ISSN 2282-6483

Alma Mater Studiorum - Università di Bologna
DEPARTMENT OF ECONOMICS

Prizes, Participation, and Gender Stereotypes in a Math Olympiad: Evidence from a Field Experiment

Federica Esposito
Valeria Maggian
Irene Mammi
Natalia Montinari
Antonio Nicolò

Quaderni - Working Paper DSE N°1235



Prizes, Participation, and Gender Stereotypes in a Math Olympiad: Evidence from a Field Experiment*

Federica Esposito† Valeria Maggian‡ Irene Mammi§ Natalia Montinari¶
Antonio Nicolò‖

Abstract

School mathematics competitions combine voluntary entry with competition in a domain central to STEM, making them an early setting in which gender gaps emerge. In this paper, we investigate whether monetary incentives increase participation and narrow the gap in a field experiment during the first stage of the Italian Math Olympiad. Thirty high schools were randomly assigned to no prize, a prize for the school's top four students, or separate prizes for the top two girls and boys. We link competition records to questionnaires collecting data on students' ability, beliefs, preferences, and non-cognitive traits completed by more than 7,700 students. In unadjusted comparisons, neither prize increases participation. Accounting for chance imbalances in school entry rules, which did not respond to treatment, reveals higher participation under the No Gender prize and a smaller, less robust increase under the Gender prize in schools where entry relies on students' initiative. Participants in treated schools score lower on average, consistent with the prizes attracting lower-scoring entrants. Despite roughly doubling girls' expected prize returns, the Gender prize does not elicit a significantly larger participation response from girls; the 8-percentage-point gap at equal measured ability persists across treatments.

Importantly, the gap ranges from about 2 points among students with weak male–science associations, measured by the Implicit Association Test, to 15 points among those with strong associations. While this relationship is descriptive, its persistence across prize schemes suggests that narrowing the participation gap may require addressing students' beliefs and gender stereotypes, rather than relying on monetary incentives alone.

Keywords: field experiment; math olympiad; gender gap; competition; monetary incentives; implicit stereotypes.

JEL codes: C93, D91, I21, J16.

*We are grateful to the Unione Matematica Italiana (UMI) and to the organizing committee of the Math Olympiad for facilitating our connections with the schools included in the project, and in particular to Luigi Amedeo Bianchi for his great support in gathering the data. Ethical approval for this study was obtained from the ethics committee of the University of Venice on 19.09.2019, Verbale n. 3/2019. This research has benefited from financial support from the Einaudi Institute for Economics and Finance (EIEF) and from Project SID 2019, titled "Mind the gender gap. On the design of affirmative actions to reduce the gender gap in the labor market", Department of Economics and Management, University of Padua. This project has also received funding from the Venice Centre in Economic and Risk Analytics for Public Policies (VERA) at the Department of Economics, Ca' Foscari University. We also thank Bianca Navoni, Ivana Pasciuta and Noè Somma for their invaluable help in implementing the project. Both Ivana Pasciuta and Noè Somma used a subset of the data collected in this study to write their master's theses. We are grateful to Fondazione Agnelli for providing us with data about the high schools and to all principals and teachers of schools involved in this research for their collaboration in data collection.

†Dept. of Economics, University of Bologna. E-mail: federica.esposito12@unibo.it

‡Dept. of Economics, Ca' Foscari University of Venice. E-mail: valeria.maggian@unive.it

§Dept. of Economics, Ca' Foscari University of Venice. E-mail: irene.mammi@unive.it

¶Dept. of Economics, University of Bologna. E-mail: natalia.montinari2@unibo.it

‖Dept. of Economics and Management, University of Padova.

Non-technical abstract

Girls take part in mathematics competitions less often than boys, even when they have similar ability. These differences emerge during school and may matter for later educational choices and careers in science, technology, engineering, and mathematics. Can monetary prizes encourage more students to enter these competitions? And can prizes reserved for girls help reduce the gender gap in participation?

We address these questions through a field experiment in 30 high schools in North-East Italy during the first stage of the 2019/2020 Italian Math Olympiad. Schools were randomly assigned to three groups. In the first group, students received no monetary prizes. In the second, prizes rewarded the four highest-scoring students in each school, regardless of gender. In the third, prizes rewarded the two highest-scoring girls and the two highest-scoring boys. Both prize schemes offered the same total amount of money, with individual rewards of €100 or €200. We combined competition records with questionnaires completed by more than 7,700 students, including both participants and non-participants. The questionnaires collected information on school grades, reasoning ability, confidence, attitudes towards competition, and gender stereotypes.

Our findings show that the role of monetary prizes depends on how schools organize participation. Some schools leave the decision to students, others encourage selected students to enter, and others enroll entire classes. A simple comparison across the three experimental groups shows no statistically clear increase in participation under either prize scheme. However, once we account for differences in schools' entry rules, participation is higher in schools offering prizes. Among students who decide for themselves whether to enter, the increase is similar under the two prize schemes: about 12 percentage points with our stricter method of linking questionnaires to competition records, and about half of that with a broader method. Prizes awarded regardless of gender show an additional and larger increase, but it comes mainly from classes enrolled as a whole. Average competition scores are lower among participants in schools offering prizes, but only where students choose whether to enter. This pattern is consistent with prizes attracting additional students who obtain lower scores, rather than reducing the performance of students who would have entered anyway. Neither prize scheme produces a clear reduction in the gender gap. Reserving two prizes for girls substantially improves their chances of winning, but we find no statistically clear evidence that girls respond more strongly than boys. Girls remain around eight percentage points less likely to participate than boys with the same measured reasoning ability. This gap does not reflect lower measured ability: girls and boys perform similarly on the reasoning test, and girls have higher school grades, including in mathematics. Yet girls tend to underestimate their position relative to classmates, while boys tend to overestimate theirs.

Gender stereotypes are closely related to the gender gap in participation. We measure how strongly students associate science with males rather than females through a word-sorting task. Stronger associations go with higher participation among boys and lower participation among girls. The estimated participation gap is about two percentage points among students with weak associations and about fifteen points among those with strong associations. This relationship is smaller but still present after accounting for ability, grades, and confidence, and the prizes do not change it. It is an association, however, and does not establish that stereotypes cause the gap.

Overall, monetary prizes may encourage participation when students are free to choose, but our results provide no clear evidence that they narrow the gender gap. They suggest that efforts to encourage girls' participation should also consider students' beliefs about their abilities and the stereotypes surrounding mathematics and science.

1 Introduction

Women remain under-represented in high-paying occupations, contributing to persistent gender gaps in careers and earnings (OECD, 2025). Part of the gap may originate well before labor-market entry, through differences in educational and career choices. In particular, choosing a more math-intensive high-school track is an important predictor of subsequent earnings (Joensen and Nielsen, 2009), while attitudes toward competition have been linked to industry choice and starting salaries among MBA students, expected and realized labor-market earnings, participation in selective entrance examinations, and entrepreneurial investment decisions (Zhang, 2013; Berge et al., 2015; Reuben et al., 2015, 2017; Buser et al., 2018). A large experimental literature suggests that gender differences in self-selection into competitive environments may therefore be one mechanism through which broader labor-market inequalities emerge: on average, women enter competitions less often than men of the same ability, and they hold less optimistic beliefs about their own performance (Niederle and Vesterlund, 2007; Croson and Gneezy, 2009).

Understanding the mechanisms through which these differences emerge during school, and whether they can be influenced by institutional design, is therefore an important step toward reducing gender disparities in STEM, with potential implications for later labor market outcomes.

School mathematics competitions provide a natural setting in which to study this question: participation is in principle voluntary, students must actively choose whether to enter a competitive environment, and the domain is mathematics. Moreover, these competitions represent an early stage of the STEM pipeline: students who participate and persist are more likely to continue into math-intensive educational tracks (Ellison and Swanson, 2018; Buser and Yuan, 2019). Importantly, they also provide a setting in which the incentives to enter competition can be directly manipulated. We exploit this feature to focus on one specific design dimension: monetary incentives. Specifically, in this paper we ask two questions. First, do monetary prizes raise participation and performance in a real math competition? Second, does the design of the prize matter for the gender gap?

We ran a field experiment during the *Giochi di Archimede*, the first, school-level stage of the 2019/2020 Italian Math Olympiad. Thirty high schools in the North-East of Italy were randomly assigned to one of three treatments: a control group with no monetary prizes, a standard prize scheme rewarding the four best-performing students within the school regardless of gender, and a gender-specific prize scheme rewarding the two best-performing girls and the two best-performing boys within the school. Under the gender-contingent prize, girls and boys sit the same competition and the Olympiad ranking remains common; what changes is prize eligibility, which is determined within gender. This scheme is the field counterpart of the affirmative action prize schemes studied in the laboratory, where they raise the entry of women into competition (Niederle and Vesterlund, 2010; Berger et al., 2020). Whether this carries over to a real competition, with real stakes and a school context, is an open empirical question.

A further strength of our design is the availability of detailed survey data. Specifically, a large sample of students in all schools completed a detailed questionnaire covering family background,

grades, cognitive ability measured using Raven's (Bilker et al., 2012), beliefs about own performance, risk and time preferences, competitiveness, and non-cognitive traits. The questionnaire was administrated to students as a separate activity from the Math Olympiad and from the prizes. We link survey responses to individual Olympiad records, allowing us to observe a rich set of individual characteristics for both students who participate in the competition and those who do not. This is a key asset of our design, as it allows us to ask not only whether incentives work, but also for whom they work.

An important feature of our setting is the way schools enroll students in the competition. Participation is open to all students, but schools organize it in different ways. In some schools, teachers encourage the best-performing students to take part; in others, entire classes are enrolled. We refer to these arrangements as the school's entry rule and record it for every school. The entry rule matters for two reasons. First, it determines how much room the prizes have to change participation: in classes enrolled as a whole, almost everyone participates and a prize can change little. Second, entry rules were unbalanced across treatments, with this imbalance being more likely to reflect chance variation in the random assignment rather than a response to the prizes (Section 4.4 shows that the entry rules did not respond to the prizes). Consistent with this interpretation, mandatory enrollment is most common in the Baseline schools, and the few changes in entry rules relative to the previous edition of the competition do not differ across treatments. We account for entry rules throughout the analysis.

We find three main results. First, there is no statistically significant average increase in participation. In the raw comparison between treated and Baseline schools, neither prize significantly raises participation. After accounting for entry rules, participation is higher in treated schools where entry relies on students' own initiative rather than on teacher invitations or class-level enrollment. The estimated increase is about 12 percentage points under the more restrictive match between questionnaire respondents and Olympiad records, and about 6 points under the broader match (the matching procedures are described in section 4.3). Second, the female participation gap remains about 8 percentage points at equal measured ability in every treatment, with no statistically significant difference in girls' and boys' responses to the prizes. The gap is not mirrored in measured ability: girls and boys have identical Raven scores, and girls have higher grades, including in mathematics. Yet girls expect to rank below their actual position in the within-class distribution of Raven scores, whereas boys expect to rank above theirs. Girls also report lower competitiveness and are less likely to have participated in the Olympiad before. Third, among participants, average scores are lower in treated schools. This happens only where students choose whether to participate, not in classes enrolled as a whole. The pattern is consistent with the prizes attracting additional students who score lower on average, rather than lowering the scores of students who would have participated anyway. In line with this reading, the effect of the prizes on total points earned, counting zero for non-participants, is small and imprecise.

The questionnaire also tells us who participates and who responds to the prizes. Participation is strongly predicted by academic background: cognitive ability, math grades and, above all, having

taken part in the Olympiad before. The response to the prizes is instead uniform: it does not vary with ability, confidence, past participation, competitiveness, or the other traits we measure, with only weak evidence of a larger response among students with higher math grades. The prizes draw entrants from a pool that is already selected on academic background, and they do not change who responds within it. The questionnaire also documents a striking pattern related to the gender gap. Students completed a Gender-Science Implicit Association Test (IAT; Greenwald et al. (1998); Nosek et al. (2002)), which measures how strongly they associate science with males rather than females. Stronger male-science associations are linked to higher participation among boys and lower participation among girls. As a result, the gender gap in participation is about 2 percentage points, and not statistically significant, among students with weak associations (the 10th percentile of the IAT score), and about 15 percentage points among students with strong associations (the 90th percentile). The association is smaller, but still present, when we control for cognitive ability, math grades, and performance beliefs. The prizes do not change it.

The paper is organized as follows. Section 2 presents the related literature. Section 3 describes the field experiment. Section 4 presents the data, the matching procedure, and the variables. Section 5 describes the empirical strategy. Section 6 reports the results, including the free margin analysis. A final section discusses the findings.

2 Related literature

Our paper contributes to the literature on gender differences in competitive entry by studying whether changes in the design of competitive environments can reduce the gender gap in participation. We do so in the context of school mathematics competitions, which also connects our study to the literature on mathematics competitions and the STEM pipeline. Finally, our rich survey data allow us to relate students' responses to a broad set of individual measures, contributing to a growing strand of research on the role of performance beliefs, implicit gender stereotypes, and other individual characteristics that have been proposed as determinants of competitive behavior.

School mathematics competitions have received increasing attention as a potential gateway into STEM. Ellison and Swanson (2018) use data from the American Mathematics Competitions and show that the gender gap among top performers is already present in ninth grade and grows during high school: girls improve less from year to year and drop out more often, especially after a near miss. Buser and Yuan (2019) find a similar pattern in the Dutch Math Olympiad, in the lab and in the field: after losing, girls are less willing than boys to compete again. Extending the focus beyond participation, Yuret (2026) examines the long-run educational and career trajectories of medalists in the European Girls' Mathematical Olympiad. Comparing female medalists with male participants in the International Mathematical Olympiad, they find persistent gender differences in subsequent educational and research careers despite similar levels of mathematical talent. Taken together, these studies show that gender differences emerge early in competitive mathematics, persist over time, and are associated with different educational and career trajectories. Our paper

addresses a complementary question. Rather than documenting the evolution or consequences of these differences, we ask whether they can be influenced at their origin by changing the design of the competitive environment. Specifically, we study whether alternative monetary incentive schemes affect students' initial decision to enter a mathematics competition and whether they reduce the gender gap in participation.

This question connects to a growing literature studying how the design of competitive environments shapes behavior. In the laboratory, Berger et al. (2020) show that selecting winners by lottery among the top performers almost closes the gender gap in tournament entry, without lowering the quality of the selected group, while Brandts et al. (2021) show that making performance goals public raises the goals people set but hurts the performance of women. In the field, experiments have instead focused on affirmative action policies implemented in incentivized mathematics contests created by the researchers. Calsamiglia et al. (2013) favor children from disadvantaged schools in a real-effort tournament, and Cotton et al. (2022) favor disadvantaged demographic groups in paid math exams. Our paper extends this line of research by studying alternative prize structures within an existing, large-scale mathematics competition. We compare a standard prize scheme with a gender-specific prize scheme, asking whether these alternative incentive structures affect students' decisions to enter the competition and whether they reduce the gender gap in participation.

Beyond estimating the average effects of alternative prize schemes, our survey data allow us to examine which students respond to these incentives. This connects our paper to a broader literature showing that competitive behavior depends not only on ability but also on individual characteristics. Recent field evidence suggests that factors such as family background, non-cognitive traits, and gender stereotypes contribute to gender differences in competitive decisions. Kang et al. (2024) show that, after narrowly failing the Chinese college entrance exam, men are twice as likely as women to retake it, and that psychological costs, parental expectations, and non-cognitive traits explain this difference. Landaud and Maurin (2020) find that in French admission contests women retry less often than men, and that this difference explains a large part of the gender gap in access to top science programs; family background matters in a similar way. Dilmaghani (2022) shows that gender stereotypes affect effort even among elite chess players. Together, these studies point to the importance of individual characteristics beyond ability in shaping competitive behavior. We build on this insight by examining whether past participation, family background, competitiveness, and gender stereotypes are associated with the initial decision to compete and with the response to monetary incentives. Closely related to these determinants, a growing body of work has examined the role of confidence and performance beliefs in shaping gender differences in competitive behavior. Specifically, women tend to report more pessimistic beliefs in male-typed domains such as mathematics, and these differences persist over time: feedback narrows the gap only temporarily (Coffman et al., 2024). At the same time, recent evidence suggests a more nuanced picture. Using a meta-analysis, Bandiera et al. (2022) show that both men and women are, on average, overconfident and that gender differences in confidence are smaller than commonly believed. Similarly, Coffman and Klinowski (2024) find that women do not seek less feedback than

men despite reporting lower confidence, while Shastry et al. (2020) show that gender differences in the attribution of negative feedback shape subsequent entry into competition. Based on these insights, our survey data allow us to examine the role of confidence and performance beliefs in students' participation decisions and responses to monetary incentives.

3 Experimental method

The Math Olympiad, organized by the *Unione Matematica Italiana* (UMI), is the oldest Italian math competition. Each year it involves about 200,000 students from grades 9 to 13 across the country. The competition is structured in three stages: the school level, the district level, and the national level. Our intervention took place at the first, school-level stage, called *Giochi di Archimede*, which was held simultaneously in all participating schools on November 21, 2019. Performance in the *Giochi di Archimede* determines access to the district stage, so the school stage is a real competition with real stakes.

We ran the experiment in high schools located in the North-East of Italy (Veneto, Emilia Romagna, Friuli Venezia Giulia, and Trentino-Alto Adige). We chose this area based on the INVALSI (Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e di Formazione - National Institute for the Evaluation of the Education and Training System) standardized tests, which indicate that it features the highest-performing students in the country and the most homogeneous performance across schools and classes. A school was eligible for the study if it satisfied three criteria: i) it took part in the Math Olympiad, ii) it had at least ten classes distributed across the five years of study, and iii) the gender ratio among enrolled students was not more unbalanced than 70-30. Eligible schools were selected through a simple random procedure and invited to participate. The final sample includes 30 schools.

Each school that agreed to participate was randomly assigned to one of three treatments:

- **"No Gender" prize treatment (NGT, 11 schools)**: within each school, the four highest-ranked students according to their performance in the *Giochi di Archimede* received a monetary reward. The first and second best-ranked students received 200 euros each, while the third and fourth best-ranked students received 100 euros each.
- **"Gender" prize treatment (GT, 10 schools)**: within each school, the female and the male student with the highest performance received 200 euros each, while the female and the male student with the second highest performance received 100 euros each. This treatment implements a within-gender prize scheme: all students sit the same competition and the Olympiad ranking is common, but prize eligibility is determined by the ranking within gender.
- **Control/Baseline (9 schools)**: in these schools there was no intervention in terms of monetary prizes.

In both the Gender and No Gender treatments, prizes were awarded at the school level: students competed only against their schoolmates, and eligibility for a prize depended on their rank within

the school. Importantly, the two schemes change the expected incentives of girls and boys by a substantial amount. Specifically, we use the scores observed in the Baseline schools to calculate the probability of winning and the expected prize for girls and boys under each scheme (Appendix Table E20). Under the No Gender scheme, a girl who enters the competition would have a 2.4 percent chance of winning a prize, compared with 5.7 percent for a boy; ; girls would receive one third of the total prize money, and in three of the nine Baseline schools no girl would rank among the top four. Under the Gender scheme the chances of winning would be 4.5 percent for girls and 4.2 percent for boys; For girls, the expected prize conditional on participation would rise from about 4 to about 7 euros, while the score required to win would fall from 74 to 65 points on the 0–100 scale. The Gender scheme thus nearly doubles girls' chance of winning and raises their expected prize conditional on participation. As we show and discuss below, however, this does not translate into a stronger participation response among girls than among boys.

Irrespective of the treatment, entry into the *Giochi di Archimede* followed the procedures adopted by each school. Specifically, in all schools, participation was open to any student wishing to take part. On top of this open-entry policy, some schools actively encouraged participation by inviting selected students, typically those with stronger mathematics performance, while others make participation mandatory for some classes, usually in math-intensive curricula. Since the entry rule chosen by a school is not random and may correlate with school characteristics, we collect it directly from the schools and account for it throughout the analysis; Section 6.2 exploits its class-level structure.

3.1 Experimental procedures

In all participating schools, we implemented the same experimental protocol; the only difference across treatments was whether monetary prizes were offered and, in the treated schools, how they were allocated. We prepared posters advertising the Math Olympiad for all participating schools and asked each school to display them at least one month before the *Giochi di Archimede*, held on November 21, 2019. While the posters in the Baseline schools advertised only the Olympiad, those in the treated schools also described the relevant prize scheme. The two prize schemes were presented to students under the names "W la Matematica" (No Gender treatment) and "Vinci con la Matematica" (Gender treatment). The posters are reproduced in Appendix G. Before the competition we also submitted a short survey to the math teacher in charge of organizing the *Giochi di Archimede* in each school, collecting the number of male and female students enrolled and the rule used to select participants for the Olympiad.

The timeline of these steps matters for the analysis, so we make it explicit. Schools that agreed to participate learned their treatment assignment through the same communication that delivered the posters and the organizational instructions, about one month before the competition, which took place on November 21, 2019. The school coordinators returned the survey reporting the entry rule within a few days of this communication, so the school-level entry rule was recorded at the very beginning of the exposure period and before students had been exposed to the posters. This

timing makes an adjustment of the entry rule to the prize schemes unlikely, although it cannot rule one out by itself; We examine this possibility in Section 4.4 by comparing each school's entry rule with the one it followed in the previous edition of the competition. The enrollment of entire classes is different: this decision is made at the class level and could still be adjusted in the weeks between the communication and the competition. Section 6.2 treats this margin separately.

In the same period, in each participating school and irrespective of treatment, students in a subset of randomly selected classes were asked to fill in a detailed questionnaire. The questionnaire was administrated as a separate activity from the Math Olympiad and from the prizes. It collected data on demographics, family background, grades, cognitive ability, beliefs, preferences, aspirations, and non-cognitive traits, described in the next section.

4 Data and variables

We use three main data sources:

1. Individual-level student data, including questionnaire responses from a subgroup of students irrespective of their participation in the Math Olympiad, as well as participation and score data for Olympiad participants;
2. Data about schools' characteristics from the teacher survey, including treatment assignment, the gender composition of the school, and the Olympiad entry rule;
3. School-level background data from the Fondazione Agnelli dataset, which covers school quality indicators for a subset of tracks[1]..

At the individual level, our data comprise two partially overlapping samples: students who completed the questionnaire and students who participated in the Math Olympiad. Because the questionnaire was administered to randomly selected classes irrespective of Olympiad participation, our questionnaire sample includes both participants and non participants. The key empirical task is therefore to match questionnaire respondents to Olympiad records. We first describe the questionnaire data in Section 4.1 and the Olympiad records in Section 4.2, and then explain the matching procedure in Section 4.3.

4.1 The questionnaire

Questionnaire administration started about one month before the competition and continued throughout the data collection period. Importantly, the questionnaire was administrated as a separate activity from the Math Olympiad and from the prizes. In each school, at least two classes per grade were randomly selected, across the study tracks offered by the school, and all students in the

[1]Fondazione Agnelli is an Italian research institute specializing in education. Its Eduscopio project evaluates upper-secondary schools using the subsequent university and employment outcomes of their graduates. See https://eduscopio.it/.

selected classes were asked to fill in the questionnaire, with written informed consent from the students themselves, if they were at least 18 years old, or from their parents, if they were minors. The questionnaire was completed online, in class, and it combined standard background questions with validated instruments from the experimental literature. In total, 8,174 students filled it in. After excluding records with missing school or gender and a small number of respondents who reported a school outside the experiment, the analysis sample contains 7,722 respondents in 497 classes of the 30 schools. Because classes were selected irrespective of the Olympiad, the sample contains both students who later took part in the competition and students who did not. The inclusion of both groups is central to our analysis, which examines not only who enters the competition but also how participation responds to the incentive schemes.

The background section collects gender, year of study, date of birth, parents' education and occupation, place of birth of the student and of the parents, language spoken at home, and number of siblings. Students also report their grades in math, Italian, and English as a foreign language at the end of the previous school year, rank the three subjects from most to least preferred, list their extra-school activities, and state their plans after high school: whether they intend to enroll at university and in which area (STEM, social sciences and humanities, health, or other).

Cognitive ability is measured with a short version of Raven's Standard Progressive Matrices. We use the nine-item form developed by Bilker et al. (2012), which is optimized to predict the score of the full 60-item test. Students had 270 seconds to solve as many of the nine items as possible, and the score ranges from 0 to 9. Right after the Raven test, we elicit beliefs about own performance: how many problems the student thinks she solved correctly, how sure she is of this answer, where she expects to rank within her class, and how sure she is of this rank. These questions give us measures of absolute and relative confidence that we use as controls and as dimensions of heterogeneity.

Preferences and non-cognitive traits are measured with validated scales. Risk and time preferences come from the survey module of Falk et al. (2023): a qualitative question on the general willingness to take risks on a 0-10 scale, and quantitative staircase items with five interdependent binary choices, between a lottery and varying sure payments for risk, and between a payment today and a larger payment in twelve months for patience. Competitiveness is measured with the revised Competitiveness Index (Harris and Houston, 2010), self-esteem with the Rosenberg scale (Rosenberg, 1965), self-efficacy with the General Self-Efficacy scale (Schwarzer and Jerusalem, 1995), and self-control with the Brief Self-Control Scale (Tangney et al., 2004). In the analysis, these measures enter as controls, added in groups, and as above-median dummies in the heterogeneity specifications.

The questionnaire also included a Gender-Science Implicit Association Test (Greenwald et al., 1998; Nosek et al., 2002). In this test, students sort words related to science and to humanities together with male and female names, and the speed of their responses reveals how strongly they associate science with the male gender. We summarize the test with the standard D-score (Greenwald et al., 2003), where positive values indicate a stronger implicit association between males and

science. The D-score is available for 7,113 respondents. Both boys and girls hold, on average, a positive D-score, and the means by gender are very close (Table 3). We use this measure in Section 6.6 to ask whether implicit stereotypes are related to the decision to enter the competition.

4.2 The Math Olympiad

The second source of data refers to students' participation and performance in the Math Olympiad. We asked each participating school to prepare a file containing records for its participants. We then combined the school files into a single dataset covering individual records for 6,424 participants in the 30 schools. Each record reports the school, the study track, the class section, the year of study, gender, date of birth, and the score. We link these Olympiad records to questionnaire respondents using the reconstructed matching procedure described below. Olympiad scores are measured using the normalized score provided in the competition data. For first- and second-year students, raw scores are rescaled from a 0–80 scale to a 0–100 scale; for later years, raw scores are already on the 0–100 scale. All score regressions use this normalized 0–100 measure and include year-of-study fixed effects.

4.3 Matching procedure

We link questionnaire respondents to Olympiad records through the following procedure. Because of privacy concerns, neither dataset contains student names, so the match is based on the combination of school, year of study, gender, and exact date of birth, with the study track and the class section used as consistency checks. Two rules protect against false positives: each questionnaire respondent can be assigned to at most one Olympiad record, and vice versa; and when several Olympiad records compete for the same questionnaire and no clear winner exists, the match is left unassigned. We classify matches by confidence level:

- **Gold**: unique matches on the full key, with consistent track and section information.
- **C1**: unique matches on school, year of study, and date of birth (and gender, when observed in both records), with compatible section but discordant or missing track information.
- **C2**: unique matches on the same core key, with track and section not used as constraints.

A handful of residual cases were resolved manually with documented rules. Appendix F describes the procedure in more detail.

Table 1 reports the number of matches by confidence level. The final match contains 2,635 gold matches, 110 C1 matches, and 161 C2 matches, for a total of 2,906 matched respondents, all inside the analysis sample. Two coverage rates help to read these figures. On the questionnaire side, 2,906 of the 7,722 respondents, about 38 percent, are matched to an Olympiad record; the remaining respondents are treated as non-participants. On the competition side, 2,906 of the 6,424 Olympiad participants, about 45 percent, are matched to a questionnaire; the others are mostly

students whose class was not selected for the questionnaire, and they do not enter the analysis. A residual concern is that some unmatched respondents did participate but could not be matched. The conservative gold-only definition used in the main analysis, the expanded gold+C1+C2 definition reported in the Appendix, and the analysis of recovered matches in Section 6.2 address this concern.

Since the two definitions give different estimates, we state here how they should be read. The two measures have errors of opposite sign. Gold matches are the only ones for which the record identity is certain, so the gold-only definition has the smallest risk of false positives but classifies some true participants as non-participants. The expanded definition recovers about 10 percent more participants and has fewer false negatives, at the cost of a higher risk of false positives. The recovered C1 and C2 matches are not evenly distributed across treatments: they are more frequent in Baseline schools than in treated schools (Section 6.2). This matters in both directions. If the recovered matches are true participants, the gold-only definition undercounts participation in the control group more than in the treated groups and overstates the differences between treatments; if they are false positives, the expanded definition overcounts control participation and understates them. The two definitions therefore provide two reference points under opposite assumptions about the recovered matches, rather than formal bounds, and we report both throughout: the gold-only estimates in the main tables and the expanded estimates in the Appendix. Appendix Table F2 asks how much differential matching error would be needed to generate the gold-only estimates. Reclassifying unmatched Baseline students at random, in the full sample the estimates would fall to zero only if 41 percent (No Gender) or 25 percent (Gender) of the unmatched Baseline students were in fact participants; on the free margin, the corresponding shares are 15 and 16 percent, against recovered C1 and C2 matches that amount to about 5 percent of the unmatched Baseline students. Reclassifying 5 percent of them as participants lowers the free-margin estimates from 0.12 and 0.13 to 0.08 and 0.09. The free-margin estimates are therefore sensitive to differential matching error of a size two or three times larger than the one we observe, while the full-sample estimates are not.

Table 1: Questionnaire–Olympiad matching status

Match status	Count	Share of questionnaires
Gold	2,635	32.2%
C1	110	1.3%
C2	161	2.0%
Unmatched	5,268	64.4%
Total	8,174	100.0%

Notes: Gold indicates strict high-confidence matches. C1 and C2 are recovered one-to-one matches with different levels of auxiliary consistency. The expanded participation definition uses Gold+C1+C2.

4.4 Treatment and selection rules

Treatment assignment is merged at the school level. Schools are classified as Baseline, No Gender, or Gender. We also merge information on the school-level selection rule governing Olympiad participation: Free, Free or Suggested, and Mixed (Free and Mandatory). This variable is important because participation mechanically depends on how schools selected or encouraged students to take part in the Math Olympiad.

Unlike the school characteristics of Appendix Table E11, the entry rules are not balanced across treatments. Seven of the nine Baseline schools have a Mixed rule, against four of eleven in the No Gender prize treatment and three of ten in the Gender prize treatment, and half of the Gender prize schools have fully free participation (Table 2). Fisher's exact test does not reject the equality of the three distributions ($p = 0.13$). An imbalance of this size is compatible with chance in a randomization over thirty schools.

Could the entry rules instead be a response to the prizes? A school that learns it has been assigned a prize has one month to react. The natural reaction is to bring more students into the competition, for instance by enrolling entire classes, so that the prize is not wasted. If this had happened, we would observe more mandatory enrollment in treated schools than in Baseline schools, and changes towards mandatory enrollment concentrated in treated schools. We observe the opposite. Mandatory enrollment is most common in the Baseline schools, which received no prize. Changes in the entry rules relative to the previous edition of the competition, which we know for 22 of the 30 schools, do not differ across treatments: six schools changed their rule, two in each treatment (Fisher's exact test, $p = 0.83$). If anything, the changes run against a strategic response. The only two schools that moved from mandatory to voluntary participation are treated schools, one in each prize treatment, while two of the four schools that moved towards mandatory enrollment are Baseline schools. Finally, the coordinators reported the entry rule within a few days of the treatment communication, before students had seen the posters. The imbalance in entry rules is therefore a chance outcome of the randomization, not a response to the prizes.

Our information on the entry rules of the previous edition comes from two sources. In October and November 2018, one year before the experiment, we surveyed the Olympiad coordinators of 462 Italian upper secondary schools about their entry practices, and twelve of our schools answered. For ten of them, the rule reported in 2018 coincides with the rule recorded in 2019. The three schools with fully free participation in 2019 already had fully free participation in 2018. Four of the six Mixed schools already made participation mandatory for some classes or tracks. The three schools with free or suggested participation reported voluntary participation in 2018, when the survey did not distinguish between free and suggested entry. The two schools that changed their rule moved from free to Mixed participation; one is a Baseline school and the other a Gender prize school. In the full 2018 survey, 77 percent of the coordinators reported that the entry rule was the same as in the previous edition, and only 4 percent reported a change. In 2026 we also asked by email the principals of the other eighteen schools how participation was organized in the 2018/19 school year, and ten of them answered. These answers were collected seven years after the experiment,

from a different respondent and from memory, so we give them less weight. They point in the same direction. Across the 22 schools for which we have either source, the rule of the previous edition coincides with, or is compatible with, the 2019 rule in 16 schools.[2]

Although the entry rules did not respond to the prizes, their imbalance affects the raw comparison between treatments, and we deal with it in three ways. Every regression controls for the entry rule. Section 6.2 removes the classes enrolled as a whole and looks only at students who decide for themselves. The Appendix re-estimates the main specifications separately by entry rule (Tables B1 to B4). A separate question concerns the enrollment of entire classes, which is decided at the class level and had more time to adjust; Section 6.2 discusses this point.

Table 2 describes how selection rules are distributed across treatment status at the school level; the student-level distribution, which differs because schools have different numbers of questionnaire respondents, is reported in Appendix Table E9.

Table 2: Schools by selection rule and treatment status

Selection rule	Baseline	No gender	Gender	Treated total	Total
Free	1	2	5	7	8
Free or suggested	1	5	2	7	8
Mixed: free and mandatory	7	4	3	7	14
Total	9	11	10	21	30

Notes: The table counts schools in the participation analysis sample. Treated total is the sum of No gender and Gender schools.

4.5 Summary statistics

We first report summary statistics separately for questionnaire respondents who participate in the Olympiad and those who do not, using the conservative gold-only matching definition (Appendix Table E3). This comparison is useful for understanding which characteristics are associated with selection into the competition before turning to the treatment regressions.

Appendix Table E10 compares questionnaire covariates across treatments. Because the randomization assigns 30 schools rather than individual students, we assess balance with regressions of each covariate on the treatment indicators, clustering standard errors at the school level and computing wild cluster bootstrap p-values. With inference at the level of the randomization, the covariates are balanced: out of 76 pairwise tests, only four return p-values below 0.10, close to what

[2]The 2018 survey was sent to the coordinators of all schools registered for the *Giochi di Archimede*. The 2026 survey included the same questions, referring to the 2018/19 school year. In both surveys, coordinators were asked whether participation was voluntary and open to all students; voluntary but restricted to certain classes; voluntary but restricted to students selected by teachers; mandatory for all students; mandatory for certain classes; or organized in another way. Among the ten 2026 answers, four schools report a rule different from the 2019 one: two moved towards a Mixed rule (one Baseline and one No Gender prize school) and two moved from mandatory to voluntary participation (one No Gender prize and one Gender prize school). The No Gender prize school that moved to a Mixed rule enrolled almost all its sampled classes in 2019; without it, the No Gender coefficient in Mixed schools falls from 0.37 to 0.31. One principal described the rule of the 2017/18 edition, and the identification of one school in the 2018 survey is probable but not certain. The answers are reported school by school in the replication material.

chance alone would produce. The clearest difference is the share of female students in the two prize treatments (0.57 in the No Gender schools against 0.49 in the Gender schools, p = 0.03), which is also the only covariate whose joint test of equality across the three treatments has a p-value below 0.10 (p = 0.09); gender enters every specification, both as a control and interacted with the treatments in the heterogeneity analyses. All specifications also control for year-of-study fixed effects, and the richer specifications add the full set of questionnaire covariates, so the estimates do not rely on exact covariate balance. Randomization is balanced at the school level: school size, gender composition, municipality population, and the Fondazione Agnelli school quality indicators do not differ across treatments (Appendix Table E11). The one exception is the distribution of entry rules, discussed in Section 4.4.

Table 3 compares boys and girls on the main questionnaire measures. Notably, boys and girls have identical Raven scores, while girls have higher grades, including in mathematics. Yet girls hold much less optimistic beliefs about their own performance: they expect fewer correct answers, they are less certain about them, and they expect a worse rank in their class. Moreover, when considering the rank a student expects against the position she actually holds in the within-class distribution of Raven scores we observe that, on average, boys expect to rank higher than they actually do, while girls expect to rank lower. The gender gap is equal to 13 percentile points, and 40 percent of boys against 56 percent of girls expect a position worse than their actual one. This pattern indicates that boys are, on average, overconfident about their rank relative to their classmates, while girls are underconfident. Girls also score lower on competitiveness, risk tolerance, self-esteem, and self-efficacy, and they are less likely to have taken part in the Olympiad before. In short, the gender gap in participation that we document below does not reflect a gap in measured ability; it is consistent instead with gaps in confidence, competitiveness, and past exposure to the competition.

Table 3: Summary statistics by gender

	Boys	Girls	Difference (G−B)	Observations
Raven score (0–9)	4.85	4.85	-0.00	7,623
Believed correct answers	5.81	5.18	-0.63***	7,619
Certainty about answers (0–10)	7.13	5.97	-1.16***	7,615
Expected rank in class (1 = best)	8.81	10.93	2.12***	7,597
Math grade	6.91	7.14	0.23***	5,997
Average grade	7.25	7.56	0.31***	5,995
Math preference rank (1 = most pref.)	1.87	1.97	0.10***	5,998
Past Olympiad participation	0.41	0.32	-0.08***	5,989
Plans a STEM field at university	0.40	0.22	-0.18***	5,799
Competitiveness	45.07	40.61	-4.45***	7,462
Willingness to take risk (0–10)	6.87	6.44	-0.43***	7,473
Risk staircase index	0.09	-0.08	-0.17***	7,461
Patience index	-0.02	0.01	0.03	7,460
Self-esteem	29.87	27.22	-2.65***	7,452
Self-efficacy	29.19	28.00	-1.18***	7,470
Self-control	42.40	43.13	0.73**	7,473
Gender-Science IAT (D-score)	0.14	0.17	0.03	7,113
Relative underconfidence	-0.07	0.06	0.13***	7,426
Participation (gold-only)	0.37	0.32	-0.05**	7,722
Olympiad score (0–100, participants)	34.91	30.32	-4.58***	2,906

Notes: Means by gender in the participation analysis sample. The difference column reports the coefficient of a regression of each variable on a female dummy, with standard errors clustered at the school level; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications). The score row refers to matched participants. Positive values of the IAT D-score indicate a stronger implicit association between males and science. Relative underconfidence is the difference between the percentile of the rank a student expects and the percentile of her actual Raven score, both computed within the class among respondents (classes with at least five respondents); positive values indicate that the student expects a worse position than the one she holds. The STEM row refers to students who report plans to enroll at university. Scale scores are missing for respondents who answered none of the underlying items. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5 Empirical strategy

We estimate linear probability models for Olympiad participation and linear models for Olympiad scores. Because Olympiad scores are observed only for students who participate, and participation itself responds to the treatments, all score regressions compare selected samples and we read them as descriptive of the pool of participants in each treatment. The main participation outcome, P_{is}^{Gold}, equals one if questionnaire respondent i in school s is matched to an Olympiad participant through a gold match. The expanded gold+C1+C2 definition is reported in the Appendix.

Let NG_s be an indicator for schools assigned to the No Gender treatment and G_s an indicator for schools assigned to the Gender treatment; Baseline is the omitted treatment group. Let S_s denote the school-level Olympiad selection-rule indicators, with Free participation as the omitted category. Let λ_y denote year-of-study fixed effects and $Female_i$ an indicator for female students. Controls are added step by step. We denote by X_{ik} the vector of controls included in specification k, which may be empty in the baseline column. We group controls into the following categories:

demographics, ability and beliefs, academic performance, past competitions, aspirations, and non-cognitive traits.

The participation specifications are:

$$P_{is}^{Gold} = \alpha_k + \beta_{1k} NG_s + \beta_{2k} G_s + S_s' \pi_k + \lambda_y + \theta_k Female_i + X_{ik}' \rho_k + \varepsilon_{isk}. \tag{1}$$

For Olympiad performance, the dependent variable is normalized Olympiad score among gold matched participants. The score specifications are:

$$Score_{is}^{Gold} = \alpha_k + \beta_{1k} NG_s + \beta_{2k} G_s + S_s' \pi_k + \lambda_y + \theta_k Female_i + X_{ik}' \rho_k + u_{isk}. \tag{2}$$

Standard errors are clustered at the school level in all specifications. Because the design has 30 clusters, significance stars in the main tables are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications); conventional clustered standard errors are reported in parentheses. Missing values in controls are left as missing; therefore, complete-case sample sizes vary across columns.

Finally, to study heterogeneity by gender, we estimate direct specifications with treatment-by-female interactions:

$$\begin{aligned} Y_{is} = {} & \alpha_k + \beta_{1k} NG_s + \beta_{2k} G_s + \theta_k Female_i \\ & + \phi_{1k}(NG_s \times Female_i) + \phi_{2k}(G_s \times Female_i) + S_s' \pi_k + \lambda_y + X_{ik}' \rho_k + e_{isk}, \end{aligned} \tag{3}$$

where Y_{is} is either gold-only participation or normalized score among gold matched participants.

We also explore whether incentives work differently for students with different baseline characteristics. We consider five characteristics H_i: above-median Raven score, above-median math grade, above-median confidence (certainty about the number of correct Raven answers, on a 0 to 10 scale), previous Olympiad participation, and above-median competitiveness. For each characteristic we estimate:

$$\begin{aligned} Y_{is} = {} & \alpha_k + \beta_{1k} NG_s + \beta_{2k} G_s + \zeta_k H_i \\ & + \phi_{1k}(NG_s \times H_i) + \phi_{2k}(G_s \times H_i) + S_s' \pi_k + \lambda_y + \theta_k Female_i + X_{ik}' \rho_k + e_{isk}. \end{aligned} \tag{4}$$

Each heterogeneity dimension is estimated in a separate regression. These specifications include the full set of questionnaire controls used in the final column of the main tables.

6 Results

This section is organized in three blocks. The first block concerns the effect of the prizes on participation. Section 6.1 reports the main participation results on the full sample, from the unadjusted comparison to the specifications with the full set of controls. Section 6.2 separates the

classes enrolled in the competition as a whole from the students who decide for themselves, and re-estimates the participation and score differences on this free margin. The second block concerns performance. Section 6.3 reports the score differences among participants, and Section 6.4 the effect of the prizes on realized performance, an outcome that does not condition on participation. The third block concerns heterogeneity. Section 6.5 asks whether the prizes work differently for girls and boys. Section 6.6 describes which students enter the competition, including the role of implicit gender stereotypes measured by the IAT. Section 6.7 asks which students respond to the prizes, and Section 6.8 closes with the specifications that split the sample by the school-level selection rule. Throughout, the main tables use the conservative gold-only matching; the corresponding expanded estimates are in Appendix A.

6.1 Participation

Table 4 reports the main participation results using only gold matches. Column (1) is the raw comparison, with the treatment indicators only. The difference from the Baseline is close to zero for the No Gender prize and negative for the Gender prize. This column mostly reflects the imbalance in entry rules documented in Table 2: Baseline schools more often enroll entire classes, which mechanically raises their participation.
Column (2) holds the entry rule constant. Participation is then 27 percentage points higher under the No Gender prize (wild cluster bootstrap $p = 0.008$) and 17 points higher under the Gender prize ($p = 0.05$). The No Gender difference is robust across specifications. The Gender difference is smaller and loses significance once academic controls shrink the complete-case sample.

In the following, we reports three sets of participation estimates.The first is the raw comparison among questionnaire respondents, in column (1) of Table 4: the difference in matched participation between students in treated and Baseline schools, with no adjustment. It follows directly from the random assignment, but it is estimated on the sampled classes and it depends on the matching between questionnaires and Olympiad records. The second is the school-level comparison on the complete administrative records of the competition, in Appendix Table E1: participants over enrolled students in each school, controlling for the entry rule. It does not depend on the sampling of classes or on the matching. These school-level differences are negative, about 13 percentage points for both prizes, while the student-level estimates that control for the same rule are positive. The sign changes because participation in the sampled classes, measured through the matching, differs from participation in the whole school: when we re-estimate the school-level regression with the share of matched questionnaire respondents as outcome, the coefficients become positive and close to the student-level ones (0.28 and 0.17). Neither the raw comparison nor the school-level comparison shows a significant increase in participation. Finally, in columns (2) to (8) of Table 4 we report the difference between treatments holding the entry rule constant. These estimates are positive and more precise. In Section 6.2 we focus on schools where participation relies on students' own initiative, rather than on teacher invitations or class-level enrollment.

Table 4: Treatment effects on Olympiad participation: Gold-only

	(1) ITT	(2) Base	(3) + Demog.	(4) + Ability	(5) + Academic	(6) + Past	(7) + Aspir.	(8) + Non-cognitive traits
No gender treatment	0.093	0.273***	0.272***	0.279***	0.291**	0.261**	0.261**	0.265**
	(0.1240)	(0.0751)	(0.0758)	(0.0737)	(0.0886)	(0.0899)	(0.0901)	(0.0917)
Gender treatment	-0.123	0.168*	0.173*	0.171*	0.154	0.126	0.129	0.127
	(0.0822)	(0.0713)	(0.0731)	(0.0724)	(0.0910)	(0.0893)	(0.0905)	(0.0919)
Selection rule		✓	✓	✓	✓	✓	✓	✓
Year FE		✓	✓	✓	✓	✓	✓	✓
Gender		✓	✓	✓	✓	✓	✓	✓
Demographics			✓	✓	✓	✓	✓	✓
Ability/beliefs				✓	✓	✓	✓	✓
Academic performance					✓	✓	✓	✓
Past competitions						✓	✓	✓
Aspirations							✓	✓
Non-cognitive traits								✓
Observations	7,722	7,722	7,679	7,592	5,993	5,987	5,987	5,900

Notes: Linear probability models; the dependent variable is gold-only Olympiad participation. Column (1) reports the unadjusted comparison with treatment indicators only; later columns add the controls indicated in the panel. Standard errors clustered at the school level (30 clusters) in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Missing values in controls are left as missing, so complete-case samples vary across columns. On the complete-case sample of column (8) (N = 5,900), the specification of column (2) gives 0.310 for the No Gender prize and 0.177 for the Gender prize, and the specification of column (1) gives 0.145 and −0.080, so the changes across columns reflect the added controls rather than the change of sample.

6.2 Mandatory classes and the free margin

In the 14 schools classified as Mixed, all students could enter the competition, but some classes are enrolled in the competition as a whole. As the class-level rule was not collected, we use the linked questionnaire and Olympiad records to capture variation in participation across classes. For each class with at least five questionnaire respondents we compute the share of respondents matched to an Olympiad record. We classify a class in a Mixed school as a high-participation class when this share is at least 70 percent. The distribution of class-level participation in Mixed schools is clearly bimodal: high-participation classes have an average participation of 88 percent, while the other classes in the same schools average 31 percent. The classification is consistent with what school coordinators declared (Table 5, Panel A).

Figure 1 shows the two distributions: in Free and Free-or-Suggested schools, class-level participation is concentrated near zero. Panel B of Table 5 shows that, within Mixed schools, high-participation classes are very unevenly distributed across treatments: they contain 75 percent of the students in No Gender schools, against 29 percent in Baseline schools and 25 percent in Gender schools. Inside these classes participation is between 83 and 89 percent in all treatments, so monetary prizes have little room to change participation there. We cannot determine whether the larger share of students in high-participation classes in No Gender schools reflects a response to the prize, with these schools being more likely schools to enroll entire classes. Because classes are classified using realized participation, which the prizes may affect, this comparison focus on differences between classes with high and lower participation but cannot identify a causal effect for classes where entry relies on students' initiative.

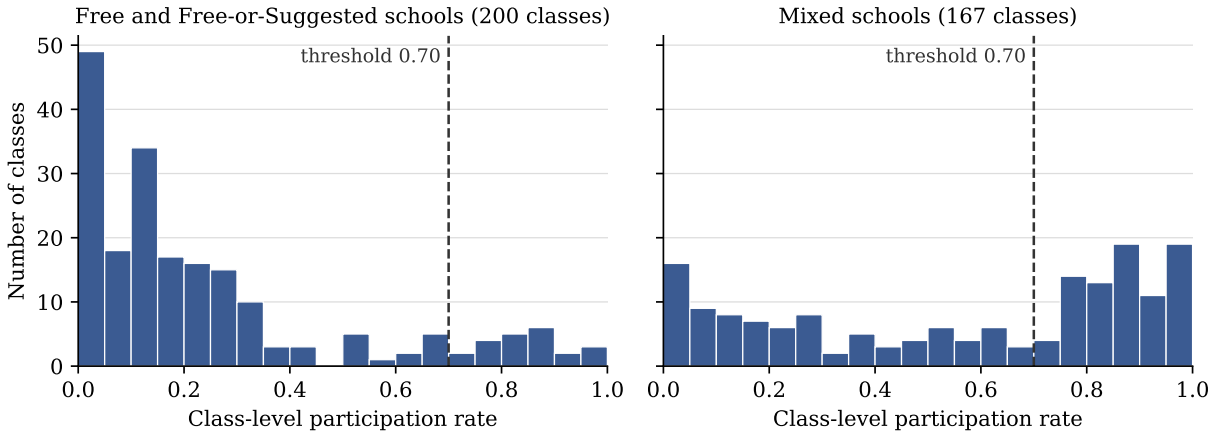

Figure 1: Distribution of class-level participation rates, by school selection rule. Classes with at least five questionnaire respondents. Participation is the share of respondents in the class matched to an Olympiad record (gold, C1 or C2). The dashed line marks the 70 percent threshold used to classify high-participation classes in Mixed schools.

Table 5: High-participation classes: validation and distribution across arms

Panel A: students in high-participation classes, by declared selection rule

	Students	In high-part. classes	Share
Free	2,594	38	1.5%
Free or suggested	1,404	230	16.4%
Mixed: free and mandatory	3,724	1,595	42.8%

Panel B: within Mixed schools, by treatment arm

	Students	Share in high-part. classes	Participation in these classes
Baseline	1,693	29.2%	0.87
No gender	1,174	75.6%	0.89
Gender	857	24.7%	0.83

Notes: A high-participation class is a class with at least five questionnaire respondents where at least 70 percent of respondents are matched to an Olympiad record (gold, C1 or C2). Panel A shows that these classes are almost absent in schools that declared free participation and concentrated in Mixed schools, consistent with the school-level declarations of the coordinators. Panel B shows the distribution of students across class types within Mixed schools and the average participation rate inside high-participation classes (gold+C1+C2 definition).

Table 6 reports the participation regressions when high-participation classes are excluded, so that the remaining sample approximates the free margin. Because the exclusion is based on realized participation, these estimates are descriptive. Only two Baseline schools have voluntary entry (Table 2).

Two main findings emerge. First, on the free margin the two prize schemes look similar: with the gold-only definition, participation is 12.0 percentage points higher under the No Gender prize and 12.9 points higher under the Gender prize in the baseline specification, and both differences are statistically significant.[3] With the full set of questionnaire controls the point estimates fall to

[3]Because the design has 30 clusters, we also compute wild cluster bootstrap p-values (Rademacher weights, 9,999

about 9 points for both treatments and lose statistical significance, partly because the complete-case sample shrinks. The result is robust to how the decomposition is built (Appendix Tables E14 and E15): the estimates are similar with a 60 or 80 percent threshold for high-participation classes, larger when schools are weighted equally, and stable when one school at a time is excluded, so no single school drives the pattern. They are smaller, about 5 to 8 points, under the expanded matching definition. Second, the large No Gender effect in the full sample comes mainly from the high-participation classes: within Mixed schools, the No Gender coefficient falls from 0.37 to about 0.15 once these classes are excluded (Appendix Table E12). Appendix Table E13 shows that treatment-by-female interactions on the free margin are close to zero in all specifications, while the female participation gap of 7 to 11 percentage points is present in every treatment.

Table 6: Participation on the free margin: excluding high-participation classes

	Gold-only			Gold+C1+C2		
	(1) Full sample	(2) Free margin	(3) Free margin + controls	(4) Full sample	(5) Free margin	(6) Free margin + controls
No gender	0.273***	0.120*	0.092	0.217**	0.053	0.038
	(0.075)	(0.057)	(0.063)	(0.084)	(0.071)	(0.071)
Gender	0.168*	0.129**	0.092	0.117	0.077	0.039
	(0.071)	(0.046)	(0.057)	(0.076)	(0.054)	(0.061)
Observations	7,722	6,127	4,941	7,722	6,127	4,941

Notes: Linear probability models. The free margin sample excludes high-participation classes in Mixed schools (classes with at least five questionnaire respondents and a matched-participation share of at least 70 percent). All specifications include the school-level selection rule, year-of-study fixed effects, and gender. Columns with controls add the full set of questionnaire controls. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The class-level split also clarifies the negative score results. Table 7 estimates the score regressions separately on the free margin and inside high-participation classes. Inside high-participation classes, where participation is near universal in every treatment, score differences across treatments are small and not statistically significant. Adding the full controls in these classes (column 6) cuts the sample by more than a third, because grades are missing for most first- and second-year students, and leaves only 9 school clusters; the point estimates become noisier, and the wild cluster bootstrap cannot distinguish them from zero (p = 0.42 and 0.32 with Webb weights). On the free margin, where the estimates are consistent with the prizes changing who participates, the negative score differences become larger and more precise: around 6 points in the baseline specification and around 7 points with controls, for both treatments. This pattern is consistent with a composition effect: the students associated with the estimated participation increase are on average weaker than those who would have participated anyway, and the average score among participants falls, while in the classes where participation is near universal the prizes have little room to change selection and we find no significant score differences. We find no evidence that the prizes reduce the performance of a given student, but we note that, without observing the students who would have

replications): 0.061 for the No Gender prize and 0.024 for the Gender prize.

participated in any case, the composition reading cannot be formally separated from an effect on individual performance. Note also that the female score gap of about 4 points is present even inside high-participation classes, where selection into participation is muted; this suggests that the score gap is not only a selection effect.

Table 7: Olympiad scores: free margin vs high-participation classes (gold-only)

	Full sample		Free margin		High-part. classes	
	(1) Base	(2) + controls	(3) Base	(4) + controls	(5) Base	(6) + controls
No gender	-3.27	-6.10**	-5.83	-7.26**	-0.03	-3.50
	(2.12)	(1.98)	(2.99)	(1.96)	(2.04)	(2.83)
Gender	-3.38	-5.93*	-5.99	-7.33***	-2.27	-8.25
	(3.23)	(2.30)	(3.51)	(1.84)	(2.87)	(2.45)
Female	-4.78***	-5.18***	-5.22***	-5.75***	-3.99***	-4.24***
	(0.95)	(0.97)	(1.04)	(1.08)	(1.24)	(1.41)
Observations	2,635	1,878	1,293	1,038	1,342	840

Notes: Dependent variable: normalized Olympiad score (0–100) among gold-only matched participants. The free margin sample excludes high-participation classes in Mixed schools; the last two columns keep only these classes. All specifications include year-of-study fixed effects and gender; specifications outside the high-participation sample also include the school-level selection rule. Columns with controls add the full set of questionnaire controls. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications). In column (6) only 9 school clusters remain: we report no significance stars for the treatment coefficients there; wild cluster bootstrap p-values with Webb weights are 0.416 for the No gender prize and 0.317 for the Gender prize. These regressions condition on participation, which itself responds to the treatments, so estimates are descriptive. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The class-level analysis also explains why results based on the expanded gold+C1+C2 definition are systematically weaker than gold-only results. Recovered C1 and C2 matches are not balanced across treatments: in free classes they cover 5.0 percent of Baseline students, against 3.5 percent in No Gender schools and 2.8 percent in Gender schools. The expanded definition therefore inflates measured participation in the control group more than in the treated groups, and this attenuates all estimated treatment effects. This does not by itself make one definition preferable: the gold-only definition limits false positives, but if it misses more true participants in the control group it can overstate the differences between treatments. We therefore report both definitions as a check on the sensitivity of the results.

Taken together, these results give a simple reading of the participation differences. Among students who decide for themselves, participation is higher under both prizes by a similar amount: about 12 percentage points with the gold-only matching and about half of that with the expanded matching (Table 6), with no difference between boys and girls. The lower average scores among participants are consistent with these additional participants being on average weaker than the students who would have participated anyway. On top of this, the No Gender prize schools show a large additional difference that runs through classes enrolled as a whole.

6.3 Olympiad scores

Because scores are observed only for participants, and participation itself responds to the treatments, the regressions in this and the following subsections compare selected samples. We therefore interpret them as descriptive of the pool of participants in each treatment, not as treatment effects on individual performance. Appendix Table E4 reports the full results for normalized Olympiad scores among gold-only matched participants. In the baseline specification, score differences across treatments are negative but not statistically significant. After adding grade controls, estimates become more negative and statistically significant; this change coincides with a substantial reduction in the complete-case sample. Section 6.2 shows that these negative differences are concentrated on the free margin and are absent inside high-participation classes, a pattern consistent with a composition effect rather than with a negative effect on individual performance. Because we do not observe the scores that non-participants would have obtained, the two explanations cannot be separated with the regressions alone. Section 6.4 therefore turns to an outcome that does not condition on participation.

6.4 Realized performance

The score regressions of Section 6.3 condition on participation, which responds to the treatments. A complementary outcome that does not condition on participation is realized performance: the normalized Olympiad score for students who participate and zero for eligible students who do not. Non-participants are assigned zero because they earned no points in the competition. This zero does not represent the score they would have obtained had they participated. Because the outcome is defined for every student in the sample, random assignment identifies the effect of the prizes on realized performance without monotonicity assumptions, exclusion restrictions, or corrections for sample selection. Like participation, this outcome relies on the matching: participants we fail to match are counted as zero. We therefore answers the following question: how much does the performance realized in the competition change once the prizes are allowed to change who enters? Our analysis combines the effect on participation and the effect on the scores of participants, and the two channels cannot be separated causally, since the second is observed on a population selected by the treatment. The same logic is used in the evaluation of training programs, where earnings including zeros for the non-employed are defined for the whole randomized population, while wages conditional on employment are not (Lee, 2009; Attanasio et al., 2011).

Table 8 reports the estimates. Columns (1) to (3) use the full analysis sample and columns (4) to (6) the free margin. Panel B expresses the coefficients in standard deviations of the outcome, whose mean in the Baseline is about 12 points on the 0–100 scale. The pattern mirrors the participation results, but the estimates are weaker. The unadjusted comparison in column (1) is close to zero for the No Gender prize and negative for the Gender prize, for the same reason discussed in Section 6.1: the chance imbalance in entry rules. Once the entry rule is controlled for, realized performance is about 7 points higher under the No Gender prize, or 0.4 standard deviations (wild bootstrap $p = 0.016$), and about 4 points higher under the Gender prize, even if the difference is imprecise ($p =$

0.12); with the full set of controls, both estimates shrink and lose significance. On the free margin, prizes increase participation by 12–13 percentage points, while estimated effects on total points earned are small and statistically insignificant (1–3 points). These results are consistent with the composition evidence of Section 6.2: prizes bring in additional, lower-scoring participants, so total points change little despite higher participation.

Table 9 adds the interaction with gender. Girls realize between 4 and 6 fewer points than boys, but the interactions between the treatments and the female indicator are small, positive, and jointly insignificant in every column (p-values between 0.24 and 0.68). Thus, although girls earn fewer points overall, we find no evidence that the prizes affect total points differently for girls and boys. This is consistent with the participation results in Section 6.5[4].

Table 8: Treatment effects on realized Olympiad performance: Gold-only

	Full sample			Free margin		
	(1) ITT	(2) Base	(3) Full controls	(4) ITT	(5) Base	(6) Full controls
Panel A: realized performance (normalized score points, 0 for non-participants)						
No gender treatment	1.492	7.077**	5.673	-0.584	2.821	1.118
	(4.000)	(2.328)	(2.976)	(3.349)	(2.121)	(2.186)
Gender treatment	-4.992	3.875	1.653	-1.664	3.012	1.072
	(2.890)	(2.213)	(2.699)	(2.944)	(1.753)	(1.950)
Panel B: same coefficients in standard deviations of the outcome						
No gender treatment	0.082	0.389**	0.312	-0.032	0.155	0.062
Gender treatment	-0.275	0.213	0.091	-0.092	0.166	0.059
Baseline mean	12.841	12.841	11.846	8.378	8.378	8.164
p-value No gender = Gender (WCB)	0.128	0.161	0.117	0.685	0.912	0.972
Selection rule		✓	✓		✓	✓
Year FE, gender		✓	✓		✓	✓
Full set of controls			✓			✓
Observations	7,722	7,722	5,900	6,127	6,127	4,941

Notes: OLS. The dependent variable is realized performance: the normalized Olympiad score (0–100) for gold-matched participants and zero for eligible students who did not participate. The outcome is defined for every student in the analysis sample, so the estimates in columns (1)–(3) are intention-to-treat effects that do not condition on participation; they combine the effect on participation and the effect on the scores of participants, which cannot be separated causally. Panel B divides the coefficients by the standard deviation of the outcome in the full analysis sample (18.17 points). Columns (4)–(6) exclude classes in Mixed schools with participation of at least 70 percent, as in Section 6.2; since this exclusion uses realized participation, these columns are descriptive. Full controls are those of column (8) of Table 4. Standard errors clustered at the school level (30 clusters) in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

6.5 Heterogeneity by gender

We estimate treatment-by-female interactions for both participation and scores. Across specifications, the interactions are not statistically significant. For participation, the estimates are similar

[4] Appendix Table E1 reports school-level estimates based on the complete administrative records of the competition. The participation numerator therefore does not depend on matching questionnaire respondents to Olympiad records. The estimates are negative and differ from the student-level estimates mainly because participation in the sampled classes differs from participation in the school as a whole.

Table 9: Realized Olympiad performance by treatment and gender: Gold-only

	Full sample		Free margin	
	(1) Base	(2) Full controls	(3) Base	(4) Full controls
No gender treatment	5.770*	4.878	1.768	0.036
	(2.557)	(3.591)	(2.725)	(2.881)
Gender treatment	2.996	1.303	2.307	0.841
	(2.334)	(3.011)	(2.164)	(2.384)
Female	-5.669**	-4.387***	-5.384*	-3.839**
	(0.915)	(1.273)	(1.166)	(1.042)
No gender × Female	2.343	1.402	1.866	1.847
	(1.230)	(1.572)	(1.690)	(1.520)
Gender × Female	1.624	0.635	1.284	0.380
	(1.081)	(1.601)	(1.282)	(1.361)
Joint test of interactions, p-value (WCB)	0.238	0.675	0.630	0.512
Observations	7,722	5,900	6,127	4,941

Notes: OLS. The dependent variable is realized performance as defined in Table 8. All columns control for the selection rule and year fixed effects; full controls are those of column (8) of Table 4. Standard errors clustered at the school level (30 clusters) in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

for male and female students.Comparing the two prizes directly, the difference between the No Gender × Female and the Gender × Female interactions is 0.029 (95 percent confidence interval −0.014 to 0.072) with the selection-rule control and 0.023 (−0.053 to 0.099) without it, so we find no difference between the two prizes in the response of girls relative to boys. Table 10 reports the 95 percent confidence intervals and the minimum detectable effects. The estimated difference in participation responses between girls and boys under the Gender prize is close to zero. However, its 95 percent confidence interval (−5 to +7 percentage points) still allows the prize to close most of the 8–9 point participation gap. We therefore cannot conclude that the gap is unchanged; equivalence within a ±5-point margin is not established. For scores, point estimates suggest more negative effects for boys, but the gender interactions are not statistically significant.

6.6 Who participates?

We next examine which students participate in the Olympiad and how participants perform. To do so, we estimate regressions for participation and scores under both matching definitions in Appendix Table E2 and in Appendix D..

Focusing on participation, evidence is threefold. First, past participation is the strongest predictor of participation: students who took part in an earlier edition are about 22 percentage points more likely to enter again. By contrast,participation in other academic competitions, such as *Certamen*, a Latin-language competition, is not associated with entry once prior Olympiad participation is controlled for. Second, math-specific performance and interest are more predictive of participation than general academic achievement: the math grade and ranking math as the favorite subject predict participation, the average grade adds little on top, and the Raven score matters

Table 10: Treatment effects by gender: Gold-only

	Participation		Score	
	Base	All controls	Base	All controls
No gender treatment	0.253**	0.246*	-5.38*	-7.21**
	(0.079)	(0.105)	(2.26)	(1.58)
Gender treatment	0.164*	0.127	-4.74	-6.39*
	(0.074)	(0.097)	(3.55)	(2.47)
Female	-0.090**	-0.076**	-7.42***	-6.51***
	(0.028)	(0.028)	(2.23)	(1.67)
No gender × Female	0.035	0.034	4.11	2.21
	(0.034)	(0.041)	(2.47)	(1.81)
95% CI	[-0.031, 0.101]	[-0.047, 0.114]	[-0.74, 8.96]	[-1.34, 5.76]
MDE (80% power)	0.095	0.115	6.93	5.07
Gender × Female	0.006	-0.001	2.72	0.94
	(0.030)	(0.036)	(2.66)	(2.21)
95% CI	[-0.052, 0.065]	[-0.072, 0.070]	[-2.49, 7.93]	[-3.39, 5.27]
MDE (80% power)	0.083	0.102	7.44	6.19
Joint WB p-value (interactions)	0.444	0.579	0.330	0.534
Selection rule	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
All controls		✓		✓
Observations	7,722	5,900	2,635	1,878

Notes: Linear models with treatment-by-female interactions. Score columns condition on participation among matched participants and are read as descriptive. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications). For the interaction terms we also report the 95 percent confidence interval (from the clustered standard errors) and the minimum detectable effect: the smallest true interaction that a 5 percent two-sided test would detect with 80 percent power given the estimated standard error (2.8 times the standard error). The joint p-value tests both interactions with the wild cluster bootstrap. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

mainly through grades. Beliefs also play a role, since students who expect to rank higher than their classmates on the Raven test are more likely to participate. Third, the gender gap remains mostly unchanged after controlling for a rich set of individual characteristics: girls are about 6.5 percentage points less likely to participate than boys with the same grades, ability, beliefs, preferences, and family background.

Aspirations point in the same direction. Students who plan to study a STEM field at university are the most likely to participate: relative to them, participation is 7 to 12 percentage points lower among students who aim at social sciences and humanities, health, or other fields, and controlling for grades, ability, math preference, and past participation reduces these differences by about half. STEM intentions are not simply another predictor of participation, but an outcome that is itself relevant to our analysis. The gender gap in STEM intentions is large: 40 percent of boys against 22 percent of girls plan to pursue a STEM field at university (Table 3), independently of our treatments.

By contrast, we find little evidence that participation is systematically associated with the non-cognitive traits measured in the questionnaire. Competitiveness, patience, self-esteem, self-efficacy, and self-control do not significantly predict entry into the Olympiad. These measures

are nevertheless informative about performance among participants: competitiveness and patience are positively associated with Olympiad scores, while self-control is negatively associated with scores. Cognitive ability, grades, and performance beliefs are also strong predictors of performance. Risk attitudes provide a less consistent pattern: greater self-reported willingness to take risks is associated with lower participation, whereas the incentivized staircase measure shows no significant relationship. Finally, a gender gap of about five points remains after accounting for the full set of observable characteristics, with girls scoring lower than boys.

The questionnaire also included a Gender–Science Implicit Association Test (IAT), which measures the strength of implicit associations between gender and science. The relationship between IAT scores and participation differs by gender: higher D-scores, indicating stronger male-science associations, are associated with higher participation among boys and lower participation among girls. Moreover, the gender gap in participation is close to zero among students with the weakest associations and about 17 percentage points among students with the strongest ones.

Table 11 reports estimates of the association between students' IAT D-scores and Olympiad participation, including whether this association differs by gender. All columns are linear probability models with the same base controls as the main tables (treatment, entry rule, and year fixed effects); they differ only in how the D-score enters. Column (1) includes the D-score alone; columns (2) to (4) add its interaction with the female indicator, on the full sample, on the free margin, and with the full set of questionnaire controls; column (5) adds a quadratic term; column (6) replaces the continuous D-score with an indicator for an above-median D-score. On average, the D-score is a weak predictor of participation: students with a stronger male-science association participate slightly more, but the coefficient is small and not significant (column 1). The association differs by gender, with opposite signs for boys and girls. When we interact the continuous D-score with gender (column 2), the female-male participation gap widens by 13.0 percentage points per point of D-score (wild cluster bootstrap $p = 0.002$). The quadratic term is insignificant ($p = 0.65$; column 5), consistent with a linear relationship.

Figure 2 plots the gender gap in participation across the D-score distribution.The median split in column 6 offers an intuitive summary: boys with strong implicit stereotypes are about 7 percentage points more likely to participate than boys with weak stereotypes, the interaction for girls of -9.3 points more than offsets this difference, and the resulting gap is about 4 percentage points among students with weak stereotypes, a gap that is itself statistically significant, and about 13 percentage points among students with strong ones (wild cluster bootstrap $p = 0.0001$).

The association is robust. It is, if anything, stronger on the free margin, where students decide for themselves (16.1 points, $p < 0.001$, column 3), and it remains present, though less precisely estimated, with the full set of questionnaire controls (7.2 points, $p = 0.05$, column 4), so it is not driven by differences in ability, grades, beliefs, or family background. Appendix Table E18 shows that it is not driven by comparisons across schools or classes either: with the D-score standardized, the female-male gap widens by 5.0 percentage points per standard deviation in the baseline specification, by 4.4 points with school fixed effects, by 3.4 points with class fixed effects, by 2.8

points with the full set of controls (p = 0.05), and by 6.2 points on the free margin. Excluding one school at a time, the coefficient stays between −4.4 and −5.7 points and its wild bootstrap p-value is below 0.05 in 29 of the 30 cases.

Finally, we examine whether the prizes work differently for students with strong and weak implicit stereotypes. As shown in Appendix Table E19, the interactions between the treatment dummies and the D-score, as well as the triple interactions with gender, are small and not jointly significant. The only exception is the Gender × Female × high-IAT term in the median-split version (column 4), which we read as suggestive given the number of interactions examined. Implicit gender stereotypes are thus strongly associated with the gender gap in participation, but we find little evidence that they shape the response to the prizes.

Table 11: Implicit stereotypes and Olympiad participation

	(1) D-score only	(2) D-score × Female	(3) D-score × Female	(4) D-score × Female	(5) Quadratic in D-score	(6) Above-median D-score
Female	-0.083***	-0.063***	-0.068***	-0.055***	-0.060***	-0.037**
	(0.010)	(0.011)	(0.012)	(0.010)	(0.011)	(0.015)
D-score	0.025	0.095***	0.112***	0.042**	0.082***	
	(0.015)	(0.021)	(0.025)	(0.017)	(0.025)	
Female × D-score		-0.130***	-0.161***	-0.072*	-0.124**	
		(0.038)	(0.037)	(0.034)	(0.044)	
D-score^2					0.042	
					(0.042)	
Female × D-score^2					-0.021	
					(0.045)	
High IAT						0.067***
						(0.015)
Female × High IAT						-0.093***
						(0.023)
Base controls	Yes	Yes	Yes	Yes	Yes	Yes
Full questionnaire controls	No	No	No	Yes	No	No
Sample	All	All	Free margin	All	All	All
Observations	7,113	7,113	5,594	5,704	7,113	7,113

Notes: Linear probability models for participation (gold-only match); the sample includes questionnaire respondents with a valid D-score. Positive D-scores indicate a stronger implicit association between males and science; High IAT is a dummy for a D-score at or above the sample median. Base controls: treatment, selection rule, and year of study. Full questionnaire controls: as in the main tables. Column (3) excludes high-participation classes. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): the p-value for Female × D-score is 0.0021 in column (2), 0.0000 in column (3), and 0.0511 in column (4); the p-value for Female × High IAT in column (6) is 0.0001. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

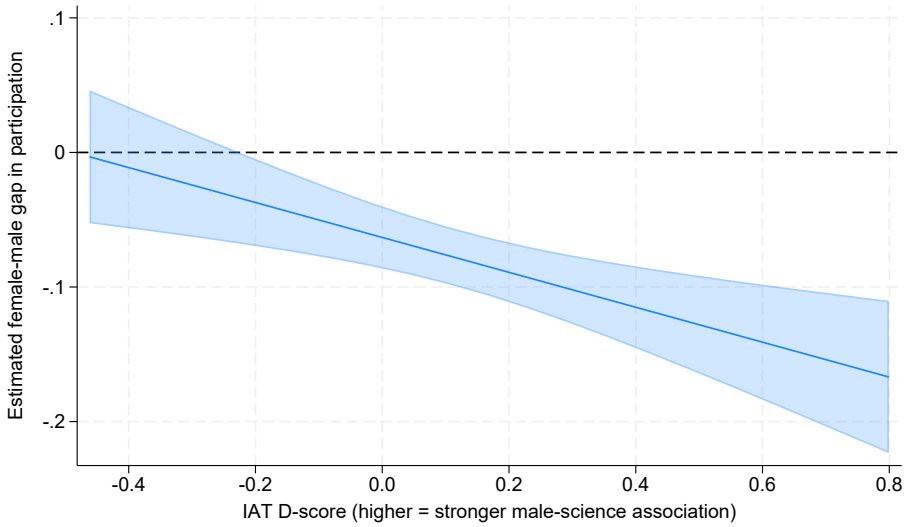

Figure 2: Estimated female-male gap in participation along the IAT D-score (5th to 95th percentile). Linear probability model interacting gender with the D-score (column 2 of Table 11), with treatment, selection-rule, and year controls; shaded area: 95 percent confidence interval, school-clustered.

6.7 Who responds to incentives?

We next explore whether our treatments differently affect participation depending on relevant individual characteristics: Raven score, math grades, confidence, previous Olympiad participation, and competitiveness. These measures allow us to assess whether incentives primarily attract students with characteristics associated with a higher propensity to participate, or broaden participation to those who would otherwise be less likely to enter the competition.

Appendix Table E5 reports these estimates for participation and Appendix Table E6 for Olympiad scores among participants. The two tables have the same structure: each column is a separate regression in which the treatment is interacted with one binary characteristic, which is equal to one when the individual is above the median of the Raven score, the math grade, the confidence measure, or the competitiveness scale, and for students who took part in the Olympiad before, respectively. Overall, the participation interaction estimates do not point to strong or systematic heterogeneity. The main positive participation effect of the No Gender treatment is present across specifications, but the interactions with Raven score, confidence, previous Olympiad participation, and competitiveness are not statistically significant. The only marginally significant interaction is between the Gender treatment and above-median math grade, suggesting a larger participation response among students with higher prior math performance. Because we examine several dimensions of heterogeneity, Appendix Table E17 reports Holm- and Bonferroni-adjusted p-values for the twelve treatment-by-characteristic interactions in participation. None is statistically significant after adjustment: the smallest wild cluster bootstrap p-value is 0.09, and all adjusted p-values equal one.

Similarly, among gold-only matched participants, the same specifications show little evidence that treatment-related differences in normalized scores vary across these characteristics (see Appendix Table E6). We find no systematic heterogeneity by family background either: treatment effects on participation and scores vary little with parental education or employment, measured separately for mothers and fathers, or with whether students have siblings (Appendix Tables E7 and E8).

Finally, focusing on the free margin, while treatments interactions with self-esteem, self-efficacy, competitiveness, risk attitudes, and patience are non significant and close to zero, we observe that beliefs about own relative performance in the Raven test play an important role. Specifically, if we define beliefs to be pessimistic (optimistic) when the student expects a rank lower (higher) than the median, we observe that the gendered prize raises participation by about 14 percentage points in most groups, but not among boys with pessimistic beliefs. This suggests that within-gender competition may be particularly relevant for girls with less optimistic beliefs, although the interaction estimates are imprecise (wild cluster bootstrap p-values of 0.13 and 0.14). Consistent with this interpretation, when underconfidence is defined based on the residual from a regression of expected rank on ability, grades, and year of study, the triple interaction is positive (0.104, SE 0.051; wild cluster bootstrap $p = 0.04$). Differently, treatment interactions are not significant when underconfidence is measured by the gap between expected and actual within-class Raven rank, suggesting a cautious approach.

6.8 Selection-rule split specifications

As a complementary analysis, we split schools according to their participation rules. The results are consistent with the class-level analysis in Section 6.2: the larger effect of the No Gender treatment is concentrated in Mixed schools, whereas the two prize schemes have similar effects in schools where participation is primarily voluntary. Pooled interaction tests indicate heterogeneity across rules for both participation (wild cluster bootstrap $p = 0.09$ and 0.05) and realized performance ($p = 0.04$ and 0.03; Appendix Table B5). In Mixed schools, No Gender raises participation by 38 percentage points ($p = 0.008$) and realized performance by about 8 points ($p = 0.06$), while Gender has no effect. Where participation is primarily voluntary, both schemes raise participation by 11–13 percentage points; estimated effects on realized performance are positive but imprecise. Full results appear in Appendix Tables B1–B4.

7 Discussion

The design of incentives in competitions can shape not only participation, but also the gender composition of entrants and the performance observed among them. Gender-specific prizes alter the expected returns to participation for girls and boys; whether this changes who enters, and whose performance is observed, depends on how students respond. Our experiment examines these questions in the Math Olympiad.

Focusing on participation, we observe that when comparing schools according to a raw criterion that does not consider differences in organizational rules when accessing the Math Olympiad, neither prize significantly increases participation. However, once we hold the entry rules constant, participation with respect to the Baseline schools is higher under the No Gender prize in every specification and, by a smaller amount, under the Gender prize. Among participants, scores are lower in treated schools only where entry relies on students' own initiative, not where participation is near universal.

Girls and boys are not significantly different in their response to either prize, even though the Gender prize roughly doubles the expected return to participation for girls.

Our interpretation of the score results is a change in the composition of participants. The lower scores among participants in treated schools appear only where students decide for themselves, and the effect of the prizes on total points earned, which counts zero for non-participants, is small. Both facts are consistent with the prizes attracting additional students who score lower on average, rather than lowering the scores of students who would have participated anyway. We cannot separate the two explanations formally, because we do not observe the scores that non-participants would have obtained. We also estimated a Heckman selection model, in which the entry rule enters the participation equation but not the score equation. It gives the same qualitative picture, but its exclusion restriction is not robust: the entry rule is not randomly assigned and may be related to school characteristics that affect scores directly. Bounds following Lee (2009) are not a viable alternative either, because the monotonicity assumption they require is hard to defend in our setting. Total points earned, which does not condition on participation, is our preferred complement to the score regressions.

Finally, we observe that the gender gap in participation is strongly related to the male-science association measured by the IAT: the gap is about 2 percentage points, and not statistically significant, among students with weak associations, and about 15 percentage points among students with strong ones. The association is smaller but still present with the full set of controls for ability, grades, and beliefs, and the prizes do not change it. This outcome relates to the null result observed when considering the effects of the Gender prize. Specifically, if the gap is related to identity and stereotypes rather than to the monetary return to participation, changing the prize alone should not be expected to close it.

By combining random assignment across 30 high schools in one edition of the Math Olympiad with detailed questionnaire data from more than 7,000 students, our study estimates the effects of (gendered) prize schemes on participation and realized performance and examines how these effects vary across students. While it provides detailed evidence on prize incentives in an established national competition, evidence from other cohorts and settings would help assess how broadly the findings apply. The questionnaire also documents associations between participation, performance beliefs, and implicit gender stereotypes, which should however be interpreted as descriptive.

8 Conclusions

This paper studies whether monetary prizes raise participation in a real mathematics competition, whether the design of the prize can reduce the gender gap in entry, and which students participate and respond to the prizes. We ran a field experiment in thirty high schools during the first stage of the Italian Math Olympiad and linked a detailed questionnaire to the competition records.

We find no statistically significant average increase in participation. Participation is higher in treated schools among students who decide for themselves whether to enter the competition, once we account for the entry rules of the schools: by about 12 percentage points with our conservative matching and by about half of that with the broader matching. Among participants, scores are lower in treated schools, a pattern consistent with the prizes attracting weaker students.

We find no statistically significant difference in the response of girls and boys. The Gender prize does not attract girls more than boys, and the female participation gap is present in every treatment, although our confidence intervals do not exclude reductions of a few percentage points. The gap does not mirror differences in measured ability: girls have the same Raven scores as boys and higher grades, including in mathematics. It lines up instead with performance beliefs and with implicit gender stereotypes. Girls tend to underestimate their rank relative to their classmates, while boys tend to overestimate theirs, and the gender gap in participation is about 2 percentage points among students with weak male-science associations and about 15 points among students with strong ones. We find no significant evidence that either prize changes these patterns.

Our results have two main implications. Monetary prizes may widen participation in math competitions where students decide for themselves, although our design does not detect an average effect across schools. And if the goal is to bring more girls in, changing the monetary return of the competition is not enough: we detect no narrowing of the gap under either prize. Interventions that act on stereotypes and beliefs, early and repeatedly, are the natural candidates, and testing them in the field is the next step.

Declaration on the use of AI

During the preparation of this work, the authors used ChatGPT (OpenAI) to assist with coding, language editing, and LaTeX formatting. The authors reviewed and edited all content and take full responsibility for the final manuscript.

References

Attanasio, O., Kugler, A., and Meghir, C. (2011). Subsidizing vocational training for disadvantaged youth in Colombia: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, 3(3):188–220.

Bandiera, O., Parekh, N., Petrongolo, B., and Rao, M. (2022). Men are from Mars, and women too: A Bayesian meta-analysis of overconfidence experiments. *Economica*, 89(S1):S38–S70.

Berge, L. I. O., Bjorvatn, K., Pires, A. J. G., and Tungodden, B. (2015). Competitive in the lab, successful in the field? *Journal of Economic Behavior & Organization*, 118:303–317.

Berger, J., Osterloh, M., and Rost, K. (2020). Focal random selection closes the gender gap in competitiveness. *Science Advances*, 6(47):eabb2142.

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., and Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven's Standard Progressive Matrices test. *Assessment*, 19(3):354–369.

Brandts, J., El Baroudi, S., Huber, S. J., and Rott, C. (2021). Gender differences in private and public goal setting. *Journal of Economic Behavior and Organization*, 192:222–247.

Buser, T., Geijtenbeek, L., and Plug, E. (2018). Sexual orientation, competitiveness and income. *Journal of Economic Behavior & Organization*, 151:191–198.

Buser, T. and Yuan, H. (2019). Do women give up competing more easily? Evidence from the lab and the Dutch Math Olympiad. *American Economic Journal: Applied Economics*, 11(3):225–252.

Calsamiglia, C., Franke, J., and Rey-Biel, P. (2013). The incentive effects of affirmative action in a real-effort tournament. *Journal of Public Economics*, 98:15–31.

Coffman, K. and Klinowski, D. (2024). Gender and preferences for performance feedback. *Management Science*. Articles in advance.

Coffman, K., Ugalde Araya, M. P., and Zafar, B. (2024). A (dynamic) investigation of stereotypes, belief-updating, and behavior. *Economic Inquiry*, 62(4):957–983.

Cotton, C. S., Hickman, B. R., and Price, J. P. (2022). Affirmative action and human capital investment: Evidence from a randomized field experiment. *Journal of Labor Economics*, 40(1):157–185.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–474.

Dilmaghani, M. (2022). Chess girls don't cry: Gender composition of games and effort in competitions among the super-elite. *Journal of Economic Psychology*, 89:102482.

Ellison, G. and Swanson, A. (2018). Dynamics of the gender gap in high math achievement. Working paper, MIT and NBER.

Falk, A., Becker, A., Dohmen, T., Huffman, D., and Sunde, U. (2023). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science*, 69(4):1935–1950.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480.

Greenwald, A. G., Nosek, B. A., and Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, 85(2):197–216.

Harris, P. B. and Houston, J. M. (2010). A reliability analysis of the revised competitiveness index. *Psychological Reports*, 106(3):870–874.

Joensen, J. S. and Nielsen, H. S. (2009). Is there a causal effect of high school math on labor market outcomes? *Journal of Human Resources*, 44(1):171–198.

Kang, L., Lei, Z., Song, Y., and Zhang, P. (2024). Gender differences in reactions to failure in high-stakes competition: Evidence from the National College Entrance Exam retakes. *Journal of Political Economy Microeconomics*, 2(2):355–397.

Landaud, F. and Maurin, E. (2020). Aim high and persevere! Competitive pressure and access gaps in top science graduate programs. PSE Working Paper 2020-79, Paris School of Economics.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071–1102.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *Quarterly Journal of Economics*, 122(3):1067–1101.

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

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002). Math = male, me = female, therefore math $\neq$ me. *Journal of Personality and Social Psychology*, 83(1):44–59.

OECD (2025). *Gender Equality in a Changing World: Taking Stock and Moving Forward*. Gender Equality at Work. OECD Publishing, Paris.

Reuben, E., Sapienza, P., and Zingales, L. (2015). Taste for competition and the gender gap among young business professionals. NBER Working Paper 21695, National Bureau of Economic Research.

Reuben, E., Wiswall, M., and Zafar, B. (2017). Preferences and biases in educational choices and labour market expectations: Evidence from a sample of elite students. *The Economic Journal*, 127(604):2153–2186.

Rosenberg, M. (1965). *Society and the Adolescent Self-Image*. Princeton University Press, Princeton, NJ.

Schwarzer, R. and Jerusalem, M. (1995). Generalized self-efficacy scale. In Weinman, J., Wright, S., and Johnston, M., editors, *Measures in Health Psychology: A User's Portfolio*, pages 35–37. NFER-Nelson, Windsor, UK.

Shastry, G. K., Shurchkov, O., and Xia, L. L. (2020). Luck or skill: How women and men react to noisy feedback. *Journal of Behavioral and Experimental Economics*, 88:101592.

Tangney, J. P., Baumeister, R. F., and Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, 72(2):271–324.

Yuret, T. (2026). Career paths of European Girls' Mathematical Olympiad (EGMO) medalists. *Gifted Child Quarterly*. Online first.

Zhang, Y. J. (2013). Can experimental economics explain competitive behavior outside the lab? Working paper, SSRN 2292929.

A Robustness using Gold+C1+C2 matches

This appendix reports the same specifications using the expanded participation definition, where gold, C1, and C2 matches all count as Olympiad participation. Table A1 is the counterpart of Table 4: linear probability models for participation with the same eight columns, from the unadjusted comparison to the full set of questionnaire controls. Table A2 is the counterpart of Appendix Table E4: normalized Olympiad scores among matched participants, with the same columns. Table A3 is the counterpart of Table 10: treatment-by-female interactions for participation and scores. As discussed in Section 4, the expanded estimates are systematically smaller than the gold-only ones.

Table A1: Treatment effects on Olympiad participation: Gold+C1+C2

	(1) ITT	(2) Base	(3) + Demog.	(4) + Ability	(5) + Academic	(6) + Past	(7) + Aspir.	(8) + Non-cognitive traits
No gender treatment	0.062	0.217**	0.216**	0.224**	0.252**	0.221*	0.221*	0.224*
	(0.1218)	(0.0840)	(0.0848)	(0.0821)	(0.0942)	(0.0945)	(0.0947)	(0.0962)
Gender treatment	-0.156*	0.117	0.121	0.118	0.111	0.080	0.083	0.080
	(0.0793)	(0.0762)	(0.0778)	(0.0772)	(0.0940)	(0.0917)	(0.0927)	(0.0939)
Selection rule		✓	✓	✓	✓	✓	✓	✓
Year FE		✓	✓	✓	✓	✓	✓	✓
Gender		✓	✓	✓	✓	✓	✓	✓
Demographics			✓	✓	✓	✓	✓	✓
Ability/beliefs				✓	✓	✓	✓	✓
Academic performance					✓	✓	✓	✓
Past competitions						✓	✓	✓
Aspirations							✓	✓
Non-cognitive traits								✓
Observations	7,722	7,722	7,679	7,592	5,993	5,987	5,987	5,900

Notes: Robustness using the expanded gold+C1+C2 matching definition. Column (1) reports the unadjusted comparison with treatment indicators only; later columns add the controls indicated in the panel. Standard errors clustered at the school level (30 clusters) in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A2: Olympiad scores among participants, by treatment: Gold+C1+C2

	(1) ITT	(2) Base	(3) + Demog.	(4) + Ability	(5) + Academic	(6) + Past	(7) + Aspir.	(8) + Non-cognitive traits
No gender treatment	-3.393	-2.576	-2.792	-2.739	-5.797**	-5.747**	-5.782**	-5.730**
	(2.0745)	(1.8671)	(1.7806)	(1.4056)	(1.6505)	(1.7379)	(1.7332)	(1.8069)
Gender treatment	-2.621	-3.178	-2.605	-2.386	-5.949**	-6.022**	-6.037**	-5.731**
	(2.7828)	(2.9258)	(2.7971)	(2.1087)	(2.1920)	(2.1517)	(2.1340)	(2.1170)
Selection rule		✓	✓	✓	✓	✓	✓	✓
Year FE		✓	✓	✓	✓	✓	✓	✓
Gender		✓	✓	✓	✓	✓	✓	✓
Demographics			✓	✓	✓	✓	✓	✓
Ability/beliefs				✓	✓	✓	✓	✓
Academic performance					✓	✓	✓	✓
Past competitions						✓	✓	✓
Aspirations							✓	✓
Non-cognitive traits								✓
Observations	2,906	2,906	2,906	2,901	2,075	2,072	2,072	2,059

Notes: Robustness using the expanded gold+C1+C2 matching definition. Column (1) reports the unadjusted comparison with treatment indicators only; later columns add the controls indicated in the panel. Standard errors clustered at the school level (30 clusters) in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Treatment effects by gender: Gold+C1+C2

	Participation		Score	
	Base	All controls	Base	All controls
No gender treatment	0.202**	0.205	-4.44*	-6.85***
	(0.081)	(0.106)	(2.01)	(1.42)
Gender treatment	0.117	0.084	-4.38	-6.23**
	(0.074)	(0.096)	(3.25)	(2.31)
Female	-0.091*	-0.075*	-6.86***	-6.48***
	(0.028)	(0.029)	(1.84)	(1.40)
No gender × Female	0.028	0.033	3.59	2.20
	(0.035)	(0.042)	(2.13)	(1.60)
95% CI	[-0.042, 0.097]	[-0.049, 0.115]	[-0.59, 7.76]	[-0.94, 5.34]
MDE (80% power)	0.099	0.118	5.96	4.48
Gender × Female	-0.001	-0.008	2.36	0.98
	(0.030)	(0.035)	(2.27)	(1.94)
95% CI	[-0.060, 0.058]	[-0.077, 0.061]	[-2.09, 6.82]	[-2.83, 4.79]
MDE (80% power)	0.085	0.099	6.36	5.44
Joint WB p-value (interactions)	0.531	0.501	0.324	0.482
Selection rule	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
All controls		✓		✓
Observations	7,722	5,900	2,906	2,059

Notes: Linear models with treatment-by-female interactions. Score columns condition on participation among matched participants and are read as descriptive. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications). For the interaction terms we also report the 95 percent confidence interval (from the clustered standard errors) and the minimum detectable effect: the smallest true interaction that a 5 percent two-sided test would detect with 80 percent power given the estimated standard error (2.8 times the standard error). The joint p-value tests both interactions with the wild cluster bootstrap. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.1 Heterogeneity robustness

This appendix reports the same theoretically motivated heterogeneity specifications using the expanded Gold+C1+C2 match definition. Tables A4 and A5 are the counterparts of Appendix Tables E5 and E6: each column interacts the treatment indicators with one characteristic H_i (above-median Raven score, math grade, confidence, or competitiveness, and previous Olympiad participation), for participation and for scores among participants. Tables A6 and A7 are the counterparts of Appendix Tables E7 and E8, with family-background indicators in place of H_i.

Table A4: Who responds to incentives? Gold+C1+C2 participation

Characteristic H_i:	Raven score above median	Math grade above median	Confidence above median	Previous Olympiad participated before	Competitiveness above median
No gender treatment	0.160	0.183	0.227*	0.257*	0.246**
	(0.105)	(0.110)	(0.094)	(0.112)	(0.092)
Gender treatment	0.065	0.025	0.086	0.089	0.099
	(0.102)	(0.103)	(0.088)	(0.090)	(0.092)
H_i (characteristic in column header)	-0.028	-0.064	0.026	0.265***	0.046
	(0.045)	(0.036)	(0.037)	(0.050)	(0.028)
No gender $\times$ H_i (characteristic in column header)	0.100*	0.075	-0.004	-0.086	-0.041
	(0.052)	(0.047)	(0.040)	(0.079)	(0.037)
Gender $\times$ H_i (characteristic in column header)	0.017	0.099*	-0.010	-0.024	-0.035
	(0.050)	(0.043)	(0.039)	(0.061)	(0.037)
Selection rule	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
Female	✓	✓	✓	✓	✓
All other controls	✓	✓	✓	✓	✓
Observations	5,900	5,900	5,900	5,900	5,900

Notes: The dependent variable is Olympiad participation using the expanded Gold+C1+C2 match definition. Each column reports a separate regression interacting treatment with the characteristic named in the column, estimated on the complete-case sample of the full specification, so the sample is constant across columns. When a heterogeneity variable is constructed from a control, that source control is omitted from that column to avoid mechanical collinearity. In the Previous Olympiad column the heterogeneity variable is past participation itself. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 4,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Heterogeneity in Olympiad scores: Gold+C1+C2

Characteristic H_i:	Raven score above median	Math grade above median	Confidence above median	Previous Olympiad participated before	Competitiveness above median
No gender treatment	-3.47	-2.34	-5.21*	-4.32**	-5.04
	(2.28)	(1.68)	(2.26)	(1.55)	(2.26)
Gender treatment	-4.72	-3.40	-6.13**	-5.35**	-5.65*
	(2.69)	(1.78)	(2.31)	(1.74)	(2.55)
H_i (characteristic in column header)	4.42*	3.50**	-0.49	2.76*	1.49
	(1.79)	(1.49)	(1.24)	(1.26)	(1.51)
No gender $\times$ H_i (characteristic in column header)	-3.10	-5.27***	-1.00	-2.58	-1.27
	(2.01)	(1.86)	(1.59)	(1.73)	(1.73)
Gender $\times$ H_i (characteristic in column header)	-1.85	-3.17	0.57	-0.90	-0.16
	(2.40)	(2.55)	(1.73)	(1.92)	(1.92)
Selection rule	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
Female	✓	✓	✓	✓	✓
All other controls	✓	✓	✓	✓	✓
Observations	2,059	2,059	2,059	2,059	2,059

Notes: The dependent variable is the normalized Olympiad score among Gold+C1+C2 matched participants (descriptive, since the sample conditions on participation). Each column reports a separate regression interacting treatment with the characteristic named in the column, estimated on the complete-case sample of the full specification, so the sample is constant across columns. When a heterogeneity variable is constructed from a control, that source control is omitted from that column to avoid mechanical collinearity. In the Previous Olympiad column the heterogeneity variable is past participation itself. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 4,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: Family-background heterogeneity: Gold+C1+C2 participation

Indicator H_i:	Father post-sec./univ. yes/no	Mother post-sec./univ. yes/no	Father employed yes/no	Mother employed yes/no	Siblings yes/no
No gender treatment	0.198	0.193	0.231**	0.211*	0.201*
	(0.101)	(0.101)	(0.096)	(0.095)	(0.087)
Gender treatment	0.063	0.062	0.150	0.077	0.060
	(0.095)	(0.093)	(0.091)	(0.090)	(0.085)
H_i (indicator in column header)	-0.064**	-0.051	0.041	-0.007	-0.028
	(0.027)	(0.040)	(0.021)	(0.020)	(0.036)
No gender $\times$ H_i (indicator in column header)	0.100**	0.092*	-0.008	0.017	0.029
	(0.039)	(0.046)	(0.048)	(0.034)	(0.059)
Gender $\times$ H_i (indicator in column header)	0.069*	0.054	-0.077*	0.003	0.025
	(0.039)	(0.047)	(0.041)	(0.033)	(0.038)
Selection rule	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
Female	✓	✓	✓	✓	✓
All other controls	✓	✓	✓	✓	✓
Observations	5,900	5,900	5,900	5,900	5,900

Notes: The dependent variable is Olympiad participation using the expanded Gold+C1+C2 match definition. Each column reports a separate regression interacting treatment with the characteristic named in the column, estimated on the complete-case sample of the full specification, so the sample is constant across columns. When a heterogeneity variable is constructed from a control, that source control is omitted from that column to avoid mechanical collinearity. In the Previous Olympiad column the heterogeneity variable is past participation itself. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 4,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: Family-background heterogeneity in Olympiad scores: Gold+C1+C2

Indicator H_i:	Father post-sec./univ. yes/no	Mother post-sec./univ. yes/no	Father employed yes/no	Mother employed yes/no	Siblings yes/no
No gender treatment	-4.02*	-3.86	-1.50	-4.74*	-7.69**
	(1.80)	(2.05)	(2.88)	(2.07)	(2.54)
Gender treatment	-5.48**	-4.74*	-4.28	-5.56*	-9.29***
	(1.99)	(2.21)	(3.44)	(2.48)	(2.71)
H_i (indicator in column header)	3.40	4.87**	3.17	0.91	-4.14*
	(1.91)	(2.12)	(1.77)	(1.71)	(1.65)
No gender $\times$ H_i (indicator in column header)	-6.26**	-5.22**	-4.64*	-1.22	2.48
	(2.11)	(2.28)	(2.17)	(1.92)	(1.74)
Gender $\times$ H_i (indicator in column header)	-1.64	-3.01	-1.64	-0.21	4.49*
	(2.37)	(2.50)	(3.18)	(1.85)	(2.14)
Selection rule	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
Female	✓	✓	✓	✓	✓
All other controls	✓	✓	✓	✓	✓
Observations	2,059	2,059	2,059	2,059	2,059

Notes: The dependent variable is the normalized Olympiad score among Gold+C1+C2 matched participants (descriptive, since the sample conditions on participation). Each column reports a separate regression interacting treatment with the characteristic named in the column, estimated on the complete-case sample of the full specification, so the sample is constant across columns. When a heterogeneity variable is constructed from a control, that source control is omitted from that column to avoid mechanical collinearity. In the Previous Olympiad column the heterogeneity variable is past participation itself. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 4,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B Selection-rule split specifications

This appendix reports additional specifications that split the sample by the school-level selection rule used for Olympiad participation. Tables B1 and B2 keep Baseline schools in the estimation sample and report coefficients for No Gender and Gender relative to Baseline. Tables B3 and B4 exclude Baseline schools and report the direct Gender-minus-No-gender comparison.

Table B1: Participation by selection-rule sample: all arms

	Base		All controls		Observations	
Selection-rule sample	No gender	Gender	No gender	Gender	Base	All controls
Gold-only						
Free + Free/Suggested	0.128	0.144	0.098	0.131	3,998	3,096
	(0.066)	(0.046)	(0.068)	(0.051)		
Mixed	0.369***	0.088	0.393***	0.009	3,724	2,804
	(0.079)	(0.075)	(0.096)	(0.104)		
Gold+C1+C2						
Free + Free/Suggested	-0.075	-0.037	-0.094	-0.057	3,998	3,096
	(0.161)	(0.149)	(0.138)	(0.129)		
Mixed	0.348***	0.056	0.385***	-0.010	3,724	2,804
	(0.077)	(0.077)	(0.091)	(0.105)		

Notes: Coefficients on the treatment indicators relative to Baseline, estimated separately in schools with primarily voluntary participation (Free and Free-or-Suggested) and in Mixed schools. All specifications include the school-level selection rule (omitted where constant within the estimation sample), year-of-study fixed effects, and gender; All controls adds the full set of questionnaire controls. Score regressions condition on participation among matched participants and are descriptive. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B2: Olympiad scores among participants by selection-rule sample: all arms

	Base		All controls		Observations	
Selection-rule sample	No gender	Gender	No gender	Gender	Base	All controls
Gold-only						
Free + Free/Suggested	11.66	9.94	5.84	4.35	678	509
	(2.37)	(2.05)	(5.02)	(5.11)		
Mixed	-4.41	-3.13	-6.97**	-5.85	1,957	1,369
	(2.19)	(3.94)	(2.09)	(2.82)		
Gold+C1+C2						
Free + Free/Suggested	3.47	0.46	-0.62	-2.39	855	638
	(2.19)	(2.50)	(2.10)	(2.31)		
Mixed	-4.16	-2.88	-6.85**	-5.50	2,051	1,421
	(1.98)	(3.77)	(2.01)	(2.69)		

Notes: Coefficients on the treatment indicators relative to Baseline, estimated separately in schools with primarily voluntary participation (Free and Free-or-Suggested) and in Mixed schools. All specifications include the school-level selection rule (omitted where constant within the estimation sample), year-of-study fixed effects, and gender; All controls adds the full set of questionnaire controls. Score regressions condition on participation among matched participants and are descriptive. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B3: Participation, Gender vs No gender by selection-rule sample (Baseline excluded)

	Gender vs No gender		Observations	
Selection-rule sample	Base	All controls	Base	All controls
Gold-only				
Free + Free/Suggested	0.020	0.036	3,800	2,954
	(0.054)	(0.046)		
Mixed	-0.288***	-0.374**	2,031	1,575
	(0.071)	(0.099)		
Gold+C1+C2				
Free + Free/Suggested	0.024	0.026	3,800	2,954
	(0.050)	(0.049)		
Mixed	-0.299***	-0.388**	2,031	1,575
	(0.069)	(0.098)		

Notes: Direct comparison between the Gender and No gender prize schemes, excluding Baseline schools. All specifications include the school-level selection rule (omitted where constant within the estimation sample), year-of-study fixed effects, and gender; All controls adds the full set of questionnaire controls. Score regressions condition on participation among matched participants and are descriptive. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B4: Olympiad scores among participants, Gender vs No gender by selection-rule sample (Baseline excluded)

	Gender vs No gender		Observations	
Selection-rule sample	Base	All controls	Base	All controls
Gold-only				
Free + Free/Suggested	-1.71	-1.68	670	502
	(2.80)	(1.64)		
Mixed	1.13	1.00	1,282	958
	(3.35)	(2.34)		
Gold+C1+C2				
Free + Free/Suggested	-3.11	-2.04	803	600
	(3.07)	(1.93)		
Mixed	1.13	1.18	1,308	978
	(3.27)	(2.28)		

Notes: Direct comparison between the Gender and No gender prize schemes, excluding Baseline schools. All specifications include the school-level selection rule (omitted where constant within the estimation sample), year-of-study fixed effects, and gender; All controls adds the full set of questionnaire controls. Score regressions condition on participation among matched participants and are descriptive. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B5: Heterogeneity by selection rule: pooled interaction specifications

	Participation (Gold-only)		Realized performance	
	(1)	(2)	(3)	(4)
No gender treatment	0.125	-0.001	4.492	0.427
	(0.065)	(0.011)	(1.993)	(0.730)
Gender treatment	0.107	0.073	3.840	2.214
	(0.032)	(0.027)	(1.406)	(0.855)
Mixed rule	0.258		10.555	
	(0.069)		(2.521)	
Free or suggested rule		-0.048		-3.239
		(0.023)		(0.757)
Mixed rule		0.239		9.283
		(0.068)		(2.313)
No gender × Mixed	0.258		3.943	
	(0.114)		(3.852)	
Gender × Mixed	-0.110		-6.324	
	(0.102)		(2.953)	
No gender × Free or suggested		0.227		7.922
		(0.093)		(2.482)
Gender × Free or suggested		0.134		5.272
		(0.032)		(0.933)
No gender × Mixed		0.385		8.006
		(0.094)		(3.339)
Gender × Mixed		-0.073		-4.652
		(0.101)		(2.731)
Joint test of interactions, p-value (WCB)	0.090	0.047	0.035	0.025
No gender effect in Mixed schools [p-value]	0.384 [0.008]	0.384 [0.008]	8.436 [0.063]	8.433 [0.063]
Gender effect in Mixed schools [p-value]	-0.003 [0.981]	-0.001 [0.995]	-2.484 [0.423]	-2.438 [0.434]
Year FE, gender, full controls	✓	✓	✓	✓
Observations	5,900	5,900	5,900	5,900

Notes: OLS on the full analysis sample. Columns (1) and (3) interact the treatment indicators with an indicator for Mixed schools (omitted: Free and Free or suggested schools); columns (2) and (4) interact them with the three selection rules (omitted: Free). Realized performance is defined in Table 8. The main treatment coefficients give the effect in the omitted rule category; the rows at the bottom report the implied effect in Mixed schools (main effect plus interaction) with its wild cluster bootstrap p-value. Full controls are those of column (8) of Table 4. Standard errors clustered at the school level (30 clusters) in parentheses; stars and joint tests are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

C Variable definitions and scales

This appendix defines the baseline variables used as controls and heterogeneity dimensions, and reports their observed ranges in the analysis sample. Table C1 gives, for each variable, the questionnaire item or instrument it comes from and how it is coded; Table C2 reports the theoretical scale of each measure and the minimum, maximum, and number of valid observations in the analysis sample. The definitions follow the variable construction used in the analysis code. Missing values are left as missing throughout the analysis.

Table C1: Definitions of baseline variables

Variable	Definition	Coding / scale
Female	Student gender indicator.	1 = female; 0 = male.
Year of study	Student's school year.	Categories 1–5.
Raven score	Number of correct answers in the questionnaire problem-solving block, based on nine Raven-style items. Within an attempted Raven block, unanswered items are scored as incorrect; if all nine items are missing, the score is missing.	0–9 correct answers.
Math grade	Self-reported grade in mathematics.	Italian 1–10 grading scale.
Average grade	Self-reported overall average grade.	Italian 1–10 grading scale.
Expected correct answers	Student's expected number of correct answers in the problem-solving block.	0–9 expected answers.
Confidence	Student's certainty about the expected number of correct answers.	0–10 scale.
Perceived class rank	Student's self-reported perceived position in class.	Rank, 1 = best.
Mathematics preference rank	Student's rank of mathematics among the listed subjects.	1 = most preferred; 3 = least preferred.
Previous Olympiad participation	Indicator for having previously participated in the Math Olympiad.	1 = yes; 0 = no.
Competitiveness	Index built from the questionnaire competitiveness items, using the original project recoding.	Higher values indicate higher competitiveness.
Risk index	Composite risk-preference measure combining the incentivized lottery staircase and self-reported willingness to take risks, using the weights adopted in the original project code.	Standardized index.
IAT D-score	Standard D-score of the Gender-Science Implicit Association Test; positive values indicate a stronger implicit association between males and science.	Approximately −2 to 2; sample range −1.24–1.40.
Self-esteem	Scale built from the ten self-esteem items in the questionnaire, with reverse coding where required.	Higher values indicate higher self-esteem.
Self-efficacy	Scale built from the ten self-efficacy items in the questionnaire. If all component items are missing, the scale is missing.	Higher values indicate higher self-efficacy.
Self-control	Scale built from the self-control items in the questionnaire, with reverse coding where required.	Higher values indicate higher self-control.

Notes: Definitions are based on the variable construction in the project do-files. All variables are measured before the Olympiad. Missing values are not imputed.

Table C2: Variable scales and ranges

Variable	N	Mean	SD	Range/categories
Female	7,722	0.533	0.499	0.000–1.000
Year of study	7,722			1, 2, 3, 4, 5
Raven score (0–9)	7,623	4.850	1.754	0.000–9.000
Math grade	5,997	7.036	1.329	1.000–10.000
Average grade	5,995	7.412	0.893	1.000–10.000
Expected correct answers	7,619	5.472	2.089	0.000–9.000
Confidence	7,615	6.514	2.252	0.000–10.000
Perceived class rank	7,597	9.944	5.819	-6.000–40.000
Previous Olympiad participation	5,989	0.362	0.481	0.000–1.000
Competitiveness	7,462	42.689	8.381	13.000–65.000
Risk index	7,461	0.000	0.766	-2.454–2.117
IAT D-score	7,113	0.157	0.382	-1.238–1.400
Self-esteem	7,452	28.453	5.209	10.000–40.000
Self-efficacy	7,470	28.554	4.064	10.000–40.000
Self-control	7,473	42.788	8.041	13.000–65.000

Notes: Descriptive statistics for the main baseline controls in the analysis sample. For continuous and dummy variables, the last column reports the observed minimum and maximum; for categorical variables, the observed categories. Missing values are left as missing: scale scores are set to missing for respondents who answered none of the underlying items, and the IAT D-score is missing when the test was not completed.

D Full regression tables

This appendix reports full coefficient tables for the main gold-only specifications. Table D1 reports every coefficient of the all-controls participation specification, column (8) of Table 4; Table D2 does the same for the all-controls score specification among participants, column (8) of Appendix Table E4. Appendix Table E2 in the main text extracts the coefficients discussed in Section 6.6 from these tables. These tables are useful for assessing which baseline characteristics predict Olympiad participation and scores.

Table D1: Full participation regression: Gold-only, all-controls specification

	(1)
Baseline	0.0000
	(.)
No gender	0.2653***
	(0.0917)
Gender	0.1273
	(0.0919)
Free	0.0000
	(.)
Free or Suggested	0.0536
	(0.0774)
Mixed - Free and Mandatory	0.3424***
	(0.0832)
year=2	0.0000
	(.)
year=3	-0.1171**
	(0.0433)
year=4	-0.1305***
	(0.0469)
year=5	-0.1498***
	(0.0430)
Female=0	0.0000
	(.)
Female=1	-0.0651***
	(0.0117)
Father's education=1	0.0000
	(.)
Father's education=2	0.0094
	(0.0191)
Father's education=3	0.0185
	(0.0140)
Father's education=4	0.0088
	(0.0160)

Father's education=5	0.0254
	(0.0268)
Mother's education=1	0.0000
	(.)
Mother's education=2	-0.0049
	(0.0243)
Mother's education=3	-0.0007
	(0.0184)
Mother's education=4	0.0128
	(0.0208)
Mother's education=5	-0.0082
	(0.0288)
Unemployed	0.0000
	(.)
White collar	-0.0265
	(0.0400)
Blue collar	-0.0174
	(0.0412)
Entrepreneur	-0.0499
	(0.0412)
Retired	-0.0423
	(0.0483)
Don't know	-0.0318
	(0.0356)
Unemployed	0.0000
	(.)
White collar	-0.0166
	(0.0169)
Blue collar	-0.0111
	(0.0168)
Entrepreneur	0.0138
	(0.0172)
Retired	0.1081
	(0.0803)
Don't know	-0.0258
	(0.0358)
Foreign student=0	0.0000
	(.)
Foreign student=1	0.0113
	(0.0238)
Father born abroad=0	0.0000
	(.)
Father born abroad=1	-0.0083

	(0.0212)
Mother born abroad=0	0.0000
	(.)
Mother born abroad=1	-0.0170
	(0.0210)
Non-Italian language at home=0	0.0000
	(.)
Non-Italian language at home=1	0.0038
	(0.0316)
Number of siblings	-0.0012
	(0.0014)
Raven score, 0-9	0.0099**
	(0.0046)
Expected correct answers	-0.0012
	(0.0028)
Confidence in expected answers	0.0031
	(0.0033)
Perceived class rank	-0.0033**
	(0.0015)
Confidence in perceived rank	-0.0019
	(0.0025)
Math grade	0.0204**
	(0.0097)
Average grade	0.0118
	(0.0108)
Mathematics preference rank	-0.0241***
	(0.0076)
Previous Certamen participation=0	0.0000
	(.)
Previous Certamen participation=1	0.0303
	(0.0299)
Previous Olympiad participation=0	0.0000
	(.)
Previous Olympiad participation=1	0.2149***
	(0.0387)
Intends to go to university=1	0.0000
	(.)
Intends to go to university=2	-0.0382
	(0.0332)
Intends to go to university=3	-0.0423
	(0.0273)
Risk index	0.0108
	(0.0128)

Willingness to take risks	-0.0129**
	(0.0053)
Patience (standardized)	0.0033
	(0.0060)
Self-esteem	0.0010
	(0.0012)
Self-efficacy	-0.0010
	(0.0021)
Competitiveness	0.0004
	(0.0009)
Self-control	0.0008
	(0.0008)
Extracurricular activity (any)	-0.0053
	(0.0101)
Weekly extracurricular activities	0.0103
	(0.0092)
Constant	-0.0907
	(0.1589)
Observations	5,900

Linear probability model, all-controls specification. Standard errors clustered at the school level in parentheses; stars from conventional clustered t-tests.

Table D2: Full score regression: Gold-only participants, all-controls specification

	(1)
Baseline	0.000
	(.)
No gender	-6.097***
	(1.982)
Gender	-5.925**
	(2.303)
Free	0.000
	(.)
Free or Suggested	-2.550
	(2.371)
Mixed - Free and Mandatory	-1.405
	(2.599)
year=2	0.000
	(.)
year=3	-13.114***
	(1.289)
year=4	-11.564***

	(1.373)
year=5	-7.310***
	(1.542)
Female=0	0.000
	(.)
Female=1	-5.178***
	(0.973)
Father's education=1	0.000
	(.)
Father's education=2	-1.099
	(1.283)
Father's education=3	1.244
	(0.959)
Father's education=4	0.591
	(1.388)
Father's education=5	3.408*
	(1.782)
Mother's education=1	0.000
	(.)
Mother's education=2	2.853*
	(1.533)
Mother's education=3	-0.186
	(1.257)
Mother's education=4	1.361
	(1.689)
Mother's education=5	-0.682
	(1.716)
Unemployed	0.000
	(.)
White collar	3.802
	(2.396)
Blue collar	5.677**
	(2.278)
Entrepreneur	4.221*
	(2.432)
Retired	4.274
	(2.649)
Don't know	4.928
	(3.269)
Unemployed	0.000
	(.)
White collar	0.764
	(0.941)

Blue collar	-0.174
	(0.878)
Entrepreneur	-0.257
	(1.153)
Retired	8.828**
	(4.069)
Don't know	-3.268
	(2.016)
Foreign student=0	0.000
	(.)
Foreign student=1	0.232
	(1.320)
Father born abroad=0	0.000
	(.)
Father born abroad=1	-0.984
	(1.258)
Mother born abroad=0	0.000
	(.)
Mother born abroad=1	-0.081
	(1.264)
Non-Italian language at home=0	0.000
	(.)
Non-Italian language at home=1	-2.418
	(1.982)
Number of siblings	-0.040
	(0.300)
Raven score, 0-9	1.653***
	(0.252)
Expected correct answers	0.361**
	(0.165)
Confidence in expected answers	0.288
	(0.170)
Perceived class rank	-0.073
	(0.059)
Confidence in perceived rank	-0.302
	(0.187)
Math grade	1.601***
	(0.563)
Average grade	2.929***
	(0.694)
Mathematics preference rank	-1.176**
	(0.449)
Previous Certamen participation=0	0.000

	(.)
Previous Certamen participation=1	-0.322
	(1.100)
Previous Olympiad participation=0	0.000
	(.)
Previous Olympiad participation=1	1.353**
	(0.617)
Intends to go to university=1	0.000
	(.)
Intends to go to university=2	1.588
	(1.734)
Intends to go to university=3	-0.921
	(1.059)
Risk index	0.842
	(0.696)
Willingness to take risks	-0.600*
	(0.301)
Patience (standardized)	1.441***
	(0.228)
Self-esteem	-0.060
	(0.059)
Self-efficacy	0.077
	(0.081)
Competitiveness	0.106**
	(0.043)
Self-control	-0.094**
	(0.037)
Extracurricular activity (any)	-0.157
	(0.534)
Weekly extracurricular activities	0.117
	(0.594)
Constant	3.272
	(7.222)
Observations	1,878

Normalized Olympiad score among gold-only matched participants. Standard errors clustered at the school level in parentheses; stars from conventional clustered t-tests.

E Additional tables

Table E1 reports school-level regressions of participation and realized performance on the treatment indicators, using the complete administrative records of the competition as numerators and the number of enrolled students as denominator. The three outcomes are participants, total normalized points, and participants with a score above the national median, all per enrolled student. Because the unit of observation is the school, inference relies on randomization inference. All coefficients are negative, and three of the six have randomization inference p-values below 0.10: relative to the Baseline, treated schools have about 13 percentage points fewer participants and about 4 fewer points per enrolled student. These regressions control for the three-category selection rule, so they are the school-level counterpart of column (2), not of column (1), of Table 4. To locate the source of the difference in sign, we re-estimate the regression of column (1) on the same 30 schools, with the same controls and the same weighting, using as outcome the share of questionnaire respondents in the school who are matched to an Olympiad record: the coefficients become 0.28 for the No Gender prize and 0.17 for the Gender prize, close to the student-level estimates. The gap between the two sets of results thus comes from the difference between participation in the sampled classes, as measured by the matching, and participation in the whole school. This difference is large in some schools. In two Baseline schools, more than 45 percent of enrolled students took part in the competition, while fewer than 10 percent of the sampled respondents are matched to an Olympiad record. We cannot separate the role of class sampling from that of matching errors concentrated in some schools.

Table E1: School-level participation and realized performance from administrative data

	(1) Participants / enrolled students	(2) Total points / enrolled students	(3) Above-median participants / enrolled students
No gender treatment	-0.133	-3.646	-0.029
	(0.095)	(2.814)	(0.040)
	[0.073]	[0.099]	[0.374]
Gender treatment	-0.130	-4.392	-0.056
	(0.093)	(2.639)	(0.033)
	[0.088]	[0.048]	[0.102]
Baseline mean	0.397	12.261	0.169
Selection rule	✓	✓	✓
Schools	30	30	30

Notes: OLS at the school level. Numerators come from the complete administrative records of the competition (all participants, not only matched questionnaire respondents); the denominator is the number of students enrolled in the school. Column (2) sums the normalized scores (0–100) of all participants; column (3) counts participants with a normalized score above the median of all participants (30.0). Heteroskedasticity-robust standard errors (HC3) in parentheses; randomization inference p-values in brackets, obtained by permuting the treatment labels across schools (5000 permutations).

This appendix reports additional descriptive statistics and free margin results referenced in the main text. Table E2 collects selected coefficients of the full specification for participation and scores under both matching definitions (Section 6.6). Table E3 compares questionnaire respondents who

participate in the Olympiad with those who do not. Table E4 reports the score regressions among gold-only participants (Section 6.3). Tables E5 to E8 report the heterogeneity specifications of Section 6.7. Table E9 gives the student-level distribution of selection rules by treatment, Table E10 the covariate balance across treatments, and Table E11 the school-level balance. Tables E12 to E15 report the free-margin results within Mixed schools, by gender, and their robustness (Section 6.2). Table E16 reports the triple interactions by underconfidence, Table E17 the multiple-testing adjustment, and Tables E18 and E19 the IAT robustness checks and the treatment-by-IAT interactions (Sections 6.6 and 6.7).

Table E20 quantifies how the two prize schemes change the expected incentives of girls and boys, by applying each scheme ex post to the realized scores of all participants in the administrative records. The Baseline schools are the natural reference, since their pool of participants did not respond to any prize.

Table E2: Who participates and who scores well: selected coefficients from the full specification

	Participation		Score (0–100)	
	(1) Gold-only	(2) Gold+C1+C2	(3) Gold-only	(4) Gold+C1+C2
Female	-0.065***	-0.067***	-5.18***	-5.17***
	(0.012)	(0.012)	(0.97)	(0.86)
Past Olympiad participation	0.215***	0.226***	1.35**	1.28**
	(0.039)	(0.037)	(0.62)	(0.55)
Past Certamen participation	0.030	0.029	-0.32	-0.39
	(0.030)	(0.029)	(1.10)	(1.02)
Raven score (0–9)	0.010**	0.008	1.65***	1.55***
	(0.005)	(0.005)	(0.25)	(0.23)
Believed correct answers	-0.001	-0.003	0.36**	0.46***
	(0.003)	(0.003)	(0.17)	(0.16)
Certainty about answers (0–10)	0.003	0.004	0.29	0.17
	(0.003)	(0.004)	(0.17)	(0.16)
Expected rank in class (1 = best)	-0.003**	-0.005***	-0.07	-0.07
	(0.002)	(0.002)	(0.06)	(0.06)
Math grade	0.020**	0.022**	1.60***	1.58***
	(0.010)	(0.010)	(0.56)	(0.57)
Average grade	0.012	0.016	2.93***	2.69***
	(0.011)	(0.010)	(0.69)	(0.68)
Math preference rank (1 = most preferred)	-0.024***	-0.025***	-1.18**	-0.95**
	(0.008)	(0.008)	(0.45)	(0.45)
Competitiveness	0.000	0.001	0.11**	0.09**
	(0.001)	(0.001)	(0.04)	(0.04)
Willingness to take risk (0–10)	-0.013**	-0.014***	-0.60*	-0.49
	(0.005)	(0.005)	(0.30)	(0.31)
Patience index	0.003	0.005	1.44***	1.25***
	(0.006)	(0.006)	(0.23)	(0.22)
Self-esteem	0.001	0.001	-0.06	-0.04
	(0.001)	(0.001)	(0.06)	(0.05)
Self-efficacy	-0.001	-0.002	0.08	0.07
	(0.002)	(0.002)	(0.08)	(0.08)
Self-control	0.001	0.001	-0.09**	-0.11***
	(0.001)	(0.001)	(0.04)	(0.04)
Observations	5,900	5,900	1,878	2,059

Notes: Selected coefficients from the full specification, which also includes treatment and selection-rule indicators, year-of-study fixed effects, parents' education and occupation, foreign-born indicators, language spoken at home, number of siblings, certainty about beliefs, college plans, and extra-school activities. Full coefficient tables are reported in Appendix D. The score columns condition on participation among matched participants; in column (4) the expanded gold+C1+C2 sample includes recovered matches whose identity is less certain, so this column should be read with caution. Standard errors clustered at the school level in parentheses; stars from conventional clustered t-tests. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E3: Summary statistics by Olympiad participation: Gold-only

	Non-participants	Participants	Test p-value	N
Female	0.554	0.494	0.024**	7722
Year of study = 1	0.158	0.277	< 0.001***	7722
Year of study = 2	0.176	0.249	0.007***	7722
Year of study = 3	0.229	0.174	0.016**	7722
Year of study = 4	0.228	0.154	0.007***	7722
Year of study = 5	0.208	0.146	0.028**	7722
Raven score (0–9)	4.694	5.144	< 0.001***	7623
Math grade	6.867	7.403	< 0.001***	5997
Average grade	7.317	7.620	< 0.001***	5995
Expected correct answers	5.356	5.692	< 0.001***	7619
Confidence	6.429	6.675	0.010**	7615
Perceived class rank	10.185	9.489	0.039**	7597
Previous Olympiad participation	0.246	0.613	< 0.001***	5989
Competitiveness	42.392	43.239	0.011**	7462
Risk index	0.015	-0.027	0.151	7461
IAT D-score	0.149	0.172	0.174	7113
Self-esteem	28.195	28.931	< 0.001***	7452
Self-efficacy	28.374	28.887	< 0.001***	7470
Self-control	42.360	43.584	< 0.001***	7473

Notes: Means by gold-only Olympiad participation. P-values come from regressions of each variable on the participation indicator with standard errors clustered at the school level (30 clusters), computed with the wild cluster bootstrap (Rademacher weights, 9,999 replications). Year-of-study categories are tested as separate indicator variables. Missing values are left as missing, so the number of observations varies across variables. Stars indicate significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E4: Olympiad scores among participants, by treatment: Gold-only

	(1) ITT	(2) Base	(3) + Demog.	(4) + Ability	(5) + Academic	(6) + Past	(7) + Aspir.	(8) + Non-cognitive traits
No gender treatment	-3.979	-3.266	-3.418	-3.227	-6.301**	-6.220**	-6.226**	-6.097**
	(2.289)	(2.118)	(1.985)	(1.541)	(1.835)	(1.940)	(1.926)	(1.982)
Gender treatment	-2.636	-3.380	-2.869	-2.568	-6.085*	-6.178*	-6.153*	-5.925*
	(3.026)	(3.231)	(3.055)	(2.246)	(2.394)	(2.361)	(2.344)	(2.303)
Selection rule		✓	✓	✓	✓	✓	✓	✓
Year FE		✓	✓	✓	✓	✓	✓	✓
Gender		✓	✓	✓	✓	✓	✓	✓
Demographics			✓	✓	✓	✓	✓	✓
Ability/beliefs				✓	✓	✓	✓	✓
Academic performance					✓	✓	✓	✓
Past competitions						✓	✓	✓
Aspirations							✓	✓
Non-cognitive traits								✓
Observations	2,635	2,635	2,635	2,631	1,888	1,886	1,886	1,878

Notes: The dependent variable is the normalized Olympiad score among gold-only matched participants. Because scores are observed only for participants, and participation responds to the treatments, these regressions compare selected samples and are read as descriptive of the pool of participants in each arm. Column (1) reports the unadjusted comparison with treatment indicators only; later columns add the controls indicated in the panel. Standard errors clustered at the school level (30 clusters) in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E5: Who responds to incentives? Gold-only participation

Characteristic H_i:	Raven score above median	Math grade above median	Confidence above median	Previous Olympiad participated before	Competitiveness above median
No gender treatment	0.223*	0.233*	0.273**	0.302**	0.282**
	(0.096)	(0.105)	(0.087)	(0.105)	(0.087)
Gender treatment	0.123	0.079	0.140	0.135	0.136
	(0.097)	(0.102)	(0.084)	(0.088)	(0.090)
H_i (characteristic in column header)	-0.009	-0.046	0.032	0.256**	0.025
	(0.035)	(0.036)	(0.036)	(0.061)	(0.032)
No gender × H_i (characteristic in column header)	0.065	0.058	-0.012	-0.096	-0.031
	(0.042)	(0.041)	(0.037)	(0.080)	(0.039)
Gender × H_i (characteristic in column header)	0.001	0.086*	-0.021	-0.020	-0.015
	(0.043)	(0.044)	(0.037)	(0.073)	(0.042)
Selection rule	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
Female	✓	✓	✓	✓	✓
All other controls	✓	✓	✓	✓	✓
Observations	5,900	5,900	5,900	5,900	5,900

Notes: The dependent variable is gold-only Olympiad participation. Each column reports a separate regression interacting treatment with the characteristic named in the column, estimated on the complete-case sample of the full specification, so the sample is constant across columns. When a heterogeneity variable is constructed from a control, that source control is omitted from that column to avoid mechanical collinearity. In the Previous Olympiad column the heterogeneity variable is past participation itself. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 4,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E6: Who responds to incentives? Olympiad scores among participants (gold-only)

Characteristic H_i:	Raven score above median	Math grade above median	Confidence above median	Previous Olympiad participated before	Competitiveness above median
No gender treatment	-2.84	-2.55	-5.65	-3.95	-5.15
	(2.43)	(1.87)	(2.50)	(1.85)	(2.29)
Gender treatment	-4.27	-4.12	-6.27*	-4.84*	-5.96
	(2.84)	(1.84)	(2.54)	(1.96)	(2.68)
H_i (characteristic in column header)	5.53**	3.80**	-0.17	3.81**	2.14
	(1.85)	(1.52)	(1.49)	(1.54)	(1.87)
No gender × H_i (characteristic in column header)	-4.44	-5.48***	-0.90	-3.83**	-1.78
	(2.11)	(1.81)	(1.81)	(1.97)	(2.08)
Gender × H_i (characteristic in column header)	-2.73	-2.54	0.45	-2.06	-0.01
	(2.41)	(2.55)	(1.96)	(2.11)	(2.29)
Selection rule	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
Female	✓	✓	✓	✓	✓
All other controls	✓	✓	✓	✓	✓
Observations	1,878	1,878	1,878	1,878	1,878

Notes: The dependent variable is the normalized Olympiad score among gold-only matched participants (descriptive, since the sample conditions on participation). Each column reports a separate regression interacting treatment with the characteristic named in the column, estimated on the complete-case sample of the full specification, so the sample is constant across columns. When a heterogeneity variable is constructed from a control, that source control is omitted from that column to avoid mechanical collinearity. In the Previous Olympiad column the heterogeneity variable is past participation itself. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 4,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E7: Family-background heterogeneity: Gold-only participation

Indicator H_i:	Father post-sec./univ. yes/no	Mother post-sec./univ. yes/no	Father employed yes/no	Mother employed yes/no	Siblings yes/no
No gender treatment	0.243**	0.242**	0.314**	0.248**	0.228**
	(0.096)	(0.095)	(0.091)	(0.093)	(0.086)
Gender treatment	0.111	0.110	0.198*	0.129	0.088
	(0.093)	(0.091)	(0.089)	(0.088)	(0.086)
H_i (indicator in column header)	-0.064**	-0.031	0.053	-0.015	-0.049
	(0.030)	(0.046)	(0.034)	(0.020)	(0.036)
No gender × H_i (indicator in column header)	0.087**	0.067	-0.054	0.023	0.047
	(0.041)	(0.053)	(0.057)	(0.036)	(0.059)
Gender × H_i (indicator in column header)	0.067	0.054	-0.078	-0.002	0.048
	(0.042)	(0.054)	(0.049)	(0.034)	(0.039)
Selection rule	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
Female	✓	✓	✓	✓	✓
All other controls	✓	✓	✓	✓	✓
Observations	5,900	5,900	5,900	5,900	5,900

Notes: The dependent variable is gold-only Olympiad participation. Each column reports a separate regression interacting treatment with the characteristic named in the column, estimated on the complete-case sample of the full specification, so the sample is constant across columns. When a heterogeneity variable is constructed from a control, that source control is omitted from that column to avoid mechanical collinearity. In the Previous Olympiad column the heterogeneity variable is past participation itself. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 4,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E8: Family-background heterogeneity: Olympiad scores among participants (gold-only)

Indicator H_i:	Father post-sec./univ. yes/no	Mother post-sec./univ. yes/no	Father employed yes/no	Mother employed yes/no	Siblings yes/no
No gender treatment	-4.44*	-3.99	-2.40	-4.67	-6.64*
	(1.90)	(2.31)	(3.48)	(2.18)	(2.60)
Gender treatment	-5.62**	-4.71	-5.81	-5.94	-7.82**
	(2.14)	(2.43)	(4.25)	(2.65)	(2.80)
H_i (indicator in column header)	3.30	5.18***	2.70	1.32	-2.48
	(2.01)	(2.15)	(2.45)	(1.95)	(1.32)
No gender × H_i (indicator in column header)	-6.10*	-5.76**	-4.05	-1.72	0.75
	(2.16)	(2.28)	(2.70)	(2.12)	(1.53)
Gender × H_i (indicator in column header)	-2.05	-3.56	-0.19	-0.01	2.44
	(2.43)	(2.44)	(3.87)	(2.07)	(1.93)
Selection rule	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
Female	✓	✓	✓	✓	✓
All other controls	✓	✓	✓	✓	✓
Observations	1,878	1,878	1,878	1,878	1,878

Notes: The dependent variable is the normalized Olympiad score among gold-only matched participants (descriptive, since the sample conditions on participation). Each column reports a separate regression interacting treatment with the characteristic named in the column, estimated on the complete-case sample of the full specification, so the sample is constant across columns. When a heterogeneity variable is constructed from a control, that source control is omitted from that column to avoid mechanical collinearity. In the Previous Olympiad column the heterogeneity variable is past participation itself. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 4,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E9: Questionnaire respondents by selection rule and treatment status

Selection rule	Baseline	No gender	Gender	Treated total	Total
Free	126	541	1,927	2,468	2,594
Free or suggested	72	891	441	1,332	1,404
Mixed: free and mandatory	1,693	1,174	857	2,031	3,724
Total	1,891	2,606	3,225	5,831	7,722

Notes: The table counts questionnaire respondents in the participation analysis sample. Treated total is the sum of respondents in No gender and Gender schools.

Table E10: Baseline characteristics by treatment arm

	Control	T1 No gender	T2 Gender	C vs T1	C vs T2	T1 vs T2	Joint	N
Female	0.552	0.572	0.491	0.656	0.204	0.028**	0.089*	7,722
Year of study = 1	0.242	0.199	0.174	0.330	0.116	0.496	0.287	7,722
Year of study = 2	0.200	0.214	0.190	0.784	0.843	0.351	0.669	7,722
Year of study = 3	0.187	0.208	0.226	0.679	0.393	0.471	0.561	7,722
Year of study = 4	0.191	0.177	0.232	0.642	0.325	0.055*	0.198	7,722
Year of study = 5	0.179	0.203	0.178	0.392	0.957	0.331	0.605	7,722
Raven score (0–9)	5.008	4.852	4.755	0.509	0.281	0.701	0.549	7,623
Math grade	6.889	7.054	7.101	0.241	0.064*	0.702	0.173	5,997
Average grade	7.352	7.467	7.403	0.318	0.582	0.568	0.605	5,995
Expected correct answers	5.450	5.427	5.521	0.922	0.774	0.581	0.864	7,619
Confidence	6.467	6.399	6.634	0.710	0.398	0.120	0.305	7,615
Perceived class rank	10.142	10.203	9.620	0.898	0.342	0.179	0.393	7,597
Previous Olympiad participation	0.331	0.404	0.346	0.448	0.861	0.473	0.706	5,989
Competitiveness	42.986	42.514	42.655	0.341	0.535	0.724	0.625	7,462
Risk index	0.008	-0.028	0.019	0.262	0.724	0.177	0.361	7,461
IAT D-score	0.156	0.148	0.165	0.778	0.687	0.594	0.876	7,113
Self-esteem	28.160	28.413	28.656	0.309	0.066*	0.193	0.162	7,452
Self-efficacy	28.680	28.492	28.529	0.379	0.436	0.819	0.649	7,470
Self-control	43.034	42.523	42.856	0.266	0.691	0.310	0.444	7,473

Notes: The first three columns report group means. Because treatment is randomized at the school level, p-values come from regressions of each covariate on treatment indicators with standard errors clustered at the school level (30 clusters) and are computed with the wild cluster bootstrap (Rademacher weights, 9,999 replications). Pairwise columns test Control vs T1, Control vs T2, and T1 vs T2; the Joint column tests the equality of the three arms. Year-of-study categories are tested as separate indicator variables. Missing values are left as missing, so the number of observations varies across variables. Stars indicate significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E11: School-level balance across treatment arms

	Baseline	No gender	Gender	F-test p	KW p	Schools
Students enrolled	833	1,060	911	0.51	0.53	30
Share of female students	0.51	0.51	0.50	0.99	0.99	30
Municipality population	69,927	95,732	54,660	0.51	0.50	30
FGA school quality index (0–100)	65.6	68.3	66.3	0.81	0.80	27
University enrollment rate	0.81	0.81	0.79	0.94	0.86	27
Share choosing STEM at university	0.56	0.55	0.56	1.00	1.00	27

Notes: Means of pre-treatment school characteristics by treatment arm, one observation per school. Students enrolled and the share of female students come from the school records collected at recruitment; municipality population refers to the school's location. The Fondazione Agnelli variables refer to the school's past graduates (2014–2017): the FGA index summarizes their university performance, the enrollment rate is the share who enrolled at university, and the STEM share is the share of enrolled graduates who chose a STEM field; these variables are missing for three schools. The F-test p-value comes from a regression of each characteristic on treatment indicators with robust standard errors; KW is the Kruskal-Wallis test. The distribution of Olympiad selection rules across arms is reported in Table 2 and is accounted for throughout the analysis.

Table E12: Participation within Mixed schools: all classes vs free classes

	Gold-only			Gold+C1+C2		
	(1) All classes	(2) Free classes	(3) Free classes + controls	(4) All classes	(5) Free classes	(6) Free classes + controls
No gender	0.369***	0.145	0.154	0.348***	0.150	0.159
	(0.079)	(0.081)	(0.070)	(0.077)	(0.084)	(0.072)
Gender	0.088	0.133*	0.058	0.056	0.121*	0.045
	(0.075)	(0.055)	(0.075)	(0.077)	(0.056)	(0.078)
Observations	3,724	2,129	1,845	3,724	2,129	1,845

Notes: Linear probability models within Mixed schools only. Free classes exclude high-participation classes. All specifications include year-of-study fixed effects and gender; columns with controls add the full set of questionnaire controls. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E13: Treatment by gender on the free margin

	Gold-only		Gold+C1+C2	
	(1) Base	(2) + controls	(3) Base	(4) + controls
No gender	0.104	0.073	0.052	0.024
	(0.067)	(0.078)	(0.075)	(0.083)
Gender	0.117*	0.088	0.076	0.040
	(0.052)	(0.065)	(0.056)	(0.066)
Female	-0.106*	-0.073**	-0.094	-0.068*
	(0.025)	(0.021)	(0.036)	(0.029)
No gender × Female	0.028	0.033	0.001	0.024
	(0.033)	(0.034)	(0.042)	(0.042)
Gender × Female	0.022	0.006	0.003	-0.004
	(0.027)	(0.028)	(0.038)	(0.033)
Observations	6,127	4,941	6,127	4,941

Notes: Linear probability models on the free margin sample (excluding high-participation classes in Mixed schools). All specifications include the school-level selection rule and year-of-study fixed effects. Columns with controls add the full set of questionnaire controls. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E14: Free-margin participation: robustness to the class classification, to the matching definition, and to school weights

	(1) Free margin (70%) Gold	(2) Free margin (70%) Gold+C1+C2	(3) Free/Free-or-Suggested schools Gold	(4) Free margin (60%) Gold	(5) Free margin (80%) Gold	(6) Free margin (70%), school wei Gold
No gender treatment	0.120*	0.053	0.172	0.092	0.160**	0.185**
	(0.057)	(0.071)	(0.077)	(0.056)	(0.062)	(0.064)
	[0.061]	[0.500]	[0.207]	[0.147]	[0.038]	[0.013]
Gender treatment	0.129**	0.077	0.115	0.117**	0.141**	0.166***
	(0.046)	(0.054)	(0.048)	(0.042)	(0.050)	(0.040)
	[0.024]	[0.198]	[0.203]	[0.039]	[0.012]	[0.003]
Baseline mean	0.210	0.261	0.040	0.193	0.251	0.210
Selection rule	✓	✓		✓	✓	✓
Year FE, gender	✓	✓	✓	✓	✓	✓
Schools	30	30	16	30	30	30
Observations	6,127	6,127	3,998	5,925	6,491	6,127

Notes: Linear probability models, baseline specification of Table 6. Columns (1), (2), (4), (5) and (6) exclude classes in Mixed schools whose realized participation (gold, C1 or C2) is at least the percentage indicated; column (3) keeps only schools that declared a Free or Free-or-Suggested rule, with no class excluded (in this group there are only two Baseline schools). Column (2) uses the expanded participation definition. Column (6) weights each student by the inverse of the number of respondents in the school, so that schools receive equal weight. Standard errors clustered at the school level in parentheses; wild cluster bootstrap p-values (Rademacher weights, 9,999 replications) in brackets: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E15: Free-margin participation: leave-one-school-out

	No gender treatment	Gender treatment
Full free-margin estimate	0.120	0.129
Range of estimates excluding one school	[0.096, 0.165]	[0.104, 0.174]
Estimates with wild bootstrap $p < 0.10$	26 of 30	30 of 30
Estimates with wild bootstrap $p < 0.05$	8 of 30	26 of 30
Excluding one Baseline school at a time		
Baseline school 1 excluded	0.165 [0.001]	0.174 [0.000]
Baseline school 2 excluded	0.109 [0.089]	0.120 [0.052]
Baseline school 6 excluded	0.125 [0.049]	0.134 [0.028]
Baseline school 7 excluded	0.127 [0.074]	0.137 [0.045]
Baseline school 18 excluded	0.118 [0.089]	0.127 [0.050]
Baseline school 22 excluded	0.118 [0.096]	0.127 [0.064]
Baseline school 23 excluded	0.107 [0.097]	0.116 [0.056]
Baseline school 32 excluded	0.120 [0.065]	0.129 [0.026]
Baseline school 34 excluded	0.096 [0.120]	0.104 [0.049]

Notes: Baseline free-margin specification of Table 6, column (1), re-estimated 30 times excluding one school at a time. Wild cluster bootstrap p-values (Rademacher weights, 999 replications) in brackets. School identifiers are the internal codes of the analysis data set.

Table E16: Calibrated underconfidence: participation and treatment interactions (free margin)

	Within-class (A)		Residual (B)	
	(1)	(2)	(3)	(4)
Female	-0.102***	-0.140***	-0.084***	-0.071*
	(0.018)	(0.044)	(0.014)	(0.035)
Underconfident	-0.027*	-0.051**	-0.077***	-0.057***
	(0.014)	(0.022)	(0.016)	(0.021)
Female × Underconfident	0.036*	0.067	0.024	-0.038
	(0.019)	(0.040)	(0.021)	(0.045)
No gender prize	0.129**	0.085	0.136**	0.127*
	(0.061)	(0.070)	(0.061)	(0.072)
Gender prize	0.126**	0.103*	0.147***	0.147***
	(0.050)	(0.056)	(0.049)	(0.052)
No gender × Underconfident		0.067*		-0.015
		(0.036)		(0.049)
Gender × Underconfident		0.015		-0.029
		(0.028)		(0.030)
No gender × Female × Underconf.		-0.085		0.038
		(0.056)		(0.059)
Gender × Female × Underconf.		-0.017		0.104*
		(0.043)		(0.051)
Observations	5,843	5,843	5,024	5,024

Notes: Linear probability models for participation (gold-only match) on the free margin, excluding high-participation (mandatory) classes. In columns (1) and (2), Underconfident equals one if the student expects a position in the class worse than her actual position in the within-class distribution of Raven scores (classes with at least five respondents). In columns (3) and (4), Underconfident equals one if the residual of the expected rank, regressed on the Raven score, the math grade, the average grade, and year of study, is positive; this sample requires valid grades and therefore excludes most first-year students. Treatment interactions with Female are included in columns (2) and (4) but not reported. All specifications include selection rule and year fixed effects. Standard errors clustered at the school level in parentheses. The wild cluster bootstrap p-value for Gender × Female × Underconfident in column (4) is 0.035 (Rademacher weights, 9,999 replications); this interaction is sensitive to the construction of the measure, as discussed in the text. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E17: Treatment × characteristic interactions in participation: multiple-testing adjustment

Interaction	Coefficient	WCB p-value	Holm p-value	Bonferroni p-value
Raven score × No gender	0.065	0.144	1.000	1.000
Raven score × Gender	0.001	0.984	1.000	1.000
Math grade × No gender	0.058	0.221	1.000	1.000
Math grade × Gender	0.086	0.087	1.000	1.000
Confidence × No gender	-0.012	0.789	1.000	1.000
Confidence × Gender	-0.021	0.627	1.000	1.000
Previous Olympiad × No gender	-0.096	0.344	1.000	1.000
Previous Olympiad × Gender	-0.020	0.821	1.000	1.000
Competitiveness × No gender	-0.031	0.468	1.000	1.000
Competitiveness × Gender	-0.015	0.750	1.000	1.000
Implicit stereotype (IAT) × No gender	-0.044	0.266	1.000	1.000
Implicit stereotype (IAT) × Gender	-0.053	0.110	1.000	1.000

Notes: Each row is the interaction between a treatment indicator and an above-median (or binary) characteristic, from the full-control participation specifications of Appendix Table E5, estimated on the complete-case sample of the full specification (N = 5,900), plus the two treatment × above-median IAT interactions, estimated with the full set of controls on the students with a valid IAT (N = 5,704). The family consists of these 12 tests. Wild cluster bootstrap p-values (Rademacher weights, 9,999 replications), and step-down Holm and Bonferroni adjusted p-values for the family.

Table E18: Implicit stereotypes and the gender gap in participation: standardized D-score and fixed effects

	(1) Base	(2) School FE	(3) Class FE	(4) Full controls	(5) Free margin
Female	-0.084***	-0.065***	-0.045***	-0.066***	-0.093***
	(0.010)	(0.010)	(0.010)	(0.012)	(0.011)
D-score (standardized)	0.036***	0.031***	0.022**	0.016**	0.043***
	(0.008)	(0.009)	(0.008)	(0.006)	(0.009)
Female × D-score (standardized)	-0.050***	-0.044***	-0.034**	-0.028*	-0.062***
	(0.014)	(0.014)	(0.013)	(0.013)	(0.014)
Wild bootstrap p-value, Female × D-score	0.002	0.004	0.016	0.051	0.000
Treatment, selection rule, year FE	✓			✓	✓
School FE		✓			
Class FE			✓		
Full controls				✓	
Observations	7,113	7,113	7,113	5,704	5,594

Notes: Linear probability models for gold-only participation. The D-score is standardized to mean zero and unit standard deviation in the sample with a valid IAT (mean 0.157, standard deviation 0.382), so the interaction coefficient is the change in the female-male participation gap per standard deviation of the male-science association. Column (2) replaces the treatment and selection-rule indicators with school fixed effects, column (3) with class fixed effects (which also absorb year of study). Column (5) restricts the sample to the free margin. Standard errors clustered at the school level in parentheses; stars are based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The estimates are descriptive associations.

Table E19: Do the prizes interact with implicit stereotypes? Treatment × IAT interactions

	Standardized D-score		Above-median D-score	
	(1)	(2)	(3)	(4)
No gender × IAT	-0.006	0.021	-0.007	0.034
	(0.016)	(0.020)	(0.037)	(0.034)
Gender × IAT	-0.007	0.017	-0.015	0.035
	(0.015)	(0.018)	(0.033)	(0.033)
No gender × Female × IAT		-0.048		-0.071
		(0.036)		(0.048)
Gender × Female × IAT		-0.052		-0.103**
		(0.029)		(0.045)
Joint test, p-value (WCB)	0.904	0.223	0.919	0.095
Observations	7,113	7,113	7,113	7,113

Notes: Linear probability models for gold-only participation, with treatment indicators, selection rule, year fixed effects, gender, and all lower-order interaction terms (not shown). IAT is the standardized D-score in columns (1)–(2) and an indicator for an above-median D-score in columns (3)–(4). The joint test refers to the two treatment × IAT interactions in columns (1) and (3) and to the two triple interactions in columns (2) and (4). Standard errors clustered at the school level in parentheses; stars based on wild cluster bootstrap p-values (Rademacher weights, 9,999 replications): * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E20: How much do the two prize schemes change the expected incentives of girls and boys? Ex post simulation on the realized scores

	Baseline schools (9)		All schools (30)	
	Girls	Boys	Girls	Boys
Participants per school	123.9	149.0	93.5	120.4
Probability of winning a prize				
No gender scheme (top 4 of the school)	0.024	0.057	0.019	0.044
Gender scheme (top 2 girls, top 2 boys)	0.045	0.042	0.042	0.031
Expected prize per participant (euros)				
No gender scheme	4.1	8.0	3.1	6.4
Gender scheme	6.7	6.3	6.3	4.7
Score needed to win a prize (lowest winning normalized score)				
No gender scheme	74.2	74.2	70.3	67.4
Gender scheme	64.5	76.0	58.2	70.3
Share of prize money to girls, No gender scheme	0.333		0.239	
Schools with no girl among the top 4	3		14	

Notes: The two prize schemes are applied ex post to the realized normalized scores of all participants in the administrative records of the competition. Under the No gender scheme the four best students of the school receive 200, 200, 100 and 100 euros; under the Gender scheme the two best girls and the two best boys receive 200 and 100 euros each. Ties are resolved with the official ranking. Probabilities and expected prizes are averages across schools of school-level ratios; the score needed to win is the average across schools of the lowest winning score, for girls the lowest score among the winning girls. The Baseline schools are the natural reference because their participation did not respond to any prize; treated schools are included in the right panel for comparison, keeping in mind that their pool of participants may have responded to the prize actually offered.

F Details of the matching procedure

This appendix describes how questionnaire respondents are linked to Olympiad records. The two sources are the questionnaire file, with 8,174 records, and the competition archive, with 6,424 individual participant records collected school by school. Neither source contains student names.

The anagraphic key combines school, year of study, gender, and exact date of birth. The study track and the class section are not used as descriptive variables but as quality constraints: a candidate pair that disagrees on track or section is not accepted at the highest confidence level. The procedure follows five rules. First, a questionnaire respondent can be assigned to at most one Olympiad record. Second, an Olympiad record can be assigned to at most one questionnaire. Third, when several Olympiad records are compatible with the same questionnaire, or several questionnaires with the same record, and no unique winner emerges from the constraints, the match is left unassigned rather than guessed. Fourth, matches are searched only within the same school. Fifth, class identifiers are built from the questionnaire data.

The confidence levels are defined as follows. Gold matches are unique matches on the full key with consistent track and section information; they come from the automated pipeline. C1 matches are unique matches on school, year of study, and date of birth, with gender compatible when observed in both records and with compatible section, but with discordant or missing track information; 108 of the 110 C1 matches follow this rule and two were resolved manually. C2 matches are unique matches on the same core key, with track and section not used as constraints; 156 of the 161 C2 matches follow this rule and five were resolved manually, in most cases because the Olympiad record was the only one left in the school compatible with the questionnaire. Table 1 in the main text reports the resulting counts.

Unmatched questionnaire respondents are treated as non-participants in the analysis. Unmatched Olympiad participants, about 55 percent of the archive, are mostly students whose class was not selected for the questionnaire; they contribute no covariate information and do not enter the analysis. Section 6.2 shows that the recovered C1 and C2 matches are not distributed evenly across treatments, which motivates the choice of the conservative gold-only definition for the main results.

As a robustness check on the matched sample, Table F1 compares matched and unmatched Olympiad participants on the variables available in the competition archive, within school. Matched participants score about 1.5 points higher, are 13 percentage points more likely to be female, and are slightly older. The higher share of girls is consistent with the sampling design: the questionnaire covered classes across all study tracks, while participants from male-dominated tracks are more likely to fall outside the sampled classes. The important result is in columns (3) and (4): none of these differences varies across treatments. The matched sample is therefore mildly selected, and we find no evidence that this selection differs across treatments, although the absence of significant interactions does not rule it out. On the questionnaire side, the comparison between matched and unmatched respondents coincides with the comparison between participants and non-participants reported in Table E3.

Table F1: Matched vs unmatched Olympiad participants

	(1) Pooled difference	(2) Difference in Baseline	(3) × No gender	(4) × Gender
Olympiad score (0–100)	1.453*	2.379	−1.815	−0.630
	(0.733)	(1.575)	(1.686)	(2.165)
Female	0.129***	0.140**	0.013	−0.065
	(0.028)	(0.055)	(0.066)	(0.070)
Year of study	0.173*	0.219	−0.208	0.170
	(0.088)	(0.136)	(0.218)	(0.188)

Notes: Each row reports regressions of the variable on a dummy for being matched to a questionnaire, in the sample of 6,424 Olympiad participants (6,417 to 6,420 with non-missing values), with school fixed effects and standard errors clustered at the school level (30 clusters). Column (1) reports the pooled matched−unmatched difference. Columns (2) to (4) come from a specification that interacts the matched dummy with the treatment arms: column (2) is the difference in the Baseline arm, columns (3) and (4) are the interactions with the two treatments. The share of participants matched to a questionnaire is 32.3 percent in Baseline schools, 50.1 percent in No gender schools, and 58.7 percent in Gender schools; these differences are mechanical, since they depend on the number of sampled classes relative to the number of participants in each school. Results are very similar when the matched dummy is restricted to gold matches (score difference 1.895, se 0.766; female difference 0.125, se 0.029). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table F2 reports the sensitivity of the participation estimates to differential matching error across treatments, discussed in Section 4.

Table F2: Sensitivity of the participation estimates to differential matching error

	Full sample		Free margin	
	No gender	Gender	No gender	Gender
Panel A: the two matching definitions				
Gold only	0.273***	0.168*	0.120*	0.129**
	(0.075)	(0.071)	(0.057)	(0.046)
Gold + C1 + C2	0.217**	0.117	0.053	0.077
	(0.084)	(0.076)	(0.071)	(0.054)
Panel B: share of unmatched Baseline students reclassified as participants				
0 percent	0.273	0.168	0.120	0.129
1 percent	0.266	0.161	0.112	0.121
2 percent	0.260	0.154	0.104	0.113
3 percent	0.253	0.148	0.096	0.105
5 percent	0.240	0.135	0.079	0.088
10 percent	0.207	0.100	0.039	0.048
Share that sets the estimate to zero	41.4 percent	25.0 percent	14.7 percent	15.8 percent
Panel C: share of gold-matched treated students reclassified as non-participants				
0 percent	0.273	0.168	0.120	0.129
5 percent	0.242	0.142	0.106	0.114
10 percent	0.212	0.118	0.090	0.098
15 percent	0.180	0.093	0.076	0.083
20 percent	0.149	0.069	0.061	0.067
Observations	7,722	7,722	6,127	6,127

Notes: Baseline participation specification (treatment indicators, selection rule, year fixed effects, gender), standard errors clustered at the school level; stars in Panel A are based on wild cluster bootstrap p-values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Panel A reports the gold-only and expanded estimates, which treat all recovered C1 and C2 matches as false positives and as true participants respectively; they show the sensitivity of the estimates to the matching definition but are not bounds on the true effect, since matching errors may differ across schools and treatments. Panel B asks how many false negatives, concentrated in the control group only, would be needed to explain the gold-only estimates: a share of the unmatched Baseline students is reclassified at random as participants (average over 50 random draws, seed 12345), and the last row reports the share that would set the estimate to zero, obtained by interpolating the simulated estimates on a grid of one percentage point. For comparison, the recovered C1 and C2 matches amount to about 5 percent of the unmatched Baseline students in free classes. Panel C performs the symmetric exercise for false positives concentrated in the treated groups only.

G Experimental materials

This appendix reproduces the posters displayed in the schools at least one month before the competition. Each school received the poster corresponding to its assigned treatment. Schools in the Control group received the poster announcing the Giochi di Archimede (Figure G1). Schools in the No Gender prize treatment received the poster of the "W la Matematica" prize, which rewards the four best students of the school (Figure G2). Schools in the Gender prize treatment received instead the poster of the "Vinci con la Matematica" prize, with separate awards for the two best girls and the two best boys (Figure G3).



Figure G1: Poster announcing the competition in the Baseline treatment.



Figure G2: Poster of the No Gender treatment ("W la Matematica"): 200 euros for the first and second ranked students of the school, 100 euros for the third and fourth.



Figure G3: Poster of the Gender treatment ("Vinci con la Matematica"): 200 and 100 euros for the first and second ranked girls, and 200 and 100 euros for the first and second ranked boys.

Quest’opera è soggetta alla licenza Creative Commons

CC BY-NC 4.0 DEED

Attribuzione - Non commerciale 4.0 Internazionale



Alma Mater Studiorum - Università di Bologna

DEPARTMENT OF ECONOMICS

Strada Maggiore 45
40125 Bologna - Italy
Tel. +39 051 2092604
Fax +39 051 2092664
http://www.dse.unibo.it