UNIVERSITA’ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI “M.FANNO”

CORSO DI LAUREA MAGISTRALE IN ECONOMICS AND FINANCE

TESI DI LAUREA

“Gender Gap in Math Competitions: Evidence from the Math Olympiad in the Italian Schools”

RELATORE:

CH.MO PROF. Antonio Nicolò

LAUREANDO: Somma Noè

MATRICOLA N. 1184936

ANNO ACCADEMICO 2019 – 2020

Il candidato dichiara che il presente lavoro è originale e non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere.
Il candidato dichiara altresì che tutti i materiali utilizzati durante la preparazione dell'elaborato sono stati indicati nel testo e nella sezione "Riferimenti bibliografici" e che le eventuali citazioni testuali sono individuabili attraverso l'esplicito richiamo alla pubblicazione originale.

The candidate declares that the present work is original and has not already been submitted, totally or in part, for the purposes of attaining an academic degree in other Italian or foreign universities. The candidate also declares that all the materials used during the preparation of the thesis have been explicitly indicated in the text and in the section "Bibliographical references" and that any textual citations can be identified through an explicit reference to the original publication.

Firma dello studente

____

I want to thank my Supervisor prof. Antonio Nicolò, prof. Natalia Montinari (University of Bologna) and prof. Valeria Maggian (University of Venice) for giving me the opportunity to work with them as research assistant in their project "Progetto Rodonea" and to present part of their field experiment in my Master's Thesis. Thank you for always being available to clarify my doubts.

Table of Contents

1. Introduction

The gender gap in labor market outcomes (careers' advancement and wage) has been for decades one of the central issues of the economic debate. Although evident and recognized in importance and magnitude, authors have suggested different explanations for this gap. Economists have supported both supply-side and demand-side explanations. Evidence from both sides confirms that many forces and circumstances (discrimination, family constraints, job and individual preferences etc.) may concur to keep women away from highly paid jobs and positions. Differences in labor market outcomes have been shown to be mainly related with STEM achievements. This confirms to be a male-dominated field in which women fail to enter and well-perform with respect to male peers. Analogously, for the gender gap in science and math achievements we find demand-side and supply-side explanations. Both experimental and nonexperimental literature has widely investigated the latter category, in particular individual preferences that could refrain females from pursuing STEM careers. In this work we mainly examine the gender gap in competitiveness that makes women fail to enter competitive contexts. The distortion of competition occurs also on their relative performance. This is confirmed in several studies and experiments. In particular, past literature shows that this distortion is mostly related to the math content of submitted tasks. This is in line with the evidence of women failing to access and underperforming in STEM fields (then in STEM careers), and this, in turn, has a strong and negative impact on women's labor market outcomes. In the literature that we are going to describe several authors find that the distortion of competition is weaker when females compete against peers of the same gender. Single-sex contests, on the contrary, seem not to have a remarkable effect on men's performance. Different explanations have different policy implications. If females particularly suffer competition against men, policies (e.g., quotas) that lead women to compete only against each other, even in male-dominated fields, could foster their performance possibly having a long-term impact. In this work we present a field experiment run during the Math Olympiad that is an extra-curricular math competition organized each year in the Italian high-schools by the Unione Matematica Italiana. In this experiment, that took place in the Italian North-East, we analyze the impact of a policy intervention on the gender-gap in participation and performance. In Chapter 2 we present a literature review. In Chapter 3 and 4 we respectively describe the experiment, our data sources and variables giving a first look at our results. In Chapter 5 we further analyze data with a regression approach and in Chapter 6 we conclude.

2. Literature Review

2.1. The Gender Gap

Recently women have managed to fill the gap both in educational outcomes and labor market participation but gender differences in labor market outcomes still persist (Goldin, Katz and Kuziemko, 2006). This disparity is reflected both in wages and career advancements (Flory et al., 2015) with horizonal and vertical job segregation playing an important role in shaping it (Altonji & Blank 1999, Bertrand & Hallock 2001) . This suggests that closing the gender gap in labor market participation is not enough to assure equity across genders since the underrepresentation of women in top positions testifies that equality of opportunities does not necessarily translate in equality of outcomes (Maggian, Montinari and Nicolò, 2017). There are different approaches in the recent literature that try to understand these differences in outcomes. The traditional approach distinguishes, dividing into these two broad categories, demand-side and supply-side explanations (we refer to the review from Azmat and Petrongolo, 2014).

On the first side, main drivers are discrimination and stereotypes represented as factors causing, in particular, job segregation keeping away women from top (and highly paid) jobs and impeding them to achieve leading roles in various contexts. In particular, Azmat and Petrongolo (2014) specify that "gender discrimination in the labor market is defined as a situation in which equally productive men and women are rewarded differently". Indeed, the crucial point here is that discrimination occurs when two or more individuals have the same ability (for example in academic career) or the same productivity (in labor markets). The principal force that causes this kind of situation are stereotypes based, in this case, on individuals' gender. Reuben, Sapienza and Zingales (2014) argue that "the stereotype of women's inferior performance on every mathematics-related task […] can lead to a decreased demand for women in STEM fields and/or a reduction in the number of women choosing to specialize in these fields". Most important, in their study this stereotype has been found affecting demand for women regardless of their actual abilities and to survive even in presence of full information about participants' past performance. Moreover, authors underline the economic relevance of this finding as the stereotype leads to a "suboptimal hiring choice", clearly "biased in favor of men". Anyway, this demand-side discouragement can be an important driver for gender differences in (academic, as we will see, and) labor market outcomes both directly keeping away females from top positions (e.g., when it materializes in discrimination) and indirectly when stereotypes are internalized by women negatively affecting their performance (phenomenon defined as “stereotype threat” by Spencer et al., 1999).

On the other side, there is a branch of the literature that proposes “supply-side explanations”. For example, that men and women are different in abilities and that they self-select in jobs due to their occupational preferences (Polachek, 1981). Some authors analyze family constraints, Dessy and Djebbari (2010) find that one explanation for within-family imbalances, in terms of outcomes, is the failure for men and women to coordinate in the timing of marriage. In the recent literature economists have examined new potential factors (still on the supply-side) that can explain the unexplained part of the gender gap, harming women’s careers and performance: psychological attributes and “individual preferences in the spheres of motivation, ambition and competitiveness” (Azmat and Petrongolo, 2014). Before examining this last factor more in detail looking at the past literature and then at our evidence, let us understand how the Gender Gap that we have described so far is linked to the fact that women and men perform differently in STEM fields.

2.1.1. Math Performance

The difference in math achievements manifests itself at different stages. In general, in OECD countries the higher propensity for boys to choose math- or science-intensive courses is already visible in the secondary school while, for example, in the U.S. this disparity persists at college level where “women are significantly less likely than men to graduate with a major in science, technology, engineering, or mathematics” (Buser, Niederle and Oosterbeek, 2014). As these authors underline, even in environments (for example high schools in the U.S.) where both boys and girl are equally likely to choose math-intensive courses and also have similar performance in mean, still women are underrepresented among the pool of best or high-achieving students in this field although this is not driven by difference in abilities across genders (Ellison and Swanson, 2010). Gender Gaps in labor market outcomes and in math performance are strictly linked and have similar patterns in recent years. As reported in Niederle and Vesterlund (2010), there are several studies supporting the evidence that math performance can well predict individuals’ future income. Or, at an earlier stage, science achievements can predict college attendance and attainments (Goldin, Katz and Kuziemko, 2006). Due to this important effect that mathematics have on careers and wages “differences in mathematics qualifications at the top of the distribution may explain a substantial part of the gender gap in income and in career outcomes more broadly” (Joensen and Nielsen, 2014) . The effect of math performance comes also from the effect on students’ propensity and readiness for STEM universities (Card and Payne, 2017). That is why the gap in math performance is causing women to be persistently underrepresented in high profitable jobs mainly related to STEM (Science, Technology, Engineering and Math) fields, "especially when excluding teaching careers" (Carlana, 2019). The same is true when including in STEM fields finance and business (Bertrand et al., 2010) . The problem , as we will see in experimental literature and as evident from what we have said so far, can be divided into two phases. First, the choice to enter math- or science-based courses. Second, once entered this careers, the performance that individuals have in this field. For the second aspect, in the recent literature there are several examples of gender gap in math performance although this gap is no longer as large as before at the mean of the distribution. Women are, in particular, underperforming at the top (Hedges and Nowell, 1995), and this is true for various math tests (Niederle and Vesterlund, 2010). There is evidence of women's lower performance also in the experimental literature that then tries to investigate potential explanations for this gender difference, we will see in next Sections. As we have said in the previous Section, on both sides (so, for what concerns women's lower probability to choose math courses and lower performance) the difference is potentially not only driven, and that is why this gender gap is an economic matter, by individuals' abilities. Joensen and Nielsen (2014) find that the underlying ability distribution for both genders in their data is equal "at least around the top decile of the distribution". Nonetheless, among these individuals the percentage of males attending math courses is much higher with respect to females. As we have done talking about labor markets experience and outcomes, we can similarly investigate (through the literature) potential "demand-side" and "supply-side" explanations causing females refraining from pursuing STEM careers and well-performing in math contexts. Here with "demand-side" we refer to the environment in which women grow up and live, while "supply-side" explanations are, as before, females' preferences.

For what concerns the "discouragement effect" that we have described in the previous Section in the case of labor markets, there is evidence of this effect in family as well as in academic environments. Women are exposed to the male-math association since they are young [1] . Beginning with gender-biased parents' expectations about daughters' and sons' math abilities (Furnham, Revees and Budhani, 2002) clearly in favor of males whose eventual success is considered "natural" opposed to women's results only coming from hard work (Yee and Eccles, 1998). Girls suffer both subtle (few models of female mathematicians and scientists) and not-so-subtle (overt-

[1] We here exploit the review from Stout et al. (2011)

discrimination) reminders that "math is for boys" and that they do not belong to the STEM world during their growth lowering their interest in pursue STEM careers. Subtle situational cues, as the low female representation (Murphy, Steele and Gross, 2007) are crucial in this sense. Moreover the subtle male-math association becomes heavier the higher is the level of education due to "the skewed gender ratio of STEM experts in academic environments" (Stout et al., 2011). Carlana (2019) shows "the importance of gender – biased environments in explaining the underconfidence of females in STEM fields". She runs an experiment in Italian schools finding a strong negative effect that teachers' gender stereotypes have on females' math performance e future academic choice, also due to the effect on women's confidence about their own math ability. From the "demand-side" point of view, there is a lot to be done but, at the same time, it is difficult to design an efficient short-term policy in order to improve the environment in which females act and perform. Still some intervention may help to reduce the problem. Joensen and Nielsen (2014) find that a change in high-school curricula and in the way they are organized can foster girls' tendency toward math-intensive courses. In particular, they exploit an exogenous variation in the cost of acquiring additional advanced mathematics courses that allows them to demonstrate two things. First, that the change in the curriculum flexibility allows students, in particular girls, to follow more advanced math-based courses. So, the way in which courses are combined results to be relevant. Second, that this has a long term positive effect on involved women's career, both in terms of earnings and career advancement (especially in high paid and male – dominated career tracks). The peculiarity of their findings is that this positive effect is asymmetric across genders since men do not benefit from this variation, mostly because high-skilled men already choose this kind of courses and mid-skilled men do not benefit. In general, due to this exogenous variation, they find evidence of the (in the case, causal) effect that additional math courses have on individuals' career for the entire ability distribution. This is important, again, to point out the importance that math performance (or in this case, the choice to specialize in this field) has in improving individuals' labor market outcomes and the relevance that has in closing the gap. Although potentially related to, or caused by, demand-side factors, other authors investigate supply side explanations. For example females' risk aversion, confidence and competitiveness. We will see in next Sections how economists have examined this preferences, in particular the attitude toward competition, its causes and consequences.

As we have argued building this parallel (and, through the evidence from literature, this causal effect that science and math performance have on labor market outcomes), the fact that there are factors causing women to underperform in math (or, in general, STEM fields) is an economic matter. A pool of (women's) talent is lost in this scenario, and managing to retrieve this talent is "crucial to sustain economic development, growth, productivity and innovation" (Joensen and Nielsen, 2014). In general, gender differences in the labor market have been shown to be detrimental from an economic perspective (Galor and Weil, 1996) .

2.2 Gender Gap and Competition

A branch of the literature has investigated, in last 20 years, the difference across genders in individual preferences and in particular in attitude toward competition. Both the effect of competition per se and of competition due to risk aversion, uncertainty, confidence and beliefs about relative performance have been investigated. As we are going to see looking at the literature, individual preferences can potentially influence both the decision to entry and the performance within competitive contexts. The point here is that the distortion of competition can vary by gender, then cause gender differences in outcomes (Niederle and Vesterlund, 2010). From the evidence emerges that it "could be a relevant trait to explain entry into fields such as sciences and mathematics, which are male dominated and viewed as competitive" (Buser, Niederle and Oosterbeek, 2014) . Fields that play, as we have argued, an important role in defining the gender gap in career advancements and earnings. Moreover, if competitiveness influences individuals' choices during adolescence, this has a potential long-term impact on females' labor market outcomes (Dreber, Von Essen and Ranehill, 2014) relative to males.

Both experimental and non-experimental literature have studied the effect of preferences on individuals' choice to entry in competitive contexts. Here we describe some important articles. In a lab experiment Niederle and Vesterlund (2007) ask participants to choose between a piece-rate incentive scheme and a tournament competition (where the relative performance is relevant in order to get compensation) in a real effort (and mathematical) task. They find a strong difference between males and females, with the latters being less likely to choose competitive environments conditional on performance in which they find no gender differences. For this reason self-selection in the competitive environment is not explained by maximization of earnings. The negative effect of competitiveness results to be robust even controlling for individuals' beliefs, risk and feedback aversion giving the opportunity to the authors to infer a negative effect purely of "taste for competition" (in addition to the gender difference in overconfidence whose effect is remarkable), although it could be that these or other preferences can instead explain the resulting gap due to the imperfection of their controls (Flory, Leibbrandt and List, 2015). Importantly, the difference is also driven by males having stronger preference for competition than the predicted one. Similar to these authors, Dreber, Von Essen and Ranehill (2014) find that in mathematical tasks males are as twice as likely than females to choose the competitive compensation scheme even controlling for performance. To understand the effect of competitiveness, Buser, Niederle and Oosterbeek (2014) run the same experiment as Niederle and Vesterlund (2007) in the Netherlands obtaining almost the same (hence expected) evidence on females' and males' choice about compensation scheme. Moreover, they use the standardized measure of competitiveness from this experiment to investigate whether this has an effect on educational choices. They also have information about participants' academic performance and perceived math ability. Authors find that, even if similar in performance to females, males are more likely to choose more prestigious and math intensive academic tracks. This choice positively correlates with the measure of competitiveness that they find explaining 20% of this gender gap in educational choice, controlling for the already cited measures of ability. The effect of competitiveness is confirmed by Zhang (2013) that similarly uses the same standardized measure finding that competitive students are more likely to take entry exams conditional on past test scores, although in her study there is no evidence of gender difference in preferences. Overconfidence and competitiveness also influence individuals' earnings expectations, potentially having an effect on job sorting and contractual negotiation. This is what Reuben et al. (2015) argue finding that these two factors (defined with the experimental measures by Niederle and Vesterlund, 2007) positively correlate with expected wages and explain the 18% of the gender gap in forecasts. Looking at actual earnings Reuben, Sapienza and Zingales (2015) find that, for MBA students involved in their experiment, competitiveness leads to higher wages (9% higher for competitive students) and can explain 10% of the realized gender gap in earnings. Outside the lab, Flory et al. (2015) run a natural field experiment to analyze job-entry decisions. They randomly assign groups of job-seekers to different compensation regimes confirming that also in these environments women tend to shy away from competition more than men, even if this gender difference in self-selection is driven by males being less averse than females to compensation based on relative performance.

As we have seen many papers find evidence of gender difference in preferences (in particular about competition). Once verified the positive correlation that attitude toward competition has with both educational choices and job-sorting, these studies demonstrate that the gender gap in competitiveness translates into the gender gap in choices (and, sequentially, in related outcomes).

Now we ask whether different preferences lead to distorted performance in competitive contexts. In case competition causes or widens the gender gap in performance, “males will tend to perform better than females in jobs that involve competition […] or when competing for promotions and advancement within their organizations” (Cotton et al, 2013) and this is relevant for highly paid and top jobs in which the relative performance is more important than the absolute one (Azmat and Petrongolo, 2014). In Section 2.2.2. we will argue that the distortion of competition is particularly linked to the math content of the tasks. This is a relevant finding. Moreover, in math contexts it results to be very large at the top of the performance distribution (Niederle and Vesterlund, 2010). It is crucial to investigate this effect from a policy perspective: allowing high-performing (in noncompetitive contexts) women to enter competitive environments could not be enough to assure equality of outcomes in presence of this distortion (Niederle and Vesterlund, 2011).

Also in this case, we can find many contributions. Gneezy et al. (2003) run an experiment dividing participants into groups and asking them to solve mazes under different compensation regimes: piece-rate, competitive and random pay. Under the first one respondents were paid for they own performance, under the second and the third one only one participant in each group is paid based respectively on relative performance (the best performer is the one who receive the compensation) and on a random selection. Comparing piece-rate and competitive pay they show that the gender gap in performance, while not significant in the first case, increases (around three times) and become significant with competitive pressure, mostly driven by men positively responding to competition. With the random pay, in which both males’ and females’ performance are similar to that under piece-rate scheme, they show that the difference found in the tournament pay is not owing to the uncertainty of compensation. Of particular interest is that in single-sex groups females as well respond to competition significantly increasing their performance and narrowing the gender gap, while males seem not to be sensitive to competitors’ gender. The peculiar suggestion from this finding is that the main concern for women could be the competition against men. We will come back on this providing further evidence. Moreover, Gneezy and Rustichini (2004) show that the cross - gender difference in response to competition is already present in a group of Israelian children aged 9 years, although in a physical real effort task and without compensation. The gap in performance is, again, present only in the competitive context.

Cotton et al. (2013) run an experiment in elementary schools submitting math tests to groups of children. They notice a stronger positive response to competition for men’s performance conditional on abilities (measured with past grades) in line with the past experimental literature. In particular, in their study high-ability females tend to underperform, the reverse instead for low ability males. They design a repeated competition showing that the males' advantage disappears after the first round arguing that the impact with competition could be the relevant trait for female children. Designing in parallel other contexts the authors show that when lowering the pressure on students, females better response to competition in each round well reflecting their abilities in the actual performance. Ors, Palomino and Peyrache (2013) in a nonexperimental study, with a real-world setting, try to examine the distortion that we are exploring. These authors analyze a very competitive setting (with large future stakes) in which students take part in an entry exam to one of the most prestigious French schools (HEC). They find that men overperform women in this high pressure environment while their performance is equal or lower (against the same students) in recent and noncompetitive academic context. The latter evidence is confirmed in the less competitive first year of the course (obviously, for admitted students). Two interesting facts have to be mentioned. First, as they notice, these results are "unlikely to be explained by a "women shying away from competition" argument" since the percentage of graduating females from science-intensive high schools is similar to that of female candidates with the same academic background. In this sense, entry and performance in competition seem still to be two distinct issues. Second, looking at the performance distribution the gender gap is more substantial at the top having a material effect on the composition of selected students. Delfgaauw et al. (2013) in their field experiment investigate the effect of competitive pressure on Dutch retailers' performance exposing to competition groups of retail stores. While observing the irrelevance of monetary incentives they find a positive effect of competition on sales growth both for female-led and male-led stores. Moreover, they find that the increase in performance is stronger when more team components have the same gender of the manager, with a symmetric effect both for men and women.

Once provided evidence of the negative impact that competitive pressure has on females' choices and relative performance (relative to males) due to their different preferences, we briefly present the Nature vs Nurture debate on the origin of these gender – specific preferences.

2.2.1 Nature or Nurture

From the literature we see that different factors lead women to stay away from competitive contexts and suffer pressure when performing. Competitiveness, risk aversion, feedback aversion, confidence potentially influence women's behavior. We have mostly explored literature about competitiveness but some authors underline the importance that other preferences have, both alone and in interacting with pure “taste for competition” (for example, risk aversion in Niederle and Yestrumskas, 2008). Other authors suggest that distributional preferences (see Dasgupta et al., 2019), inclination for cooperation and egalitarianism (Kuhn and Villeval, 2011) can explain women’s response to competition. Gender difference in beliefs could be a relevant factor as the distortion of competition is mainly encountered in stereotypical – male tasks (Niederle and Vesterlund, 2011).

In general, there are several contributions trying to find evidence of gender specific preferences. There is also a wide debate about the origin of gender differences: the “Nature vs Nurture” debate. On the first side, part of the literature has tried to show a causal link between biological factors and females’ outcomes and preferences (for a review see respectively Azmat and Petrongolo, 2014, and Croson and Gneezy, 2009). On the nurture side, many authors sustain that causality comes from sociocultural factors. The main two approach are, first, to demonstrate that difference among males and females do not exist at very young age but are “learned” from the environment where individuals grow up and, second, to show that females from opposed cultures or backgrounds behave differently.

Sutter and Rutzler (2010) and Gneezy and Rustichini (2004) find that male and female children respond to competition as adults, this is not true for Cardenas et al. (2012) who study Swedish and Colombian children aged around 10 (or for 3-years-old children in US in Savikhin, 2011). For what concerns the second approach, females from single sex schools have shown to be more similar to boys in terms of risk aversion and competitiveness (Booth and Nolen, 2012, and Booth and Nolen, 2009), and this clearly confirms that circumstances in which individuals grow up are relevant in shaping their preferences. Gneezy et al. (2009) in an important experiment try to solve this debate. Similarly to other studies that we have described, they ask participants to face a physical task choosing to perform either under a piece-rate or a competitive compensation regime. Nevertheless, they run this experiment with children from opposed patriarchal and matrilineal societies: respectively, the Maasai society in Tanzania and the Khasi society in India. In the Maasai society, with a patriarchal organization, results replicate those we have seen so far: males compete more than females. The suggestive finding is that in the matrilineal society, where children have completely a different culture (and different examples) about the role of women, the classic finding is reversed: in this context females are those who show a stronger preference for the competitive regime. Once established the effect of socialization on individuals’ inclinations, Andersen et al. (2013) wonder at what age this effect takes place. Similarly to the previous study, they compare children and adolescents from patriarchal and matrilineal societies in India. They find that while 7-years-old children show no gender differences in both kind of societies, this is still true only for 15-years-old teenagers in the matrilineal society. In the opposed one, men becomes more competitive relative to women around puberty.

There is evidence in the literature that both nature and nurture influence individual preferences. The fact that nurture plays a still relevant role crucially confirms that there is room for intervention in order to manipulate or, at least, better address these preferences (Niederle and Vesterlund, 2011).

2.2.2 Math Tasks

In the articles that we have described some authors show that relevant findings depend on which task is performed. For instance Dreber, Von Essen e Ranehill (2014) reveal that the gender gap in competition entry does not endure when submitting verbal tasks and that math tasks are relevant only for women, where these are "two different tasks that differ in associated stereotypes". Cotton et al. (2013) in their experiment in elementary schools show that the first-round negative impact of competitive pressure on female children's performance is not in place when switching to questions about arts and language. In other studies, actually, task does not seem to matter for very young students (previously described, Dreber et al., 2011 and Cardenas et al.,2012). In another important paper Shurchkov (2011) finds no gender difference in performance in verbal task while also in her study males better perform in math task. Interestingly, in a low pressure setting this gap shrinks and in the verbal task females even outperform males. This testifies that there are other forces than simple gender-specific taste for competition or, at least, that this preference is driven by other factors as stereotypes. Math task are typically associated to men. Past literature shows that the gender gap in math is smaller in contexts where gender stereotypes are less pronounced. Moreover, experiments demonstrate that individuals are underconfident in tasks typically associated to the opposite gender, this is true also for men (Carlana, 2019). This corroborates the hypothesis that females' underperformance are subtly undermined by erroneous gender-tasks associations that self-fulfill through the effect on individuals' behavior. Undoubtedly, as we have shown are "more male associated tasks that are important for labor market outcomes" (Dreber, Von Essen and Ranehill, 2014) and from a policy perspective it is clear that this dislike for competition, wherever it comes from, needs to be better addressed. Niederle and Vesterlund (2010) suggest some other potential explanations for females' sensitiveness to math contexts and tasks. For example, the peculiarity of math questions whose answers are either correct or wrong and therefore "in contrast to verbal test scores, math test scores may better predict actual rank as well as future relative performance". Another reason could be that the proportion of men that choose and specialize in this field with respect to women is much higher, then both actual and expected competitors are mainly males. Therefore, the real issue could be the competition against men. We think that the second explanation potentially has both a direct and indirect impact as the unbalanced presence of men in parallel contributes to strengthen field-specific stereotypes. Additionally, as we have argued, also the direct impact of "competition against men" could be owing to the "stereotype threat". In next Section we further explore the latter argumentation from these authors to understand possible policy implications.

2.2.3 Competing Against Men or Women?

We have argued that confidence plays a major role when choosing the compensation scheme. One circumstance in which both beliefs about relative performance and attitude toward competition can be altered is the single sex competition (Niederle and Vesterlund, 2011). Huguet and Regner (2007) investigate the effect of "stereotype threat" finding that in a mathematical in-classroom test (presumably submitted to measure math abilities) women underperform in mixed-gender groups while not in same-gender groups. The gender composition seems to matter also in other studies, for example in the article by Delfgaauw et al. (2013), where teams of retailers responded better to competition in this condition. If we come back to Gneezy, Niederle and Rustichini (2003), they show that the same is true when individuals compete against each other: the distortion of competition causing female underperforming disappears in single-sex tournaments. The gender gap narrows in this environment also due to men responding less to the competitive compensation scheme. Not only performance, also probability of entry increases in same-gender groups conditional on performance (Booth and Nolen, 2012).

Niederle et al. (2013) analyze the effect of a second-stage "soft quota" in which at least one winner of the competition has to be female, with no restrictions on the second winners' characteristics. In the first stage participants need to choose the compensation regime as in the other experiments. Once chosen the tournament, they know that, at least for females, the competition becomes gender specific within the group they are selected in. Females need only to outperform participants of the same gender in order to get the reward. The authors find that their affirmative action increases the probability of entry for females while weakly discouraging males' tournament choice with both effects being wider than "predicted by the change in probability of winning". Moreover, they compare the tournaments with and without the quota to assess eventual policy costs (e.g., reverse discrimination toward males that would have gained the compensation absent the action). They evaluate these costs both in the tournament with the actual quota and in the other tournament taking realized performance and applying a fictious "ex-post quota". The fact that policy costs emerge only in the second case demonstrates that the quota is efficient only in case of prior announcement, giving the opportunity to females to change in confidence and beliefs about own relative performance. The affirmative action, notably, fosters high-achieving females' tournament entry while narrowing the gap in confidence (as beliefs within gender are different) . From the previous comparison, as the authors underline, we see that "indirect effects that occur through self-selection into competitions" are crucial in order to counterbalance eventual costs that cancel out when the distortion of mixed gender competition is substantial. Maggian, Montinari and Nicolò (2017) run a lab experiment to study the effect of quotas on women's career advancement. In this experiment they investigate the optimal timing for the introduction of a gender quota, possibly without any loss of efficiency, designing three types of treatment: a quota at entry level, at top level and at both stages. They find that the second and third treatments (a quota both at the top and at each stage of a career) successfully foster high performing women's participation, where the first one turns to be "more advisable" specifically "under a principle of minimum interference". Important, they observe the negative effect of a quota at entry-level that adversely affects women's confidence causing them not to well perform at higher stages of their career. As we have seen, these kind of policies are sometimes accused to lower the efficiency (in terms of winners' performance) and to create reverse discrimination against males. This is not supported by evidence in the studies that we have described. On the contrary, they are needed when distortions as direct discrimination, stereotypes and their effect on individuals' behavior induce a sub-optimal selection (Niederle et al., 2013) among winners of a competition or hired workers.

Some authors argue that to increase effectiveness of policies the timing is crucial. Preference-manipulating policies should be optimally run when preferences become gender-specific due to socialization (e.g. around puberty, Andersen et al., 2013). Undoubtedly, in the long term these kind of policies can be beneficial, but in the short run affirmative actions can be proven to improve efficiency and favor equity. Moreover, policies that foster the presence of women among high or best-achieving mathematicians, engineers and scientists potentially have long-term positive effects on female' perception of own abilities and potential in these fields relaxing societal constraints coming from well-rooted stereotypes (Stout et al., 2011). Finally, in past literature other contributions suggest that other policies can change females' preferences about competition as performance feedback, that can improve women's confidence (Dreber, Von Essen and Ranehill, 2014), or teamwork (Flory et al., 2014, review from Azmat and Petrongolo, 2014).

3. The Field Experiment: Progetto Rodonea

This Field Experiment has been run during the Math Olympiad in randomly selected schools in the Italian North – East. The Math Olympiad, organized by "Unione Matematica Italiana" is the Italian oldest math competition that involve each year about 200'000 students from the whole country. It is an individual competition divided into three phases: school level, district level and national level. In parallel, it includes team-based contests. This experiment took place in the first phase where the competition was at school level, i.e. in order to win one student had to overperform only participants from the same school (where the competition physically took place). This phase is called "Giochi di Archimede" and took place on the 21st of November 2019 simultaneously in involved schools. Although sometimes with different criteria and different quotas per school, participants in higher levels are chosen based on previous level ranking. That is why not only the 1st place is relevant for students in order to consider successful their participation. Moreover participants within schools are selected mainly based on three criteria: free, "suggested" (where professors invite presumably best-achieving students in math subjects to enjoy the competition) and "mandatory" participation. In the latter case students are obliged to participate mostly based either on meritocracy or belonging to specific (math-intensive) school curricula. However, the chosen criterion within the school is likely to be nonrandom and correlate with school-specific characteristics as prestige, reputation, attitude towards these extra-curricular competitions etc. The second and third criteria can generally imply the first one since among the pool of remaining (i.e., not-invited or not-obliged) students, pupils can ask to enjoy the competition. Though, it is likely to be rare and whenever different criteria coexist the intentional exclusion probably discourages not-selected students.

The experiment began about one month before the Olympiad when randomly selected students within the (in turn, randomly selected) schools where asked to fill a survey in order to collect data (age, family and academic background, self-reported math grade etc..) about them. Most important for our analysis, in this survey it was included a widely used and recognized test to measure students' non-cognitive abilities: the Raven Test. We will exploit the score in this test in order to identify high ability students to investigate the eventual heterogeneous effect of the intervention along the ability distribution. Once selected, schools were (again randomly) divided into three treatment groups with different incentive schemes:

1. In the first group, independently on gender or other individual characteristics, the four students at the top of each single school ranking received a monetary reward. The first and second ranked student were rewarded with 200 euros each. The third and fourth classified participants received 100 euros. From now on, we will call this group "No Gender".
2. In the second group, both the first man and the first woman were rewarded with the highest pay (200 euros). So, virtually, in order to get the reward, it became a within-gender competition. The same division for the other prizes, both second male and second female student received 100 euros. We call this the "Gender" group.
3. The control group with no incentives, i.e. no intervention in the experiment. We call this group "Baseline".

Competitions and incentive schemes are at school level. In order to get the reward each student had to rank at the top in her/his own school. Most important, the ranking-based prizes were announced well before the competition in the involved schools. As we have seen in the literature, affirmative actions have shown to need prior announcements in order to be effective allowing individuals to react adapting or modifying their preferences or expectations (and related behavior) .

This interventions are supposed to foster both participation and performance within the competition. In order to better evaluate the effect of these treatments we will exploit the control group where we can observe outcomes absent the action. In particular, since from the previous Sections we have seen that women behave differently in single sex competitions, we want to understand whether the "Gender" treatment has a positive effect in either narrowing or closing the cross-gender participation and performance gap (we will first control for realized gaps both in the Control group and the No Gender group). So, we will deepen these two comparisons. First, the effect of monetary incentives (or higher pressure in an already competitive environment) comparing treatment groups with the control one. Second, we will investigate whether the intervention with gender-based rewards has a different effect on participants' outcomes.

4. Data and Variables

In this Section we present our data sources and sample. We exploit four main sources:

1. The survey submitted to students in involved school;
2. Individual-level data about participation and Math Olympiad performance;
3. Data about schools' characteristics;
4. Fondazione Agnelli[2]'s dataset.

Most important, the survey was randomly submitted to students within involved school, that is why we do not have information from this source for each math Olympiad participant. While we collected data about school characteristics for each single school, the Fondazione Agnelli's dataset does not cover every track (or, curriculum) within schools. We have two main samples: students that filled the questionnaire and students that enjoined the math competition. We will look, in particular, at the second group to detect eventual gap in performance and evidence of realized treatment effect with a cross-sample comparison. In order to exploit other data sources we will deepen the analysis controlling our results' robustness with a regression approach trying to solve the selection problem. Therefore, at the end, for our analysis we will study the so-called "matched" group for which we have complete informations (i.e. students that both filled the survey and participated at the Math Olympiad). Exploiting all sources we will further restrict the analysis to the group whose involved tracks are included in the Fondazione Agnelli's dataset. The latter is important to check whether the effect that we see is driven by schools' quality or characteristics.

We will look at data comparing distributions of variables across different groups. In order to understand if these groups are similar (or if the mean of considered variables in distinct groups is significantly different) we will exploit two tests: the T-Test and the Kruskal-Wallis Test with the latter being a generalization of the Mann-Whitney Test Method for multivalued variables (i.e., for variables that define two or more than two groups). The null hypothesis for the T-Test is that the mean of considered variables is equal in both groups while for the Kruskal-Wallis Test is that samples belong to the same population (i.e., the mean is not significantly different across defined groups). For both tests we use the version for clustered data since in our sample observations are divided into clusters (schools) whose intra-correlation should be taken into account. When computing means and running tests, for variables that refer to schools we have one observation for

[2] I want to thank Fondazione Agnelli (https://www.fondazioneagnelli.it) for allowing me to use their data.

each single school, the same for tracks within schools (e.g., variables defined based on Fondazione Agnelli's dataset).

Let us see into detail these sources, our derived variables, the sample's characteristics and data about performance. For the sake or readability we will include in the text only the most significant tables. You will find other tables at the end of this work (p.50).

4.1. The Survey

The first source is our survey. In this survey we collected, about one month before the competition, informations about randomly selected students within involved schools. Both timing and randomness are relevant, in particular for the No Gender and Gender groups, in order not to influence participants when filling the survey. Indeed, the questionnaire and the intervention with incentive schemes were presented as separated issues. For the scope of this analysis, we will exploit data about the Raven Test scores, test that was included in the survey in order to measure students' non-cognitive abilities. This is the only students' objective ability measure at our disposal. "The Raven's Standard Progressive Matrices (RSPM) instrument is a multiple-choice test used to assess mental ability associated with abstract reasoning" (Bilker et al., 2012). The original test from Raven (1938) includes 60-items in which the respondent has to choose the right answer mainly among 6 or 8 choices. Other versions of this test are available. In our Survey we used the version from Bilker et al. (2012, version A) that includes 9 items. Authors have studied the predicted power (predictive of the 60-items test score) of smaller variants of the test optimizing first the number of items then identifying the right questions to include in the best-predictive 9-items version.

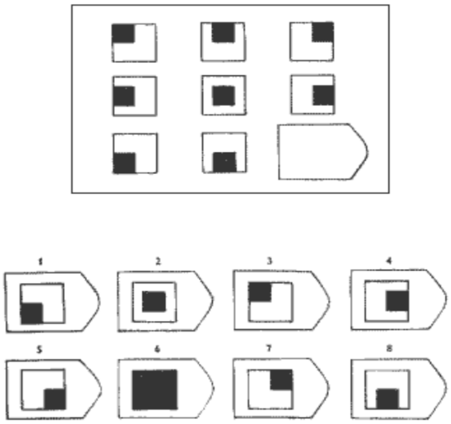

Figure 1. Example of item from Raven (1938)

For each right answer respondents received one point, so the minimum and the maximum of the Raven Test score were respectively 0 and 9. Students had 270 seconds to answer correctly as many questions as they could. After this time they could no longer change or add responses being redirected to the next phase of the questionnaire. Respondents were aware of the limited time but there wasn't any timer on their screen. This should have prevented desperate or random answer. In Figure 1 we show one question as an example, where the right answer is the number 5.

Based on the Raven Test Score we have identified top students, defined as student whose score was above a certain threshold in the score distribution at school level. We have established three thresholds: the 50th , 75th and 90th percentiles in order to identify respectively students belonging to the within-school (among students who filled the survey) top 50, 25 and 10 of the distribution. From an economic perspective it is important to understand how the treatments' effect, if any, is distributed along the ability distribution. We have seen in the reported literature that mainly top-skilled women's outcomes are distorted by competitive contexts. In general, efficiency-optimizing policies designed to encourage participation and foster performance of specific groups of people are most often meant for top abilities or (at least potentially) high achieving subjects that fail to enter appropriate context and to well-reflect their abilities in realized performance. That is why also in this case we will investigate whether our intervention can manage to boost top females' performance increasing, in turn, their probability of winning the competition and access upper levels.

In Table 11 (p.50) we see that the survey was submitted to students coming from 29 schools: 8 belonging to the Baseline group, 11 and 10 respectively for the No Gender and the Gender group. In next paragraphs we will look at info on schools. In the second Section of this table (Section "Info on students") we see that in the entire sample as well as in the top 10, 25 and 50 both males and females are well represented (never below the 40% of the subsamples), even when separately considering different treatment groups. In the same Table we report within-group means of Raven Test scores for the entire sample of students that filled out the survey (i.e., the entire score distribution) and for top students, also dividing by gender. In the last column we report the p-values of the Kruskal-Wallis Test for clustered data that tests the hypothesis that the mean of our variables is not statistically different across treatment groups. High p-values testify that both when considering entire groups and single genders, we cannot reject the hypothesis that they belong to the same population in terms of abilities (measured with the Raven Test Score). This is true both for the entire sample of involved students and for students at the top of the distribution. It is important to clarify that students are statistically equally skilled across groups, so we do not expect to see one group performing better than the others absent the action (i.e., if we observe different performance it is not due to different abilities across treatment groups).

As underlined in the previous Section, in order to exploit informations about students' abilities to investigate heterogenous treatment effects along the ability distribution we will restrict the analysis to the "matched" subsample (students that both filled the survey and enjoined the math competition). We are interested in understand two things: first, if the just-explained evidence of treatment groups belonging to the same population holds in this subsample; second, whether the matched group is statistically different from the remaining pool of students or it can be considered representative of the entire sample. Table 12 (p.51) is similar to the previous one but considers only the matched group. Still, both when considering within-group averages either dividing by gender or not we cannot reject the Kruskal Wallis Test's null hypothesis. This is true for the entire score distribution as well as for top students. For what concerns the comparison between "matched" and "not-matched" students we look at Table 14 (p.52) where we report within-group means of test scores for various sub-groups. We can notice that when considering the entire sample of students that filled the survey matched individuals have significantly higher scores. This means that, even if the survey was randomly submitted in the schools, we casually selected better student based on these scores. This difference holds, in particular, for females. Anyway, looking at the top of the distribution (top 10, 25 and 50) we cannot reject the T-Test's null hypothesis of equality of means for matched and not-matched individuals. High p-values suggest that the difference across these two subsamples is mostly driven by disparity at the bottom of the score distribution. Hence, looking at this evidence from our survey we have run significant randomness checks. Students, in terms of abilities are well distributed across treatment groups and looking at Olimpyad scores we will examine a representative group of students, in particular top students, thanks to the randomization when submitting the questionnaire.

4.2. The Math Olympiad

The second source are data about participation and individual's performance at the Math Olympiad. In the following table we report data about schools and students for the sample of individuals for which we have Math Olympiad scores. Still, looking at the second Section "Info on students" we see that both males and females are well represented in the sample of 6060 participants. In the third section "Olympiads" we report data about participation within the 29 involved schools and test whether there is a significant difference across treatment groups. Within-group averages are computed considering one observation for each single school. While there is no cross-subsample significant difference in the number of total and female participants, in the Gender group the percentage of female participants with respect to the total number of females within school is significantly lower. This difference does not hold for male participants.

Table 1. Variables by Treatment

	Baseline	No gender	Gender	K-Wallis [c] (p-value)
SAMPLE OF INDIVIDUALS WITH SCORES				
INFO ON SCHOOLS [d]				
Number of participating schools	9	11	9	
Average number of females in school	444 (51%)	547 (51%)	464 (51%)	0.7754
Average number of students in school	833	1060	935	0.5410
Olympiads selection criteria [a]:				
Free	1	2	4	
Suggested	1	5	2	
Mandatory	7	4	3	
Average number of tracks per school	3.67	3.73	4.44	0.3364
Track Specific [b]:				
Females' Matriculation Rate (University)	86%	89%	82%	0.0836
Males' Matriculation Rate (University)	85%	86%	81%	0.5873
Females choosing STEM fields	48%	48%	41%	0.2805
Males choosing STEM fields	64%	59%	61%	0.9786
FGA Index	68.44	71.87	68.44	0.2765
INFO ON STUDENTS				
Number of participating students	2344	2404	1302	
Males	1281 (54.6%)	1330 (55.3%)	782 (60.1%)	
Females	1063 (45.4%)	1074 (44.7%)	520 (39.9%)	
Olympiads selection criteria [a]:				
Free	20 (0.8%)	250 (10.4%)	417 (32%)	
Suggested	58 (2.5%)	542 (22.5%)	317 (24.4%)	
Mandatory	2266 (96.7%)	1612 (67.1%)	568 (43.6%)	
OLYMPIADS				
Total Participants	2344	2404	1302	
Average number of female participants in school	118 (46%)	98 (43%)	57 (41%)	0.4211
Average number of participants in school	260	219	144	0.5333
Female participants vs Females in school	37%	19%	14%	0.0464
Male participants vs Males in school	41%	28%	22%	0.1155
Participants in school vs Students in school	38%	22%	17%	0.0665
Olympiad Score (Average)	32.53 (16.33)	31.99 (14.49)	31.96 (15.27)	0.9562
Olympiad Score Females (Average)	30.57 (14.81)	29.84 (12.81)	29.82 (13.86)	0.9603
Olympiad Score Males (Average)	34.16 (17.32)	33.72 (15.50)	33.39 (15.98)	0.9137

Percentages and standard deviations in parentheses.

a. It is the criterion that professors in the school use to select Olympiads participants. The participation can be on a voluntary basis (i.e. "Free"). On the other hand participants can be chosen by professors ("Suggested") that invite best students to participate, or it can be that some classes or best students are forced to participate based on merits or other characteristics ("Mandatory"). Also these two criteria in some cases could imply that not invited or forced students can join the competition but it is likely to be rare.

b. These data refer to single tracks within each school and come from Fondazione Agnelli. Data are not available for every track but almost the entire sample is covered. Numbers are within sample averages.

c. The null hypothesis is that means are equal across samples (i.e. samples belong to the same population). When comparing Olympiad Scores and Track Specific data it is a Kruskal-Wallis test for clustered data. Clusters: Schools.

d. Observations are single schools and single tracks both for averages and for tests.

Important to verify is whether this women's lower participation is due to the treatment's effect or to other factors. In the first case the incentive scheme based on students' gender could discourage females' entry in the competition, this would not be in line with the reported literature. In the second case selection criteria or other factors could negatively influence females' participation and we should control for this in our regression analysis since there could be a negative effect also on observed Olympiad score. We want to point out two important facts:

- First, also the fact that participation is not significantly different across treatment groups should be investigated since this does not necessarily means that monetary incentives do not encourage participation. As before, there are other factors potentially influencing participation for both genders limiting or even counter-balancing any treatment effect. Moreover, this effects could be heterogenous across genders regardless of the type of incentive scheme. For example, schools that select best students could tend to choose more males if they think that men are better prone to competition with respect to females or simply because of stereotypes about students' perceived math ability. When examining treatment effects we have to take into account for these residual factors since they potentially alter our evaluation of results in particular, as we will see, in groups where students are almost never free to choose whether to enjoy the competition. We will deepen this issue in Section 4.3 ;
- Second, the Kruskal-Wallis Test allows us to compare at the same time all three groups, and this is useful to understand whether they can be considered homogeneous based on observed variables. The limit of this Test is that it says nothing about cross-groups pair comparison. From the previous table we see that all variables suggest that in the Gender group the participation is lower both considering the absolute number and the percentage of students within the schools enjoying the competition. Our test verifies that (except for the percentage of female participants vs the number of women within the schools) these differences are not significant. But, since means are very different, it could be that in a pair-comparison the Gender group results to be very different from, for example, the Baseline group. Again, we should then examine whether this is caused by the incentive scheme. In next Sections we will report only the significant pair-comparisons that will give us useful insights for our analysis.

Other variables shown in Table 1 will be described in proper paragraphs. Now we look at individual scores at the Math Olympiad.

4.2.1. Olympiad Scores

Italian high-school courses last 5 year. The test in the competition is different for the first two years (called “biennio”) and the other three years (called “triennio”). In particular, students attending the first two years of the course have to answer 16 problems, while older students have a test with 20 questions. In order to make scores comparable, those of the first group of students are multiplied by 1.25. That is why in next tables sometimes we will call them “Normalized Scores”. This could be an issue when comparing scores of students from different classes but we will address it controlling for this difference in our regression analysis. For what concerns scores, students receive 5 points for each right answer, 1 or 0 points respectively for blank and wrong responses. This should discourage random answering. As for the Raven Test Score, we want to verify whether “matched” students can be considered a representative group of the whole sample of participants in terms of Math Olympiad scores. This is important as we will look at this narrow group to understand if top students have differently reacted to incentive schemes. For this purpose we look at Table 13 (p.52) that similarly to Table 14 (p.52) reports averages of scores both for matched and not-matched individuals. When looking at the whole sample as well as single treatment groups averages are not significantly different between matched and not-matched students. The same holds when considering single genders. This means that if we restrict the analysis to the sample of matched individuals we can potentially extend our results to the whole sample. This narrow sample is not statistically different from the entire sample that filled the survey in terms of abilities (see Section 4.1) and is not different from the sample for which we have data about Math Olympiad performance (in terms of Olympiad scores). The matched subsample comes from the intersection of these two broader samples and for these reasons it can be considered a valid representative group of both of them.

We want to look now at Math Olympiad scores to investigate first evidence from summary statistics. In the Table 2 (below) we report within (treatment) group Olympiad score averages both for the entire sample of participating students and for top students. Remember that top students are defined looking at the score distribution of individuals that filled the survey for who we have the Raven Test scores. Testing the equality of means with the Kruskal-Wallis Test, we can see that in the entire sample averages are very similar (p = 0.9562) even when separately considering males and females. This means that in this first look we cannot find any treatment effect since there is no variation in average scores across different groups. Students’ performance are not significantly

Table 2. Difference across samples in performance at the Math Olympiads

Sample	Subsample	N° Obs	Baseline	No Gender	Gender	P-Value (K-Wallis [a])
Entire with Score [b]		6060	32.53	31.99	31.96	0.9562
Entire with Score [b]	Females	2664	30.57	29.84	29.82	0.9603
Entire with Score [b]	Males	3396	34.16	33.71	33.39	0.9137
Top 10%		490	46.92	36.64	40.23	0.1654
Top 10%	Females	245	41.30	32.33	37.54	0.1502
Top 10%	Males	245	52.64	41.96	42.21	0.2141
Top 25%		986	40.98	34.94	37.05	0.4202
Top 25%	Females	496	35.96	31.95	33.59	0.6970
Top 25%	Males	490	45.81	38.93	39.59	0.3848
Top 50%		1676	37.74	33.29	35.05	0.5405
Top 50%	Females	865	33.74	30.90	31.97	0.8205
Top 50%	Males	811	41.61	36.54	37.54	0.5553

a. The null hypothesis is that means are equal across samples (i.e. samples belong to the same population). It is a Kruskal-Wallis test for clustered data. Clusters: Schools.

b. The Sample includes all individuals that participated at the Math Olympiad.

higher in subsamples that received monetary rewards, neither in the No Gender nor in the Gender group. If we look at the top of the ability distribution we find the same evidence although the scores become very different, in particular in the top 10. We will deeply investigate this first evidence's robustness controlling for differences across treatment groups. Indeed, these differences could hide eventual treatment effects because of the incomparability of treatment groups.

4.2.2. The Observed Gap

Before looking at other variables and include them in our regression analysis, we want to check for any realized gender gap in math performance along the ability distribution. In Table 3, we report males' and females' average score in the entire sample for which we have scores, separately reporting also within-treatment-group averages. We do the same restricting the sample to the matched group, looking at top students. We then report the difference and test the null hypothesis of equality of means to understand whether this difference is statistically significant. In the top panel of our table we see that the gap in performance is present in the whole sample as well as in each single group, but we cannot reject the hypothesis that means are equal across genders. Interestingly, in other panels we see that among high-skilled student males overperform females with the gender gap being wider and significant at the top of the ability distribution when considering all three groups together. Males significantly ($p < 0.05$) perform better than females both in the top 10, 25 and 50 respectively with a positive difference in average performance of 8.65, 7.57 and 6.34. However, looking at within-group differences we cannot infer any treatment effect in widening or narrowing this gender gap since the cross-gender difference in performance is never statistically significant in the single treatment groups. Also in this case, we will check for robustness of these results in next Sections.

What remains interesting in this first look are two things. First, the gender gap in performance seems to be wider and significant at the top of the ability distribution. Second, in the top 10 the gap shrinks in the Gender group compared to the other groups, although we cannot hypothesize any treatment effect since the gender gap is not significant neither in the Baseline nor in the No Gender group. In the previous Section we have shown that students' (even considering single genders) realized performance are statistically not different in means across treatment groups. We have found now that this evidence holds also comparing within-groups gender gaps. The latter comparison takes into account that even if treatment groups are different in characteristics, males and females come from the same schools sharing the same trends. However, in each group there are several schools. Still, we have to control for other factors influencing participation and performance to understand if "ceteris paribus" treatments had any significant effect on females' relative outcomes. That is, we do not only care about absolute effects on participation and performance (that we have seen in the previous Section) but also about the heterogenous effect across genders, especially at the top of the distribution were the gap seems to be wider.

Table 3 . Gender Gap in performance at the Math Olympiads

Sample	Subsample	N° Obs	Average: Males	Average: Females	Difference	P-Value (T-Test [a])
Entire with Score		6060	33.81	30.13	3.67	0.1564
Entire with Score	Baseline	2344	34.16	30.57	3.58	0.2996
Entire with Score	No Gender	2404	33.72	29.84	3.87	0.3150
Entire with Score	Gender	1302	33.39	29.81	3.57	0.5550
Top 10%		490	44.50	35.84	8.65	0.0319
Top 10%	Baseline	113	52.64	41.30	11.34	0.1629
Top 10%	No Gender	219	41.96	32.33	9.63	0.1072
Top 10%	Gender	158	42.21	37.54	4.67	0.4586
Top 25%		986	40.95	33.37	7.57	0.0251
Top 25%	Baseline	247	45.81	35.96	9.84	0.3146
Top 25%	No Gender	420	38.93	31.95	6.98	0.0948
Top 25%	Gender	319	39.59	33.59	6.00	0.0937
Top 50%		1676	38.18	31.84	6.34	0.0225
Top 50%	Baseline	407	41.61	33.74	7.87	0.3282
Top 50%	No Gender	759	36.54	30.90	5.64	0.2867
Top 50%	Gender	510	37.54	31.97	5.57	0.1876

a. T-Test for clustered data. Clusters: Schools.

4.3. School Data

In this experiment we had a two-way randomization. First, schools were randomly included in the experiment among the pool of schools located in the Italian North-East that comprehends four regions: Veneto, Emilia Romagna, Friuli-Venezia Giulia and Trentino-Alto Adige. Second, selected schools were randomly assigned to a treatment group. The fist randomization is necessary to avoid to include in the experiment "special" schools in order to be able to generalize findings that need not to be strictly linked to the selection procedure (i.e., to be externally valid). The second randomization is important to assure that schools in different groups are similar and single schools' characteristics (e.g., quality, reputation, teaching methods etc.) are not correlated with the type of treatment. We know that the randomization procedure allows us to meet the condition of exogeneity for our treatment variable. Indeed, this permits to investigate treatments' effect simply looking at summary statistics (in particular, means). Unfortunately, the randomization process needs to be verified because characteristics could not be successfully balanced across treatment groups making them not directly comparable. We exploit our data sources to investigate this balancing verifying that control variables effectively have not any statistical relationship with the treatment variable. Once verified the randomization procedure we will include these variables in regressions in order to control for differences across selected schools. In order to simplify our analysis we do this exercise only for schools included in the "matched" sample (i.e., schools in which both the questionnaire and the survey were submitted and in which there are matched students). Indeed, this is the sample we are interested in.

In Table 4 we report the distribution of schools among treatments divided by region. At the bottom of this table we report the p-value of the Fisher's Test. Under the null hypothesis there is no significant relationship between the treatment variable and the variable "region". We cannot reject this hypothesis, hence the 27 selected schools are well-randomized across regions.

Table 4. Distribution of Schools Among Treatments by Region

	Baseline	No gender	Gender	Total
Veneto	2	2	3	7
Emilia Romagna	1	5	3	9
Friuli-Venezia Giulia	4	4	1	9
Trentino - Alto Adige	0	0	2	2
Total	7	11	9	27
Fisher's exact = 0.322				

Before the competition took place we submitted a survey to one math professor (the one that was in charge of the organization of the within-school competition) for each involved school to collect some data about own school's characteristics. Most important for this analysis, we asked the overall number of students (asking to specify the distribution of pupils among genders) in the school and the way the math Olympiad participants were selected in the school. In order to understand whether schools are homogeneous across treatments we look at within group averages for the number of students. Restricting the analysis at the matched group we exploit the top panel of Table 12 (p. 51, Info on Schools). Differences between within-group averages are not statistically significant neither in the number of students nor in the number (and percentage) of females. For what concerns selection, as seen before, we have defined three criteria: free, suggested and mandatory participation. In order the treatment to have a positive effect on students' participation the first criterion is needed since the monetary incentive could not change students' choice in an environment in which they actually do not choose. Moreover, the selection criterion is potentially correlated with other schools' characteristics as prestige, quality, reputation etc. That is, it can capture other factors that influence also students' realized performance (not only competition entry). In Figure 2 we show the distribution of schools among selection criteria for each treatment group. We reported this distribution also in table form (see Table 18, p.54). In Table 18 the p-value of the Fisher's Test suggests that there is not significant relationship between the type of treatment and the selection criterion (although this comparison includes all three groups together). Also this characteristic results to be well balanced across treatment groups. Coming back to Figure 2 we can say something about this distribution. The first relevant fact that we want to underline is that

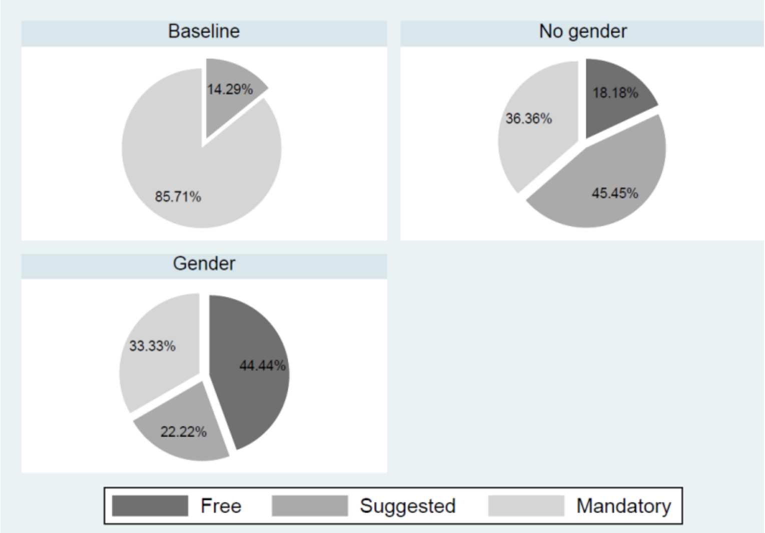

Figure 2. Distribution of Schools among Selection Criteria (by Treatment)

in the schools assigned to the Baseline group none of them allows students to freely enter the competition. On the contrary the vast majority of schools (6 out of 7) use mandatory participation. Second, the Gender group is the one that includes the highest number (44.44%) of school with the latter criterion. Also the No Gender group seems not to be well-balanced since the majority of schools chose the criterion "Suggested". Said this, we are interested in cross-groups balancing since our need is the comparison between groups (so, that the type of treatment is not correlated with the selection criteria). Even if the randomization worked in this case, these facts suggest that the distribution of schools among selection criteria could hardly compromise the cross-treatment-groups comparison. We will come back on this, exploiting evidence from the regression approach, in the last part of our analysis.

Within schools students attend different tracks (or, curricula). These tracks can differ in several characteristics. Some curricula are more math-intensive (e.g., Liceo Scientifico), others focus more on arts and language (e.g., Liceo Artistico, Liceo Classico, Liceo Linguistico etc.). Tracks differ also in the way they prepare students to university or to the labor markets (academic vs vocational tracks). Although schools are most often specialized in some curricula, they always include different tracks that share same school trends but could heavily differ in quality and already cited characteristics. For the scope of this analysis, the math intensity surely represent the most important feature. In our tables (1, 11, 12) we report for each sample the within-group averages of curricula present in involved schools. This says something about the type of school and the degree of specialization. Looking at Table 12 (p.51), we cannot reject the hypothesis that schools included in the matched sample are homogenous across treatment groups in the number of offered curricula (Kuskal-Wallis p-value: 0.2938). However, this evidence still holds when looking at the sample of schools in which the survey was submitted (Table 1, p.24) and schools for which we have data about Math Olympiad outcomes (Table 11 p.50). Not surprisingly since for the major part they include the same schools. Our main concern is that Olympiad participants (and then matched students) are equally distributed among academic curricula across different treatment groups. In Table 16 (p.53) we show this distribution. The Chi2 Test (with a p-value equal to 0.0000) shows that there is a significant relationship between tracks and treatments. Clearly, this is a relevant issue since having more students attending math-intensive courses in a particular group could hide or simply alter observed treatment effects. For example, females in the No Gender group could overperform that included in the gender-based treatment just because they belong to different tracks. Interpreting higher average performance in the first group as "female disliking single-gender competitions” would be wrong. In Table 17 (p.53) we show the gender composition of single tracks. Here we notice that the type of curriculum is significantly related (Chi2 Test p-value = 0.0000) to the variable “gender”. This means that some tracks are male-dominated (e.g. Liceo Scientifico – Scienze Applicate that offers a strong preparation in math), in other tracks the reverse holds (e.g. Liceo Classico). From the previous and the latter evidence, we deduct that in our regression approach we should control for track-specific effects including dummies for each curriculum.

4.3.1 Other Sources: Fondazione Agnelli

The difference across tracks can potentially alter our analysis. Including dummies for each track allows us to control for cross-treatment-group differences in the distribution of students among curricula. What remains uncontrolled for are single tracks’ characteristics. Same curricula offered in different schools share almost the same program but can still differ. Until now, we have not mentioned the most important factor that influence students’ readiness for math competitions and realized performance: school quality. We need a measure of quality since, as for already defined variables, the unbalanced distribution of students among high-quality schools is another potential threat for interpretation of results. In order to solve these two issues (difference in same curricula characteristics across schools and quality of education) we exploit data from the Fondazione Agnelli’s Dataset (that covers almost our entire “matched” sample).

Fondazione Agnelli (from now on, FGA) collects data about tracks within each single school. In particular, they have informations about past years tracks’ performance. We think that this approach is really effective since, as argued before, considering the entire school would cause altered evaluations due to strong differences across curricula offered in the same school. For these reasons, in this paragraph we will not refer to involved schools but to single tracks within these schools (in particular, schools included in the matched sample). From the FGA’ s dataset we first exploit data about past students’ matriculation (at University) rates represented as the percentage of students that successfully applied for any university after completed the high school. Data refer to within-track averages in the period 2014-2017. This permits to compare single tracks’ academic attitude and quality. This is relevant for two reasons. First, in our sample also vocational tracks are involved (e.g., Istituto Professionale). Second, matriculation rates are likely to strongly correlate with school quality. We have data about matriculation rates divided by gender, this allows to

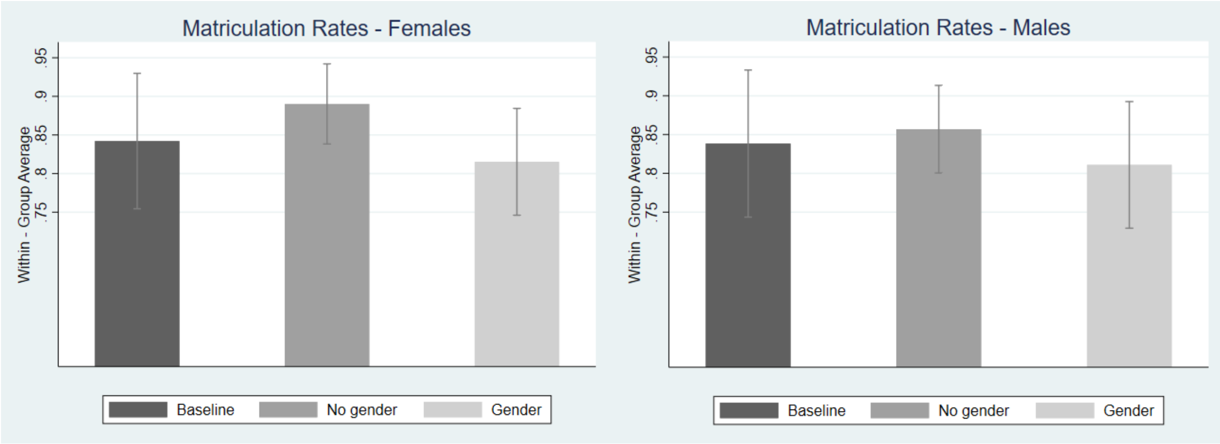

Figure 2. Within-groups average matriculation rates by gender. Unit: single track. Data source: FGA.

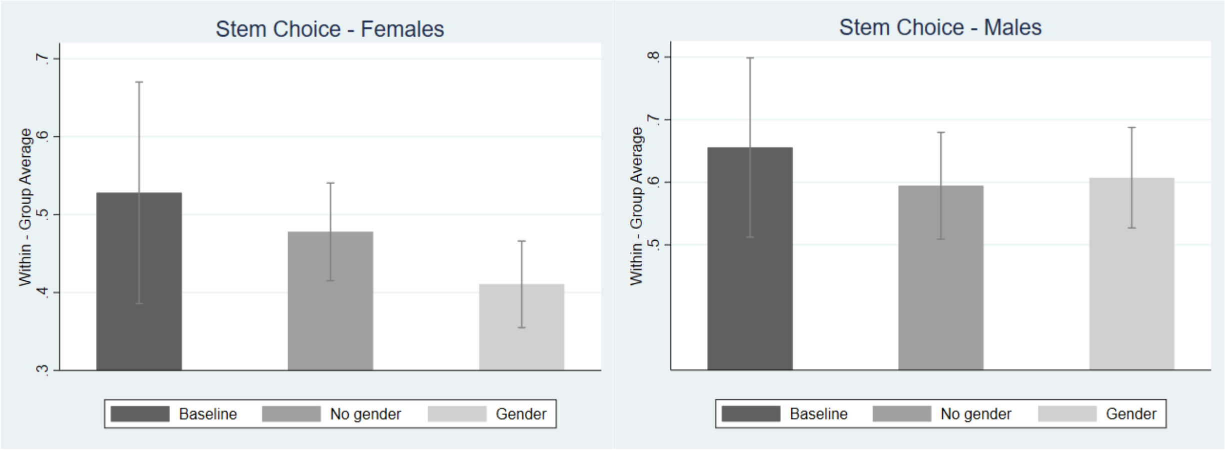

Figure 3. Within-groups average proportion of student choosing STEM careers by gender. Unit: single track. Data source: FGA.

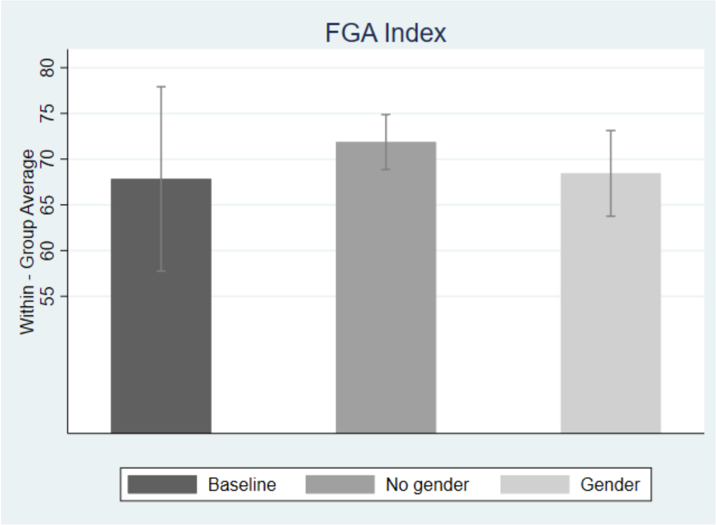

Figure 4. FGA Index - within group averages. Unit: single track. Data source: FGA.

evaluate more deeply tracks in our sample. Moreover, gender-specific data potentially capture other forces influencing students' attitude toward math competitions. In Figure 2 we show within-treatment-group averages with confidence interval bars that help to compare our three group. In this figure, as well as in the other similar ones that we are going to show, confidence interval bars take into account that groups of tracks belong to the same school (i.e., they control for intra-clusters correlation). We report these averages also in table form (Table 12, p. 51) and test the Kruskal-Wallis Test' null hypothesis. Within group averages are very similar and there is not significant difference across groups both in males' and females' matriculation rates (although slightly lower in the Gender group). In Figure 3 (and Table 12, p.51) we show the within-group average percentage of males and females that chose STEM (Science, Technology, Engineering and Mathematics) fields among those who successfully applied for any University. With this variable that we call "STEM choice" we also control for math intensity in academic programs of tracks involved in our sample. Looking at confidence interval bars and p-values in Table 12 we see that also in this case our three treatment groups are not significantly different. Further examining the reported figure we want to underline two facts. First, also in this case the Gender group has lower means, in particular for females. This is a potential threat since in the group in which we want to foster women's performance with the gender-based incentive scheme, the average number of females choosing the STEM field is lower (although the difference is not significant in the cross-groups comparison). Second, it is evident that, while the average matriculation rates in the previous figure (Figure 2) seem to be almost equal across genders, the percentage of males choosing STEM-based careers results to be significantly higher than that of females. Of course averages refer to tracks involved in our matched sample but this is certainly in line with evidence from the past literature described in Chapter 2. In order to have a synthetic measure of tracks' quality we finally exploit the FGA index. This index is computed by FGA that assigns a score from 0 to 100 to each single track based on students' (that got high school diploma in that track) performance in the University career in the period 2014-2017. Data about these performance include grades (weighted for the number of credits associated to respective courses) and obtained credits in the first year of course at University were these two informations have equal weights (50:50). In particular, the index[3] is a standardized measure that takes into account for heterogeneity across Universities and courses and for this reason it makes scores comparable across tracks.

[3] For further explanations see FGA's website (https://eduscopio.it/dati-e-metodologia)

We will see that the effect of schools' quality on students' participation could be either positive or negative. On the contrary, for what concerns the effect of tracks' quality on students' performance we certainly expect to find a positive effect just because pupils are better trained. In Figure 4 (and Table 12, p.51) we show averages within treatment groups and also in this case we find that, thanks to randomization, students are well distributed among schools with different FGA Index scores.

We want to clarify that shown averages do not exactly reflect those present in the FGA's dataset. First, because in this dataset data were not ready to use for our analysis' purpose. This is true both for matriculation rates and STEM choices. In the latter case we identified fields to include in the broader category "STEM". Adding up single careers data (e.g., math, statistics, engineering etc) we obtained the percentage of students choosing STEM fields (for single genders). Second, there is little difference in the way we have collected data. When collecting data about tracks we had a single item called "Istituto Tecnico" while they make a further distinction between "Istituto Tecnico – Economico" and "Istituto Tecnico – Tecnologico". For this reason, for schools in which both kind of curricula were included we simply computed means for all cited variables.

5. Regression Analysis

In this Chapter we deepen our analysis with a regression approach. In previous Sections we have first looked at realized outcomes in our experiment in terms of participation and performance. In particular, we have investigated these outcomes in several subgroups both examining cross-gender differences and analyzing how these results vary along the students' ability distribution. Most important, we have seen how participation and performance vary across treatment groups in order to understand whether incentive schemes have differently (or not) influenced students' outcomes. From this first look we have found evidence of lower participation for females in the Gender Group and of a positive gender gap in performance in favor of males, in particular among the pool of high-ability students. No difference in performance across treatment groups has emerged. The interpretation of these results hardly depends on the characteristics of considered subsamples. Looking at variable at our disposal we have seen that our three treatment groups differ in some factors, although the most significant difference is in the distribution of students among school curricula. This is a relevant difference since tracks differ in math intensity and other important factors. In a perfect randomization scenario it would not be necessary to further analyze outcomes since the treatment variable would be perfectly exogenous, samples would be balanced and the interpretation of results exploiting differences across means would be unbiased. Since randomized groups result to be not perfectly balanced and different in some characteristics we want now to check the robustness of previous results analyzing the effect of our treatments. In case the treatment variable is correlated with observed variables, the inclusion of the latters in our regressions allows to eliminate (or, at least, reduce) the bias coming from this correlation since the treatment variable would no longer be correlated with the error. We exploit this approach to investigate the effect of the two incentive schemes both on participation and on students' performance. The Baseline group permits to analyze the cited effects comparing the No Gender and the Gender group with a "no intervention" scenario represented from the control group itself.

5.1. Participation

In this Section we focus on the first outcome variable: participation. Here we want to understand whether incentive schemes have caused participation to vary across treatment groups. We can exploit already shown variables to check whether eventual differences in participation are driven by other factors (or schools' characteristics). In Table 5 we report results of this regression where

Table 5. Treatments' effect on the percentage of participants

	(1)	(2)	(3)	(4)
	OLS	OLS	OLS	OLS
Percentage of Participants in School [1]				
No Gender [2]	-0.101	0.0248	0.0195	0.0582
	(0.0671)	(0.0539)	(0.0511)	(0.0524)
Gender [2]	-0.197***	-0.0650	-0.0641	-0.0339
	(0.0408)	(0.0531)	(0.0464)	(0.0572)
FGA		-0.00346	-0.00448*	-0.00604**
		(0.00216)	(0.00206)	(0.00170)
Females in School		0.000163	0.000160	-0.0000286
		(0.000116)	(0.000128)	(0.0000997)
Students in School		-0.000136	-0.000112	-0.0000153
		(0.0000810)	(0.0000786)	(0.0000699)
Selection[3]: Suggested		0.0767	0.0831	0.0549
		(0.0654)	(0.0493)	(0.0348)
Selection[3]: Mandatory		0.213***	0.208***	0.216***
		(0.0413)	(0.0395)	(0.0324)
Constant	0.356***	0.388**	0.366*	0.503**
	(0.0303)	(0.118)	(0.177)	(0.161)
Track Dummies	No	Yes	Yes	Yes
Matriculation Rates [4]	No	No	Yes	Yes
Stem Choice [4]	No	No	Yes	Yes
Location [5]	No	No	No	Yes
Observations	10132	8447	8391	8391
R^2	0.263	0.653	0.684	0.774

1. The dependent variable is the percentage of students in the school participating at the Math Olympiad.
2. Base Category: Control group "Baseline".
3. Base Category : "Free";
4. Track Specific variables defined by gender (included both for males and females);
5. Includes regional dummies and the number of inhabitants of the city where the school is located.

Standard errors in parentheses. . Standard errors adjusted for 27 clusters in school.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

the dependent variable is the percentage of participants in the school. This is our main regression for the first outcome variable since we do not want to analyze the treatment effect on the absolute number of participants that does not take into account the different school size (in terms of number of students). As you can notice from the bottom panels we add in each column variables to control for any difference across treatment groups. Among the set of controls we include variables described in Chapter 4. We further add only the "number of inhabitants of the city where the school is located". This because we think that the variable "region" in our regression does not perfectly describe the location of involved schools. In the same region schools could be located either in big cities/centers or in small towns. In this sense, the location can, for example, differently influence school's quality or financial resources. The location can also capture the different attitude in the school toward students and teaching methods. Coming back to our regression, the observations are students coming from the 27 considered schools, and standard errors are adjusted for the intra-clusters correlation (that is, the regression takes into account that observations in our sample are divided in clusters). In this regression we exploit the whole sample at our disposal, including both students that filled the survey and students that enjoined the math competition. In the top panel we show the coefficients of most relevant variables. In the first specification (column 1) only the Gender treatment has a negative and significant effect. This effect is no longer significant when adding controls in our regression. Remember that the base category for the treatment variable is the Baseline group. Hence, the observed treatment effects are with respect to the no-intervention scenario. The FGA index, our proxy for school quality, has a negative and significant effect both in the third and the forth columns and is robust to the introduction of control variables. This means that the higher the quality of the school the lower is the percentage of participating students (with respect to the total number of students in the school). The negative effect of the FGA index could capture the different approach that high quality schools have with these extra-curricular competitions. For example, high quality schools could allow only best student to participate in order to obtain higher average performance in the competition. When adding controls in our regression we can notice that the "Selection Criteria" is the most relevant variable driving cross-schools (and, in turn, cross-groups) differences in the percentage participation. This, in pair with the not-significant coefficients of the treatment variable, suggests that realized differences do not depend on the introduction of the incentive schemes. Therefore, looking at the second outcome variable (performance) it is necessary to control both for participation and selection criteria when analyzing treatment effects on students' score at the Math Olympiad. The cross-groups difference in selection and participation could hardly bias interpretation of results and policy outcomes. At the end of this work (Appendix A, p.56) we study the same treatment effects using as dependent variable the individual participation. In particular, it is a dummy that takes 1 if the student enjoined the math competition, 0 otherwise. In the greater part of our sample students seem not to be free to choose whether to participate or not. In this scenario it is hard for monetary incentives to motivate students or change their participation choice. For this reason, we suggest that the interpretation of results from the probit model in Appendix A should be simply interpreted as the effect that treatments have on individual probability to enjoy the competition. Even if regression outcomes remains hard to interpret, we add this Appendix to show that also changing the dependent variable, the main factor influencing students' participation is the adopted selection criterion. From the same regression it emerges that there is a positive and significant gender gap in participation in favor of males. Also in this regression the coefficient of the FGA index variable remains negative and significant.

5.2. Treatments' Effect on Score

Interventions with monetary incentives in our experiment seem not to have significantly affected participation at the Math Olympiad, neither at school nor at individual level. As we have argued, the way in which participants are selected in involved schools indicates that there was little room for treatments to have any significant effect. In this paragraph we investigate the effect that the No Gender and the Gender treatment had on the second outcome variable: individual scores. We include in our regression model (OLS) all control variables at our disposal, as in the previous Section. In Table 6 we include in the analyzed sample the Baseline, the Gender and the No Gender groups, where the first one is the base category of the treatment variable. Since from now on we analyze individual scores, we include in the set of controls also dummies for the attended class (i.e., year of the course that goes from 1 to 5). It is important to consider the attended class for two reasons: first, the Math Olympiad test is different for the first two years (and normalizing individual

Table 6. Treatment Effect on Score

	(1)	(2)	(3)	(4)	(5)
Normalized Score	OLS	OLS	OLS	OLS	OLS
No gender	0.292	-3.381*	-3.733	-4.450*	-4.425*
	(2.080)	(1.511)	(1.835)	(1.703)	(1.714)
Gender	0.0820	-0.136	-0.783	-5.529*	-5.119**
	(2.455)	(1.520)	(1.499)	(2.228)	(1.831)
Female	-3.719***	-3.563***	-3.517***	-4.215***	-4.506***
	(0.807)	(0.580)	(0.570)	(0.915)	(0.954)
FGA		0.212*	0.248*	0.204*	0.275*
		(0.0883)	(0.109)	(0.0907)	(0.124)
Female # No gender				1.686	2.052
				(1.090)	(1.111)
Female # Gender				0.242	0.449
				(1.657)	(1.721)
Constant	32.04***	12.02*	12.09*	25.63***	29.12**
	(1.374)	(5.271)	(5.539)	(6.242)	(9.238)
Class Dummies [1]	Yes	Yes	Yes	Yes	Yes
Track Dummies	No	Yes	Yes	Yes	Yes
Location [2]	No	Yes	Yes	Yes	Yes
Info on School [3]	No	No	Yes	Yes	Yes
Participation [4]	No	No	No	Yes	Yes
Selection	No	No	No	Yes	Yes
Matriculation Rates [5]	No	No	No	No	Yes
Stem Choice [5]	No	No	No	No	Yes
Observations	6048	5094	5094	5094	5076
R^2	0.073	0.137	0.138	0.165	0.167

1. Dummies for Years of the course (1 to 5). ;
2. Includes regional dummies and the number of inhabitants of the city where the school is located.
3. Includes the number of females and the total number of students in school;
4. Includes the percentage of participants in school with respect to the total number of students and the percentage of females with respect to the total number of participants in school;
5. Track Specific variables defined by gender (included both for males and females);

Standard errors in parentheses. Standard errors adjusted for 27 clusters in school.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

scores could not be enough to make scores comparable across students attending different classes); second, the attended class influences participants' preparedness in math topics (it also includes the information "age" that we do not add in the set of controls to avoid collinearity). As underlined in the previous Section, we include in our controls both the number of participants in the school and the selection criterion adopted. We control also for the percentage of female participants since the selection criteria could have an heterogenous effects across genders (and simply because schools significantly differ in the number of female participants). In columns (4) and (5), where we control first for selection and participation and then for track-specific past performance, the coefficients of the treatment variable are negative and significant. This evidence suggests that monetary incentives had a negative impact on individuals performance. We will come back on this evidence in Section 5.3 questioning the effective comparability of the treatment groups with the control one. As evident from the Chapter 4, the gender gap in performance is negative and significant (coefficient of the dummy female). Being female has almost the same effect, in magnitude, of our interventions. The coefficient of the FGA Index, as expected, is positive: the higher the quality of the school the better are individual performance. As you can see in our Table, we have interacted the treatment variable with the dummy "female" to investigate any gender-specific effect of our treatments. In Chapter 2 we have shown that in the past literature there is evidence of heterogenous effect across genders of single-gender competitions. This gender-specific effect is captured from these interactions. Neither the No Gender nor the Gender treatment has a significant impact on females' Olympiad scores. This evidence is in line with the one found in Chapter 4 when looking at the cross-groups difference in mean performance.

The peculiarity of our intervention (a quota on the top) suggests that the treatments potentially have a significant effect on students in the upper tail of the ability distribution as they represent the portion of students that can more reasonably expect to get the compensation. In order to investigate these effects, in next regressions we separately compare first the No Gender and then the Gender group with the Baseline. In this regressions we include a dummy "Top10" that takes 1 if the student is in the top ten of the school-level ability distribution (based on the Raven Test Scores) among those who filled the survey. Exploiting this information we restrict the analysis to the "matched group". It is necessary because among students that did not fill the survey there could be potential top-ability pupils that we cannot compare due to this missing information. In Table 7 we report results of the comparison between the No Gender and the Baseline group. The dependent variable is "Normalized Score". We include the dummy "Top10" and let it to interact both with the dummy

Table 7. Treatment Effect (Comparison Baseline – No Gender) on Score of Top 10 at School Level

	(1)	(2)	(3)	(4)	(5)
Normalized Score	OLS	OLS	OLS	OLS	OLS
Female	-3.738*	-5.465*	-5.556*	-5.733*	-5.309*
	(1.681)	(2.009)	(2.199)	(2.212)	(2.177)
No Gender	0.125	-5.883	-5.387	-4.941	-2.392
	(1.969)	(2.926)	(2.709)	(3.017)	(2.135)
Female # No Gender	0.0603	3.905	4.078	4.244	3.716
	(1.895)	(2.130)	(2.332)	(2.327)	(2.268)
FGA		0.184*	0.230**	0.254	0.250
		(0.0833)	(0.0752)	(0.131)	(0.181)
Top10		12.67***	17.13***	16.98***	16.83***
		(1.655)	(1.629)	(1.586)	(1.663)
Female # Top10		-6.449**	-5.975	-5.639	-5.366
		(1.780)	(3.975)	(4.004)	(4.022)
No Gender # Top10			-6.989**	-6.941**	-6.947**
			(2.346)	(2.217)	(2.243)
Female # No Gender #			-0.0289	-0.532	-0.668
Top10			(4.478)	(4.417)	(4.441)
Constant	32.18***	17.25**	20.76***	34.64***	25.52
	(1.263)	(5.748)	(4.832)	(8.522)	(14.44)
Class Dummies [1]	Yes	Yes	Yes	Yes	Yes
Track Dummies	No	Yes	Yes	Yes	Yes
Location [2]	No	Yes	Yes	Yes	Yes
Info on School [3]	No	No	Yes	Yes	Yes
Participation [4]	No	No	No	Yes	Yes
Selection	No	No	No	Yes	Yes
Matriculation Rates [5]	No	No	No	No	Yes
Stem Choice [5]	No	No	No	No	Yes
Observations	4747	1666	1666	1666	1666
R^2	0.062	0.206	0.215	0.223	0.233

1. Dummies for each Year of the course (1 to 5);
2. Includes regional dummies and the number of inhabitants of the city where the school is located.
3. Includes the number of females and the total number of students in school;
4. Includes the percentage of participants in school with respect to the total number of students and the percentage of females with respect to the total number of participants in school;
5. Track Specific variables defined by gender (included both for males and females);

Standard errors in parentheses. Standard errors adjusted for 18 clusters in school.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

"female" and the treatment variable (where we consider only the treatment "No Gender" with a compensation scheme that does not take into account ranked students' gender). We finally include the triple interaction between the female dummy, the treatment variable and the dummy "Top10". The dummy "female" still has a negative and significant coefficient in each column confirming the gender gap in performance, even if it is not specific of the upper tail in the ability distribution (coefficient of the interaction "Female # Top10" not significant). Controlling for heterogenous effects along the ability distribution the coefficient of the treatment variable is no longer significant, suggesting that the treatment did not influence all students' scores. As expected the dummy "Top10" has a positive and significant effect: being an high-ability students strongly influences realized score in the math competition. This evidence proves that the Raven Test score, also in our sample, is a good proxy for students' abilities. The No Gender treatment has a significant and negative effect on high-ability individuals' scores. This means that the effect that we found in the previous regression is mainly located at the top of the ability distribution. Indeed, coefficients of the interaction "No Gender # Top10" are significant in each column. The triple interaction allows to understand whether the interaction between the first two variables (Female and No Gender), hence the gender-specific treatment effect, is present among high-skilled students (for who the dummy Top10 is equal to 1). The coefficient for this interaction is never significant in different specifications (that the differ in the number of control variables included). We do the same exercise in Table 19 (p.54) looking at the top 25 of the within-school ability distribution. We find almost the same evidence. A positive and significant coefficient of the dummy "Top25" in column (3), (4) and (5). Hence, being in the Top25 of the ability distribution as a positive effect on own performance. The negative and significant coefficient of the interaction "Female # Top25", in pair with a lower coefficient for the dummy "Female" with respect to the previous table, suggests that the gender gap is even wider among students in the top 25. This is the main difference with the Table 7 since the treatment still has no effect but at the top of the ability distribution ("No Gender # Top25"). There is not significant gender-specific treatment effect neither in the whole sample (Female # No Gender) nor among top students (Female # No Gender # Top25).

In Table 8 we compare the Baseline and the Gender group adding, as for the No Gender group in previous analysis, specific effects for students in the Top10. The Gender treatment is the one in which we expected the most significant gender-specific effect on females' performance, in line with the evidence from literature that single-sex contests lower the distortion of competition on women's outcomes. We have to underline that in this case females are still competing against men, what is gender specific are monetary incentives for top-ranked students (incentive scheme explained in Section 3.). Evidence from this regression is in line with the one found in the No Gender treatment. In this subsample the gender gap is no longer significant when including in the regression the whole set of controls (column 5). Neither the gap is significant among high-ability student (Female # Top10). Monetary incentives seem to have also in this case negative and significant effect on students' performance, in particular when controlling for participation and selection (column 4). The negative effect is stronger for high-skilled students (Gender # Top10). Also in this case the effect is never gender specific along the ability distribution (see Female # Gender and Female # Gender # Top10), although the coefficient of the triple interaction is notable in magnitude. Also for the Gender group we have reported in the Section "Tables" (p.50) the same regression that considers students in the top 25. We found the same evidence as Table 8, the only

Table 8. Treatment Effect (Comparison Baseline - Gender) on Score of Top 10 at School Level

	(1)	(2)	(3)	(4)	(5)
Normalized Score	OLS	OLS	OLS	OLS	OLS
Female	-3.676*	-5.899*	-5.118*	-4.698*	-4.245
	(1.680)	(2.068)	(2.300)	(2.175)	(2.208)
Gender	-0.284	-1.997	0.0836	-7.217*	-8.796**
	(2.520)	(2.604)	(2.508)	(2.716)	(2.219)
Female # Gender	0.0767	0.753	-0.443	-0.758	-1.266
	(2.097)	(2.448)	(2.959)	(2.856)	(2.869)
FGA		0.140	0.224**	-0.0717	-0.0944
		(0.0744)	(0.0594)	(0.153)	(0.148)
Top10		11.57***	16.82***	16.73***	16.53***
		(1.653)	(1.617)	(1.627)	(1.632)
Female # Top10		-2.209	-6.578	-6.074	-6.055
		(2.576)	(4.031)	(4.120)	(4.026)
Gender # Top10			-9.039***	-9.140***	-9.345***
			(1.810)	(1.783)	(1.841)
Female # Gender #			7.223	6.596	7.139
Top10			(4.915)	(4.821)	(4.708)
Constant	31.23***	19.12***	17.42***	66.46**	37.60
	(1.259)	(4.491)	(3.927)	(18.26)	(21.44)
Class Dummies [1]	Yes	Yes	Yes	Yes	Yes
Track Dummies	No	Yes	Yes	Yes	Yes
Location [2]	No	Yes	Yes	Yes	Yes
Info on School [3]	No	No	Yes	Yes	Yes
Participation [4]	No	No	No	Yes	Yes
Selection	No	No	No	Yes	Yes
Matriculation Rates [5]	No	No	No	No	Yes
Stem Choice [5]	No	No	No	No	Yes
Observations	3644	1173	1173	1173	1173
R^2	0.076	0.277	0.287	0.293	0.302

1. Dummies for each Year of the course (1 to 5);
2. Includes regional dummies and the number of inhabitants of the city where the school is located.
3. Includes the number of females and the total number of students in school;
4. Includes the percentage of participants in school with respect to the total number of students and the percentage of females with respect to the total number of participants in school;
5. Track Specific variables defined by gender (included both for males and females);

Standard errors in parentheses. Standard errors adjusted for 15 clusters in school.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

difference is the significant gender gap among high-skilled students, the same we found when considering top 10 and top 25 students in the No Gender group.

5.3. Control the Control Group

The evidence that we found investigating the effect of monetary incentives on students' outcomes at the Math Olympiad is in some way counterintuitive. Indeed, prizes seem to negatively affect both males' and females' performance with no gender-specific effects. This evidence holds when comparing the Baseline group both with the No Gender and the Gender group. The fact that these two treatments, even if different in the compensation scheme, have the same significant effect (equal across genders and more relevant for top students when considering either the top 10% or the top 25% of the ability distribution) suggests that students may dislike monetary prizes. One possible explanation could be that teenagers feel high pressure when competing for these prizes reacting with lower performance. Still this explanation results to be counterintuitive. In this Section we provide another explanation for this finding.

Until now we have compared together all three groups to understand whether they are homogenous in terms of characteristics. We have noticed that they are not perfectly balanced. Thanks to the regression approach we have controlled for differences across groups. Here we argue that the regression approach, even if rich in terms of control variables, could not efficiently investigate treatment effects when groups are not compatible. In our dataset we have a large set of variables, here we report the ones that, in our opinion, cause the argued incompatibility across groups. In Chapter 4 we used the Kruskal Wallis Test to test the hypothesis that all three groups are, at the same time, not significantly different in means (of considered variables). The main drawback of this type of randomization check is that it says nothing about the pair comparison across subsamples. In Table 9 below we report the pair comparison (for example B-NG compares means of the No Gender and the Baseline group) with a T-Test where the null hypothesis is that means are equal across groups (it is a clustered T-Test in the bottom panel were we compare individual data, the clusters are the schools). In the top Panel it is clear that the Gender group and the Baseline have significantly different participation rates. This is true for the subsample of females (p = 0.0036), of males (p = 0.0067) and entire schools (p = 0.0009). In Section 5.1 we have shown that treatments had no room to influence participation both at school and individual level and that the selection criterion adopted within schools is the most relevant factor determining participation at the Math Olympiad, at least in our sample. In Figure 2 (p. 31) you can notice that more than the 85% of schools included in the control group use the “mandatory” criterion. In none of them students are free to choose whether to enter the competition. This makes impossible the comparison, for example, across groups in which participation is always free. As we have argued, selection criteria can capture other school characteristics other than define the quality of students participating at the competition. Not surprisingly, the noted different in the distribution of schools among selection criteria here reflects (in Table 9) in a significantly different participation across groups, being much higher in the Baseline. In particular, the Gender group is very different from the control one. Also the No Gender group results to be different from the Baseline when considering within group averages of participation rates both for females (p = 0.0266) and the entire schools (p = 0.0009). While the Baseline group seems to be the special one, in the last column we

Table 9 . Comparison of samples

	Baseline	No gender	Gender	T-Test [a] (p-value) B–NG	T-Test [a] (p-value) B-G	T-Test [a] (p-value) NG-G
SAMPLE WITH SCORES [c] - PARTICIPATION [e]						
Average number of female participants in school	145 (48%)	98 (43%)	57 (41%)	0.2562	0.0175	0.1549
Average number of participants in school	316	219	144	0.2652	0.0183	0.2370
Female participants vs Females in school	34%	19%	14%	0.0266	0.0036	0.3370
Male participants vs Males in school	40%	28%	22%	0.2297	0.0067	0.5065
Participants in school vs Students in school	36%	22%	17%	0.0495	0.0009	0.4272
MATCHED DATA WITH SURVEY [d]						
TOP [b] 10						
Olympiad Score (Average)	46.92	36.64	40.23	0.0411	0.1685	0.3010
Olympiad Score Females (Average)	41.30	32.33	37.54	0.0572	0.4674	0.0849
Olympiad Score Males (Average)	52.64	41.96	42.21	0.0368	0.0381	0.9506

B : Baseline, NG : No Gender, G: Gender.

a. When comparing Olympiad Score it is a T-Test for clustered data. Clusters: Schools.
b. Top students are defined based on their Raven Test Score in the Survey. In particular they belong to the top X if their score is above the top X threshold (100 - X percentile) of the score distribution at school level. The distribution includes all scores of students that filled out the questionnaire.
c. The subsample includes all participants at the Math Olympiad.
d. The subsample includes Olympiad participants that filled out the questionnaire. The potential respondents were randomly selected
e. Observations are single schools.

can notice that the two treatment groups with monetary incentives are very similar and within group averages are never significantly different for every considered variable (or subsample). The control group is, in our opinion, not comparable to the others because of the missing variability of selection criteria in favor of the mandatory participation. Selection criteria have an effect on performance per se, but influence also participation at school and individual level.

5.4. The “Gender” Effect

In this Section we compare the two homogenous groups: No Gender and Gender group. Indeed, for reasons explained in the previous Section, we think it is reasonable to take away the control group that seems to be very different from these two. Here we want to investigate whether the single-sex competition for monetary compensations can foster students’ and, in particular, females’ performance. Also in this case we exploit the set of control variables seen in previous regressions. In Table 10 we report the regression outcomes in which we include five regressions that differ in the number of controls. Examining the effect of splitting prizes among genders, we want still to understand whether there was an heterogenous effect along the ability distribution and then include the dummy “Top10” letting it to interact both with the treatment variable and the dummy “female”. Finally, we include the triple interaction to examine whether there is a gender-specific treatment

Table 10. Treatment Effect (Comparison No Gender - Gender) on Score of Top 10 at School Level

	(1)	(2)	(3)	(4)	(5)
Normalized Score	OLS	OLS	OLS	OLS	OLS
Female	-3.711***	-2.354*	-1.944*	-1.871*	-2.029*
	(0.899)	(0.887)	(0.810)	(0.811)	(0.842)
Gender	-0.225	4.992***	5.508***	3.925	3.177*
	(2.536)	(1.125)	(1.191)	(2.160)	(1.151)
Female # Gender	-0.150	-1.692	-2.955	-3.035	-2.960
	(1.560)	(1.715)	(1.883)	(1.938)	(1.977)
FGA		0.259*	0.285*	0.281**	0.125
		(0.0906)	(0.102)	(0.0936)	(0.0885)
Top10		9.861***	10.77***	10.43***	9.973***
		(1.175)	(1.820)	(1.653)	(1.534)
Female # Top10		-4.317*	-6.538**	-6.436**	-5.941**
		(1.710)	(2.152)	(1.903)	(1.874)
Gender # Top10			-1.954	-2.311	-2.108
			(2.157)	(2.056)	(1.948)
Female # Gender # Top10			5.887	6.372*	6.335*
			(3.280)	(2.990)	(3.006)
Constant	33.50***	6.546	5.684	9.530	3.344
	(1.731)	(6.106)	(6.508)	(8.626)	(5.742)
Class Dummies [1]	Yes	Yes	Yes	Yes	Yes
Track Dummies	No	Yes	Yes	Yes	Yes
Location [2]	No	Yes	Yes	Yes	Yes
Info on School [3]	No	No	Yes	Yes	Yes
Participation [4]	No	No	No	Yes	Yes
Selection	No	No	No	Yes	Yes
Matriculation Rates [5]	No	No	No	No	Yes
Stem Choice [5]	No	No	No	No	Yes
Observations	3705	1763	1763	1763	1763
R^2	0.092	0.207	0.209	0.223	0.230

1. Dummies for each Year of the course (1 to 5);
2. Includes regional dummies and the number of inhabitants of the city where the school is located.
3. Includes the number of females and the total number of students in school;
4. Includes the percentage of participants in school with respect to the total number of students and the percentage of females with respect to the total number of participants in school;
5. Track Specific variables defined by gender (included both for males and females);

Standard errors in parentheses. Standard errors adjusted for 19 clusters in school.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

effect among top-skilled students. We here show only the regressions that investigate the treatment effect on top ten students' performance, since there is not remarkable evidence for the top 25. Still, we control for participation and selection criteria. The "Female" dummy's coefficient is negative and significant in each regression, showing also in this case that there is a strong realized gender gap in performance. The Gender treatment seems to have a positive and significant effect on performance (columns (2), (3) and (5)) although it has no gender specific effect on females (see interaction "Female # Gender"). The standardized measure of school quality (FGA index) has not significant effect when including all controls and, however, the effect is weak in magnitude in all columns. Differently from the regressions in previous paragraphs but in line with the reported literature, the gender gap in performance is stronger and significant at the top of the ability distribution (Female # Top10). The positive effect of the Gender treatment is not specifically present among top students (coefficients of the interaction “Gender # Top10” are never significant). The main interaction (triple interaction “Female # Gender # Top10”) allows to understand whether our intervention had the desired effect in narrowing the gender gap in performance fostering high ability females’ relative performance. As noted above, the Gender treatment had an overall positive effect on students’ performance with no gender-specific effect. The latter effect is significant and robust to the introduction of all controls. This may suggest that all students may prefer to perform against same-gender competitors. However, when looking at top students we notice that among these participants the effect is heterogenous across genders. The triple interaction has positive and significant coefficients in column (4) and (5). The gender-based reward has a significant effect in narrowing the gender gap in performance among students in the top ten improving females’ performance. This is clear also from the fact that the interaction Female # Top10 is lower when introducing together the triple interaction and controls for selection and participation. Taking away the control group, that we have shown to be not comparable to the other ones, we finally find evidence of a significant effect of our policy intervention, that efficiently improved women’s performance with gender-based rewards exploiting individual preferences toward single-sex competitions.

6. Final Discussion

We have analyzed the effect of two policy interventions in a field experiment. In randomly selected schools we introduced monetary compensations for top-ranked students in the Math Olympiad's competition. The purpose of these two interventions was to analyze the effect that different treatments can have in fostering females' participation and improving their performance in a stereotyped-male task trying to manipulate the distortion of competition widely discussed in past literature. In particular, the No Gender treatment allowed us to investigate and isolate the effect of monetary incentives on students' outcomes, while the Gender treatment was intended to examine whether gender-based rewards could have heterogenous effects across genders giving the opportunity to women to narrow or close the gap with their male peers. We also investigated whether these interventions had a remarkable effect on high-ability students since top-performing females seem to be those who mostly fail to enter and well-perform in competitive contexts. The latter evidence clearly is an economic (and equity) matter. Exploiting data about a subsample where there was no policy intervention, we had the possibility to examine policy effects comparing treatment groups with an "absent the action" scenario. Monetary prizes had no effects on participation neither at school nor at individual level. Unfortunately the participation was mostly driven by other factors, giving small room to incentive schemes to change students' entry decision. Our data suggest that when looking at individual scores, ceteris paribus, monetary incentives had a negative and significant effect on students' performance with no gender-specific effect. In Section 5.3 we have argued that these results can be hardly driven by the control group being not comparable with the treatment ones. The missing variability in selection criteria and the scrupulous selection of participants in schools randomly included in the Baseline group make this subsample very special. Even without isolating the effect of monetary incentives per se we compared the No Gender and the Gender group to understand whether the introduction of a single-sex competition for monetary prizes had any remarkable effect on females' relative performance. The Gender treatment had the desired effect in particular on performance of females in the top ten of the school-level ability distributions. This findings are in line with the literature showing that females may prefer same-gender competitions. We think that results from this experiment, due to the peculiarity of the designed treatments, give us other suggestions. Monetary incentives were different across subsamples. In the No Gender group students had to overperform the entire pool of participants in order to get the reward. In the Gender group, instead, participants had to overperform only peers of the same gender. Even if different in rewards, both groups shared the fact that, regardless of the compensation, the math Olympiad remained a mixed-gender competition. In the real world it is hard to imagine quotas in some contexts (e.g., entry exams, recruiting) and making females to compete only against individuals of the same gender could be not applicable. We suggest that gender-based rewards (even if nonmonetary) could potentially foster women’s performance also in mixed gender competitions, helping them to reduce the gap with their male peers.

Tables

Table 11. Variables by Treatment

	Baseline	No gender	Gender	K-Wallis [c] (p-value)
SAMPLE WITH SURVEY [d]				
INFO ON SCHOOLS [e]				
Number of participating schools	8	11	10	
Average number of females in school	495 (53%)	547 (51%)	464 (51%)	0.8680
Average number of students in school	922	1060	935	0.7484
Average number of tracks per school	3.87	3.73	4.2	0.6587
Track Specific [b] :				
Females' Matriculation Rate (University)	86%	89%	82%	0.0836
Males' Matriculation Rate (University)	85%	86%	81%	0.5873
Females choosing STEM fields	48%	49%	41%	0.2805
Males choosing STEM fields	64%	59%	61%	0.9786
FGA Index	68.44	71.43	68.44	0.2765
INFO ON STUDENTS				
ENTIRE SAMPLE [d]				
Number of participating students	1597	2369	2864	
Males	709 (44.4%)	1007 (42.5%)	1463 (51.1%)	
Females	888 (55.6%)	1362 (57.5%)	1401 (48.9%)	
Raven Test Score (Average)	4.99 (1.66)	4.89 (1.75)	4.80 (1.73)	0.6719
Raven Test Score Females (Average)	4.86 (1.66)	4.92 (1.70)	4.80 (1.70)	0.8457
Raven Test Score Males (Average)	5.17 (1.65)	4.84 (1.82)	4.80 (1.76)	0.1862
TOP [a] 10				
Number of students in the Top10	303	380	502	
Males	147 (48.5%)	159 (41.8%)	271 (54%)	
Females	156 (51.5%)	221 (58.2%)	231 (46%)	
Raven Test Score (Average)	7.31 (0.65)	7.24 (0.87)	7.38 (0.63)	0.6884
Raven Test Score Females (Average)	7.24 (0.64)	7.14 (0.84)	7.39 (0.64)	0.3184
Raven Test Score Males (Average)	7.39 (0.66)	7.39 (0.89)	7.38 (0.62)	0.9951
TOP [a] 25				
Number of students in the Top25	654	806	1167	
Males	310 (47.4%)	330 (40.9%)	601 (51.5%)	
Females	344 (52.6%)	476 (59.1%)	566 (48.5%)	
Raven Test Score (Average)	6.58 (0.86)	6.63 (1.00)	6.38 (1.07)	0.6798
Raven Test Score Females (Average)	6.50 (0.87)	6.56 (0.96)	6.36 (1.05)	0.7712
Raven Test Score Males (Average)	6.67 (0.84)	6.73 (1.06)	6.41 (1.09)	0.5137
TOP [a] 50				
Number of students in the Top50	1039	1463	1850	
Males	477 (45.9%)	590 (40.3%)	928 (50.2%)	
Females	562 (54.1%)	873 (59.7%)	922 (49.8%)	
Raven Test Score (Average)	5.95 (1.08)	5.91 (1.20)	5.75 (1.24)	0.6936
Raven Test Score Females (Average)	5.85 (1.09)	5.85 (1.18)	5.71 (1.22)	0.8048
Raven Test Score Males (Average)	6.08 (1.07)	6.01 (1.23)	5.79 (1.27)	0.4830

Percentages and standard deviations in parentheses.

a. Top students are defined based on their Raven Test Score in the Survey. In particular they belong to the top X if their score is above the top X threshold (100 - X percentile) of the score distribution at school level. The distribution includes all scores of students that filled out the questionnaire.

b. These data refer to single tracks within each school and come from Fondazione Agnelli. Data are not available for every track but almost the entire sample is covered.

c. The null hypothesis is that means are equal across samples (i.e. samples belong to the same population).When comparing Track-Specific data and Raven Test Scores it is a Kruskal-Wallis test for clustered data. Clusters: Schools.

d. The sample includes all individuals that filled out the questionnaire irrespectively from participation in the Olympiad.

e. Observations are single schools and single tracks both for averages and for tests

Table 12. Variables by Treatment

	Baseline	No gender	Gender	K-Wallis [e] (p-value)
MATCHED DATA WITH SURVEY [d]				
INFO ON SCHOOLS				
Number of participating schools	7	11	9	
Average number of females in school	495 (52%)	547 (51%)	464 (51%)	0.8576
Average number of students in school	938	1060	935	0.8132
Average number of tracks per school	3.57	3.73	4.4	0.2938
Track Specific [c]:				
Females' Matriculation Rate (University)	84%	89%	82%	0.0783
Males' Matriculation Rate (University)	84%	86%	81%	0.7109
Females choosing STEM fields	53%	48%	41%	0.1860
Males choosing STEM fields	66%	59%	61%	0.9104
FGA Index	67.84	71.87	68.44	0.3199
INFO ON STUDENTS				
Total Matched Individuals	621	1170	757	
Males	309 (49.7%)	526 (45%)	433 (57.2%)	
Females	312 (50.3%)	644 (55%)	324 (42.8%)	
Olympiads selection criteria [a]:				
Free	0 (0 %)	52 (4.4%)	232 (30.6%)	
Suggested	22 (3.5%)	241 (20.6%)	134 (17.7%)	
Mandatory	599 (96.5%)	877 (75%)	391 (51.7%)	
Raven Test Score (Average)	5.01 (1.62)	5.30 (1.68)	5.11 (1.71)	0.2558
Raven Test Score Females (Average)	4.90 (1.62)	5.35 (1.58)	5.14 (1.68)	0.1178
Raven Test Score Males (Average)	5.13 (1.61)	5.23 (1.79)	5.08 (1.72)	0.8130
Olympiad Score (Average)	34.95 (18.02)	31.57 (14.20)	33.22 (15.21)	0.6518
Olympiad Score Females (Average)	31.07 (15.78)	29.75 (12.74)	30.58 (13.39)	0.9695
Olympiad Score Males (Average)	38.86 (19.28)	33.80 (15.53)	35.19 (16.18)	0.4494
TOP [b] 10				
Number of students in the Top10	113	219	158	
Males	56 (49.6%)	98 (44.7%)	91 (57.6%)	
Females	57 (50.4%)	121 (55.3%)	67 (42.4%)	
Raven Test Score (Average)	7.34 (0.65)	7.53 (0.77)	7.51 (0.65)	0.5263
Raven Test Score Females (Average)	7.18 (0.57)	7.43 (0.72)	7.55 (0.68)	0.1228
Raven Test Score Males (Average)	7.50 (0.69)	7.64 (0.83)	7.48 (0.62)	0.7529
Olympiad Score (Average)	46.92 (20.86)	36.64 (15.57)	40.23 (16.33)	0.1654
Olympiad Score Females (Average)	41.30 (19.41)	32.33 (13.53)	37.54 (15.41)	0.1502
Olympiad Score Males (Average)	52.64 (20.89)	41.96 (16.32)	42.21 (16.78)	0.2141
TOP [b] 25				
Number of students in the Top25	247	420	315	
Males	126 (51%)	180 (42.9%)	184 (57.7%)	
Females	121 (49%)	240 (57.1%)	135 (42.3%)	
Raven Test Score (Average)	6.60 (0.86)	6.99 (0.90)	6.69 (0.96)	0.1862
Raven Test Score Females (Average)	6.51 (0.83)	6.88 (0.86)	6.72 (0.98)	0.3521
Raven Test Score Males (Average)	6.68 (0.89)	7.12 (0.93)	6.67 (0.94)	0.0775
Olympiad Score (Average)	40.98 (20.54)	34.94 (14.76)	37.05 (15.92)	0.4202
Olympiad Score Females (Average)	35.96 (18.19)	31.95 (12.72)	33.59 (14.27)	0.6970
Olympiad Score Males (Average)	45.81 (21.55)	38.93 (16.30)	39.59 (16.61)	0.3848
TOP [b] 50				
Number of students in the Top50	407	759	519	
Males	207 (50.9%)	322 (42.4%)	282 (54.3%)	
Females	200 (49.1%)	437 (57.6%)	228 (45.7%)	
Raven Test Score (Average)	5.94 (1.08)	6.26 (1.12)	6.01 (1.18)	0.1337
Raven Test Score Females (Average)	5.85 (1.07)	6.19 (1.08)	5.95 (1.22)	0.1291
Raven Test Score Males (Average)	6.02 (1.08)	6.35 (1.18)	6.06 (1.15)	0.1904
Olympiad Score (Average)	37.74 (18.93)	33.29 (14.65)	35.05 (15.54)	0.5405
Olympiad Score Females (Average)	33.74 (16.58)	30.90 (13.24)	31.97 (13.48)	0.8205
Olympiad Score Males (Average)	41.61 (20.25)	36.54 (15.83)	37.54 (16.63)	0.5553

Percentages and standard deviations in parentheses.

a. It is the criterion that professors in the school use to select Olympiads participants. The participation can be on a voluntary basis (i.e.

"Free"). On the other hand participants can be chosen by professors ("Suggested") that invite best students to participate, or it can be that some classes or best students are forced to participate based on merits or other characteristics ("Mandatory"). Also these two criteria in some cases could imply that not invited or forced students can join the competition but it is likely to be rare.

b. Top students are defined based on their Raven Test Score in the Survey. In particular they belong to the top X if their score is above the top X threshold (100 - X percentile) of the score distribution at school level. The distribution includes all scores of students that filled out the questionnaire.

c. These data refer to single tracks within each school and come from Fondazione Agnelli. Data are not available for every track but almost the entire sample is covered. Numbers are within sample averages.

d. The subsample includes Olympiad participants that filled out the questionnaire. The potential respondents were randomly selected.

e. The null hypothesis is that means are equal across samples (i.e. samples belong to the same population. When comparing Track-Specific data, Raven Test and Olympiad Scores it is a Kruskal-Wallis test for clustered data. Clusters: Schools.

Table 13. Difference in Olympiad Scores – Sample with scores

Variable	Sample	Subsample	N° Obs: Not Matched	N° Obs: Matched	Average: Not Matched	Average: Matched	P-Value (T-Test [a])
Score [c]	Entire [b]		3512	2548	31.70	32.88	0.5994
Score [c]	Entire [b]	Females	1384	1280	30	30.28	0.9010
Score [c]	Entire [b]	Males	2128	1268	32.79	35.51	0.3011
Score [c]	Baseline		1723	621	31.66	34.95	0.4708
Score [c]	Baseline	Females	751	312	30.37	31.07	0.8841
Score [c]	Baseline	Males	972	309	32.66	38.86	0.2890
Score [c]	No Gender		1234	1170	32.38	31.57	0.8503
Score [c]	No Gender	Females	430	644	29.98	29.75	0.9491
Score [c]	No Gender	Males	804	526	33.66	33.80	0.9788
Score [c]	Gender		545	757	30.22	33.22	0.4574
Score [c]	Gender	Females	196	324	28.55	30.58	0.6913
Score [c]	Gender	Males	349	433	31.15	35.20	0.4852

a. T-Test for clustered data. Clusters: Schools.

b. Includes the whole sample with Scores independently on the treatment.

c. Normalized score.

Table 14. Difference in Raven Test Scores – Sample with survey

Variable	Sample	Subsample	N° Obs: Not Matched	N° Obs: Matched	Average: Not Matched	Average: Matched	P-Value (T-Test [a])
Raven Score	Entire [b]		4282	2584	4.70	5.17	0.0459
Raven Score	Entire [b]	Females	2371	1280	4.68	5.20	0.0245
Raven Score	Entire [b]	Males	1911	1268	4.72	5.15	0.0547
Raven Score	Top 10%		695	490	7.21	7.48	0.1903
Raven Score	Top 10%	Females	363	245	7.16	7.40	0.2439
Raven Score	Top 10%	Males	332	245	7.26	7.55	0.1769
Raven Score	Top 25%		1641	986	6.34	6.80	0.0813
Raven Score	Top 25%	Females	890	496	6.30	6.75	0.0849
Raven Score	Top 25%	Males	751	490	6.38	6.84	0.0842
Raven Score	Top 50%		2676	1676	5.70	6.10	0.0743
Raven Score	Top 50%	Females	1492	865	5.65	6.05	0.0946
Raven Score	Top 50%	Males	1184	811	5.76	6.17	0.1169

a. T-Test for clustered data. Clusters: Schools.

b. Includes the whole sample with the Raven Test Score (the sample of individuals that filled out the questionnaire).

Table 15. Distribution of Students Among Treatments by Region

	Baseline	No gender	Gender	Total
Veneto	825 (35.2%)	1066 (44.3%)	474 (36,4%)	2365
Emilia Romagna	491 (21%)	859 (35.7%)	429 (33%)	1779
Friuli-Venezia Giulia	1028 (43.8%)	479 (20%)	154 (11.8%)	1661
Trentino - Alto Adige	0	0	245 (18.8%)	245
Total	2344	2404	1302	6050

Table 16. Distribution of Participants Among Treatments by Track

	Baseline	No gender	Gender	Total
Altro	48	0	0	48
Istituto Professionale	0	15	0	15
Istituto Tecnico	244	184	255	683
Liceo Artistico	0	0	10	10
Liceo Classico	72	205	45	322
Liceo Linguistico	2	23	26	51
Liceo Musicale	109	0	0	109
Liceo Scientifico	619	1296	360	2275
Liceo Scientifico - Indirizzo Sportivo	334	0	25	359
Liceo Scientifico - Scienze Applicate	607	626	546	1779
Liceo Scienze Umane	212	34	29	275
Liceo Scienze Umane - Economico Sociale	71	21	6	98
Total	2318	2404	1302	6024
Pearson P = 0.000				

Table 17. Distribution of Participants Among Gender Types by Track

	Male	Female	Total
Altro	29	19	48
Istituto Professionale	7	8	15
Istituto Tecnico	518	165	683
Liceo Artistico	5	5	10
Liceo Classico	85	237	322
Liceo Linguistico	12	39	51
Liceo Musicale	64	45	109
Liceo Scientifico	1136	1139	2275
Liceo Scientifico - Indirizzo Sportivo	221	138	359
Liceo Scientifico - Scienze Applicate	1234	545	1779
Liceo Scienze Umane	28	247	275
Liceo Scienze Umane - Economico Sociale	39	59	98
Total	3378	2646	6024
Pearson P = 0.000			

Table 18. Distribution of Schools Among Selection Criteria - Treatment

	Baseline	No gender	Gender	Total
Free	0	2	4	6
Suggested	1	5	2	8
Mandatory	6	4	3	13
Total	7	11	9	27
Fisher's exact = 0.134				

Table 19. Treatment Effect (Comparison Baseline – No Gender) on Score of Top 25 at School Level

Normalized Score	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) OLS
Female	-3.738*	-4.479*	-4.111*	-4.427*	-4.235*
	(1.681)	(1.863)	(1.796)	(1.792)	(1.742)
No Gender	0.125	-4.782	-3.626	-3.699	-1.567
	(1.969)	(2.767)	(2.569)	(2.995)	(2.305)
Female # No Gender	0.0603	3.165	2.722	2.954	2.639
	(1.895)	(2.027)	(2.061)	(2.062)	(1.993)
FGA		0.166	0.201*	0.244	0.281
		(0.0870)	(0.0863)	(0.138)	(0.183)
Top25		9.747***	12.46***	12.25***	11.81***
		(1.213)	(1.128)	(1.152)	(1.095)
Female # Top25		-4.573***	-5.165**	-4.752**	-4.242*
		(0.994)	(1.504)	(1.537)	(1.594)
No Gender # Top25			-4.416*	-4.518*	-4.161*
			(1.766)	(1.771)	(1.687)
Female # No Gender # Top25			1.326	1.117	0.804
			(1.891)	(1.848)	(1.870)
Constant	32.18***	15.91*	18.00**	31.07**	15.10
	(1.263)	(6.106)	(5.504)	(8.935)	(14.77)
Class Dummies [1]	Yes	Yes	Yes	Yes	Yes
Track Dummies	No	Yes	Yes	Yes	Yes
Location [2]	No	Yes	Yes	Yes	Yes
Info on School [3]	No	No	Yes	Yes	Yes
Participation [4]	No	No	No	Yes	Yes
Selection	No	No	No	Yes	Yes
Matriculation Rates [5]	No	No	No	No	Yes
Stem Choice [5]	No	No	No	No	Yes
Observations	4747	1666	1666	1666	1666
R^2	0.062	0.202	0.206	0.214	0.223

1. Dummies for each Year of the course (1 to 5);
2. Includes regional dummies and the number of inhabitants of the city where the school is located.
3. Includes the number of females and the total number of students in school;
4. Includes the percentage of participants in school with respect to the total number of students and the percentage of females with respect to the total number of participants in school;
5. Track Specific variables defined by gender (included both for males and females);

Standard errors in parentheses. Standard errors adjusted for 18 clusters in school.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 20. Treatment Effect (Comparison Baseline - Gender) on Score of Top 25 at School Level

	(1)	(2)	(3)	(4)	(5)
Normalized Score	OLS	OLS	OLS	OLS	OLS
Female	-3.676*	-4.744*	-3.862	-3.561	-3.192
	(1.680)	(1.862)	(1.882)	(1.786)	(1.805)
Gender	-0.284	-1.555	1.021	-7.305*	-10.56***
	(2.520)	(2.445)	(2.045)	(2.721)	(1.990)
Female # Gender	0.0767	0.559	-0.806	-1.039	-1.405
	(2.097)	(2.451)	(2.681)	(2.661)	(2.642)
FGA		0.137	0.210**	-0.0391	-0.0268
		(0.0798)	(0.0692)	(0.145)	(0.115)
Top25		9.202***	12.20***	12.01***	11.78***
		(1.086)	(1.057)	(1.012)	(1.011)
Female # Top25		-3.060*	-5.131**	-4.776*	-4.715*
		(1.141)	(1.693)	(1.666)	(1.605)
Gender # Top25			-5.461**	-5.509**	-5.380**
			(1.430)	(1.432)	(1.453)
Female # Gender # Top25			3.590	3.255	3.496
			(2.087)	(2.053)	(2.034)
Constant	31.23***	15.72**	14.14*	56.53**	12.61
	(1.259)	(5.228)	(4.823)	(18.30)	(17.44)
Class Dummies [1]	Yes	Yes	Yes	Yes	Yes
Track Dummies	No	Yes	Yes	Yes	Yes
Location [2]	No	Yes	Yes	Yes	Yes
Info on School [3]	No	No	Yes	Yes	Yes
Participation [4]	No	No	No	Yes	Yes
Selection	No	No	No	Yes	Yes
Matriculation Rates [5]	No	No	No	No	Yes
Stem Choice [5]	No	No	No	No	Yes
Observations	3644	1173	1173	1173	1173
R^2	0.076	0.267	0.274	0.280	0.288

1. Dummies for each Year of the course (1 to 5);
2. Includes regional dummies and the number of inhabitants of the city where the school is located.
3. Includes the number of females and the total number of students in school;
4. Includes the percentage of participants in school with respect to the total number of students and the percentage of females with respect to the total number of participants in school;
5. Track Specific variables defined by gender (included both for males and females);

Standard errors in parentheses. Standard errors adjusted for 15 clusters in school.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix A: Individual Participation

Probit Model - Treatment effect on individual participation

	(1)	(2)	(3)	(4)	(5)
Participation [1]	Probit[7]	Probit[7]	Probit[7]	Probit[7]	Probit[7]
No gender [2]	-0.00790	0.194^{***}	0.190^{***}	0.180^{***}	0.131
	(0.0916)	(0.0429)	(0.0531)	(0.0449)	(0.0786)
Gender [2]	-0.263^{***}	0.0133	0.00920	0.0152	-0.0403
	(0.0698)	(0.0575)	(0.0649)	(0.0585)	(0.0661)
Female	-0.116^{***}	-0.112^{***}	-0.118	-0.1000	-0.114^{*}
	(0.0210)	(0.0215)	(0.0608)	(0.0588)	(0.0554)
FGA		-0.00702^{*}	-0.00701^{*}	-0.00858^{**}	-0.00610^{*}
		(0.00288)	(0.00288)	(0.00283)	(0.00250)
Selection[3]: Suggested		0.179^{*}	0.179^{*}	0.158^{**}	0.187^{***}
		(0.0802)	(0.0801)	(0.0518)	(0.0477)
Selection[3]: Mandatory		0.374^{***}	0.373^{***}	0.346^{***}	0.357^{***}
		(0.0487)	(0.0486)	(0.0472)	(0.0402)
Female # No gender			0.00896	0.000697	0.0191
			(0.0617)	(0.0615)	(0.0612)
Female # Gender			0.00792	-0.0170	0.00958
			(0.0597)	(0.0592)	(0.0576)
Class Dummies [4]	Yes	Yes	Yes	Yes	Yes
Track Dummies	No	Yes	Yes	Yes	Yes
Matriculation Rates [5]	No	No	No	Yes	Yes
Stem Choice [5]	No	No	No	Yes	Yes
Location [6]	No	No	No	No	Yes
Observations	10342	8447	8447	8391	8391
Pseudo R^2	0.112	0.261	0.261	0.269	0.282

1. The dependent variable is a dummy for participation equal to 1 if the student enjoined the competition.
2. Base Category : Control group “Baseline”;
3. Base Category : “Free”;
4. Dummies for each Year of the course (1 to 5);
5. Track Specific variables defined by gender (included both for males and females);
6. Includes regional dummies and the number of inhabitants of the city where the school is located.
7. Marginal effects are reported.

Standard errors in parentheses. Standard errors adjusted for 27 clusters in school.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In this Appendix we want to show the also at individual level the treatments with monetary incentives had no effect on students’ probability to enter the competition. In the Table above we show results of the probit regression where the dependent variable is a dummy that takes 1 if the student participated at the Math Olympiad, 0 otherwise. Although the No Gender treatment has a positive and significant effect on individual participation in column (2), (3) and (4), the coefficient is not significant in the last column (where we add the full set of controls) for neither of the treatments. The negative and significant coefficient of the dummy “female” shows that the gender gap is present also when considering participation. The school quality is still negatively related to the probability to participate, here we confirm what we argued in Section 5.1. Finally, also in these regressions the selection criterion adopted within the attended school results to be the most relevant factor. Since our interventions needed the “free” participation in order to change individual decision to enjoy the Olympiad, it was hard for monetary incentives to have a considerable impact.

Bibliography

Altonji, J.G. and Blank, R.M., 1999. Race and gender in the labor market. Handbook of labor economics, 3, pp.3143-3259.

Andersen, S., Ertac, S., Gneezy, U., List, J.A. and Maximiano, S., 2013. Gender, competitiveness, and socialization at a young age: Evidence from a matrilineal and a patriarchal society. Review of Economics and Statistics, 95(4), pp.1438-1443.

Azmat, G. and Petrongolo, B., 2014. Gender and the labor market: What have we learned from field and lab experiments?. Labour Economics, 30, pp.32-40.

Bertrand, M., Goldin, C. and Katz, L.F., 2010. Dynamics of the gender gap for young professionals in the financial and corporate sectors. American economic journal: applied economics, 2(3), pp.228-55.

Bertrand, M. and Hallock, K.F., 2001. The gender gap in top corporate jobs. ILR Review, 55(1), pp.3-21.

Booth, A.L. and Nolen, P., 2012. Gender differences in risk behaviour: does nurture matter?. The economic journal, 122(558), pp.F56-F78.

Booth, A. and Nolen, P., 2012. Choosing to compete: How different are girls and boys?. Journal of Economic Behavior & Organization, 81(2), pp.542-555.

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), pp.354-369.

Buser, T., Niederle, M. and Oosterbeek, H., 2014. Gender, competitiveness, and career choices. The Quarterly Journal of Economics, 129(3), pp.1409-1447.

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

Cárdenas, J.C., Dreber, A., Von Essen, E. and Ranehill, E., 2012. Gender differences in competitiveness and risk taking: Comparing children in Colombia and Sweden. Journal of Economic Behavior & Organization, 83(1), pp.11-23.

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

Cotton, C., McIntyre, F. and Price, J., 2013. Gender differences in repeated competition: Evidence from school math contests. Journal of Economic Behavior & Organization, 86, pp.52-66.

Croson, R. and Gneezy, U., 2009. Gender differences in preferences. Journal of Economic literature, 47(2), pp.448-74.

Dasgupta, U., Mani, S., Sharma, S. and Singhal, S., 2019. Can gender differences in distributional preferences explain gender gaps in competition?. Journal of Economic Psychology, 70, pp.1-11.

Delfgaauw, J., Dur, R., Sol, J. and Verbeke, W., 2013. Tournament incentives in the field: Gender differences in the workplace. Journal of Labor Economics, 31(2), pp.305-326.

Dessy, S. and Djebbari, H., 2010. High-powered careers and marriage: can women have it all?. The BE Journal of Economic Analysis & Policy, 10(1).

Dreber, A., von Essen, E. and Ranehill, E., 2014. Gender and competition in adolescence: task matters. Experimental Economics, 17(1), pp.154-172.

Ellison, G. and Swanson, A., 2010. The gender gap in secondary school mathematics at high achievement levels: Evidence from the American Mathematics Competitions. Journal of Economic Perspectives, 24(2), pp.109-28.

Flory, J.A., Leibbrandt, A. and List, J.A., 2015. Do competitive workplaces deter female workers? A large-scale natural field experiment on job entry decisions. The Review of Economic Studies, 82(1), pp.122-155.

Furnham, A., Reeves, E. and Budhani, S., 2002. Parents think their sons are brighter than their daughters: Sex differences in parental self-estimations and estimations of their children's multiple intelligences. The Journal of genetic psychology, 163(1), pp.24-39.

Galor, O. and Weil, D.N., 1993. The gender gap, fertility, and growth (No. w4550). National Bureau of Economic Research.

Gneezy, U., Leonard, K.L. and List, J.A., 2009. Gender differences in competition: Evidence from a matrilineal and a patriarchal society. Econometrica, 77(5), pp.1637-1664.

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

Gneezy, U. and Rustichini, A., 2004. Gender and competition at a young age. American Economic Review, 94(2), pp.377-381.

Goldin, C., Katz, L.F. and Kuziemko, I., 2006. The homecoming of American college women: The reversal of the college gender gap. Journal of Economic perspectives, 20(4), pp.133-156.

Hedges, L.V. and Nowell, A., 1995. Sex differences in mental test scores, variability, and numbers of high-scoring individuals. Science, 269(5220), pp.41-45.

Huguet, P. and Regner, I., 2007. Stereotype threat among schoolgirls in quasi-ordinary classroom circumstances. Journal of educational psychology, 99(3), p.545.

Joensen, J.S. and Nielsen, H.S., 2014. Math and gender: Heterogeneity in causes and consequences of math. Economic Journal.

Kuhn Peter, Marie-Claire Villeval. 2011. Do Women Prefer a Co-operative Work Environment ? IZA Discussion Paper 5999, Bonn, and GATE WP 11-27

Maggian, V., Montinari, N. and Nicolo, A., 2017. Do Quotas Help Women to Climb the Career Ladder. A Laboratory Experiment. GATE—Lyon Saint-Etienne, WP, p.1724.

Murphy, M.C., Steele, C.M. and Gross, J.J., 2007. Signaling threat: How situational cues affect women in math, science, and engineering settings. Psychological science, 18(10), pp.879-885.

Niederle, M. and Vesterlund, L., 2007. Do women shy away from competition? Do men compete too much?. The quarterly journal of economics, 122(3), pp.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), pp.129-44.

Niederle, M. and Vesterlund, L., 2011. Gender and competition. Annu. Rev. Econ., 3(1), pp.601-630.

Niederle, M. and Yestrumskas, A.H., 2008. Gender differences in seeking challenges: The role of institutions (No. w13922). National Bureau of Economic Research.

Niederle, M., Segal, C. and Vesterlund, L., 2013. How costly is diversity? Affirmative action in light of gender differences in competitiveness. Management Science, 59(1), pp.1-16.

Ors, E., Palomino, F. and Peyrache, E., 2013. Performance gender gap: does competition matter?. Journal of Labor Economics, 31(3), pp.443-499.

Polachek, S.W., 1981. Occupational self-selection: A human capital approach to sex differences in occupational structure. The review of Economics and Statistics, pp.60-69.

Raven, J.C. and JH Court, 1938. Raven's progressive matrices. Los Angeles, CA: Western Psychological Services.

Reuben, E., Sapienza, P. and Zingales, L., 2014. How stereotypes impair women's careers in science. Proceedings of the National Academy of Sciences, 111(12), pp.4403-4408.

Reuben, E., Sapienza, P. and Zingales, L., 2015. Taste for competition and the gender gap among young business professionals (No. w21695). National Bureau of Economic Research.

Savikhin, A.C., 2011. Is there a gender gap in preschoolers' competitiveness? An experiment in the US. Mimeo.

Shurchkov, O., 2012. Under pressure: gender differences in output quality and quantity under competition and time constraints. Journal of the European Economic Association, 10(5), pp.1189-1213.

Spencer, S.J., Steele, C.M. and Quinn, D.M., 1999. Stereotype threat and women's math performance. Journal of experimental social psychology, 35(1), pp.4-28.

Stout, J.G., Dasgupta, N., Hunsinger, M. and McManus, M.A., 2011. STEMing the tide: using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (STEM). Journal of personality and social psychology, 100(2), p.255.

Sutter, M., & Rützler, D. (2010). Gender differences in competition emerge early in life. IZA Discussion Paper 5015.

Yee, D.K. and Eccles, J.S., 1988. Parent perceptions and attributions for children's math achievement. Sex Roles, 19(5-6), pp.317-333.

Zhang, Y.J., 2013. Can experimental economics explain competitive behavior outside the lab?. Available at SSRN 2292929.