (0.152 for girls, −0.009 for boys, and 0.088 overall).[31] In case of one-sides partial compliance, as in our setup, this is equal to the ATT. Instead, there is little evidence that the program has any effect on the performance of boys.

The effect we observed for girls is quite large in magnitude for educational interventions. By means of comparison, Bloom et al. (2008) report that the average annual gain in math tests between grade 2 and 3 of primary school is 0.89 standard deviations. Bloom (2008) shows that decreasing class size by 10 children (from 22 to 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 math gender gap. In the control group, the gender gap in math is 0.324, while in the treated group it is 0.221, implying a reduction of 31.7 % in the treated group with respect to the control group. To account for differences in the pre-test, we compute the reduction in math gender gap 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 coefficient estimates from Eq. (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

[23] We were able to collect information about the teachers' characteristics in 49 out of 50 classes (one teacher refused to provide consent for data processing). To avoid losing an entire (control) class, we do not include teachers' characteristics in the estimations at the class level. Teachers' characteristics are used in the balance tests.

[24] We also estimated a model where imputing missing values, rather than using indicator variables signalling missing.

[25] During regular sessions, the instructors 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 classroom activities. These tests were then sent by mail to the research team. Of the 57 children absent from the post-test, 35 children took the deferred session. As it was impossible to have full control over this process, we chose 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 referrals for special education (Anderson 1997, Wehmeyer and Schwartz 2001). This finding supports the decision to also include SEND children in the analysis.

[27] Complete results are presented in Table A.6 in Appendix A.

[28] If we do not control for pre-test scores, the effect for girls is 0.131 ($p < 0.10$) and for boys 0.023 (not significant); thus, controlling for the pre-test scores only helps improving the estimates. Results are available from the authors upon request.

[29] In Appendix B, we present the main and the heterogeneous results using the latent ability estimated with IRT models as a dependent variable rather than the standardized test-score. The results are confirmed and are similar in magnitude.

[30] Using imputed missing values instead of dummy variables for missing control variables confirms the results. The results are also confirmed when we exclude all migrant boys or first-generation migrant boys, which leads to a slight imbalance between treated and control children. Both results are available on request from the authors.

[31] Results are available from the authors upon request.

Table 4
Main results: effects of the treatment.

	Post-test scores controlling for pre-test scores			Post-test scores controlling for pre-test, family background and class variables		
Variable	Overall (1)	Girls (2)	Boys (3)	Overall (4)	Girls (5)	Boys (6)
Treatment	0.076**	0.152***	−0.028	0.083**	0.142**	−0.009
	(0.030)	(0.053)	(0.045)	(0.033)	(0.055)	(0.046)
Pre-test score	0.763***	0.744***	0.784***	0.739***	0.737***	0.748***
	(0.023)	(0.033)	(0.026)	(0.025)	(0.035)	(0.033)
Girl	−0.090*			−0.097**		
	(0.048)			(0.047)		
Constant	−0.001	−0.091**	0.008	0.163	−0.194	0.290
	(0.065)	(0.038)	(0.109)	(0.157)	(0.225)	(0.249)
Chi2 (girls = boys)		5.57**			4.14**	
Observations	*888*	*448*	*440*	*888*	*448*	*440*
R-squared	0.611	0.599	0.630	0.616	0.603	0.641
School FE	YES	YES	YES	YES	YES	YES
Addit. controls				YES	YES	YES

Notes: Standardized test scores. Standard errors clustered at the class level in parentheses. Sample (d) (children present at the pre- and post-test). Additional controls include SEND (special education needs and disability) dummy broad definition (children with any form of special education needs or disability), parental education (high-educated parents: 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 4–6) are available in Table A.6. *** p <0.01, ** p < 0.05, * p < 0.1.

children. Secondly, we compare each counterfactual math gender gap with the corresponding observed value. The actual math gender gap for the control group is 0.324, and the counterfactual one for this group had they been treated is 0.170, implying a reduction of 47.5 %. The actual math gender gap for the treated group is 0.221, and the counterfactual one for this group had they not been treated is 0.369, implying a reduction of 40.1 %.

4.2. Heterogeneity in treatment effects

Table 5 describes the estimates of a model with an interaction term between treatment and prior achievement. We confirm that the intervention has little effect for boys, regardless of pre-test scores. Instead, we find that the treatment is more effective on well-performing girls. For each additional unit in standardized pre-test scores, the treatment effect increases by 0.127 post-test score units. We can appreciate how the treatment effect varies with pre-test scores and the corresponding confidence intervals by inspecting Fig. 3. For instance, the point estimate of the treatment effect is close to zero for girls whose pre-test scores are 1 standard deviation below the average, while for girls who are 2 standard deviations above the average, the treatment effect is around 0.4 (=0.155+2·0.127). The effect is statistically significant for girls with pre-test scores exceeding −0.2 s.d., which is slightly below the girls' average pre-test score (−0.09 s.d.).[32]

Table 5
Heterogeneous effects of the treatment by prior achievement level.

Variable	Overall (1)	Girls (2)	Boys (3)
Treatment	0.081**	0.155***	−0.013
	(0.033)	(0.053)	(0.048)
Pre-test score	0.719***	0.679***	0.735***
	(0.038)	(0.050)	(0.041)
Treatment* Pre-test score	0.062	0.127*	0.028
	(0.048)	(0.064)	(0.058)
Constant	0.139	−0.159	0.292
	(0.159)	(0.224)	(0.251)
Treatment: Chi2 (girls = boys)		5.05**	
Treatment*Pre-test score: Chi2 (girls = boys)		1.66	
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 parentheses. 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 (high-educated parents: 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.

We then analyse how treatment affects children with different parental education by including an interaction term between treatment and parental education for the overall sample and then separately for boys and girls.[33] The results are reported in Table 6.[34] Once again, we find no treatment effects for boys. Instead, we observe that in terms of point estimates, girls with low-educated parents benefit most from the treatment; however, the difference between girls with low and with high-educated parents is not statistically significant.

Overall, we observe that MATL labs improve the math skills of girls, and in particular, well-performing girls (and to some extent of girls with low-educated parents). These findings are not fully consistent with previous research. Two best-evidence review papers by Slavin and co-authors analysing the effect of different active and cooperative math learning interventions (Slavin & 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 attending lengthy active learning math programs. MATL is a short-term program, and we speculate that also the skills of low-performing girls might improve if the intervention were implemented over a longer period of time. More generally, further investigation is needed to shed light on why the intervention in the present forms is not capable of improving the performance of boys and less performing children.

4.3. Robustness checks

We replicate the main analyses on different samples. The results are

[32] As a robustness check, we replicated the analysis by interacting the treatment variable with pre-test quintiles instead of pre-test as 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 approximately linear.

[33] We define as "low education" situations where neither parent has tertiary education qualifications and as "high education" situations where at least one parent has a tertiary degree.

[34] Full estimates are available from the authors upon request.

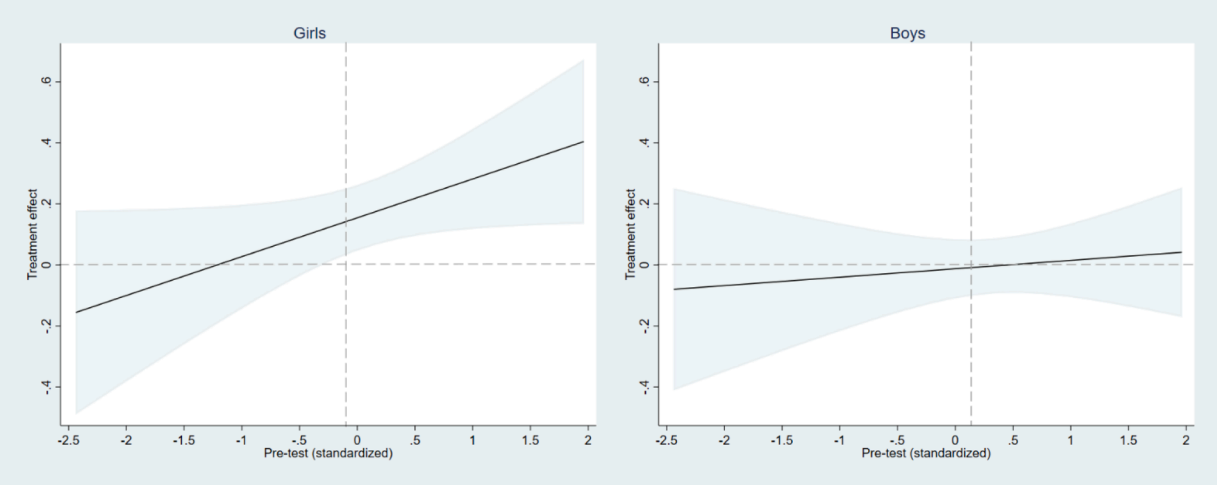

Fig. 3. Treatment effect by prior achievement levels.
Notes: Effect of the treatment by pre-test scores for boys and girls (estimates from regression in Table 4), with 95 % confidence intervals. Sample (d), 888 observations. The dashed horizontal line represents a zero-treatment effect, whereas the dashed vertical line represents the pre-test score mean for girls and boys respectively.

Table 6
Heterogeneous effects of the treatment by parents' education.

	Overall (1)	Girls (2)	Boys (3)
Treatment	0.060	0.182**	−0.075
	(0.051)	(0.072)	(0.068)
Treatment* high-educated parents	0.026	−0.099	0.119
	(0.096)	(0.133)	(0.148)
Observations	*888*	*448*	*440*
R-squared	0.616	0.604	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 parentheses. 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), migratory background (migrant I generation, II generation, information missing), class size and time schedule. The interaction between treatment and parents' education missing is also controlled for. Full results are available upon request.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

reported in Table 7. First, we exclude from the analysis 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 from the pre-test. Fourth, we include the children who were absent from the post-test but were given a post-test on a deferred date.[35] 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 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.[36]

We also perform a different type of robustness check. Given the possible presence of classical measurement error in the pre-test scores, we estimate the main model for girls and boys by adjusting for measurement error.[37] As no measure of reliability was available, we used a coefficient of 0.80 (which is considered a common measure in the literature and indicates a good value of reliability). The results are confirmed, although the effect for girls is slightly smaller. Results are available in the Appendix (Table A.7).

5. Possible channels

The MATL intervention has proven to be effective on girls. We now explore the potential channels through which the program might have improved girls' math skills. The program could improve abilities by increasing problem-solving competences, 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 in some dimensions but not others, or by improving children's skills in dealing with 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 according to whether the children underwent treatment

[35] In the pre-analysis plan (PAP) we had decided to: exclude SEND children; include the post-test taken in the deferred session; include children who were absent from the pre-test by marking them with a missing dummy. We subsequently decided to proceed differently in the core analysis, but the choices specified in the PAP are presented here as robustness checks.

[36] Using imputed missing values instead of dummy variables for missing pre-test scores confirms the results (available from the authors on request).

[37] We adjusted for classical measurement error in the pre-test control variable. We did not adjust for measurement error in the post-test scores, as this does not introduce bias into the estimates. The reliability coefficient is a function of measurement error and can be conceived as the correlation coefficient between the test scores in two different applications of the testing process (Livingston 2018).

Table 7
Robustness checks.

	Post-test scores excluding children with certified special educational needs or disabilities			Post-test scores excluding children with any special educational needs or disabilities			Post-test scores including pre-test score missing dummy			Post-test score including children sitting the post-test at deferred session		
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	0.093**	0.144***	0.008	0.111***	0.159***	0.017	0.110***	0.165***	0.035	0.074**	0.118**	−0.002
	(0.035)	(0.053)	(0.051)	(0.037)	(0.053)	(0.054)	(0.037)	(0.056)	(0.047)	(0.032)	(0.050)	(0.046)
Pre-test scores	0.764***	0.740***	0.771***	0.769***	0.734***	0.786***	0.733***	0.716***	0.731***	0.744***	0.737***	0.739***
	(0.027)	(0.036)	(0.033)	(0.026)	(0.034)	(0.034)	(0.029)	(0.037)	(0.034)	(0.026)	(0.035)	(0.033)
Pre-test sc. missing							−0.069	−0.195	0.078			
							(0.097)	(0.128)	(0.151)			
Constant	0.032	−0.228	0.092	0.090	−0.034	0.152	−0.012	−0.419	0.262	0.153	−0.055	0.242
	(0.174)	(0.213)	(0.309)	(0.159)	(0.194)	(0.338)	(0.185)	(0.261)	(0.234)	(0.152)	(0.204)	(0.271)
Chi2 (girls = boys)			3.42*			3.96**			3.51*		2.84*	
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 parentheses. 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 (high-educated parents: at least one parent with 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

or not. Finally, we analyse if treated children are more likely than controls not to leave some items blank. Apart from 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, by improving attitudes towards math, or by reducing the chances to leave questions unanswered. At the moment, this leads us to infer that MATL worked by directly improving girls' general math skills.

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

We analyse whether the treatment has a differential impact by item format, cognitive dimension, or level of difficulty of the single test 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.[38]

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 Eq. (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.

The results are reported in Table 8. These models were estimated separately for boys and girls, controlling for pre-test scores and school fixed effects.[39] For each group of items, we tested the equality of the treatment coefficients across item categories.[40]

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

[38] The classification is available upon request from the authors.

[39] Since the test-scores in this section are based on the answers to just a 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 for appropriate comparisons, the estimate of the treatment effect from the comparable all-items model is reported in the first panel of Table 8.

[40] As reported in Table 8, 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 level of the class. The results are very similar and available upon request.

Table 8
Treatment effect by type of item.

		Girls		Boys	
All items	Outcome	Treatm.	S.E.	Treatm.	S.E.
	Post-test score	0.152**	0.059	−0.028	0.061
DIFFICULTY	**Outcome**	**Treatm.**	**S.E.**	**Treatm.**	**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*	**Chi2**	*p*
	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.**	**S.E.**	**Treatm.**	**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*	**Chi2**	*p*
	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.**	**S.E.**	**Treatm.**	**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*	**Chi2**	*p*
	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 is available upon request. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

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.

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 their own abilities, and higher levels of anxiety and stress (Mullis et al., 2008; Else-Quest et al., 2010; Hill et al., 2016; OECD, 2016; Di Tommaso et al., 2021). Attitudes are a key factor to understanding 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.

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.[41] 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.

Table 9
Treatment effect on attitudes towards mathematics.

Variable	Attitudes (1)	Attitudes (2)
Girls	−0.750*	−0.831**
	(0.388)	(0.375)
Treatment effect on boys	−0.474	−0.477
	(0.301)	(0.298)
Treatment effect on girls	−0.495	−0.486
	(0.358)	(0.350)
Constant	16.500***	16.094***
	(0.222)	(0.555)
Observations	*882*	*882*
R-squared	0.053	0.072
School FE	YES	YES
Additional controls	NO	YES

Notes: Standard errors clustered at the class level in parentheses. 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 built as the sum of these 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Consistent with the existing literature, we observe a sizable gender gap in attitudes in favour of boys (Table A.8 in the Appendix). We do not find effects of the treatment on the attitudes of boys or girls (for both, there is a small negative effect, but the estimates are very imprecise and never statistically significant) (Table 9).[42]

We may conclude that the success of MATL on girls' math skills was not mediated by a positive change in their attitudes towards math. This was a surprising finding. However, if the concept of what mathematics is, is grounded on traditional teaching practices and already heavily rooted in children's minds, it may be difficult to change. This would be especially true for a short intervention delivered by an external teacher rather than by the child's familiar classroom teacher. Longer programs may have more of 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 tendency to leave questions unanswered.[43] 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).

We use two models to estimate the effect of MATL on the tendency to leave items blank: 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. In addition to the treatment variable, we include

[41] The English translation of the full questionnaire is available as Appendix C (C.3).

[42] We also perform the analysis using the first component delivered by principal component analysis as a dependent variable and obtain very similar results.

[43] Girls are more likely to leave omitted answers than boys in multiple-choice tests, even when omitted questions and wrong answer are given the same score – as in our setting (Iriberri and Rey-Biel 2021).

the usual controls, school fixed effects, and the corresponding blank item indicator in the pre-test. We find a negative and significant effect of the treatment on the number of non-response items (Table 10). 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 of leaving at least 2 items blank, the average marginal effect of the treatment is −0.082. Hence there is evidence that MATL is effective in reducing non-response, although the effect is small.

When analysing 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 math gender gap 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.[44] Overall, these results do not support the hypothesis that MATL improves girls' performance by reducing the tendency to leave questions unanswered and suggests that the observed change is due to a real improvement in girls' math skills.

6. Discussing alternative mechanisms

We now turn our attention to discussing alternative mechanisms that might explain the positive estimates of the program effects. Potential issues relate to the characteristics of those who delivered the intervention, their awareness of the gender perspective, and the design of the test. As we do not have data to test directly these alternative channels, we support our claims with the existing evidence that is available, where possible from multiple sources or systematic reviews.

6.1. Teacher characteristics and teacher quality

The intervention was delivered by young female college graduates with a degree in mathematics education. One might argue that the gains made may have resulted from having "better teachers", rather than from the teaching methodology itself. The main difficulty with this line of reasoning, however, is that it fails to explain why the intervention only affected girls. Also, despite their recent academic training, the instructors' hands-on experience working with children pales in comparison to that of the teachers' years – and sometimes decades – of classroom experience.[45]

Nevertheless, let us assume that the external instructors are much better than the teachers in our experimental classes (for example, are at the 95° percentile of the teacher quality distribution, as compared to the 50° percentile of the class teachers). We may elaborate on the estimates provided in the existing literature on teacher value-added (Chetty et al., 2014; Hanushek et al., 2019) according to which a one standard deviation improvement in the teacher effectiveness increases test scores by 0.14–0.15 s.d. in math per school year. Rescaling these figures for the time spent in class by our external instructors (15 h), we would end up with an increase in test scores due to being a better teacher of only approximately one eighth (0.019 s.d.) of the estimated effect of our intervention on girls (0.14 s.d.).[46]

A similar argument can be made if we considered the observed gains as due to a role model effect of the instructors delivering the intervention. Identifying the channels through which a role model can have a positive effect in this setting is not easy. The existing literature on role models in education has mainly focused on teacher/student gender, particularly in STEM related subjects. The idea is that students may perform better when assigned to a same-sex teacher if they identify themselves with such a role model (Paredes, 2014). This literature, however, does not apply to our case-study, because both the instructors and the regular teachers are females. Nevertheless, one could exploit the evidence from this literature to infer the size of the role model effect in our setting.[47] Different studies find a beneficial effect of having a same-gender teachers, but overall, the evidence is mixed and not conclusive, especially in primary school (Coenen et al., 2018; de Gendre et al., 2023). Altogether, the existing studies find an effect of at most 0.05 s.d. in a full school year (Dee, 2007; Sansone, 2017; Coenen et al., 2018; de Gendre et al., 2023). Rescaling this effect, we would find a very small figure for a short-term intervention like ours, i.e. 0.003 s.d.

In conclusion, we do not believe that the substantial increase in girls' math test scores can be attributed to the specific characteristics of the instructors. Even if we were to add up the role model effect and better quality of the instructors, they would explain a maximum of 0.022 s.d. (15 percent) of the 0.14 s.d. increase in girls' math skills.

6.2. Awareness of the goal of the intervention

A possible concern is related to the teachers' and instructors' awareness of the intervention's ultimate goal of reducing the gender gap in math. This raises the question as to whether the driving force of the intervention's effects lies in the teaching methodology itself or in the fact that the practitioners are more "gender aware".

In fact, the schools had to be 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 out by the regional authorities. Yet, the teachers of both the treated and the control classes were aware of the gender perspective, and there are no major reasons to expect a difference between the two groups. Moreover, the teachers, who were also asked not to reveal the goal of the project to the children, were not actively involved in the laboratories but were merely

[44] 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 of getting the answer right 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 −0.14 (meaning that treatment makes the number of blank items decrease by 0.14), even if the probability of giving the correct answer by chance was equal to 1 (obviously far from truth), we would end up with an increase of 0.14 correct answers (on a 20-item 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 greater.

[45] On average, teachers in our sample have 14 years of experience in teaching mathematics.

[46] According to Chetty et al. (2014), who provide robust estimates for the United States, a one standard deviation improvement in the teacher value added increases test scores by 0.14 s.d. in math per school year. Hanushek et al. (2019), find similar results at the international level: a one standard deviation increase in teacher cognitive skills is associated with 0.15 s.d. higher student performance in math (0.10 s.d. in reading). In Italy one school year corresponds to 200 days. A math teacher usually spends in class 6 hours per week doing math, or 240 hours per year. Hence, being exposed to a teacher in the 95° rather than in the 50° percentile of the distribution (a 2 standard deviations difference) could explain an increase in students' performance by about (2*0.15/240)*15=0.019 s.d..

[47] A role model effect may be exercised by tutors who are significantly younger than the average regular teacher. The perceived similarity of the role model to the self may have positive effects (Gladstone and Cimpian 2021), although, while it is true that younger teachers may be perceived as more similar (or less distant) from students than older regular teachers, this may not be the case for young children. Nevertheless, this can be considered as a possible alternative channel that may have contributed to the estimated positive effect of the intervention for girls.

Table 10
Treatment effect on blank items.

	OLS			LOGISTIC		
	Overall	Boys	Girls	Overall	Boys	Girls
Variables	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	−0.146**	−0.142*	−0.137*	0.284***	0.298***	0.223**
	(0.061)	(0.077)	(0.072)	(0.101)	(0.113)	(0.161)
Gender	0.008			0.799		
	(0.054)			(0.173)		
N. of blank items at pre-test	0.138***	0.146**	0.115***			
	(0.041)	(0.057)	(0.039)			
Pre-test score std.	−0.037	−0.028	−0.056	1.009	0.916	1.183
	(0.038)	(0.055)	(0.042)	(0.167)	(0.188)	(0.357)
At least 2 blank items pre-test				5.579***	3.955***	7.307***
				(1.650)	(1.741)	(4.749)
Constant	0.070	−0.260	0.441	0.043	0.257	0.000
	(0.243)	(0.282)	(0.369)	(0.114)	(0.636)	(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 parentheses. 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 (high-educated parents: at least one parent with 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

observing.

The instructors were also fully aware of the aim of their work. This was inevitable, as the inclusive participation of all children is a distinctive element of the program, which emphasizes the importance of conducting activities promoting the active participation of the entire class. To some extent, the instructors' awareness of the gender perspective of the program may have contributed to improving the girls' performance more than that of the boys.[48]

In this light, we acknowledge that the program has two elements that cannot be disentangled. Future work aimed at evaluating a scale-up of the intervention should consider implementation of two parallel programs: one like the current one, which combines the teaching methodology and gender awareness, and the other with only the teaching component. This would be challenging to implement, however, because it would require a deliberate decision to provide incomplete information about the program to the school boards and regional authorities endorsing the project.

6.3. Design of pre- and post-test

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 about the appropriateness of using assessments designed by the program developers, as such measures have been found to overstate program impacts (Pellegrini et al., 2018). Nonetheless, we believe that our results hold true. First, the tests were standardized and scored blindly by the instructors, leaving no room for conscious or unconscious bias in grading, either in terms of gender or experimental group. Second, they were conceived as comprehensive measures of abilities in numeracy. Even if there were some bias, we would expect it to influence the results of both boys and girls. Yet this is not the case in our experiment, where the results of the treated and control groups differ only for girls.

7. Medium-term impact

Here we present the analysis of medium-term effects of the intervention, which were not included in the original pre-analysis plan due to an initial lack of funding. Because of high attrition and of school closure due to the Covid-19 pandemic, what follows should be interpreted only as suggestive evidence.

Our intervention took place in the spring 2019 and the short-term impact was evaluated by administering a test to the children of the treatment and control groups approximately one month after the intervention. To assess the medium-term effects, we planned to carry out a new evaluation in spring 2020. However, schools in Italy remained closed until the end of the school year due to the Covid-19 pandemic. We therefore had to postpone the evaluation until October 2020, about 18 months after the intervention, when the students were just starting fifth grade. By then, the first wave of the pandemic was almost over, and schools had reopened, at least until the next surge shut them down again. Schools participating in the RCT were asked for permission to administer the test to all the children involved in the experiment. Just 14 of the original 25 schools participating agreed to take part in the medium-term evaluation. This substantial attrition may have been caused by the headmasters' concerns over letting outsiders onto school premises. We excluded three of the 14 schools from the statistical analysis because they only involved one class (treated or control). The final sample thus consists of 429 children.

Overall, the characteristics of the children in schools participating in the medium-term evaluation are comparable to those in schools that opted out. Treated and control children are still comparable in terms of pre-test scores and socio-demographic characteristics, although mild differences exist in terms of time schedule (Table A.9).

Table A.10 shows that the intervention still had a positive effect even 18 months after its implementation. For girls, the size of the medium-term impact is 0.146 s.d. The point estimate for boys is also positive and relatively large, although not statistically significant. The fact that the positive effects does not wear off is undoubtedly a positive result, and even surprising, given the short duration of the project. It is possible that the teachers picked up some of the instructors' strategies and started implementing them themselves later on, to the benefit of both girls and boys.

In conclusion, the medium-term results provide further support for our previous findings. However, these additional results need to be interpreted with caution. First, although the balance tests do not reveal substantial selection bias in terms of observable characteristics in the follow-up, there was strong attrition between the two assessments. Second, the estimated effects relate to a period of extended school closings. It has been shown that the Covid-19 pandemic had a negative effect on children's learning (Contini et al., 2022) and we cannot control if treated and control children are balanced in terms of absences or class closure due to Covid-19).[49]

8. External validity

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

To examine whether and how participating units differ from the regional and national levels, we exploit data from the second grade INVALSI standardized national achievement test held during the previous scholastic year 2017–18, and compare individual and family characteristics of the children in the experimental classes (treated and control) with the child population at large.[50] The results show that the children in the experimental classes perform better on both the math and Italian INVALSI tests than children at the regional and national level (Table A.11). It may be noticed that the gender gap in math is larger in the participating classes: this is consistent with the common finding that girls lag behind boys in math test scores particularly among well performers. The educational level of the parents and the proportion of children who attended kindergarten are also higher in the experimental group.

Taken together, 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.

9. Conclusions

There are many studies on the gender gap in mathematics, but

[48] An altered behaviour may also be triggered by the so-called "Hawthorne effect", whereby subjects of an experimental study who are aware of the aim of the experiment may inadvertently change their behaviour, potentially biasing the results.

[49] In particolar, Contini et al. (2022), investigating the same schools in Torino, show that in October 2020 children in grade 4 had experienced a loss in math of about 0.19 s.d. because of the Covid-19 pandemic. Among children with low-educated parents the learning deficit was larger for girls (−0.3 s.d.) and pupils with high initial abilities (up to −0.5 s.d.).

[50] With the schools' consent, we obtained INVALSI test scores in math and Italian, oral marks in math and Italian, as well as the experimental class averages of pupils' childcare attendance, and mothers' and fathers' education levels. To analyse regional and national test scores, we analysed the representative sample of classes where the test was administered under external supervision (to reduce cheating).

research on the role of teaching methods is lacking. To our knowledge, this is the first rigorous study to determine whether there is a link between teaching methodologies and gender differences in math skills. Given the concern and commitment that many countries and the international community have shown toward the gender gap in mathematics and women's careers in STEM subjects, it is somewhat surprising that so little attention has been paid so far to the role played by teaching methodologies in tackling these problems.

We implement a teaching methodology aimed at improving primary children's mathematical skills. The approach, grounded in active and cooperative learning practices, provided 15 h of math laboratories (MATL) focusing on peer interaction, the sharing of ideas, students' engagement, problem posing, and problem solving. We evaluate the methodology using a randomised controlled trial conducted in the province of Torino, involving 50 third grade classes in 25 schools, and 1044 students. The teaching practices employed in the MATL intervention could reduce the gender gap in mathematics because they incorporate specific features that appear to foster girls' learning and reduce anxiety: (i) no pressure and competition for performance on tasks; (ii) positive attitudes toward mistakes, valued as an opportunity to learn; (iii) use of a narrative context; (iv) support for balanced participation.

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. Consequently, the intervention reduced the gender gap in mathematics by somewhere in the range of 40 % to 47.5 %. In addition, we found that girls with high pre-test scores and girls with low educated parents benefit the most. Our results are encouraging and suggest that properly designed teaching methodologies may improve math performance among girls.

The laboratories were conducted in the class by postgraduate students in math education; thus, they were well implemented. This does not imply that we would still find a positive outcome if regular schoolteachers were conducting the lab activities. Thus, our key finding is that these active learning methodologies for teaching mathematics have the potential to reduce the gender gap in math, when properly run. Further research is needed to design effective teachers' training and evaluate its impact on children's math skills if schoolteachers delivered the intervention, to provide more definitive evidence. If teachers embraced the methodology and applied it widely across the curriculum, the intervention effect might be even larger and more lasting. To carry out such an evaluation, the randomised controlled trial should be expanded by increasing the sample size, including different regions or countries, extending the evaluation to a longer period, and including the assessment of longer-term effects.

Our paper is the first evaluation of the causal effect of a teaching methodology on the gender gap in mathematic. In terms of policy implications, given the positive effect for girls and the null effect for boys, our findings – if confirmed at scale – suggest that training mathematics teachers in active and collaborative practices would imply a substantial overall improvement over the current situation.

CRediT authorship contribution statement

Maria Laura Di Tommaso: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Dalit Contini:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Dalila De Rosa:** Data curation, Formal analysis, Software. **Francesca Ferrara:** Investigation, Methodology. **Daniela Piazzalunga:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Ornella Robutti:** Investigation, Methodology.

Declaration of competing interest

None.

Data availability

The authors do not have permission to share data.

Acknowledgments

We gratefully acknowledge financial support from the University of Torino and the Fondazione Compagnia di San Paolo (Progetto di Ateneo2016 "Tackling the gender gap in mathematics in Piedmont", MATHGAP – website https://sites.google.com/view/mathgendergap/). The project has received the ethical approval by the Ethics Committee of Collegio Carlo Alberto on July 29, 2021. Fondazione Agnelli, the Regional Board of Education in Piedmont (Ufficio Scolastico Regionale), and the Centro Servizi Didattici of Torino Città Metropolitana were partners of the projects. 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. We also thank the instructors: Isabella Boasso, Laura De Conti, Serena Gallipoli, Federica Lucco-Castello; the administrative manager: Silvia D'Incau, the external consultant: Ketty Savioli. We are also grateful to Davide Azzolini, Simone Balestra, Nicola Bazoli, Giorgio Bolondi, Camilla Borgna, Ylenia Brilli, Pietro Di Martino, Chiara Giberti, Stefania Marcassa, Ignacio Monzon, Pauline Morault, Juan Morales, Simone Moriconi, Chiara Pronzato, Enrico Rettore, Claudia Senik, Giuseppe Sorrenti, 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 the pupils who actively took part in the program. The trial has been registered with the AEA RCT Registry: AEARCTR-0003651 (Contini, D., Di Tommaso, M.L., Piazzalunga, D. (2018). "Tackling the Gender Gap in Mathematics in Italy", AEA RCT Registry. December 10. https://doi.org/10.1257/rct.3651-1.0).

The paper uses confidential data collected in collaboration with the schools participating in the project. Replication materials will be provided upon request. The authors commit to preserve data and code for a period of no less than five years following the publication of the manuscript, and to provide assistance to requests for clarification and replication.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.econedurev.2024.102538.

References

Alan, S., & Ertac, S. (2019). Mitigating the gender gap in the willingness to compete: Evidence from a randomized field experiment. *Journal of the European Economic Association, 17*(4), 1147–1185.

Alan, S., Ertac, S., & Mumcu, I. (2018). Gender stereotypes in the classroom and effects on achievement. *Review of Economics and Statistics, 100*(5), 876–890.

Anderson, K. (1997). Gender bias and special education referrals. *Annals of Dyslexia, 47*, 151–162.

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

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 English. In D. Lyn, M. B. Bussi, G. A. Jones, R. A. Lesh, B. Sriraman, & D. Tirosh (Eds.), *Handbook of international research in mathematics education*. Abingdon: Routledge.

Arzarello, F., & Robutti, O. (2010). Multimodality in multi-representational environments. *ZDM : The International Journal on Mathematics Education, 42*(7), 715–731.

Battisti, M., Fedorets, A., & Kinne, L. (2023). *Cognitive skills among adults: An impeding factor for gender convergence?*. *IZA DP* 16134.

Berlinski, S., & Busso, M. (2017). Challenges in educational reform: An experiment on active learning in mathematics. *Economics Letters, 156*, 172–175.

Bietenbeck, J. (2014). Teaching practices and cognitive skills. *Labour Economics, 30*, 143–153.

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 P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *The sage handbook of social research methods*. London: SAGE Publications Ltd.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*(4), 289–328.

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 (Chicago, Ill.), 55*(1), 143–152.

Boaler, J. (2016). *Mathematical mindsets: Unleashing students' potential through creative math, inspiring messages and innovative teaching*. Hoboken, New Jersey: Jossey-Bass.

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.

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. *Quarterly Journal of Economics, 134*(3), 1163–1224.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593–2632.

Coenen, J., Cornelisz, I., Groot, W., Maassen van den Brink, H., & Van Klaveren, C. (2018). Teacher characteristics and their effects on student test scores: A systematic review. *Journal of Economic Surveys, 32*(3), 848–877.

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*. https://doi.org/10.1257/rct.3651-1.0. *AEA RCT Registry*. December 10.

Contini, D., Di Tommaso, M. L., Muratori, C., Piazzalunga, D., & Schiavon, S. (2022). Who lost the most? Mathematics achievement during the COVID-19 pandemic. *B.E. Journal of Economic Analysis & Policy, 22*(2), 399–408.

De Gendre, A., Feld, J., Salamanca, N., & Zölitz, U. (2023). *University of Zurich Working Paper 438*.

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. (2019). Understanding gender differences in STEM: Evidence from college applications. *Economics of Education Review, 72*, 219–238.

Delaney, J. M., & Devereux, P. J. (2020). Math matters! The importance of mathematical and verbal skills for degree performance. *Economics Letters, 186*, Article 108850.

Di Tommaso, M. L., Maccagnan, A., & Mendolia, S. (2021). Going beyond test scores: The gender gap in Italian children's mathematical capability. *Feminist Economics, 27*(3), 161–187.

Dossi, G., Figlio, D., Giuliano, P., & Sapienza, P. (2021). Born in the family: Preferences for boys and the gender gap in math. *Journal of Economic Behavior & Organization, 183*, 175–188.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. New York: Ballantine Books.

Dweck, C. S. (2007). 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* (pp. 47–55). Washington DC: American Psychological Association.

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), 109–128.

Ellison, G., & Swanson, A. (2023). Dynamics of the gender gap in high math achievement. *Journal of Human Resources, 58*(5), 1679–1711.

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.

Ferrara, F., & Ferrari, G. (2020). Reanimating tools in mathematical activity. *International Journal of Mathematical Education in Science and Technology, 51*(2), 307–323.

Ferrara, F., Ferrari, G., Robutti, O., Contini, D., & Di Tommaso, M. L. (2021). When gender matters: A study of gender differences in mathematics. In M. Inprasitha, N. Changsri, & M. Boonsena (Eds.), *2. Proceedings of the 44th Conference of the International Group for the Psychology of Mathematics Education* (pp. 419–426). PME.

Francesconi, M., & Parey, M. (2018). Early gender gaps among university graduates. *European Economic Review, 109*, 63–82.

Fryer, R. G., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics, 2*(2), 210–240.

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*, Article 101978.

Gladstone, J. R., & Cimpian, A. (2021). Which role models are effective for which students? A systematic review and four recommendations for maximizing the effectiveness of role models in STEM. *International Journal of STEM Education, 8*, 59.

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *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 (New York, N.Y.), 320*(5880), 1164–1165.

Gutierrez, A., & Boero, P. (2006). *Handbook of research on the psychology of mathematics education. Past, present and future*. Rotterdam: Sense Publishers.

Hanushek, E. A., Piopiunik, M., & Wiederhold, S. (2019). The value of smarter teachers international evidence on teacher cognitive skills and student performance. *Journal of Human Resources, 54*(4), 857–899.

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

Ho, H. Z., Senturk, D., Lam, A. G., Zimmer, J. M., Hong, S., Okamoto, Y., et al. (2000). The affective and cognitive dimensions of math anxiety: A cross-national study. *Journal for Research in Mathematics Education, 31*(3), 362–379.

INVALSI. (2018). *The Invalsi tests according to Invalsi*. Rome: INVALSI. Available online at: https://invalsi-areaprove.cineca.it/docs/2018/INVALSI_tests_according_to_INVALSI.pdf.

Iriberri, N., & Rey-Biel, P. (2021). *Brave* boys and *play-it-safe* girls: Gender differences in willingness to guess in a large scale natural field experiment. *European Economic Review, 131*, Article 103603.

Lave, J., & Wenger, E. (1991). *Situated learning. legitimate peripheral participation*. Cambridge: University of Cambridge Press.

Lavy, V. (2016). What makes an effective teacher? Quasi-experimental evidence. *CESifo Economic Studies, 62*(1), 88–125.

Lippmann, Q., & Senik, C. (2018). Math, girls and socialism. *Journal of Comparative Economics, 46*(3), 874–888.

Livingston, S. A. (2018). Test reliability – basic concepts. *Research Memorandum*. ETS RM–18-01. Available online at: https://www.ets.org/Media/Research/pdf/RM-18-01.pdf.

Machin, S., & Puhani, P. A. (2003). Subject of degree and the gender wage differential: Evidence from the UK and Germany. *Economics Letters, 79*(3), 393–400.

Mullis, I. V. S., Martin, M. O., Foy, P., Olson, J. F., Preuschoff, C., Erberber, E., et al. (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.

Nemirovsky, R., Borba, M., Dimattia, C., Arzarello, F., Robutti, O., Schnepp, M., et al. (2004). Introduction. PME Special issue: Bodily activity and imagination in mathematics learning. *Educational Studies in Mathematics, 57*, 303–321.

Nicoletti, C., Sevilla, A., & Tonei, V. (2022). *Gender stereotypes in the family*. Institute of Labor Economics. IZA DP, 15773.

Niederle, M., & Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives, 24*(2), 129–144.

Niederle, M., & Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics, 3*(1), 601–630.

Nollenberger, N., Rodríguez-Planas, N., & Sevilla, A. (2016). The math gender gap: The role of culture. *American Economic Review, 106*(5), 257–261.

OECD. (2009). *Creating effective teaching and learning environments. First results from Talis*. Paris: OECD Publishing.

OECD. (2015). *The ABC of gender equality in education. aptitude, behaviour, confidence*. Paris: OECD Publishing.

OECD. (2016). *PISA 2015 results (Volume I): Excellence and equity in education*. Paris: OECD Publishing.

OECD. (2019). *PISA 2018 results (Volume II): Where all students can succeed*. Paris: OECD Publishing.

OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. Paris: OECD Publishing.

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.

Paredes, V. (2014). A teacher like me or a student like me? Role model versus teacher bias effect. *Economics of Education Review, 39*, 38–49.

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*. 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.
Puzio, K., & Colby, G. T. (2013). Cooperative learning and literacy: A meta-analytic review. *Journal of Research on Educational Effectiveness, 6*(4), 339–360.
Sansone, D. (2017). Why does teacher gender matter? *Economics of Education Review, 61*, 9–18.
Schwerdt, G., & Wuppermann, A. C. (2011). Is traditional teaching really all that bad? A within-student between-subject approach. *Economics of Education Review, 30*(2), 365–379.
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.
Thompson, P. W. (2014). Constructivism in mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education*. Dordrecht: Springer.
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*. Washington, DC: Institute of Education Sciences.
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.