Feb, 2026
GRINS DISCUSSION PAPER SERIES DP N° 84/2026 ISSN 3035-5567

GRINS

Visuo-spatial abilities and gender gap in mathematics

DP N° 84/2026

Authors:
Adriana Di Liberto, Ludovica Giua, Giovanni Piumatti, Barbara Romano

Feb, 2026
GRINS DISCUSSION PAPER SERIES DP N° 84/2026 **ISSN 3035-5567**

Visuo-spatial abilities and gender gap in mathematics

Adriana Di Liberto, Ludovica Giua, Giovanni Piumatti, Barbara Romano

KEYWORDS

#Visuo-spatial abilities #mathematics #gender gap #bricks #education #RCT

JEL CODE

I21, I24, J16

ACKNOWLEDGEMENTS

This study was funded by the European Union - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

CITE THIS WORK

Author(s): Adriana Di Liberto, Ludovica Giua, Giovanni Piumatti, Barbara Romano. Title: Visuo-spatial abilities and gender gap in mathematics. Publication Date: 2026.

This paper evaluates an educational intervention aimed at improving visuo-spatial and mathematical skills among primary school students and at reducing the gender gap in learning mathematics through teacher training and classroom use of building bricks. Using a randomized controlled trial on a sample of Italian schools, we find that treatment effects are stronger with longer exposure and in-person teacher training, and that the intervention's effectiveness on students' outcomes is mediated by teachers' improvements in spatial orientation skills. Gender differences emerge, with girls benefiting mainly in visuo-spatial abilities and boys in mathematics.

Visuo-spatial abilities and gender gap in mathematics*

FEBRUARY 9, 2026

Adriana Di Liberto (Università di Cagliari),
Ludovica Giua (Università di Cagliari),
Giovanni Piumatti (Fondazione Agnelli), and
Barbara Romano (Fondazione Agnelli)

Abstract. This paper evaluates an educational intervention aimed at improving visuo-spatial and mathematical skills among primary school students and at reducing the gender gap in learning mathematics through teacher training and classroom use of building bricks. Using a randomized controlled trial on a sample of Italian schools, we find that treatment effects are stronger with longer exposure and in-person teacher training, and that the intervention's effectiveness on students' outcomes is mediated by teachers' improvements in spatial orientation skills. Gender differences emerge, with girls benefiting mainly in visuo-spatial abilities and boys in mathematics.

Keywords. Visuo-spatial abilities; mathematics; gender gap; bricks; education; RCT.

JEL codes. I21; I24; J16.

* The research is funded by the European Union - Next Generation EU, in the framework of the GRINS - Growing Resilient, Inclusive and Sustainable Project (GRINS PE00000018 – CUP E63C22002140007). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union and the European Union cannot be held responsible for them. The project is also promoted by Exor and Fondazione Agnelli, with the scientific contribution of Politecnico di Torino and support from The LEGO Foundation. The RCT is registered at the AEA RCT Registry (AEARCTR-0010914). We thank seminar participants at the CRENoS Workshop (Cagliari, 2025), the Socio-Emotional Skills and Education Workshop (Trento, 2025), SIE Conference (Napoli, 2025) and the Spoke 3 GRINS Final Wokshop (Napoli, 2026). We are grateful to Marco Caliendo, Elif Kubilay, Matti Sarvimäki and Thomas Siedler for comments and discussions.

1. INTRODUCTION

Male students tend, on average, to outperform their female peers in mathematics test scores. A large body of research attributes this gap not to innate ability differences, but to disparities in experiences and in the expectations that families, teachers, and peers hold regarding boys' and girls' abilities (OECD, 2015). Consistent with this interpretation, cross-country evidence shows that gender gaps in mathematics and science are substantially smaller in more gender-equal societies (Machin & Pekkarinen, 2008; Guiso et al., 2008; Else-Quest, Hyde & Linn, 2010; Dossi et al., 2021), suggesting that cultural norms and social environments play a key role in shaping educational outcomes.

Early-life experiences are particularly important for cognitive development, and gender gaps may emerge when children have unequal access to cognitively enriching activities (Carneiro & Heckman, 2003; Currie & Almond, 2011). One key channel operates through the development of visuo-spatial abilities. They play an important role in perceiving spatial relations, improving visual memory, forming mental representations of objects in space, understanding abstract concepts, and developing numerical abilities. Core components include mental rotation (i.e., the ability to mentally rotate representations of two- or three-dimensional objects) and spatial orientation (i.e., the ability to describe or classify spatial dimensions). Boys often receive more opportunities than girls to engage in spatially intensive play, partly because parents select different toys for different genders. The existing literature documents a strong link between childhood visuo-spatial skills and later quantitative reasoning and mathematics achievement (Uttal et al., 2013; Barnes and Raghubar, 2014; Mix et al., 2021).

Visuo-spatial abilities can also be fostered through construction play, which enhances problem-solving skills, eye–hand coordination, spatial awareness, and numerical abilities (Wolfgang et al., 2001; Nath et al., 2014). Beyond cognitive skills, non-cognitive factors such as mathematics self-efficacy and math anxiety play an important role in shaping performance and exhibit pronounced gender differences (Stobart et al., 1992; Murphy, 1982; Ferguson, 2015; Alan et al., 2019). Finally, the school environment can either amplify or mitigate these disparities: teacher stereotypes influence classroom interactions, and biased expectations have been shown to translate into differential practices and feedback toward male and female students (Alan et al., 2018; Carlana, 2019; Lavy et al., 2024; Martinot et al., 2025).

This paper evaluates the effectiveness of an educational intervention aimed at enhancing visuo-spatial and mathematical skills among primary school students through hands-on activities using building bricks. A central objective of the intervention is to strengthen spatial reasoning at an early age and to reduce gender gaps in mathematics by supporting girls' learning and alleviating math-related anxiety relative to boys. The project also includes a targeted teacher training component designed to promote the adoption of innovative teaching practices in the classroom.

Our primary research question is whether students in treated classrooms exhibit higher levels of visuo-spatial and mathematical skills at the end of the school year compared to students in control classrooms. In addition to cognitive outcomes, we examine changes in students' mathematics self-efficacy and math anxiety, with a particular focus on gender heterogeneity. To identify causal effects, the intervention was evaluated through a randomized controlled trial (RCT).

The empirical analysis indicates that the impact of the intervention on students' outcomes is mediated by improvements in teachers' own skills following the training. Moreover, students exposed to the treatment from the beginning of the school year experience slightly larger gains, underscoring the importance of treatment duration. We also find that in-person teacher training is more effective than remote training.

We document meaningful gender differences in treatment effects. Girls primarily benefit in terms of visuo-spatial skills, while boys experience larger improvements in mathematical competencies. In addition, boys significantly increase their use of building bricks at home following the intervention, a behavioral response that may generate longer-term cognitive benefits.

This study relates to several strands of literature. First, it contributes to research emphasizing the importance of teacher quality for student achievement (Rivkin et al., 2005; Chetty et al., 2014; Coenen et al., 2018; Hanushek et al., 2019). While this literature has not reached a consensus on how to define effective teaching or on scalable policies to improve teacher quality (Bietenbeck, 2014; Schwerdt & Wuppermann, 2011), our results highlight the potential of innovative pedagogical approaches that can be implemented using existing human resources. Second, we contribute to the literature on teacher bias, which documents systematic gender differences in teachers' expectations and their consequences for student outcomes (Dee, 2007; Carlana, 2019; Alesina et al. 2024). Finally, our findings speak to the relatively limited literature examining how teaching methodologies interact with gender differences in STEM-related outcomes (Di Tommaso et al., 2024).

The remainder of the paper is organized as follows. Section 2 describes the experimental design and empirical strategy. Section 3 presents the data. Section 4 discusses the results for teachers and students, and Section 5 concludes.

2. EXPERIMENTAL DESIGN AND EMPIRICAL STRATEGY

Matabì is a project aimed at developing the visuo-spatial skills of third- and fourth-grade primary school students through specific teacher training and the use of Lego Duplo® bricks in the classroom. Primary schools from various Italian regions participated in the project. Within each school, a group of mathematics teachers for third and fourth grade voluntarily took part in the project. The first edition of the project took place during the 2022–23 school year and involved five primary schools in the city of Turin. The second edition, held in the following school year, involved 11 primary schools located in Piedmont, Sicily, and Campania.

The impact evaluation of the project was designed as a randomized controlled trial (RCT). The RCT design involves assigning the treatment to a portion of the sample (treatment group) and comparing the outcomes with another portion of the sample that participated in the experiment but did not receive the treatment (control group). In this context, each teacher participating in the project was randomly assigned to one of the groups, using stratified randomization within each school and grade. As a result, all students in the same class belong to the group to which their teacher was assigned.

The experimental design differs slightly between the first and second editions (see Figure 1). In the first edition (2022–23), the design included three groups: one treated from the beginning of the school year, one treated starting from the second half of the school year, and one untreated group serving as the control group. Assignment to the three groups occurred with equal probability. Thus, each group corresponds to one-third of the 2022–23 edition's sample. In the second edition (2023–24), the design included two groups: one treated from the beginning of the school year and one untreated (control group). In this second edition, randomization assigned 60% of the teachers to the treatment group and 40% to the control group.

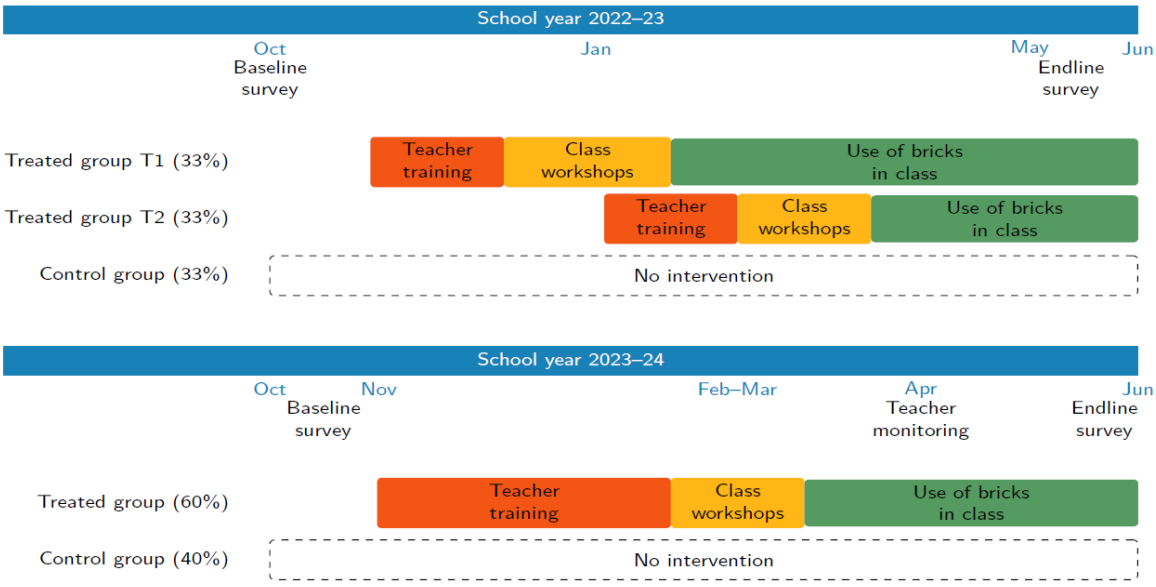
The treatment consists of a training program for third- and fourth-grade mathematics teachers and three classroom workshops.

Teacher training included a cycle of theoretical lectures on visuo-spatial skills: for the first treated group in 2022–23, the training was delivered in person during the first semester; for the groups treated later in the 2022–23 school year and in 2023–24, the training was delivered remotely.

The classroom workshops involved the use of Duplo® bricks with students. In the first workshop, the Matabì instructor led the activity with the teacher observing; in the second, the teacher led the workshop with the instructor present; and in the third, the teacher worked independently. From the first workshop onward, teachers kept the bricks and could use them freely during curricular hours,

whether for mathematics or other subjects. As a result, teachers treated earlier in 2022–23 had potential access to the materials for a longer period than those treated later in the year.

Figure 1. Timeline of the project



During the 2022–23 school year, all participating students completed three paper-based questionnaires: baseline (October 2022), midline (January 2023), and endline (June 2023). Teachers completed a baseline questionnaire on their demographic and professional background. Treated teachers were also asked to complete four additional monitoring questionnaires (December, January, February, and June) to track kit usage. However, declining response rates limited the quality of this information. By January, nearly half (44% of the 9 respondents out of 15 treated teachers) had completed the third workshop and the full training path.

In 2023–24, data collection procedures were adjusted both to reflect changes in the experimental design and to gather more detailed information from all teachers, including those in the control group. All teachers and students completed two questionnaires: one before the treatment began (baseline, November 2023) and one after the end of the intervention (endline, May/June 2024). In addition, in April 2024, an extra monitoring questionnaire was administered to treated teachers. Of the 38 teachers assigned to treatment, 34 responded. Among them, 68% had already completed the third Matabì workshop. In the previous four weeks: 41% had used the Matabì kit with project worksheets at least 3–4 times, 35% had used it for independently structured mathematics activities, 34% had used it for activities in other subjects, 38% had used it for short concentration exercises or "spaced learning" activities. On average, teachers reported about 5 hours of total kit usage in the previous four weeks.

The data collected in the 2022–23 edition measure the effect of the treatment on students' visuo-spatial skills, mathematical cognitive abilities, self-perceived efficacy in studying mathematics, and mathematics anxiety. The 2023–24 data, in addition to measuring effects on these four student outcomes, also allow for the evaluation of the treatment’s impact on teachers' visuo-spatial skills.

Since the sample was selected through randomization, the analysis can rely on endline data, comparing outcomes between students whose mathematics teacher received the training and those whose teacher did not. Given the nature of the intervention, the estimated effects on students reflect a combination of indirect effects (through teacher training) and direct effects (through the use of the bricks in the classroom).

The estimation methodology is consistent across both editions. The analysis uses a linear regression model estimated by ordinary least squares (OLS), and accounts for the fact that treatment was randomized at the class level by clustering standard errors at the class level:

$$y_{icgs} = \beta T_{cgs} + \gamma X_{icgs} + \delta_g + \rho_s + \varepsilon_{icgs}.$$

Here, y_{icgs} is the outcome measured at endline of student i in class c, grade g and school s, while T_{cgs} is a dummy variable identifying the treatment. All regressions include school (ρ_s) and grade fixed effects (δ_g), which capture unobserved characteristics shared by students attending the same school and grade (third or fourth). The full model specification also controls for the student's gender and migration background, as well as baseline scores on maths tests (X_{icgs}). In the second edition we are also able to include a dummy for whether the student plays with bricks at home at baseline. ε_{icgs} is the error term.

3. DESCRIPTIVE STATISTICS AND BALANCING

Teachers

The first edition of the project (school year 2022–23) involved five schools in the city of Turin and included 48 teachers and 53 classes in the analysis, 24 of which were third-grade classes (45%) and 29 fourth-grade classes (55%). The second edition (school year 2023–24) involved 11 schools located in Piedmont, Sicily, and Campania, with a total of 64 teachers and 83 classes, 59 of which were third-grade classes (71%) and 24 fourth-grade classes (29%). The higher number of classes compared to teachers is due to some teachers working across multiple classes.

In the first edition, the 48 teachers were evenly divided into three groups of 16: one treated from the beginning of the school year, one treated in the second half of the school year, and one untreated control group. In the second edition, 38 teachers were assigned to the treatment group (starting from the beginning of the school year) and 26 to the control group.

The data show that, consistent with the national average for primary schools (96%, source: Ministry of Education, 2022–23), the teaching staff is predominantly female (98% in the 2022–23 edition and 92% in the 2023–24 edition). The average age of participating teachers falls between 40 and 50 years, with two-thirds of the sample aged between 40 and 60.

In both editions, a questionnaire was administered to all teachers prior to training to collect demographic information and details about professional experience. Thanks to random assignment, the characteristics of participating teachers are expected to be balanced between the treatment and control groups. Below, we report the comparison of observable characteristics between treated and control teachers, as a preliminary check of randomization quality and to correctly interpret the causal estimates from the RCT.

Years of work experience show no statistically significant differences between groups: in the first edition, control group teachers had on average 17 years of tenure, while those treated from the beginning and those treated mid-year had an average of 19.5 and 19 years, respectively. In the second edition, the average experience was higher, with 23.6 years among treated teachers and 19.3 years among controls. In all cases, differences between groups are not statistically significant. Similarly, the average number of hours dedicated to teaching mathematics per week is comparable across groups: between 6.9 and 7.1 hours in the first edition, and 7.4 hours (control) and 6.9 hours (treatment) in the second edition. The proportion of teachers reporting the use of specific teaching methods (e.g., Bortolato or Montessori) and/or participation in other projects or experiments is also statistically equivalent between groups, around 40% in both editions.

Thus, random assignment successfully ensured a good balance of observable teacher characteristics at baseline, an essential condition for reliable causal inference. Age distribution, years

of experience, and the tendency to adopt alternative teaching methods show no statistically significant differences between treatment and control groups, confirming the validity of the experimental design.

In addition to demographic data, in the 2023–24 teacher questionnaire we collected further relevant information for the analysis. One such measure is the teachers' perceived self-efficacy in teaching mathematics. This is assessed through responses to 16 questions covering various dimensions of teaching practice, including perceived ability to: motivate students, adapt teaching to individual needs, use effective assessment strategies, and foster critical thinking in the classroom. Each item is scored on a scale from 0 to 1. An overall self-efficacy index is calculated as the simple average of the individual scores. The average self-efficacy is very similar between treated and control teachers (0.798 for the treatment group and 0.822 for the control group), and the difference is not statistically significant.

Two additional fundamental characteristics of teachers collected at baseline in the 2023–24 edition are their visuo-spatial abilities and implicit gender associations. To measure teachers' visuo-spatial abilities, two tests were administered: one assessing mental rotation and one assessing spatial orientation. The first measures the ability to mentally rotate objects in space, while the second evaluates the ability to navigate an environment and mentally represent spatial relationships. Since the training provided to teachers focused more extensively on spatial orientation, the corresponding test is particularly relevant for evaluating any potential effects of the intervention.

Implicit gender associations were measured using the Implicit Association Test (IAT), which assesses the strength of automatic associations between concepts (e.g., gender, ethnicity) and attributes (e.g., positive, negative). The IAT is widely used in the literature to detect potential stereotypes and unconscious biases. In this project, teachers completed the IAT to measure the possible presence of an unconscious bias associating scientific disciplines more strongly with boys and humanities disciplines with girls, and vice versa.

The value of collecting this information lies in the fact that, although the Matabì project focuses on the importance of students' visuo-spatial abilities for learning mathematics and on potential gender differences, the treatment's influence on students is only indirect. The directly treated individuals are the teachers, who, if assigned to the treatment group, receive specific training on visuo-spatial skills and information regarding gender differences in STEM fields. In this experiment, collecting data on these teacher characteristics is crucial, as recent research investigating the causes of gender gaps in STEM competencies — typically favouring male students — has highlighted the significant role of teachers' implicit gender biases (e.g., Carlana, 2019).

Descriptive statistics suggest that at baseline, the two groups are statistically equivalent in terms of these characteristics. The average score on the mental rotation test is slightly higher among treated teachers (3.17) compared to the control group (2.8), but the difference is not statistically significant. Conversely, on the spatial orientation test, the treated group scored slightly lower on average (1.83) than the control group (2.24), but again, the difference is not statistically significant.

In addition to the visuo-spatial tests, all teachers completed the IAT to assess automatic associations between gender and academic disciplines. On average, both groups display a slight tendency to associate science with the male gender and the arts with the female gender, with no significant differences between treated and control groups. The distribution of scores suggests wide individual heterogeneity, but, as with previous variables, no systematic imbalances between groups emerge.

In summary, the analysis of all observed teacher characteristics in the treatment and control groups indicates that the random assignment in this study successfully produced well-balanced groups.

Students

Student information and performance data were collected through questionnaires. In the first edition, all questionnaires were administered in paper format, whereas in the second edition, a digital format

(tablet or computer, depending on the school's available resources) was adopted. The main demographic characteristics examined include gender and migration background, the latter defined based on students' responses to the question regarding the language spoken at home.

We consider four outcome variables. Two relate to students' cognitive skills: visuo-spatial ability and mathematical competence. Two additional outcome variables, which may also be influenced by the treatment, pertain to non-cognitive domains that the literature identifies as important factors in mathematics learning: perceived self-efficacy in studying mathematics and math anxiety. Students' outcome variables are always standardized to have a mean of 0 and a standard deviation of 1, allowing for direct comparisons across students from different classes. In the case of visuo-spatial abilities, the same test was administered to students in both third and fourth grades. As expected, given that visuo-spatial skills naturally improve with age, we observe that in the 2022–23 school year third-grade students answered 43% of the questions correctly on average, while fourth-grade students answered 51%. In the 2023–24 school year, these percentages were 42% and 45%, respectively. For this reason, scores are standardized separately by grade level to account for skill differences due to age, thus ensuring comparability between third- and fourth-grade students.

Since certain aspects of the data collection process and measurement of variables differ between the two editions, in this section we present the descriptive analysis separately for the 2022–23 and 2023–24 school years. As with teachers, the student analysis begins with a verification of balance between the groups involved in the experiment in terms of individual characteristics measured at baseline. Below, we present the balance tests for demographic and outcome variables for the sample collected in the 2022–23 school year, followed by those for the 2023–24 sample.

Students' characteristics: 2022-23 edition

The final sample of the 2022–23 edition of Matabì consists of 863 students observed at both baseline (October 2022) and endline (June 2023). Slightly less than 9% of the sample (74 students) holds a certified BES status (Special Educational Needs).

As described in Section 2, teachers and their students were randomly assigned into three groups. The treatment was delivered to a first group of 291 students starting from the first semester and to a second group of 285 students starting from the second semester. The control group comprises the remaining 33% of the sample (287 students). Among all students, 363 were enrolled in third grade, while a larger number — 500 students (58% of the total sample) — were enrolled in fourth grade.

Table 1 summarizes the results of the comparison of observed baseline characteristics across the two treated groups (in the first and second semester) and the control group. The first three columns of the table report mean values for the share of students enrolled in third grade, the share of female students, and the proportion of students with a migration background, respectively for first-semester treated students, second-semester treated students, and the control group. The last four columns report the differences between the mean values observed in each treatment group and the control group, along with the corresponding p-value. As previously seen in the case of teachers, this test indicates whether the treatment and control groups can be considered balanced, which is a key requirement for the subsequent impact evaluation analysis using a Randomized Controlled Trial design.

Table 1. Balance of demographic characteristics at baseline

	Treatment 1st semester	Treatment 2nd semester	Control	Treatment 1st semester – Control		Treatment 2nd semester – Control	
	Mean	Mean	Mean	Difference	p-score	Difference	p-score
Grade 3	0.454	0.347	0.460	-0.006	0.879	-0.113	0.006
Female	0.515	0.523	0.509	0.007	0.871	0.014	0.736
Migration background	0.316	0.326	0.279	0.037	0.326	0.048	0.216

The results suggest that the group treated in the first semester is very similar to the control group. In contrast, the p-values reported in the last column indicate that the group treated in the second semester differs from the control group in terms of a lower number of third-grade students. It is important to note that the final estimation model accounts for this slight imbalance between groups at baseline.

As for the outcome variables, during the baseline data collection before the start of the intervention we measured students' math skills, perceived self-efficacy in mathematics, and math anxiety. The test on visuo-spatial abilities, however, was only administered at endline.

The variable measuring mathematical skills is based on two sets of questions covering both mathematical and geometrical domains. The type and number of questions included in this test vary between third and fourth grades and across survey rounds. For this reason, the math skill variable was first normalized to express the percentage of correct answers and then standardized (mean zero, standard deviation one) separately for each survey wave (baseline and endline).

Self-efficacy in mathematics is measured by calculating an index derived from the students' responses to five questions exploring their perceptions of their own mathematical abilities. Responses are on a scale from 1 ("Strongly disagree") to 4 ("Strongly agree"). The final score is the average of the responses to the five questions, which represents the composite measure of perceived self-efficacy in the subject. The indicator measuring math anxiety is based on a similar criterion. Again, the score assigned to responses ranges from 1 to 4, and the index is the average of the score obtained.

Table 2 explores the balance of outcome variables at baseline, presenting the results of an OLS regression analysis that examines the relationship between the treatment and each of the three outcome variables, conditioned on the baseline values of demographic characteristics. The randomization structure is accounted for by clustering standard errors at the class level.

Column 1 reports the results related to the score in the maths test. The conditional mean differences highlight the disadvantage for students treated in the second semester compared to the control group. However, although the value is large, when accounting for all observed characteristics, the coefficient for the group treated in the second semester is not statistically significant. A similar result is observed for the other two outcome variables (columns 2 and 3): the non-significant coefficients for the variables identifying students in both the first and second treatment groups indicate that the conditional mean differences between the two treated groups and the control group are not statistically different from zero.

The gender indicator confirms results already reported by extensive literature on the topic. Column 1 shows that, on average, female students score about 0.09 standard deviations lower than male students on the math test, and this gender gap is statistically significant. This is an expected result that confirms the significant gap between boys and girls in mathematics. Column 2 also suggests lower levels of self-efficacy for females, while column 3 highlights that female students experience higher levels of math anxiety, almost one-third of a standard deviation higher than male students. Again, these results align with expectations given the randomization and extensive literature on the topic.

Unlike the other outcome variables, in the 2022-23 edition the students' visuo-spatial skills were not measured at baseline but only at the endline. The variable that captures the visuo-spatial skills score is based on 20 questions: 10 related to mental rotation and 10 to spatial orientation. The overall score is calculated as the percentage of correct answers.

Table 2. Balancing of outcomes measured at baseline

	(1)	(2)	(3)
Dependent variables	Mathematics	Self-efficacy in math	Math anxiety

Treatment in 1st semester	0.018	0.004	0.013
	(0.114)	(0.122)	(0.074)
Treatment in 2nd semester	-0.156	0.046	0.108
	(0.107)	(0.131)	(0.079)
Female	-0.091*	-0.156**	0.304***
	(0.052)	(0.071)	(0.078)
Grade 3	1.104***	0.247***	0.107*
	(0.100)	(0.091)	(0.061)
Migration background	-0.291***	0.173*	0.084
	(0.062)	(0.092)	(0.083)
Constant	-0.282***	-0.089	-0.267***
	(0.087)	(0.122)	(0.072)
Observations	863	777	863
R-squared	0.391	0.037	0.061
School Fixed Effects	Yes	Yes	Yes
Note: Robust standard errors in parentheses, clustered at class-level; ***p<0.01, **p<0.05, *p<0.1.			

Students characteristics: 2023-24 edition

The final sample for the 2023-24 edition of Matabì consists of 974 students, of whom 63 have a BES certification (6%). As in the previous edition, the students are from both third-grade (713 students) and fourth-grade (261 students) primary school classes. Unlike the previous edition, there is only one treatment group in this case. The randomization split assigned 550 students to the treatment group and 424 students to the control group.

Table 3 shows the results of the comparison between the treatment and control groups on demographic characteristics at baseline. In the second edition of the Matabì project, students were also asked about the frequency of playing with building blocks at home. The first two columns of the table report the average values for the percentage of students in third grade, the proportion of female students, the number of students with a migratory background, and the frequency of playing with building blocks at home for the treatment group and the control group, respectively. The last two columns report the differences between the average values observed in the treatment group and the control group and the corresponding p-value. About three-quarters of the students are in third grade and half are female and both characteristics are balanced between the treatment and control groups. As for migratory background, there is a very small and non-significant difference: 12.9% of the students in the treatment group speak a language other than Italian, compared to 11.8% in the control group. Finally, the percentage of students who report playing with bricks at home (often or sometimes) is slightly higher in the treatment group (68.9%) compared to the control group (65.1%), but this difference is also not statistically significant.

Table 3. Balancing of demographic characteristics at baseline

	Treatment	Control	Treatment - control	
	Mean	Mean	Difference	p-score
Grade 3	0.751	0.708	0.043	0.130
Female	0.471	0.517	-0.046	0.158
Migration background	0.129	0.118	0.011	0.601
Use of bricks at home	0.689	0.651	0.038	0.209

Unlike the 2022-23 edition, the outcome variables considered in the baseline analysis for the 2023-24 edition are four and include, in addition to mathematical cognitive skills, self-efficacy and anxiety, also visuo-spatial skills. This latter indicator is constructed from 13 items: 5 related to mental rotation

and 8 to spatial orientation. The index is calculated by averaging the correct answers out of the total, considering missing answers as incorrect.

As in the 2022-23 edition, mathematical skills are measured through math and geometry questions, with a structure that varies between different surveys and school grades. At baseline, the test for third-grade students included 20 math questions and 4 geometry questions; at endline, 12 questions were proposed for each subject. For fourth-grade students, the baseline questionnaire included 12 math questions and 12 geometry questions, while at endline, students faced 15 math questions and 22 geometry questions. For each survey, the final variable is the number of correct answers, subsequently standardized for comparability between grades.

Mathematical self-efficacy is measured through an index constructed from eight questions (instead of five, as in the 2022-23 survey) that assess the student's perception of their own mathematical abilities. The possible answers range from 1 ("Not at all") to 5 ("Very much"). The final score is the average of the responses to the eight questions, representing a synthetic measure of perceived self-efficacy in the subject.

Finally, math anxiety is an index that quantifies the level of anxiety perceived by students in the mathematical context. This is measured through nine questions, which investigate the level of anxiety, agitation, or worry of students in various situations related to studying math: such as facing a test, completing a task independently, or listening to an explanation from a teacher or peer. Again, the responses follow a scale ranging from 1 ("Not at all") to 5 ("Very much"), and the final score is calculated as the average of the values assigned to each question.

The estimates presented in Table 4 allow analysing the balance of outcomes at baseline, taking into account the students' demographic characteristics (gender, migratory background, whether they play with bricks at home, grade, and school) and the structure of randomization at the class level through clustering standard errors at the class level. As expected, in the presence of random treatment assignment, the coefficient associated with the treatment variable is not statistically significant in any of the four columns, suggesting that there are no systematic differences between the two groups.

In the second column, third-grade students show a significantly higher score in maths compared to fourth-grade students. This result could stem from the fact that the questionnaire administered to the third-grade students was relatively easier than that of the fourth-grade students. An interesting result emerges from the coefficients associated with the variable indicating whether the student plays with bricks at home: the coefficient is positive and significant in columns 1, 2, and 3, indicating that these students tend to score higher in visuo-spatial skills, mathematics, and self-efficacy in maths. This further reinforces the hypothesis of an association between playing with bricks and the development of cognitive skills.

Table 4. Balancing of outcomes measured at baseline

	(1)	(2)	(3)	(4)
Dependent variables	Visuo-spatial abilities	Mathematics	Self-efficacy in math	Math anxiety
Treatment	0.108	0.123	-0.047	-0.107
	(0.079)	(0.092)	(0.064)	(0.079)
Female	-0.168**	-0.059	-0.331***	-0.122*
	(0.073)	(0.070)	(0.057)	(0.066)
Grade 3	-0.133	0.412***	0.048	-0.037
	(0.111)	(0.126)	(0.083)	(0.106)
Migration background	-0.030	-0.030	-0.144	0.166
	(0.123)	(0.098)	(0.105)	(0.102)
Use of bricks at home	0.183**	0.184***	0.203***	-0.038
	(0.072)	(0.065)	(0.068)	(0.071)
Constant	-0.001	-0.462***	0.024	0.240**
	(0.114)	(0.124)	(0.093)	(0.115)
Observations	987	974	986	979

R-squared	0.040	0.101	0.062	0.020
School Fixed Effects	Sì	Sì	Sì	Sì
Nota: Robust standard error in parentheses, clustered at class level; *** p<0.01, ** p<0.05, * p<0.1.				

4. RESULTS

Effect on teachers

In the second edition of Matabì (academic year 2023-24), we collected information to estimate whether the training had an impact on the teachers' visuo-spatial skills. For this purpose, two specific tests were administered to measure the baseline visuo-spatial abilities of both the teachers assigned to the treatment group and those assigned to the control group: the first focuses on spatial orientation, and the second on mental rotation. Both tests were administered before the start of the treatment (baseline) and at the end (endline), in order to detect any changes in the teachers' visuo-spatial skills.

It is important to emphasize that the training provided to the teachers focused primarily on developing skills related to spatial orientation. This implies that, in interpreting the results, the test on spatial orientation is more relevant than the test on mental rotation. If the treatment has a direct and significant impact on teachers' visuo-spatial skills, we expect an improvement in the scores of both tests, with a more pronounced change in the spatial orientation test score, considering the emphasis placed on this skill in the Matabì teacher training.

To quantify this improvement, we create three distinct indicators that measure the change in teachers' visuo-spatial skills, measured as the difference (delta) between the scores obtained by each teacher at endline and those obtained at baseline. The first indicator refers to the sum of the scores from both tests, while the other two separately report the results from each test.

Table 5 presents the results of estimates from a simple OLS model that allows us to analyze the effect of treatment group assignment on overall visuo-spatial skills (column 1) and on specific skills related to spatial orientation (column 2) and mental rotation (column 3). The dependent variable measures the difference between the score obtained on the tests by each teacher at baseline and at endline. The variable of interest used to estimate the treatment effect is a dummy variable equal to one if the teacher received the training and zero otherwise. Column 1 indicates that teachers in the treatment group show an improvement of about 1.6 points compared to teachers in the control group, but this difference is not statistically significant. The coefficient in column 3, which refers to the mental rotation test, also does not significantly differ from zero and is smaller. However, if we consider the improvement of teachers solely in the area of spatial orientation (column 2), the difference between the treatment and control groups is statistically significant and equals 1.1 points.

Table 5. Effect of treatment on teachers' visuo-spatial skills

	(1)	(2)	(3)
	Gain in visuo-spatial abilities of teachers		
	Spatial orientation and mental rotation tests	Spatial orientation test	Mental rotation test
Treatment	1.564	1.116*	0.448
	(0.938)	(0.560)	(0.756)
Observations	49	49	49
R-squared	0.401	0.276	0.357
School fixed effects	Yes	Yes	Yes
Note: Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1.			

Effect on students

The results presented in the previous subsection suggest a positive effect of the treatment on teachers' visuo-spatial skills. However, in this intervention, the primary beneficiaries are the students, although the estimated effect consists of the indirect impact of the training provided to teachers on the use of bricks in the classroom, which affects various student outcome variables. Thus, students' outcomes are less directly related to the intervention and may be more difficult to move (Kraft, 2020).

The main evaluation questions of the analysis aim to examine: 1) whether students in classes that participated in the Matabì project in the treatment group show higher levels of visuo-spatial and mathematical skills at the end of the year compared to students in the control group, and 2) whether the effects differ between male and female students in the treatment group.

To answer these questions, we explore the impact of Matabì on four main student outcomes: visuo-spatial skills, mathematical skills, perceived self-efficacy in mathematics, and anxiety towards mathematics. The variable of interest used to estimate the treatment effect is constructed as a dummy variable equal to one if the student is in a class taught by a teacher who participated in the training, and zero otherwise.

The effects are examined separately for the first and second editions of the program. In what follows, we present the main results. Further details are available in the Appendix outcome (Tables A1—A4 for the 2022-23 edition and Tables A5—A9 for the 2023-24 edition), which also shows several sensitivity checks to test the robustness of the estimated effect by including additional control variables. In addition to the treatment variables, the main estimates are conditioned on various important student characteristics such as gender, migration background, and baseline mathematical skills, as well as grade and school of attendance. The analysis for the 2023-24 edition also controls for the propensity to play with the bricks at home before the project started. Furthermore, the full model introduces an additional variable calculated as an interaction between the treatment identification variable and the student's gender indicator, allowing us to estimate the treatment effect separately for male and female students.

Effects on students: 2022-23 edition

Table 6 shows the average treatment effect separately for male and female students across the four outcomes considered: visuo-spatial abilities, mathematics, self-efficacy in mathematics, and math anxiety. The reported results are obtained from a regression model that includes some baseline student characteristics collected in the 2022-23 school year (gender, migration background, and mathematical skills score), as well as grade and school fixed effects.

Table 6. Effect of the treatment on students' outcomes

	(1)	(2)	(3)	(4)
Dependent variables	Visuo-spatial abilities	Mathematics	Self-efficacy in math	Math anxiety
Total effect of the treatment in 1st semester				
Males	0.228*	0.038	0.021	-0.097
	(0.118)	(0.098)	(0.128)	(0.124)
Females	0.129	0.040	0.170	-0.061
	(0.106)	(0.124)	(0.129)	(0.094)
Total effect of the treatment in 2nd semester				
Males	0.045	-0.211	0.092	-0.121
	(0.118)	(0.151)	(0.173)	(0.117)
Females	0.208*	-0.015	0.140	-0.078
	(0.111)	(0.120)	(0.161)	(0.114)
Note: Robust standard errors in parentheses, clustered at class level; *** p<0.01, ** p<0.05, * p<0.1. All models control for gender, migration background, score in mathematics at the baseline, grade, and school fixed effects.				

As for the overall treatment effect on visuo-spatial skills, the coefficient is positive and significant for male students treated in the first semester (a little over 0.20 standard deviations), while the coefficient for female students is lower and not significant. In contrast, for students treated in the second semester, we observe a positive and significant effect only for female students (around 0.20 standard deviations), while the coefficient for male students is not significantly different from zero.

The comparison between the first and second semesters highlights that coefficients are slightly larger for students treated from the beginning of the school year, thus being exposed to the treatment for a longer period compared to their peers. This suggests that the duration of the treatment is an important factor in determining its effectiveness. Additionally, it should be noted that the training for teachers in the first semester was conducted in person, while those trained in the second semester participated remotely. Therefore, it is possible that the mode of training delivery may have influenced the treatment's effectiveness. The impact evaluation for the 2023-24 edition, where teacher training was conducted online but the exposure to the treatment lasted the entire school year, will provide further insights into this. The results in Table A1 in the Appendix also suggest that the positive effect is mainly observed among third-grade students treated in the first semester (column 6).

Although in nearly all cases the sign of the estimated coefficients is as expected, Table 6 also highlights that the treatment did not have significant effects on the other outcome variables analysed. Possible explanations are that the treatment did not last long enough or that measurable effects on these outcomes arise later.

Effects on students: 2023-24 edition

Table 7 presents the main results for the four outcome variables divided by gender. Recall that in the second edition there was a single treatment during the school year. Overall, the results do not show any significant effects of the treatment on any of the four outcomes considered. The absence of an effect is robust to the inclusion of control variables and the use of different model specifications. In other words, this initial analysis seems to suggest that the intervention did not have a substantial impact on students' visuo-spatial abilities, mathematics, self-efficacy in mathematics, or math anxiety. However, it is important to remember that the effect of the treatment on students is mediated by the teachers, as they are the ones who receive the training directly. Therefore, it is possible that some specific characteristics of the teachers may influence the effectiveness of the treatment, and it is important to further investigate this possibility.

Table 7. Effect of the treatment on students' outcomes

	(1)	(2)	(3)	(4)
Dependent variables	Visuo-spatial abilities	Mathematics	Self-efficacy in math	Math anxiety
Total effect of the treatment				
Males	-0.055	0.067	0.008	0.006
	(0.095)	(0.083)	(0.092)	(0.112)
Females	-0.004	-0.103	0.036	0.011
	(0.085)	(0.104)	(0.092)	(0.090)
Note: Robust standard errors in parentheses, clustered at class level; *** p<0.01, ** p<0.05, * p<0.1. All models control for gender, migration background, score in mathematics at the baseline, propensity to play with bricks at home at the baseline, grade, and school fixed effects.				

Thanks to the information collected in the second edition, the empirical analysis explores the presence of heterogeneous effects, using estimation models that allow identifying whether the impact of the treatment on the students differs based on individual characteristics of the teachers. All the results presented below are based on the main specification of the model. In this case, instead of the interaction between the treatment variable and the gender variable, we introduce an interaction

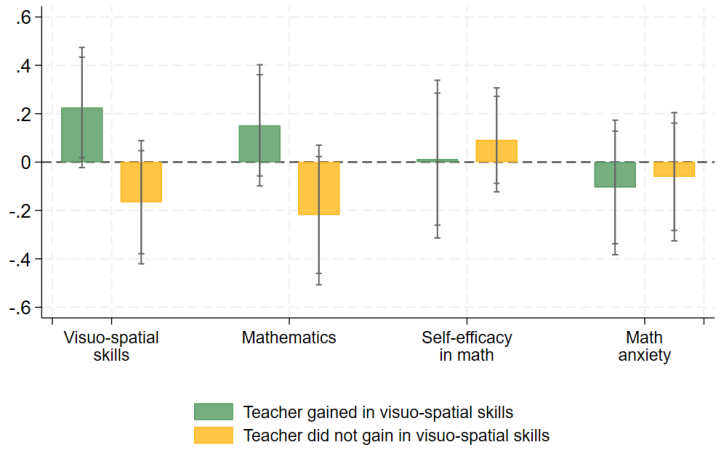
between the treatment variable and a measure of heterogeneity, which identifies an important characteristic of the teacher.

The first aspect to consider is whether the effect of the treatment on students varies depending on the improvement in the teachers' visuo-spatial skills. In other words, we consider whether students who had teachers whose scores in the visuo-spatial test improved due to the treatment benefited more from the intervention. It is likely that teachers who did not improve were less effective in their teaching or made less use of the building blocks. In this case, we can also expect that the experience of this treatment subgroup was more similar to that of the control group, resulting in smaller effect sizes. In this model specification, the focus is on the interaction between the treatment and an indicator identifying teachers whose scores in the visuo-spatial tests increased between baseline and endline.

The results presented in Figure 2 show that the estimated coefficient for this interaction is 0.226, suggesting a positive effect on the visuo-spatial skills of students whose teachers improved their own skills. For students assigned to teachers who did not show improvement or experienced a deterioration in their visuo-spatial skills, the estimated coefficient is -0.166, a result suggesting a possible negative relationship, although not statistically significant. For mathematics, the coefficient for students in classes taught by teachers who improved their visuo-spatial skills is 0.152, while the one associated with teachers who worsened is -0.219. In this case, the results are not statistically significant, although they suggest a similar direction to those on visuo-spatial skills. As for self-efficacy in mathematics and anxiety about mathematics, the estimated coefficients are of negligible magnitude and lack statistical significance. However, these last two outcome variables measure specific personality traits of the students, and it is plausible that these may require more time before any changes can be observed.

Furthermore, as suggested by Figure 3, the positive effect observed on students seems to be primarily driven by the improvement in the teachers' scores on the spatial orientation test (shown on the left), while no significant effects are found associated with the scores on the mental rotation test (shown on the right). This finding is consistent with what emerged in the analysis of the effects on the teachers, which highlighted how the training received by them focused predominantly on spatial orientation skills rather than mental rotation.

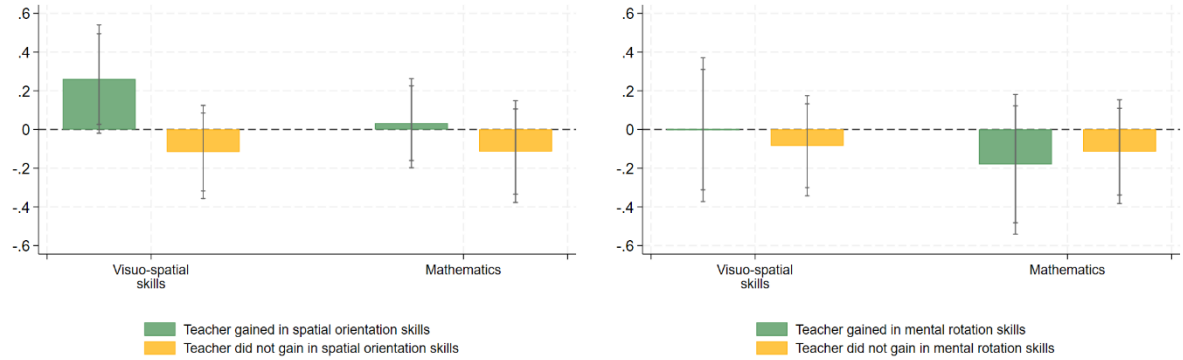
Figure 2. Effect of the treatment on the students, by gain in visuo-spatial skills of the teacher



Further exploring the possibility of differential effects by gender, the results suggest that the impact of the improvement in teachers' visuo-spatial skills varies between male and female students. Specifically, for female students assigned to teachers who showed an improvement in their scores, a positive effect of nearly 0.25 standard deviations is observed on their visuo-spatial skills. For male students, on the other hand, a positive and statistically significant impact of 0.417 standard deviations is observed on their mathematical scores.

Overall, these results suggest that while female students seem to benefit primarily in terms of visuo-spatial skills, male students show an improvement in cognitive skills. This heterogeneity between males and females could be explained by the presence of threshold effects and non-linear returns of visuo-spatial skills on cognitive development in mathematics, with the threshold being already surpassed on average by boys but not by girls. The data collected before the intervention began indicate that boys, compared to girls, use the blocks more frequently at home and have higher baseline visuo-spatial and mathematical skills. In this case, the result seems to suggest that girls may require a greater intensity of the treatment or a longer duration before seeing effects on their mathematical skills as well.

Figure 3. Effect of the treatment on the students, by gain in spatial orientation and mental rotation skills of the teacher



In addition to the improvement of teachers, we also examine whether other activities conducted in the classrooms may have played a role in the effectiveness of the intervention. In recent years, Italian schools have been involved in an increasing number of educational projects in very diverse learning areas that could have influenced the results. The examination of the heterogeneity of the treatment effect in relation to the adoption of other teaching methods or participation in alternative projects by the teachers suggests the presence of a possible displacement effect, i.e., a negative interaction between the treatment and adherence to other programs or educational strategies. This effect is observable only on visuo-spatial skills scores, while there is no effect on mathematical skills.

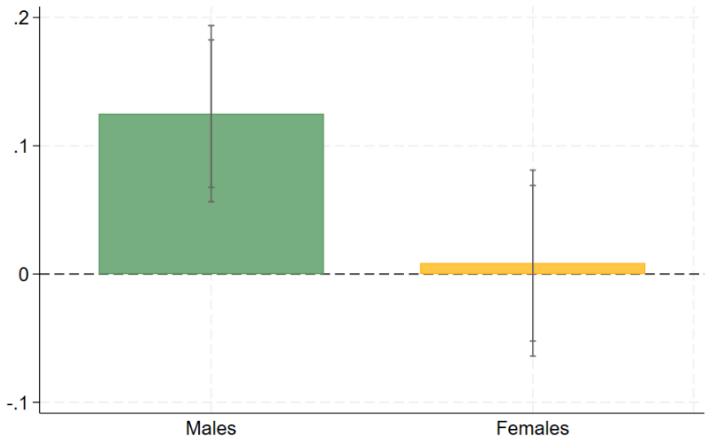
Further analysis exploits teachers' results on the Implicit Association Test (IAT) measuring implicit gender biases. We estimate heterogeneous treatment effects separately for teachers who exhibit an automatic association (slight, moderate, or strong) between Male–Science and Female–Arts, those who display no automatic association between gender and science, and those with an automatic association (slight, moderate, or strong) between Male–Arts and Female–Science.

The results indicate a positive effect of the intervention on female students taught by treated teachers who initially exhibited an implicit Male–Science and Female–Arts association. One possible interpretation is that, in contexts where unconscious gender stereotypes were stronger, the intervention may have increased teachers' awareness of differential classroom behaviors between boys and girls, or led teachers to devote greater attention to girls during the use of the building blocks. This finding is consistent with Alesina et al. (2024), who document a reduction in immigrant–native bias when teachers are made aware of their own stereotypes. However, given data limitations, we refrain from drawing stronger conclusions and leave further investigation of these mechanisms to future research. Although no significant differences are observed among teachers regarding these characteristics, we also considered in the heterogeneity analysis: the teachers' perceived self-efficacy, years of service, and the number of hours per week dedicated to mathematics. In all these cases, no significant differential effects seem to emerge.

Last, the information collected on students allows to explore one of the mechanisms that, as expected, links teacher training to student outcome variables. Figure 4 shows the effect of the treatment on the use of the bricks at home. The estimate is also conditioned on the children's habits

of using the bricks before the intervention, meaning that the effect reflects a relative change within individual participants, accounting for initial differences. The analysis reveals a positive and significant treatment effect, but only for boys. For them, the results indicate that the program increases the likelihood of playing with the bricks at home compared to the control group, with an increase of 12.5 percentage points. To better understand this value, consider that at baseline, 70% of boys reported playing with the bricks at home, compared to 64% of girls, with a statistically significant difference of 6 percentage points. The estimates obtained imply that, thanks to the treatment, the likelihood that boys play with the bricks increases by almost 18% compared to the baseline.

Figure 4. Effect of the treatment on the use of bricks at home



In this context, the positive effect for boys is particularly relevant because it offers insights into the potential impacts of this change on cognitive development pathways. If the treatment indeed increased the use of the bricks, and since the use of bricks is positively correlated with cognitive outcomes, it is plausible that this variation could have long-term positive effects on visuo-spatial and mathematical skills. The absence of similar effects for girls could instead be related to the lower availability of bricks at home, which in many families might be limited, or to their lower propensity to use them. It is not possible to distinguish between these hypotheses here, but this may have reduced the opportunities for girls to benefit from the intervention.

5. CONCLUSIONS

This paper provides experimental evidence on the effectiveness of an educational intervention designed to strengthen visuo-spatial and mathematical skills in primary school students through construction play activities and complementary teacher training.

The empirical analysis suggests that the impact of the Matabì program on students' visuo-spatial and mathematical skills is mediated by the effectiveness of the teacher training and by the frequency with which students use the bricks.

In the program’s first edition (school year 2022-23), classes treated from the very start of the year exhibited slightly larger effects than those that joined only in the second semester. This evidence suggests the importance of the treatment duration in determining its effectiveness. This result may also reflect the in-person teacher training delivered in the first semester, suggesting that the mode of delivery matters.

The 2023-24 edition, in which teacher training moved online and lasted the entire school year, provides further insights. Overall treatment effects are negligible, consistent with the view that online training is less effective than face-to-face instruction. However, students whose teacher registered gains on the visuo-spatial test, especially on spatial-orientation items, showed positive impacts. This

suggests that teachers' enhanced abilities in these areas have a direct benefit on students, consistent with the fact that the treatment effect on the students is mediated by the teachers.

Future research based on administrative standardized test scores will provide additional insights into the effectiveness of the program in shaping students' cognitive skills and narrowing the gender gap in mathematics.

REFERENCES

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.

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.

Alesina, A., Carlana, M., La Ferrara, E., & Pinotti, P. (2024). Revealing stereotypes: Evidence from immigrants in schools. American Economic Review, 114(7), 1916–1948.

Barnes, M. A., & Raghubar, K. P. (2014). Mathematics development and difficulties: The role of visual–spatial perception and other cognitive skills. Pediatric Blood & Cancer, 61(7), 1195–1202.

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

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. The Quarterly Journal of Economics, 134(3), 1163–1224.

Carneiro, P. M., & Heckman, J. J. (2003). Human capital policy. NBER Working Paper No. 9495.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. American Economic Review, 104(9), 2633–2679.

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.

Currie, J., & Almond, D. (2011). Human capital development before age five. In O. Ashenfelter & D. Card (Eds.), Handbook of Labor Economics (Vol. 4, pp. 1315–1486). Elsevier.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. Journal of Human resources, 42(3), 528-554.

Di Tommaso, M. L., Contini, D., De Rosa, D., Ferrara, F., Piazzalunga, D., & Robutti, O. (2024). Tackling the gender gap in mathematics with active learning methodologies. Economics of Education Review, 100, 102538.

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.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), 103–127. https://doi.org/10.1037/a0018053

Ferguson, A. M., Maloney, E. A., Fugelsang, J., & Risko, E. F. (2015). On the relation between math and spatial ability: The case of math anxiety. Learning and Individual Differences, 39, 1–12. https://doi.org/10.1016/j.lindif.2015.02.007

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

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.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. Educational Researcher, 49(4), 241–253.

Lavy, V., & Megalokonomou, R. (2024). The short- and the long-run impact of gender-biased teachers. American Economic Journal: Applied Economics, 16(2), 176–218.

Machin, S., & Pekkarinen, T. (2008). Global sex differences in test score variability. Science, 322(5906), 1331–1332. https://doi.org/10.1126/science.1162573

Martinot, P., Colnet, B., Breda, T., Sultan, J., Touitou, L., Huguet, P., Spelke, E., Dehaene-Lambertz, G., Bressoux, P., & Dehaene, S. (2025). Rapid emergence of a maths gender gap in first grade. Nature, 643, 1020–1029.

Murphy, R. (1982). Sex differences in objective test performance. British Journal of Educational Psychology, 52(2), 213–219. https://doi.org/10.1111/bjep.1982.52.issue-2

OECD. (2015). The ABC of gender equality in education (PISA). OECD Publishing.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. Econometrica, 73(2), 417–458.

Stobart, G., Elwood, J., & Quinlan, M. (1992). Gender bias in examinations: How equal are the opportunities? British Educational Research Journal, 18(3), 261–276.

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.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. Psychological Bulletin, 139(2), 352–402.

Wolfgang, C. H., Stannard, L. L., & Jones, I. D. (2001). Block play performance among preschoolers as a predictor of later school achievement in mathematics. Early Child Development and Care, 166(1), 33–41.

APPENDIX

Full results: 2022-23 edition

Table A 1. Effect of the treatment on the visuo-spatial abilities of students

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment at 1st semester	0.177**	0.177**	0.172	0.178*	0.228*	0.372**	0.013
	(0.073)	(0.073)	(0.104)	(0.101)	(0.118)	(0.137)	(0.178)
Treatment at 2nd semester	0.041	0.042	0.127	0.132	0.045	0.204	-0.017
	(0.069)	(0.069)	(0.087)	(0.088)	(0.118)	(0.219)	(0.164)
Grade 3		0.006	-0.562***	-0.514***	-0.511***		
		(0.059)	(0.087)	(0.082)	(0.080)		
Female				-0.389***	-0.410***	-0.535***	-0.270*
				(0.058)	(0.084)	(0.096)	(0.132)
Treatment at 1st sem. * Female					-0.099	-0.109	-0.136
					(0.101)	(0.114)	(0.158)
Treatment at 2nd sem. * Female					0.162	0.253	-0.001
					(0.149)	(0.316)	(0.169)
Migration background (baseline)				-0.149*	-0.146*	-0.193	-0.081
				(0.075)	(0.073)	(0.119)	(0.100)
Maths score (baseline)			0.515***	0.489***	0.484***	0.414***	0.604***
			(0.046)	(0.043)	(0.044)	(0.061)	(0.054)
Observations	863	863	863	863	863	363	500
R-squared	0.080	0.080	0.246	0.288	0.291	0.290	0.330
Grade	3&4	3&4	3&4	3&4	3&4	3	4
School fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Robust standard errors in parentheses, clustered at class level; *** p<0.01, ** p<0.05, * p<0.1. Variables “migration background” and “maths score” are measured at baseline.							

Table A 2. Effect of the treatment on the score in mathematics of students

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment at 1st semester	0.053	0.044	0.039	0.039	0.038	-0.028	0.008
	(0.143)	(0.104)	(0.099)	(0.096)	(0.098)	(0.116)	(0.160)
Treatment at 2nd semester	-0.149	-0.209*	-0.110	-0.108	-0.211	-0.151	-0.196
	(0.147)	(0.105)	(0.106)	(0.106)	(0.151)	(0.179)	(0.213)
Grade 3		-0.530***	-1.192***	-1.170***	-1.166***		
		(0.097)	(0.109)	(0.107)	(0.106)		
Female				-0.240***	-0.306***	-0.384***	-0.213
				(0.062)	(0.099)	(0.123)	(0.149)
Treatment at 1st sem. * Female					0.002	0.160	-0.154
					(0.120)	(0.151)	(0.168)
Treatment at 2nd sem. * Female					0.196	0.259	0.089
					(0.169)	(0.210)	(0.234)
Migration background (baseline)				0.013	0.017	-0.004	0.115
				(0.063)	(0.064)	(0.073)	(0.112)
Maths score (baseline)			0.600***	0.592***	0.588***	0.416***	0.795***
			(0.046)	(0.045)	(0.045)	(0.040)	(0.050)
Observations	863	863	863	863	863	363	500
R-squared	0.123	0.187	0.413	0.427	0.429	0.365	0.433
Grade	3&4	3&4	3&4	3&4	3&4	3	4
School fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Robust standard errors in parentheses, clustered at class level; *** p<0.01, ** p<0.05, * p<0.1. Variables “migration background” and “maths score” are measured at baseline.							

Table A 3. Effect of the treatment on the students’ self-efficacy in mathematics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)

Treatment at 1st semester	0.099	0.105	0.103	0.097	0.021	0.079	-0.109
	(0.127)	(0.116)	(0.110)	(0.111)	(0.128)	(0.200)	(0.153)
Treatment at 2nd semester	0.029	0.068	0.119	0.116	0.092	0.240	0.100
	(0.152)	(0.136)	(0.136)	(0.138)	(0.173)	(0.183)	(0.212)
Grade 3		0.309***	0.001	0.001	0.000		
		(0.094)	(0.109)	(0.108)	(0.108)		
Female				-0.151**	-0.216**	-0.133	-0.286***
				(0.070)	(0.100)	(0.150)	(0.089)
Treatment at 1st sem. * Female					0.150	-0.087	0.337**
					(0.129)	(0.187)	(0.122)
Treatment at 2nd sem. * Female					0.048	-0.286	0.201
					(0.188)	(0.236)	(0.226)
Migration background (baseline)				0.206***	0.208***	0.184	0.288***
				(0.075)	(0.075)	(0.133)	(0.094)
Maths score (baseline)			0.283***	0.293***	0.294***	0.213***	0.409***
			(0.042)	(0.043)	(0.042)	(0.066)	(0.059)
Observations	825	825	825	825	825	344	481
R-squared	0.039	0.061	0.110	0.123	0.124	0.104	0.156
Grade	3&4	3&4	3&4	3&4	3&4	3	4
School fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Robust standard errors in parentheses, clustered at class level; *** p<0.01, ** p<0.05, * p<0.1. Variables “migration background” and “maths score” are measured at baseline.							

Table A 4. Effect of the treatment on the students’ math anxiety

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment at 1st semester	-0.076	-0.076	-0.074	-0.079	-0.097	-0.083	-0.115
	(0.084)	(0.084)	(0.084)	(0.082)	(0.124)	(0.164)	(0.154)
Treatment at 2nd semester	-0.056	-0.059	-0.095	-0.099	-0.121	-0.137	-0.139
	(0.085)	(0.086)	(0.087)	(0.086)	(0.117)	(0.155)	(0.170)
Grade 3		-0.026	0.217**	0.178**	0.179**		
		(0.068)	(0.081)	(0.079)	(0.079)		
Female				0.317***	0.290***	0.360**	0.226
				(0.061)	(0.108)	(0.127)	(0.170)
Treatment at 1st sem. * Female					0.036	0.058	0.035
					(0.144)	(0.180)	(0.218)
Treatment at 2nd sem. * Female					0.043	0.045	0.094
					(0.154)	(0.250)	(0.211)
Migration background (baseline)				0.110	0.112	0.197	0.065
				(0.107)	(0.107)	(0.158)	(0.155)
Maths score (baseline)			-0.220***	-0.199***	-0.200***	-0.179***	-0.247***
			(0.042)	(0.042)	(0.041)	(0.054)	(0.064)
Observations	863	863	863	863	863	363	500
R-squared	0.037	0.038	0.068	0.095	0.095	0.116	0.097
Grade	3&4	3&4	3&4	3&4	3&4	3	4
School fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Robust standard errors in parentheses, clustered at class level; *** p<0.01, ** p<0.05, * p<0.1. Variables “migration background” and “maths score” are measured at baseline.							

Full results: 2023-24 edition

Table A 5. Effect of the treatment on the visuo-spatial abilities of students

	(1)	(2)	(3)	(4)	(5)	(6)	(7)

Treatment	0.073	0.035	-0.009	-0.029	-0.055	-0.085	-0.043
	(0.090)	(0.079)	(0.069)	(0.067)	(0.095)	(0.082)	(0.101)
Grade 3		-0.126	-0.260***	-0.244***	-0.243***		
		(0.097)	(0.082)	(0.081)	(0.081)		
Female				-0.201***	-0.230**	-0.270***	-0.025
				(0.060)	(0.088)	(0.069)	(0.110)
Treatment * Female					0.050		
					(0.121)		
Migration background (baseline)				-0.092	-0.091	-0.070	-0.056
				(0.098)	(0.099)	(0.122)	(0.127)
Play w\ bricks at home (baseline)				0.206***	0.207***	0.191***	0.232*
				(0.051)	(0.052)	(0.056)	(0.120)
Maths score (baseline)			0.327***	0.314***	0.313***	0.315***	0.323***
			(0.029)	(0.027)	(0.028)	(0.032)	(0.053)
Observations	974	974	974	974	974	713	261
R-squared	0.001	0.049	0.147	0.168	0.168	0.171	0.224
Grade	3&4	3&4	3&4	3&4	3&4	3 only	4 only
School fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Robust standard errors in parentheses, clustered at class level; *** p<0.01, ** p<0.05, * p<0.1. Variables “migration background”, “play w\ bricks at home” and “maths score” are measured at baseline.							

Table A 6. Effect of the treatment on the score in mathematics of students

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	0.068	0.057	-0.003	-0.018	0.067	-0.052	-0.063
	(0.111)	(0.088)	(0.072)	(0.072)	(0.083)	(0.088)	(0.141)
Grade 3		0.170	-0.017	-0.007	-0.010		
		(0.110)	(0.092)	(0.091)	(0.091)		
Female				-0.181***	-0.085	-0.272***	0.018
				(0.060)	(0.094)	(0.068)	(0.109)
Treatment * Female					-0.170		
					(0.119)		
Migration background (baseline)				-0.131	-0.133	-0.061	-0.217
				(0.091)	(0.091)	(0.109)	(0.152)
Play w\ bricks at home (baseline)				0.119*	0.115*	0.121	0.091
				(0.061)	(0.061)	(0.076)	(0.096)
Maths score (baseline)			0.456***	0.447***	0.449***	0.482***	0.370***
			(0.032)	(0.031)	(0.031)	(0.039)	(0.039)
Observations	971	971	971	971	971	710	261
R-squared	0.001	0.088	0.276	0.290	0.292	0.276	0.373
Grade	3&4	3&4	3&4	3&4	3&4	3 only	4 only
School fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Robust standard errors in parentheses, clustered at class level; *** p<0.01, ** p<0.05, * p<0.1. Variables “migration background”, “play w\ bricks at home” and “maths score” are measured at baseline.							

Table A 7. Effect of the treatment on the students’ self-efficacy in mathematics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	0.039	0.072	0.043	0.022	0.008	0.100	-0.188
	(0.086)	(0.070)	(0.074)	(0.073)	(0.092)	(0.085)	(0.125)
Grade 3		0.275***	0.181*	0.197*	0.198*		
		(0.092)	(0.098)	(0.100)	(0.100)		
Female				-0.300***	-0.316***	-0.338***	-0.096
				(0.056)	(0.085)	(0.061)	(0.111)

Treatment * Female					0.029		
					(0.112)		
Migration background (baseline)				0.033	0.034	0.084	-0.125
				(0.118)	(0.117)	(0.144)	(0.231)
Play w\ bricks at home (baseline)				0.083	0.084	0.111	0.068
				(0.082)	(0.082)	(0.101)	(0.130)
Maths score (baseline)			0.227***	0.218***	0.218***	0.172***	0.385***
			(0.035)	(0.035)	(0.035)	(0.048)	(0.033)
Observations	972	972	972	972	972	711	261
R-squared	0.000	0.038	0.083	0.106	0.106	0.090	0.184
Grade	3&4	3&4	3&4	3&4	3&4	3 only	4 only
School fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Robust standard errors in parentheses, clustered at class level; *** p<0.01, ** p<0.05, * p<0.1. Variables “migration background”, “play w\ bricks at home” and “maths score” are measured at baseline.							

Table A 8. Effect of the treatment on the students’ math anxiety

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	-0.002	-0.002	0.015	0.009	0.006	0.065	0.007
	(0.099)	(0.077)	(0.077)	(0.077)	(0.112)	(0.095)	(0.126)
Grade 3		-0.043	0.007	0.016	0.017		
		(0.096)	(0.096)	(0.094)	(0.094)		
Female				-0.106	-0.109	-0.149*	0.042
				(0.064)	(0.104)	(0.076)	(0.115)
Treatment * Female					0.005		
					(0.133)		
Migration background (baseline)				0.332***	0.332***	0.331***	0.372*
				(0.088)	(0.089)	(0.104)	(0.189)
Play w\ bricks at home (baseline)				-0.050	-0.050	-0.061	0.038
				(0.078)	(0.079)	(0.090)	(0.172)
Maths score (baseline)			-0.123***	-0.122***	-0.122***	-0.145***	-0.062
			(0.034)	(0.033)	(0.033)	(0.036)	(0.071)
Observations	970	970	970	970	970	710	260
R-squared	0.000	0.065	0.079	0.091	0.091	0.101	0.123
Grade	3&4	3&4	3&4	3&4	3&4	3 only	4 only
School fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Robust standard errors in parentheses, clustered at class level; *** p<0.01, ** p<0.05, * p<0.1. Variables “migration background”, “play w\ bricks at home” and “maths score” are measured at baseline.							

Table A 9. Effect of the treatment on the students’ propensity to play with the bricks at home

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	0.078*	0.087***	0.081**	0.067***	0.125***	0.052*	0.101
	(0.040)	(0.031)	(0.031)	(0.025)	(0.035)	(0.028)	(0.063)
Grade 3		0.002	-0.014	0.000	-0.002		
		(0.049)	(0.048)	(0.041)	(0.041)		
Female				0.055**	0.121***	0.052*	0.061
				(0.025)	(0.037)	(0.030)	(0.045)
Treatment * Female					-0.117**		
					(0.051)		

Migration background (baseline)				-0.052	-0.053	-0.028	-0.120
				(0.046)	(0.045)	(0.051)	(0.109)
Play w\ bricks at home (baseline)				0.369***	0.366***	0.370***	0.377***
				(0.037)	(0.037)	(0.045)	(0.060)
Maths score (baseline)			0.041**	0.026*	0.027*	0.018	0.033
			(0.016)	(0.015)	(0.015)	(0.016)	(0.031)
Observations	974	974	974	974	974	713	261
R-squared	0.007	0.054	0.061	0.205	0.209	0.213	0.210
Grade	3&4	3&4	3&4	3&4	3&4	3 only	4 only
School fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Robust standard errors in parentheses, clustered at class level; *** p<0.01, ** p<0.05, * p<0.1. Variables “migration background”, “play w\ bricks at home” and “maths score” are measured at baseline.							