Do quotas help women to climb the career ladder? A laboratory experiment

Valeria Maggian[*], Natalia Montinari[†], Antonio Nicolò[‡§]

Abstract: Women are less likely to enroll in selective or scientific courses, applying for promotions and are particularly underrepresented in both middle and top positions in the STEM field. Quota are often advocated as an instrument to reduce this gender gap, but it remains unclear at what step of the career ladder they more effectively foster women's reaching the top. Many factors may affect quota's success or failure, such as how they would in turn affect teamwork and trust between members within organizations or coordination in applying for promotions. In this paper, by means of a laboratory experiment implementing a two-stage tournament, we evaluate the impact of three different interventions in affecting individual decision to climb the career ladder, abstracting away from other possible confounding factors. We find that, compared with no intervention, a gender quota introduced in the initial stage of competition was ineffective in encouraging women to compete for the top; quotas introduced in the final stage of competition or at both stages increased women's willingness to compete for the top, without distorting the performance of the winners.

Keywords: Affirmative action, competition, multi-stage tournament, gender gap, laboratory experiments.
JEL: D03, C91, J24

[*] **Corresponding Author**: Ca` Foscari University of Venice, Department of Economics, Cannaregio 873, Fondamenta San Giobbe, 30121, Venice, Italy. E-mail: valeria.maggian@unive.it. Phone: +39 041 234 9150.
[†] Dipartimento di Scienze Economiche, Università degli Studi di Bologna, Piazza Scaravilli 2, 40126, Bologna, Italy.
[‡] Dipartimento di Scienze Economiche e Manageriali, Università degli Studi di Padova, via del Santo 33, 35123, Padova, Italy.
[§] School of Social Sciences University of Manchester, M13 9PL Manchester, UK.

Do quotas help women to climb the career ladder? A laboratory experiment

1. Introduction

Women are less likely to enroll in selective or scientific courses (Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, 2017; National Science Foundation, 2009) and to apply for top positions in competitive environment in the first place, especially in the STEM field (National Research Council, 2009, Kulis et al., 2002). Differences between women's and men's careers often emerge at an early stage and have long-lasting effects, with women also experiencing fewer opportunities for career advancement. Gender inequality comes at a cost. Not taking advantage of the skills of highly qualified women constitutes a waste of talent and, consequently, a loss of economic growth potential (Cavalcanti and Tavares, 2016).

Gender quotas are often advocated as possible solutions to reduce the gender gap. While their effectiveness also depends on how they would in turn affect teamwork and collaboration between members within organizations, as well as coordination when applying for promotions, in this paper we focus on individual decision making, isolating the effect of quotas on the decision to compete to the top. We aim to study, by means of a laboratory experiment, how gender quotas at different stages of a multi-stage tournament impact women's and men's choosing to compete, their performance, and beliefs about their own relative ranking in a clean environment, abstracting away from other possible confounding factors affecting the appointment of women to leadership positions (see, for example, Born et al., 2018; Reuben and Timko, 2017). Multi-stage tournaments resemble a career ladder because individuals are called upon to make sequential decisions, and early-stage decisions affect future opportunities for advancement, such as competing for a place in selective educational institutions, or for promotion within organizations.

Depending on the sector considered, the gender gap is indeed observed at every stage of the career ladder, or only at the top. A prominent example of the gender gap perpetuating throughout the entire career path is found in the STEM fields. Not only are young women less likely to graduate in engineering and computer science (in many countries, the proportion of women in these fields of study is far less than half; see, e.g. UNESCO, 2017) but also, once graduated, women are less apt to work as professional in these fields. Data from a subset of OECD countries has indicated that among graduates with science degrees, 71% of men, but only 43% of women, work as professionals in

physics, mathematics and engineering (Flabbi and Tejada, 2012). Moreover, despite the fact that enrollments in STEM disciplines has been rising in recent years, women academics in the US, for example, hold less than 20% of combined tenured and tenure-track posts in physics, chemistry and computer science (Williams et al., 2017), and they are more likely than men to quit these careers (Fouad et al., 2017). Furthermore, evidence of a gender gap in application for research grant support, considered as a major key to success in order to get ahead in research careers, was found across disciplines in several studies (Ley and Hamilton, 2008; Marsh, Jayasinghe and Bond, 2008; Waisbren et al., 2008)[1], even in absence of any gender difference in the grant awarding outcome, conditional on participation. This evidence suggests that individual decision making when applying for research grants is an important determinant of the observed gender gap.

In other fields, while women are well-represented in early career stages, very few are seen to attain top positions. In business schools, for example, men and women are equally present, yet in 2018 women held only 4.8% of Fortune 500 CEO roles.[2]

Different environments may, therefore, require that policy interventions aimed at closing the gender gap be made at different career stages. Whether affirmative action through the timed introduction of gender quotas would differently influence the choices women make at particular stages of their career is, however, far from clear.

If gender quotas were implemented at entry level, an early sorting of women may be avoided, with positive spillovers at later stages, but they might also reinforce women's under-confidence or self-stereotyping (i.e. activating a stereotype threat), impairing their chances of career advancement. Affirmative action at senior levels may have a cascade effect at lower levels and create role models for future generations, but may also be inadequate incentive for women to make positive early career decisions (Bertrand et al., 2014). Different strategies have been adopted in different countries and job sectors: quotas introduced in the early stages of a career are often invoked in traditionally male-dominated occupational sectors such as science and/or technology. Interventions at senior levels have been implemented in Sweden, where the government has introduced voluntary quotas for universities hiring full professors. The "Cascade Model" introduced in German universities in 2012 is a quota system applying at all levels of academic careers to ultimately increase the number of women at the highest level (Wallon et al., 2015).

Assessing the effect of policy interventions at different stages of a career is difficult in empirical studies using cross-country analysis or panel data because the collected data may suffer endogeneity problems: the introduction of such interventions may be related to changes in social attitudes towards

[1] For a more complete analysis see also European Commission (2009): The Gender Challenge in Research Funding: Assessing the European National Scenes. Luxembourg: OPOCE.

[2] Retrieved on August 27th, 2018, from http://fortune.com/fortune500.

women or changes made by women themselves. Additionally, empirical studies may lack randomization, since women may self-select into career paths and/or organizations depending on the affirmative action possibly in place. Our laboratory-based economic experiment, while only purporting to represent a stylized version of the complexity of the labor market, maximizes the internal validity of our study, and provides some preliminary evidence on this largely unexplored question. More specifically, we examine the effect of a gender quota in a two-stage tournament at (i) the entry level, (ii) the top level, or (iii) both levels of competition by implementing an experiment built upon the seminal study by Niederle and Vesterlund (2007). Participants in our experiment perform a simple arithmetic task in four separate stages. In stage one, for each correct answer they are paid according to a piece rate payment scheme, and in stage two they participate in a competitive winner-takes-all tournament. In stages three and four, which measure individual willingness to get to the top, participants are asked to choose either to compete in the first and the second stages, respectively, of a multi-stage tournament, or if they prefer to opt out from the tournament and be paid according to a piece rate.

Our results are summarized as follows. First, in our Control treatment, where gender quotas are absent, male participants were found to be significantly more willing than their female counterparts to compete in the two-stage tournament. A gender quota applied at the entry level eliminated the gender gap in the willingness to compete only at the entry level of the multistage tournament, while a gender quota at the top level (or at both levels) increased the proportion of women competing in both stages. Second, when we restricted our attention to the best-performing participants, we found that the entry level gender quota induced fewer of them to compete at the top level of the multistage tournament compared to gender quota implementation at the top, or both levels. Third, compared to our Control treatment, a gender quota at the entry level negatively impacted the best-performing women's confidence in their own ability, reducing their willingness to compete at the top level. By contrast, quotas at the top or both levels did not reduce the confidence of the best-performing women. Finally, irrespective of the stage of introduction, gender quotas did not distort efficiency as measured by the performance of individuals who reached the top.

It is important to note that our experimental measure of competitiveness, first introduced by Niederle and Vesterlund (2007), has been proved to be a good predictor of entrepreneurs' investment choices (Berge et al., 2015), participation in a competitive high school entry exam (Zhang, 2012) and of both students' choice of the college track and dropping out decision, particularly for girls (Almås et al., 2016). Most interestingly, Buser et al. (2014) showed that male high school students in the Netherlands, experimentally measured to be more competitive than their female schoolmates, were also more likely to choose more prestigious mathematics- and science-intensive academic tracks. The

same result is found for students' choices of high school Math-intensive specialization in Switzerland (Buser et al., 2017). In the same vein, Reuben et al. (2015) provide evidence that the taste for competition experimentally measured in a sample of high-ability MBA graduates at the Booth School of Business at the University of Chicago explained subsequent gender differences in earnings and industry choice. They also found that competitive individuals were more likely to be working in high-paying industries nine years later, suggesting that the relation between taste for competition and earnings, measured in the laboratory, persists in the long run, and may contribute to explaining gender differences in subsequent career paths.

Several studies have identified multiple explanations for the existence of the gender gap in top positions. Discrimination and gender stereotypes (Altonji and Blank, 1999; Goldin and Rouse, 2000) as well as different trade-offs made by men and women in formulating their career and family plans (Galor and Weil, 1996; Erosa et al., 2002; Dessy and Djebbary, 2010) may hinder women's career progression. Women have also been found to perceive themselves as less qualified to run for political office (Lawless and Fox, 2005), expressing lower career-entry and career-peak pay expectations (Bylsma and Major, 1992; Lyons and Schweitzer, 2014) and lower expectations of successfully fulfilling the demands of occupational success compared to men (Barbulescu and Bidwell, 2013). Researchers have also found evidence of women's weaker inclinations to participate in competition due to different distributional preferences, higher risk aversion, and lower levels of self-confidence compared to men (Niederle and Vesterlund, 2007; Eckel and Grossman, 2008; Bartling et al., 2009; Borghans et al., 2009; Croson and Gneezy, 2009; Balafoutas et al., 2012; Azmat et al., 2016).

Closing the gender gap is now one of the main objectives of the political arena. In light of the relevance of policy interventions aimed at balancing the participation of men and women in the labor market, recent studies have examined the effect of affirmative action policies on efficiency and on increasing women's participation in competitive environments in the laboratory. Balafoutas and Sutter (2012) found that affirmative action practices encourage women to enter competition more often, without negatively affecting efficiency. Similar results were obtained by Niederle et al. (2013), who observed that gender quotas do not negatively affect the overall performance of people selected, while achieving a more diverse set of winners. In their recent field experiment investigating affirmative actions in Colombia, Ibañez et al. (2015) found that the gains of attracting female applicants far outweighed the loss of male applicants. Previous studies have, however, modeled job markets as single-stage tournaments, with subjects choosing whether or not to enter a one-shot, isolated contest. In this setting, it is not possible to analyze the effects of a gender quota on women's choices and performance on the career ladder; in the real-world case, the choice of entering a competitive process entails several stages of advancement to career peak. Czibor and Dominiguez-

Martinez (2018) analyze in a laboratory experiment the impact of gender quotas introduced in the final stage of the competition and found them to be effective in increasing the representation of high-ability women at the top. Our experimental design, however, allows us to assess the effects of affirmative actions at different stages of a tournament, providing first evidence on the possible drawbacks of quotas implemented at an early stage of a competitive (stereotyped) environment. We provide clean results about the impact of placing gender quotas at different stages on women's willingness to compete, ruling out, by design, confounding factors such as women's beliefs about future discrimination and changes in attitudes towards women, as well as strategic concerns. The rest of the paper is organized as follows: Section 2 presents the study's experimental design; Section 3 provides our results on the effectiveness of gender quotas in the different stages of competition; and Section 4 concludes.

2. Experimental Design

The experiment was divided into five stages.[3] The experimental task in stages 1-4 was to add as many sets of three three-digit numbers and two decimal numbers as possible within 4 minutes. Each experimental session comprised of 24 participants: 12 males and 12 females who were randomly assigned to two equally sized groups, exactly balanced with respect to gender.[4]

In stage 1 (*piece rate*), subjects performed the task under a piece rate incentive scheme and received €0.50 for each correct calculation. In stage 2 (*compulsory tournament*), the 12 group members, 6 males and 6 females, competed against each other. The four members who solved the most calculations were paid €1.50 per correct answer. The other eight group members received nothing. In stages 3 and 4 *(multistage tournament)* subjects decided whether they wanted to be paid under a piece rate or a tournament scheme. Therefore, in stage 3 the first choice occurred (*choice 1*): if a piece rate was chosen, then subjects had the same incentives as in stage 1. If the tournament scheme was chosen, the subject's performance in the task was compared to the other group members' performance in stage 2, and a subject's payoff was equal to €1.50 per correct answer only if his/her performance was better than the performance of the fourth best performer in stage 2; otherwise it was equal to zero.[5] In stage

[3] After stage 5, the experimental sessions included a second part, in which we measured the possible spillovers of gender quotas on subsequent dishonest behavior using a variation of the die-under-the-cup task (Shalvi et al., 2011). Results from this part of the experiment were discussed in Maggian and Montinari (2017).

[4] Participants were informed about the random assignment into groups and the gender composition of each group only before the implementation of stage 2.

[5] For example, suppose that in the group [A-L] individual D is the only participant who decides to enter competition in stage 3, while all other group members decide to be paid according to the piece rate payment scheme; suppose also that the four best performers in stage 2 are ranked as follows: B, A, F and G. Finally, suppose stage 3 is randomly selected for payment at the end of the experiment for all participants. Monetary payoffs are therefore determined as follows: participant D gets 1.50 for each stage 3 correct answer if in stage 3 he gave more correct answers than participant G in stage 2, zero otherwise; all other participants belonging to the group A-L are getting 0.50 euro for each correct answer they gave in stage 3, since they chose the piece rate payment scheme to be applied to their performance in stage 3.

4 participants who chose to compete in stage 3 could make their second choice (*choice 2*): if a piece rate was chosen, then subjects had the same incentives as in stage 1. If the tournament was chosen, the subject's payoff was equal to €3 per correct answer only when his/her performance was better than the performance of the second best performer in stage 2. In stages 3 and 4, the ex-ante win probability was thus equal to 4/12=0.33 and 2/12=0.17, respectively. Note that, in stage 3, subjects chose whether to compete in a tournament or to be paid under a piece rate compensation scheme being aware that, in stage 4, only those who had chosen to compete in stage 3 could decide whether to continue to compete in the forthcoming tournament or to apply a piece rate payment to their performance.[6] Specifically, in stage 4 every participant was aware that: i) the piece rate payment scheme was applied *by default* to all those subjects who chose to be paid under a piece rate compensation scheme already in stage 3, since they forfeited the opportunity to make any subsequent choice about the payment scheme to be applied to their performance in stage 4; ii) the piece rate scheme was also applied to all those participants who chose to be paid under a tournament compensation scheme in both stage 3 and stage 4, but were not among the winners of stage 3; iii) the tournament payment scheme was thus only applied to those who had chosen to compete in stage 3 and in stage 4, conditionally on their being among the winners of the tournament in stage 3.[7]

Finally, in stage 4, before performing the task, those subjects who had chosen competition in both stage 3 and 4 were informed about whether they had won or lost the competition in stage 3, but not about their relative ranking. This was done in order to allow subjects' performances to be comparable across stages. In all stages, indeed, participants performed the tasks being aware of the payment scheme applied to their performance.

Figure 1 contains an overview of the first four stages of the experiment.

[6] Instructions about stage 2 were given only once stage 1 was completed. Once stage 2 was completed, instructions about both stages 3 and 4 were given at the same time. While instructions about stage 1 and 2 were identical across treatments, before making their decision to enter the competition or not in stage 3 participants were fully informed about treatment variations in both stage 3 and stage 4.

[7] In stages 2, 3 and 4, ties between participants participating in tournaments were broken randomly.

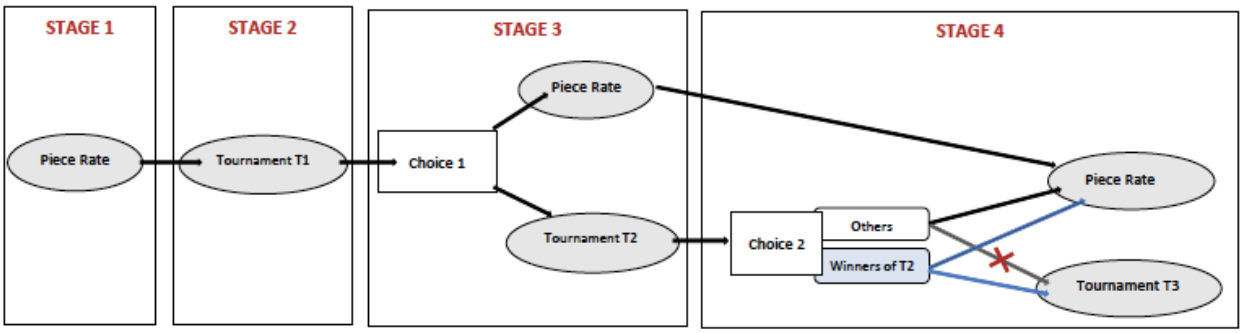


Figure 1. Overview of stages 1-4 of the experiment.
Note. In stage 1(2), all participants face a PR(T) without possibility of opting out. In stage 3, participants experience one of the two payment schemes (either the Tournament (T2) or the Piece Rate) depending on their Choice 1. In stage 4, participants who choose Tournament T2 (before knowing if they are among the winners of T2) can choose between Piece Rate and Tournament T3. They are, however, informed that only winners in the tournament T2 of stage 3 can actually participate in the tournament T3 in stage 4; if this condition is not satisfied, the Piece Rate is applied in stage 4, even if TournamentT3 was their Choice 2.

In our experiment, as in Balafoutas and Sutter (2012) and Niederle et al. (2013), if a subject chooses to compete in stage 3 and in stage 4, his performance is compared to all other group members' performance in stage 2[8]. Therefore, in our design, each subject's decision to compete does not depend on his beliefs about other subjects' decisions to compete, eliminating the impact of strategic concerns. This design greatly simplifies subjects' decisions but, admittedly, it does not allow a subject's decision to enter competition to impose externalities on others or to affect other's decision to choose the tournament, mimicking the coordination problem people often encounter in many real-world scenarios.

To investigate which policy intervention is more likely to positively affect women's top achievements (i.e. in stage 4, the most competitive tournament), we implemented four treatments. In the No-Quota treatment (NQ), winners of the tournaments in stages 3 and 4 are respectively the four and two group members with the most correct calculations, regardless of their gender. In the Quota-at-initial-stage treatment (Q1), there are at least two women among the four winners of the stage 3 tournament, meaning that the two best-performing women are winners, irrespective of their ranking within their group. Similarly, in the Quota-in-final-stage treatment (Q2), the best-performing woman is always one of the two winners in stage 4. In the Quota-in-both-stage treatment (Q1Q2), a gender quota is introduced in both stages, 3 and 4, following the rules described above.

Before performing the task in stage 4, once all participants had submitted their choice of the payment scheme to be applied in stage 4 (but before being told whether they were among the winners of the stage 3 competition), we elicited the beliefs of all subjects regarding their relative rank in stage 2.

[8] The alternative option (i.e. comparing the performance of the participants competing in stage 4 to the performance of all other group members choosing to compete in stage 3) would have implied that participants hold beliefs about the number of group members competing/winning in stage 3, reintroducing strategic concern in our experiment.

Subjects had to indicate their expected rank both within the whole group of twelve members and within the sub-group composed only of members of their own gender (i.e. six members). Correct guesses were rewarded €1 each, and feedback was given only at the end of the experiment.

In stage 5, we elicited subjects' risk preferences (Crosetto and Filippin, 2013), their attitudes toward competition (Houston et al., 2002; Harris and Houston, 2010) and some socio-demographic information. [9]

In each stage, all subjects were informed about the absolute number of correctly solved exercises. While participants did not receive any feedback on their relative ranking until the very end of the experiment,[10] those individuals who chose competition in both stage 3 and 4 were informed immediately before performing the task in stage 4 whether they were among the winners of stage 3 competition or not.[11]

2.1 Experimental Procedures

The experiment was conducted using z-Tree (Fischbacher, 2007) at GATE-LAB, the experimental laboratory of GATE (Lyon, France). Subjects were undergraduate students mostly from the fields of Management (39.3%) and Engineering (38%), recruited via the HROOT software (Bock et al., 2014).[12] From September to December 2016, 384 subjects (50% female) participated, divided among four sessions of 24 subjects per treatment (16 sessions in total). Sessions were randomly assigned to treatments so that all participants within the same session were assigned to the same treatment and none participated in more than one treatment. In each session, on arrival at the lab subjects were randomly assigned to two groups of 12, each composed of six men and six women.[13]

The duration of each session was about 95 minutes and the average payment was 17 euro, including a show-up and participation fee of five euro. To avoid wealth effects, one stage among stages 1 to 4 of Part 1 was randomly selected for payment at the end of the experiment and added to the earnings of stage 5 of Part 1 and to the earnings of Part 2.[14]

[9] In section A.1, we provide a more detailed description of the construction of our measure of individual attitudes towards competition and statistics relating to its distribution among participants.

[10] This was done so that participants did not condition their choices based on previous competition outcomes.

[11] This information did not affect individuals' decision making, since it was provided just before performing the stage 4 task, and after their choice of payment scheme to be applied in stage 4 had already been made.

[12] In section A.1 in the Appendix, we provide a more complete description of our sample, including a comparison of the characteristics of the participants assigned to our different treatments.

[13] Details of the recruitment procedures applied to achieve exact gender balance at the session level are provided in the Online Appendix, section B.1.

[14] The experiment was conducted in French. An English version of the experimental instruction is provided in the Online Appendix, section B.2.

3. Results

In this section, we present our results. First, we describe the effect of gender quotas on women's decisions to compete at the top in the multi-stage tournament (Section 3.1). Then, we focus on how gender quotas affected both willingness to compete in stage 4 and participants' beliefs about their own relative ranking of high-ability participants (in Section 3.2 and Section 3.3, respectively). Finally, we consider the impact of gender quotas on the performance of winners in the different treatments (Section 3.4).

When comparing the means of an ordinal continuous variable for two independent groups, such as our subjects' performance by gender, we performed a two-sample Mann–Whitney test as well as a two-sample t-test, reporting in the main text only the results of the Mann-Whitney, unless the two tests gave different results. When investigating whether a relationship existed between two categorical variables, for example when comparing participants' choices to compete and their gender, we performed a Pearson χ^2 test (or a Fisher exact test, if one of our cells had less than 5 observations). We also ran a set of Probit regressions to take account of any heterogeneity at the individual level, such as in participants' fields of study and beliefs.

The choice of implementing non-parametric tests in our analysis was made to deal with our small sample size. The interpretation of statistically significant estimates from small samples requires caution since, when the true effect is also small, a statistically significant estimate is likely to have occurred by chance (type I error) because the estimates have to be very large for statistical significance to be claimed (Gelman and Carlin, 2014). To address this concern, we performed two robustness checks. First, we implemented a retrospective power calculation suggested by Gelman and Carlin (2014) and further developed by Lu et al. (2019).[15] Second, to take our multiple hypotheses testing framework into account, we implemented a multiple comparison p-value adjustment, accounting for false discovery rate (Benjamini and Hochberg, 1995), in light of the fact that we were initially interested in the effect of gender quota introduction and did not know ex-ante which quotas would be most effective in positively affecting women's willingness to compete at the top. Both robustness checks are extensively discussed in section A.4 of the Appendix.

Finally, we adopted the convention of deeming significant only the test reporting a p-value below 5%.

[15] We utilized these tools, since the standard power formula used to calculate power ex-post gives much noisier and larger results than when used for ex-ante power calculations (Gelman, 2019; Hoening and Heisey, 2001).

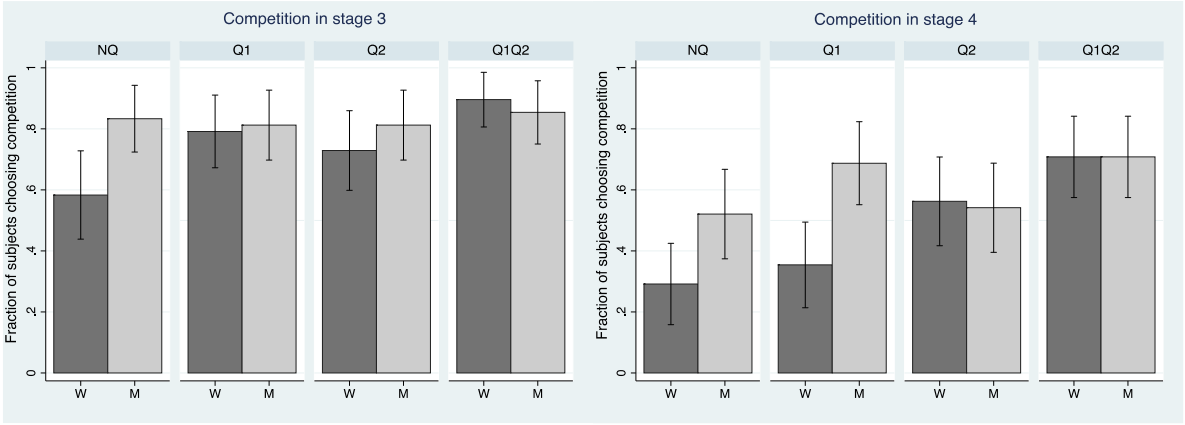
3.1 Decision to compete in presence of gender quotas

In line with similar previous studies (Balafoutas and Sutter, 2012; Niederle et al., 2013), in the compulsory tournament of stage 2 we found that the men performed, on average, better than women, but this difference was not statistically significant (except in Q2, Mann-Whitney test, p=0.016, MW henceforth).[16] Nevertheless, men were more willing to enter competition than women in the NQ treatment. Moreover, gender quotas differently affected the gender gap in willingness to compete, depending on the stage at which they were set, as described below.

Result 1.

In the NQ treatment male participants are significantly more willing to enter competition than female participants in both stage 3 and stage 4. Compared to the NQ treatment:

- *a gender quota in stage 3 (Q1 treatment) eliminates the gender gap in the willingness to compete in stage 3 but not in stage 4;*
- *a gender quota in stage 4 (Q2 treatment) slightly increases the proportion of women competing in stage 3 and eliminates the gender gap in stage 4;*
- *gender quotas in both stage 3 and stage 4 (Q1Q2 treatment) eliminate the gender gap in both stages.*



Panel (a) **Panel (b)**

Figure 2. Decision to compete in the multi-stage tournament by gender and treatment.

Note. Fraction of men and women who chose to compete in stage 3 (panel a) and stage 4 (panel b), across all treatments (N=384 participants, 48 men and 48 women per treatment). The bars show, for each treatment and gender, the proportion of participants (between 0 and 1) who chose the tournament in each stage of the multi-stage tournament. Error bars, mean ± SEM.

Support for result 1 can be found in Figures 2, 3 and in Tables 1-3. As shown in Figure 2, in the NQ treatment, our benchmark, we observed a gender gap in the willingness to compete in both stages 3 and 4 of the experiment. With respect to the benchmark, the introduction of a quota in stage 3 (Q1 treatment), reduced the observed gender gap but had no positive effect in terms of closing the gender gap in stage 4. The introduction of a quota in stage 4 (treatment Q2), slightly increased the proportion

[16] Treatment variations were announced after stage 2. More details of the average performance of men and women in stage 1 (piece rate) and stage 2 (compulsory tournament) are provided in Section A.2 of the Appendix.

of women choosing competition with respect to the NQ treatment in stage 3 and boosted women`s participation in the tournament of stage 4 such that gender differences were no longer significant. In treatment Q1Q2, a significant positive effect on women`s willingness to compete with respect to the NQ treatment was observed, with no difference between men and women. Results of a set of Pearson χ^2 tests on the gender gap in the willingness to compete in each treatment are reported in Table 1. The fraction of men who chose to compete in stage 3 and 4 did not vary across treatments, ($\chi^2(3)$ tests, in both stages: $p>0.125$).

Table 1. Gender gap in the willingness to compete in the multi-stage tournament.

Treatment	**NQ**		**Q1**		**Q2**		**Q1Q2**	
	Women	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**
Stage 3	58.33 (28/48)	83.33 (40/48)	79.17 (38/48)	81.25 (39/48)	72.92 (35/48)	81.25 (39/48)	89.58 (43/48)	85.42 (41/48)
	χ^2 (1)=7.261 p=0.007		χ^2 (1)=0.06 p=0.798		χ^2 (1)=0.944 p=0.331		χ^2 (1)=0.381 p=0.537	
Stage 4	29.17 (14/48)	52.08 (25/48)	35.42 (17/48)	68.75 (33/48)	56.25 (27/48)	54.17 (26/48)	70.83 (34/48)	70.83 (34/48)
	$\chi^2(1)$=5.23 p=0.022		$\chi^2(1)$=10.69 p=0.001		$\chi^2(1)$=0.042 p=0.837		$\chi^2(1)$=0.000 p=1.000	

Note. In each treatment, both for stage 3 and stage 4, the reported frequencies and Pearson χ^2 tests refer to N=96.

Our findings suggest that introducing a quota in stage 3 had the expected direct effect of increasing women's willingness to compete in the tournament at the lowest level, but it did not push them to compete in stage 4.

Figure 3 sheds further light on this phenomenon: it illustrates the fraction of participants who competed in stage 4, conditional on having competed in stage 3. The percentage of women who chose to continue to compete in stage 4 declined from 50% in NQ to 45% in Q1, while it increased to 77.14% in Q2, and 79.07% in Q1Q2. The percentage of men increased from 63% in NQ to 85% in Q1, 77.14% in Q2, and 82.93% in Q1Q2.[17] When considering the final stage of competition, in Q2 treatment the percentage of women competing increased significantly compared with both the NQ and Q1 treatment. Moreover, while a gender quota introduced in an initial stage did not increase the percentage of women who competed in stage 4, it increased the percentage of men compared to the NQ treatment. Differently, a gender quota in the final stage boosted women's willingness to compete in stage 4 without reducing men's participation, while the effect on men associated with the quota in

[17] The difference in the willingness to compete between men and women in stage 4 is not significant in all treatments except treatment Q1, according to a set of χ^2 tests. NQ: $\chi^2(1)=1.052$, $p=0.305$; Q1: χ^2 (1)=13.443, p=0.000; Q2: χ^2 (1)= 0.996, p=0.318; Q1Q2: χ^2 (1)=0.203, p=0.653.

an early stage seemed to persist in Q1Q2. The results of a set of χ^2 tests on the willingness to compete across treatments are reported in Table 2.

As a robustness check, we first implemented a retrospective power analysis on our results,[18] indicating that the probability of falsely finding results of the opposite sign (type S error) is almost 0 for all the statistically significant estimates, while the expected magnitude of overestimation (type M error) is at most 1.65, suggesting that our estimates are at most 1.65 times as large as the true effect size, and usually less than 1.5 times as large as the true effect size. As a second robustness check, we implemented a multiple comparison p-value adjustment so that, when adjusting our outcomes for false discovery rate, statistical significances were mostly unchanged. Therefore, our estimates were only marginally affected by the limitation associated with our small sample size.

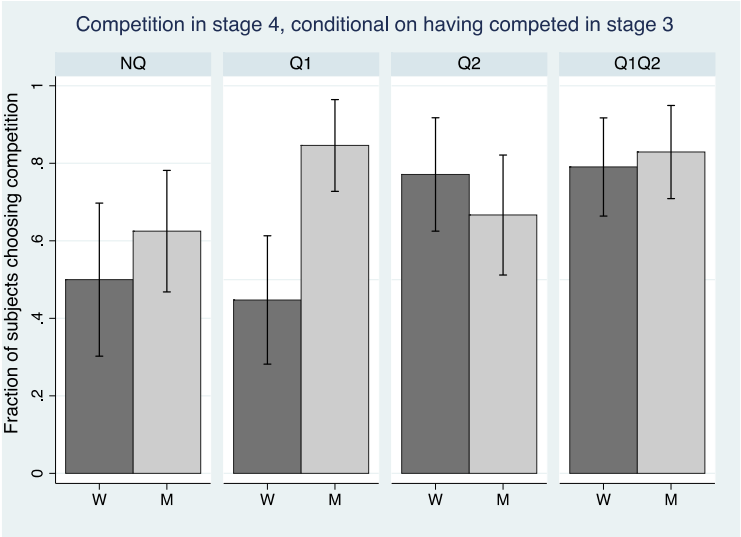


Figure 3. Competition in stage 4.

Note. Fraction of men and women who chose to compete in stage 4, across all treatments, conditional on having chosen the competition in stage 3 (N=303 participants, NQ: 68; Q1: 77; Q2: 74 and Q1Q2: 84). The bars show, for each treatment and gender, the proportion of participants (between 0 and 1) who chose the tournament in stage 4. Error bars, mean ± SEM.

[18] In section A.4 of the Appendix, we report the results of i) a retrospective power calculation using standard error estimates (see Figure A5), and ii) a multiple comparison p-value adjustment (see Table A5). Specifically, concerning the retrospective power calculation we employed t-tests to obtain standard errors and then together with theoretically plausible possible effects to calculate the probability of falsely finding results of the opposite sign (type S error) and expected magnitude of overestimation (type M error). For multiple comparison p-value adjustments, we replicated Table 2 after adjusting for false discovery rate.

Table 2. Treatments' effects on the willingness to compete in the multi-stage tournament.

Treatment Comparison	Women	Men
NQ vs Q1	$\chi^2(1)$=0.179 p=0.672, N=66	$\chi^2(1)$=4.949, p=0.026, N=79
NQ vs Q2	$\chi^2(1)$=5.043, p=0.025, N=63	$\chi^2(1)$=0.150, p=0.699, N=79
NQ vs Q1Q2	$\chi^2(1)$=6.543, p=0.011, N=71	$\chi^2(1)$=4.270, p=0.039, N=81
Q1 vs Q2	$\chi^2(1)$=7.991, p=0.005, N=73	$\chi^2(1)$=3.410, p=0.065, N=78
Q1 vs Q1Q2	$\chi^2(1)$=10.197, p=0.001, N=81	$\chi^2(1)$=0.042, p=0.838, N=80
Q2 vs Q1Q2	$\chi^2(1)$=0.042, p=0.838, N=78	$\chi^2(1)$=2.818, p=0.093, N=80

Note. In each treatment the proportion tests refer to the number of participants who chose to compete in stage 4, conditional on having chosen the competition in stage 3, N=303 participants, NQ: 68; Q1: 77; Q2: 74 and Q1Q2: 84.

In order to provide a formal analysis of the determinants of participants' willingness to compete in stage 4 and the effects of alternative policy interventions, while controlling for performance in previous stages (i.e. the compulsory piece rate and tournament)[19] as well as for other individual characteristics (e.g. field of study, beliefs about own performance, etc.), we ran a number of probit regressions, which confirmed all of our main findings.

More specifically, we estimated a Probit model of the form:

$$\Pr(choice_i = 1) = \Phi\left(\alpha + \beta female_i + \sum_j \gamma_i\, policy_i + \sum_j \delta_j\, female_i \times policy_j + C'\zeta + \epsilon_i\right) \quad (1)$$

The decision of individual *i* whether or not to compete in stage 4 (choice=1 if a subject chooses competition in stage 4, 0 otherwise) was regressed on gender, the three treatments Q1, Q2 and Q1Q2, the interaction between gender and each of the treatments, and a vector of controls denoted by C'. The vector of controls comprised the participants' performances in the compulsory piece rate of stage 1 (*Performance in stage 1)* and in the compulsory tournament of stage 2 (*Performance in stage 2*), the beliefs about own performance in stage 2 (*Belief on Performance in stage 2*) and willingness to take risk (*Willingness to Take Risk*). We also controlled for competitive attitude by including a shorter version of the Competitiveness Index and for the participants' fields of study, being Management, Engineering or Others (Psychology, Economics, Sociology, Medical Science, Languages, etc.). More specifically, in our regression model we included a set of dummy variables, with Engineering being the reference category. $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. In Table 3, we report the results from estimating (1) based on different sets of controls.

In column (1) we control only for the gender and the different treatments. We find that, with respect to the NQ treatment, both Q2 and Q1Q2 treatments have a positive and significant effect on the choice

[19] In Section A.3 of the Appendix we present both a numerical and a graphical representation of the proportion of men and women who decided to compete in stage 4, conditional on having competed in stage 3, divided into quartiles based on their performance in stage 2.

to compete in stage 4, while Q1 treatment has no significant effect. A significant gender gap is also found, indicated by the negative and significant coefficient of the dummy female. In column (2) we add the participants' performances in both stage 1 and in stage 2: only the coefficient of the participants' performances in the compulsory tournament is positive and significant, while the other results are unchanged with respect to column (1). In column (3) we add beliefs about own relative performance in stage 2. The coefficient of this variable has a negative and significant effect, suggesting that the higher (i.e. the worse) a participant's belief about his or her own relative ranking, the lower the probability of their continuing to compete. Looking at the other independent variables, we observe that the gender dummy is not more significant (p=0.1), suggesting that the greater part of the gender gap can be explained by the different beliefs men and women hold about their own relative ranking. We also find that the variable accounting for performance in stage 2 does not achieve significance, suggesting that what matters in explaining participants' decisions to compete is not their performance per se, but rather their beliefs about their ability. In columns (4) and (5) we added to the regression model the interaction terms between the treatment and gender, respectively controlling for the participants' beliefs about their ability, and not. In both columns (4) and (5), we observe that being female in the NQ treatment diminishes the probability of competing in stage 4, as suggested by the significant coefficient of the *female* variable. Moreover, the significant coefficients of the interaction terms "*female x Q2"* and "*female x Q1Q2"* indicate the extent to which the disparity between men and women in entering into competition in stage 4 changes in the Q2 and Q1Q2 treatments, respectively, with respect to the NQ treatment. Interestingly, "*female x Q1"* is never significant, suggesting that the gender gap in competition in stage 4 does not vary when the Q1 and the NQ treatments are compared. When comparing the results of column (4) and (5), our results are unchanged, the only difference being that the role of beliefs is evident once more, since participants' performances in stage 2 are no more significant when controlling for them. Finally, in column (6) we observe that the more competitive participants are, the more likely they will compete, as expected (competitiveness is captured by the competitiveness index). In columns (4) and (5) none of the coefficients of the treatment variables is significant, showing that the males' behavior is unaffected by the treatment.

Table 3. Probit regressions: Choice of competition in the second tournament of the multi-stage competition (stage 4).

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Estimation Method: Probit Regression, Marginal Effects.						
Dependent Variable: Choice of competition in stage 4						
Independent Variables						
Female	-.126**	-.111**	-.092	-.249**	-.248**	-.174
	(.053)	(.055)	(.055)	(.101)	(.101)	(.108)
Q1 treatment	.109	.104	.110	.130	.111	.131
	(.070)	(.071)	(.071)	(.103)	(.104)	(.105)
Q2 treatment	.134*	.140*	.150**	-.019	-.020	-.021
	(.070)	(.070)	(.070)	(.104)	(.104)	(.103)
Q1Q2 treatment	.290***	.292***	.283***	.168	.155	.183*
	(.063)	(.063)	(.064)	(.099)	(.100)	(.101)
Performance in stage 1 (compulsory piece rate)	-	-.004	-.002	-.006	-.005	-.001
		(.017)	(.017)	(.017)	(.018)	(.018)
Performance in stage 2 (compulsory tournament)	-	.053***	.021	.058***	.027	.020
		(.016)	(.019)	(.016)	(.020)	(.020)
Belief on Performance in stage 2 (higher value=lower rank)	-	-	-.035***	-	-.034***	-.033***
			(.012)		(.012)	(.012)
Female x Q1 treatment	-	-	-	-.045	-.002	-.011
				(.153)	(.153)	(.157)
Female x Q2 treatment	-	-	-	.295**	.309**	.338***
				(.109)	(.106)	(.100)
Female x Q1Q2 treatment	-	-	-	.260*	.263*	.214
				(.120)	(.121)	(.133)
Competitiveness	-	-	-	-	-	.014***
						(.003)
Willingness to Take Risk	-	-	-	-	-	.002
						(.002)
Study: Management	-.127**	-.115*	-.104*	-.112*	-.105*	-.136**
	(.059)	(.060)	(.061)	(.061)	(.062)	(.063)
Study: Others	-.080	-.058	-.041	-.052	-.040	-.074
	(.072)	(.074)	(.074)	(.075)	(.076)	(.078)
N. of observations	384	384	384	384	384	384
Log pseudolikelihood	-249.309	-243.017	-238.8302	-238.079	-234.279	-223.432
Wald chi2	28.19***	40.90***	46.91***	49.80***	56.36***	76.77***
Pseudo R2	0.057	0.081	0.097	0.100	0.114	0.155

Note: Table 3 presents the marginal effects from a series of Probit regressions. The dependent variable is the choice to compete in stage 4. Robust Standard errors appear in parentheses. ***, ** and * indicate significance at the 1% level, 5% level and 10% level, respectively.

Overall, the results in columns (1) to (6) of Table 3 suggest that the beliefs about own ability can partially explain the difference in the choice of competing in stage 4. However, in models (1)-(2) and (4)-(5) the dummy for the gender difference is still negative and significant, while the statistically

significant residual gender gap in entry decisions disappears in column (6), when controlling for the Competitiveness Index.

Apart from treatment Q1, whose coefficient never reaches statistical significance, the other treatments` coefficients have a positive and significant impact on women's choices.

Finally, when computing the robustness checks on Table 3's outcome, by implementing a power analysis and by adjusting our p-values for multiple comparison, our outcomes confirm that our estimates were not obtained by chance.[20]

The above analysis shows that a gender quota set in an initial stage only temporarily increases women's willingness to compete, thus not encouraging them to compete in stage 4. On the contrary, in Q2, a quota in the final stage has no effects on the initial stage compared to the NQ treatment, but it does induce those women who decided to compete in stage 3 to keep competing because they now have more chance of winning. Introducing a quota in both stages 3 and 4 has the same effect of Q2 on women`s participation in the last tournament.

3.2 Decision to compete of best performing participants

In the analysis of the effectiveness of gender quotas, high-ability participants deserve special attention. As documented by previous studies (Balafoutas and Sutter, 2012; Niederle et al., 2013), high-ability women shy away from competitions that they could potentially win, and the introduction of a gender quota in a one-shot tournament shows effective results in encouraging their participation. The main effect of gender quotas in a one-shot tournament is, therefore, linked to a change in the pool of competitors: high-ability women decide to compete, so that the quota is shown not to be detrimental to overall efficiency.

To shed light on the effect of gender quotas placed in different stages of a multi-stage tournament, we focus our attention on high-ability male and female participants, identifying as best-performing those who gave five or more correct answers in stage 4. This threshold, computed by means of numerical simulations as in Niederle et al. (2013), makes the choice of competing payoff-maximizing in all treatments (see section A.5 of Appendix for details and a discussion of the procedure used for identifying the threshold. Section A.5 also contains a robustness analysis identifying as best-

[20] Figure A5 in section A.5 of the Appendix contains a graphical representation of the estimation of the type S and type M errors for Table 3's coefficients estimates: we find that type S error probability is almost 0 for all statistically significant estimates, while type M error rate is at most 1.65, suggesting that our estimates are at most 1.65 times as large as the true effect size and usually less than 1.5 times as large as the true effect size. Table A6 in Section A.4 in the Appendix, replicates Table 3 after adjusting for false discovery rate (Benjiamini and Hochberg 1995). In all models, statistical significances are mostly unaffected by multiple comparison correction.

performing those who gave five or more correct answers in stage 2[21]. Results are virtually unchanged).

A set of χ^2 tests suggests that the proportion of best-performing participants does not vary across treatments (χ^2 (3) =5.086, p= 0.166). When considering the proportion of best-performing participants by gender, we observe that for male participants, this proportion does not vary across treatments ($\chi^2(3)=0.508$, p=0.917), while significant variations are found for females ($\chi^2(3)=12.755$, p=0.005). The difference is driven by treatments Q1 and Q1Q2 in which there is a significantly higher percentage of best performers compared to the NQ treatment (NQ vs Q1: p=0.004, NQ vs Q1Q2: p=0.040).[22]

Our findings regarding best-performing participants are summarized in Result 2.

Result 2.

In the NQ treatment, our data suggest that best-performing male participants are more willing to enter competition than best-performing female participants.

In the Q1 treatment, a gender quota in stage 3 only does not eliminate the gender gap inducing less best-performing female participants to compete in stage 4.

In the Q2 and in the Q1Q2 treatments, a gender quota in stage 4 eliminates the gender gap in stage 4.

Support for result 2 is provided in Figure 4 and Table 4. In Figure 4 we consider only 196 participants identified as best-performing. We observe that, in the NQ treatment, best-performing females appear to be less likely to access competition at the top than best-performing males (50% (N=8/16) vs 62.96% (N=17/27), even if the difference does not achieve statistical significance $\chi^2(1)= 0.694$, p=0.405. Focusing on the Q1 treatment, it can be noted that in Q1 the fraction of best-performing women entering competition in stage 4 is lower than the proportion of best-performing men ($\chi^2(1)=8.905$, p=0.003, N=54), providing additional evidence of the potential negative effects of an initial stage gender quota.

[21] The reason why we decided not to use the subjects' performance in stage 2 for the identification of best-performing participants in our analysis is that subjects' performance in stage 2 is not a good predictor of their performance in the following stages. The misclassification does not depend on subjects` choice to compete in stage 3 and in stage 4, both overall and separately for gender. See the Appendix for a more detailed discussion.

[22] A more detailed analysis of the distribution of best performers and other variables of interest in our sample is provided in Table A2 of the Appendix.

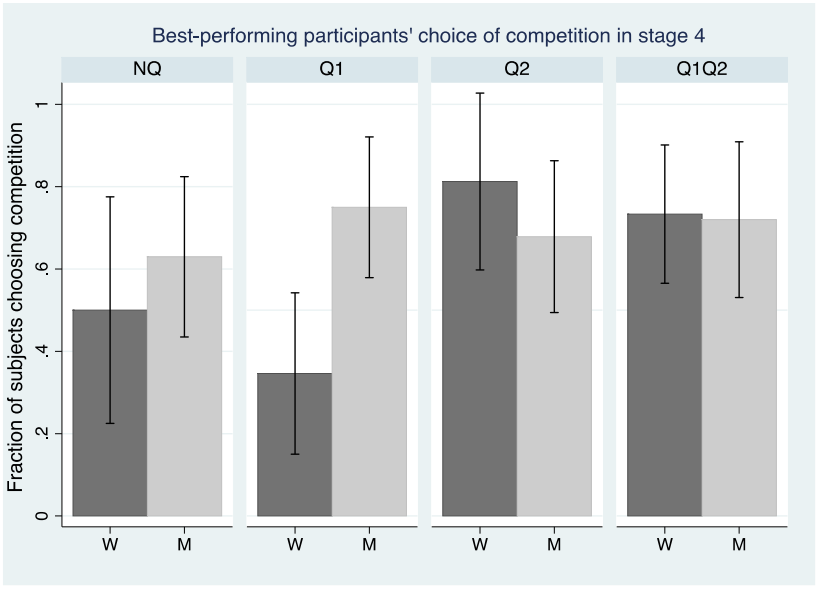


Figure 4. Competition in stage 4 by best-performing participants.
Note. Fraction of best-performing female and male participants who chose to compete in stage 4, across all treatments (N=196 participants; NQT: 43; Q1T: 54; Q2T: 44 and Q1Q2T: 55). The bars show, for each treatment and gender, the proportion of best-performing participants (between 0 and 1) who chose the tournament in both stages of the multi-stage tournament. Error bars, mean ± SEM.

Taking Q1 as the benchmark, in Q2 and Q1Q2 the proportion of best-performing women who choose to compete significantly increases (Q1 vs. Q2: Fisher's exact test, p=0.005, N=42; Q1 vs. Q1Q2:$\chi^2(1)$=8.449, p=0.007 N=56), while the choice of best-performing men is unaffected by the treatments ($\chi^2(3)$= 1.052, p=0.789 N=108). Moreover, 59.09% (13/22) of best-performing women chose not to compete after stage 3 in Q1, while only 18.75% (3/16) took the same decision in Q2, and 21.43% (6/28) in Q1Q2.

However, conducting the above-described analysis on the subset of 196 best-performing participants suffers from the limitations associated with a small sample size. To tackle this problem, in Table 4 we present the results of a set of Probit regressions identical to those presented in models 4-6 of Table 3, which consider the whole sample of 384 participants, the only difference being that in the set of the explanatory variables a dummy identifying best-performing participants is included as a regressor. Differently from the tests implemented so far, the regression analysis allows us a better estimate of the effects of our treatment variations on best-performers' willingness to compete at the top, while controlling for individual characteristics. First, we observe that, in absence of any gender quotas, being a best-performing female significantly decreases the probability of accessing competition at the top level with respect to being a best-performing male (conditional marginal effect of female (in NQ) and best-performers = -.222**, Robust Standard errors=.092). This result also holds when controlling for participants' beliefs (column 2, conditional marginal effects = -.219**, Robust Standard errors=.092), while the gender gap disappears when the participants' competitive index and risk attitude is also included in the analysis (column 3; conditional marginal effects =-.151, Robust

Standard errors=.102). Second, in the Q1 treatment, the gender quota in stage 3 does not close the gender gap in stage 4: best-performing women are less likely to continue competing (both in models 1 and 2, conditional marginal effects = -.230**, Robust Standard errors=.106; while in model 3 conditional marginal effects = -.157, Robust Standard errors=.111).

Table 4. Probit regressions: Choice of competition in the second tournament of the multi-stage competition (stage 4).

	Model 1	**Model 2**	**Model 3**
Independent Variables			
Female	-.229**	-.226**	-.154
	(.102)	(.102)	(.108)
Q1 treatment	.130	.112	.131
	(.103)	(.104)	(.104)
Q2 treatment	-.022	-.021	-.022
	(.103)	(.103)	(.103)
Q1Q2 treatment	.174*	.162	.188*
	(.099)	(.101)	(.100)
Performance in stage 1	-.018	-.017	-.009
(compulsory piece rate)	(.018)	(.018)	(.019)
Performance in stage 2	.045**	.012	.007
(compulsory tournament)	(.018)	(.021)	(.021)
Best performers	.141**	.148**	.133**
	(.063)	(.064)	(.066)
Belief on Performance in stage 2	-	-.035***	-.035***
(higher value=lower rank)		(.012)	(.013)
Female x Q1 treatment	-.074	-.028	-.038
	(.154)	(.155)	(.159)
Female x Q2 treatment	.296**	.310**	.338***
	(.108)	(.106)	(.100)
Female x Q1Q2 treatment	.222	.223	.177
	(.130)	(.131)	(.141)
Competitiveness	-	-	.014***
			(.003)
Willingness to Take Risk	-	-	.002
			(.002)
Study: Management	-.112*	-.105*	-.135**
	(.062)	(.063)	(.063)
Study: Others	-.054	-.044	-.075
	(.076)	(.076)	(.078)
N. of observation	384	384	384
Log pseudolikelihood	-235.649	-231.621	-221.401
Wald chi2	52.13***	59.32***	80.33***
Pseudo R2	0.109	0.124	0.163

Robust Standard errors appear in parentheses. ***, ** and * indicate significance at the 1% level, 5% level and 10% level, respectively.

3.3 The impact of affirmative action on participants' beliefs about own relative ability

In this section, to elucidate the potential mechanism underlying the negative effect of Q1 treatment on the decision of best-performing women not to continue competing in stage 4, we focus on how men's and women's beliefs about their relative ability vary across treatments and how these affect their decisions to compete in stage 4.

All participants were asked to guess their ranking in the compulsory stage 2 tournament, once having submitted their choice of the payment scheme to be applied in stage 4. Importantly, the beliefs elicitation phase was implemented before participants performed the task in stage 4 and before being told whether they were among the winners of the stage 3 competition (when competition was previously chosen as a payment scheme to be applied in stage 3).[23] Our findings are summarized in Result 3.

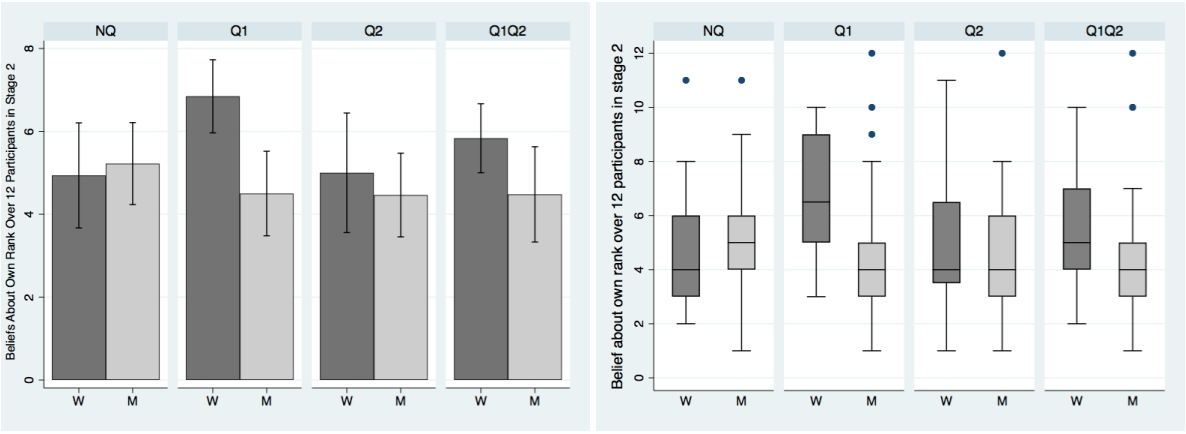
Result 3.

Compared to the NQ treatment, in the Q1 treatment the presence of a quota in stage 3 negatively impacts best-performing women's confidence in their own ability, reducing their willingness to compete in stage 4. Differently, in Q2 and Q1Q2 treatments, the presence of quotas does not change best-performing women's confidence.

Support for result 3 is provided by Figures 5 and 6. Panel (a) of Figure 5 reports the average beliefs of best-performing participants about their ranking position in stage 2 (in the compulsory tournament with 12 participants) while panel (b) reports the entire distribution of such beliefs by gender in each treatment.

From Figure 5 it can be noted how in Q1 the overall distribution of beliefs of best-performing women is shifted toward higher (i.e. worse) ranks compared to the other treatments: best-performing women have more pessimistic beliefs about their relative ability. Statistical tests support the graphical evidence. In Q1, best-performing women rank themselves on average two positions lower compared to both the NQ and Q2 treatments, while comparison with the Q1Q2 treatment leads to almost no significant differences (6.85/12 in Q1 vs. 4.94/12 in NQ, 5/12 in Q2, 5.83/12 in Q1Q2; Mann-Whitney test (MWT henceforth), $z=2.67$, $p=0.008$, $z=2.49$, $p=0.013$ and $z=1.66$ $p=0.097$ respectively).

[23] We decided to elicit beliefs about the participant's rank in the compulsory tournament of stage 2 because beliefs about the rank in the following stages necessarily depend on the participant's choice to compete in these stages, so that this additional information should be elicited as well. With respect to the timing of the belief elicitation phase, we decided to implement it after the participant submitted his/her choice of payment scheme to be applied in stages 3 and 4, in order to avoid any effect of the beliefs' elicitation procedure on individuals' decisions. Therefore, participants' beliefs are affected by the treatments, since we ask participants' beliefs about their own ranking in stage 2 only once instructions about both stages 3 and 4 have been explained.



Panel (a) **Panel (b)**

Figure 5. Beliefs about own ranking over 12 participants in stage 2.

Note. Panel (a) reports the average beliefs of best-performing female and male participants about their relative ranking over a group of 12 participants in stage 2 (compulsory tournament). The bars show, for each treatment and gender, the average rank guessed by participants (between 1 and 12, where 1 identifies the best performance and 12 identifies the worst performance). Error bars, mean ± SEM.
Panel (b) reports the distribution of beliefs. In each panel, the line inside the box represents the median; the lower and upper borders of the box represent the 25th (Q1) and 75th (Q1Q2) percentiles, while the ends of the whiskers represent the most extreme values within Q1Q2+1.5(Q1Q2-Q1) and Q1-1.5*(Q1Q2-Q1), respectively). Dots represent outliers.
In both panels we have N=196 participants; NQT: 43; Q1T: 54; Q2T: 44 and Q1Q2T: 55.

By contrast, we observe that the beliefs of best-performing men are not affected by the treatments. As for the decision to continue to compete, in order to limit the problem of a reduced sample size, we complement our analysis by both replicating Figure 5 and implementing a set of Tobit estimations on our full sample of 384 participants. In all regression models, the dependent variable is the self-reported beliefs about own ranking in the tournament of stage 2 (ranging from 1 to 12, where 1(12) is associated to the best (worst) rank) while the set of independent variables included in the analysis include: the gender of the participant, the choice to compete in stage 4, the number of given correct answers and the participant's actual rank in the compulsory treatment, a set of dummies accounting for treatments and their interactions with gender, and a dummy accounting for being a best-performing participant. Our main findings are confirmed, with the coefficient associated with women reporting a worse expected rank than males being significant for the Q1 treatment, but not for the Q2 and Q1Q2 treatments. More detailed results can be found in Section A.6 of the Appendix.

In Figure 6, beliefs are grouped in quartiles[24] to additionally investigate how both their distribution and the decision to compete in stage 4 are affected by the treatment manipulation. However, since

[24] Participants who expect to be ranked in position 1, 2 or 3 are assigned to the first quartile (Q1); those who expect to be ranked in position 4, 5 or 6 are assigned to the second quartile (Q2); those who expect to be ranked in position 7, 8 or 9 are assigned to the third quartile (Q1Q2); finally, those who expect to be ranked in position 10, 11 or 12 are assigned to the fourth quartile (Q4).

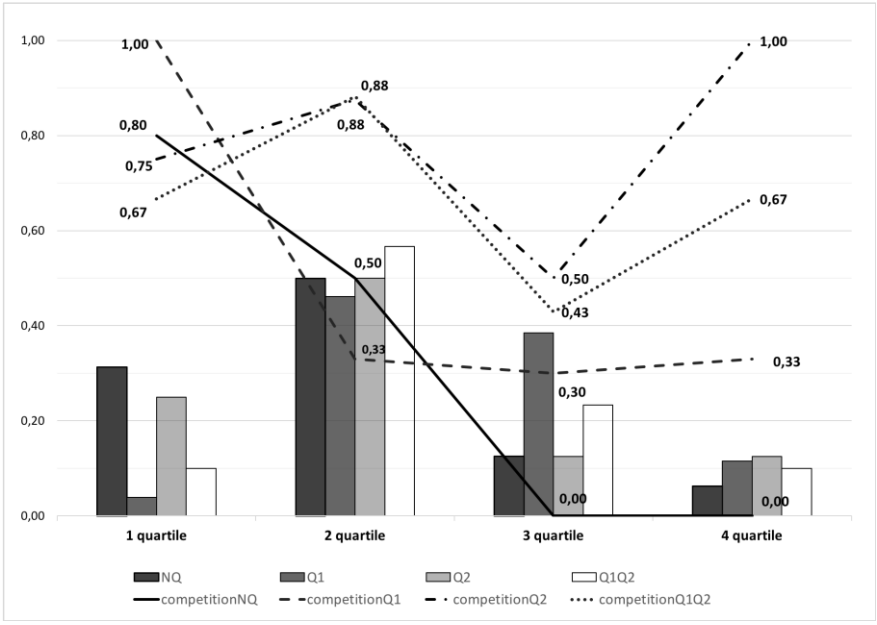
when slicing our data into quartiles we further reduce our simple size, we acknowledge that, while our outcomes are consistent with those presented so far, and also with the results reported by Bracha et al. (2019), they should be considered as only indicative of a possible direction for future research. Panel (a) refers to women, panel (b) to men. Looking at the best-performing women, the bars in panel (a) of Figure 6 report the distribution of beliefs in each treatment by quartile; the lines represent the proportion of women deciding to compete in stage 4 depending on the quartile of their beliefs. In the beliefs distribution, note how treatment Q1 induces a shift from the first to the third quartile, compared to the NQ and Q2 treatments. Only one of the 26 best-performing women in Q1 expects to be ranked in the first quartile, while ten beyond those 26 believe they would be ranked in the third quartile, between the 6th and 9th positions. Note also that a similar shift, although less pronounced, is observed in treatment Q1Q2. Regarding the choice to compete depending on beliefs, it can be seen how differences between treatments are more pronounced for the best-performing female participants who expect to be ranked in the second quartile: in Q1 significantly fewer women decide to compete compared to the Q2 and the Q1Q2 treatments, according to a set of Fisher's exact test (Q1 vs Q2, p= 0.028; Q1 vs Q1Q2, p=0.005).[25] Consider now the best-performing men, whose beliefs and decisions to compete are reported in panel (b) of Figure 6. From the distribution of beliefs, it can be noted that, overall, best-performing men are unaffected by the introduction of a quota.[26] Similarly, in the choice to compete, differences are not statistically significant across treatments.

By comparing panels (a) and (b) we focus on gender differences. Figure 6 confirms what is already highlighted in Figure 5: most best-performing men believe themselves to be ranked in the first and second quartile— not the case for best-performing women, most of whom believe themselves to be ranked in the second or third quartiles, a difference more evident in Q1 and Q1Q2 treatments. Of the decision to compete in stage 4, it is notable how, in each treatment and overall, gender differences in willingness to compete in stage 4 are not significant for participants who expect to be ranked in the first, third and fourth quartiles.[27] Among participants who believe they would rank in the second quartile, however, we find a significant gender difference in Q1 with fewer best-performing women entering the competition compared to men (Fisher's exact test, p=0.009), although this is not the case in the other treatments (NQ: p=0.685; Q2: p=0.177; Q1Q2: p=0.283).

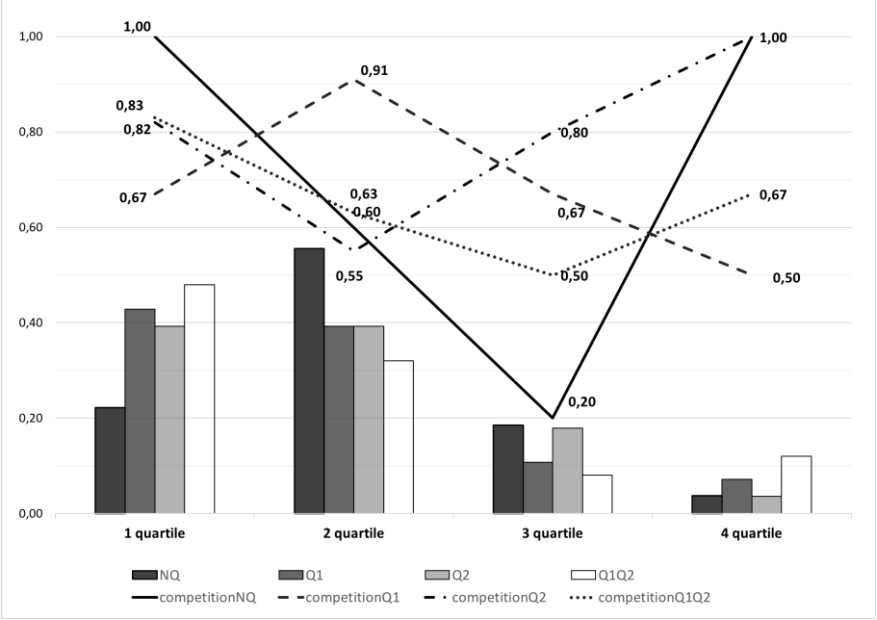
[25] Fisher exact tests for the participants in the second quartile: NQ vs Q2, p= 0.282; NQ vs Q1Q2, p= 0.059; NQ vs Q1, p=0.648; Q2 vs Q1Q2, p=1.000.

[26] Once gender quotas are introduced, the fraction of men expecting to be ranked in the first (second) quartile increases (decreases). These differences do not, however, achieve significance according to a set of pairwise Fisher exact tests (all p-values > 0.262).

[27] Fisher exact tests: first quartile: NQ: p=0.455; Q1: p= 1.000; Q2: p=1.000; Q1Q2: p=0.516; third quartile: NQ: p= 1.000; Q1: p=0.510; Q2: p=1.000; Q1Q2: p=1.00; fourth quartile: NQ: p=1.000; Q1: p= 1.000; Q2: p= 0.333; Q1Q2: p= 1.000.



Panel (a)



Panel (b)

Figure 6. Distribution of best-performing participants' beliefs and decisions to compete.

Note. Panel a (b) reports the distribution of beliefs of best-performing female (male) participants about their relative ranking over a group of 12 participants in stage 2 (compulsory tournament) and the fraction of best-performing female participants who decide to compete in stage 4. For each treatment, beliefs are grouped in quartiles (expected ranks 1-3 correspond to the first quartile; expected ranks 4-6 correspond to the second quartile; expected ranks 7-9 correspond to the third quartile; expected ranks 10-12 correspond to the fourth quartile). Panel (a): N=88; NQT: 16; Q1T: 26; Q2T: 16 and Q1Q2T: 30. Panel (b) : N=108; NQT: 27; Q1T: 28; Q2T: 28 and Q1Q2T: 25.

Our findings suggest that setting a quota at an initial stage of the career ladder (stage 3) may act as a gender stereotype cue, affecting best-performing women`s beliefs about their own relative ranking and discouraging their continued competing.

3.4 Efficiency

In a further analysis, we investigated the effect of gender quotas on the performance of the winners of the competition in stage 4 to evaluate the impact of such policies on efficiency. Our findings are summarized in Result 4.

Result 4. *Taking NQ treatment as a benchmark, there is no difference in the average performance of the winners of stage 4, implying that the introduction of a gender quota does not harm efficiency.*

Support for result 4 can be found in Figure 7.

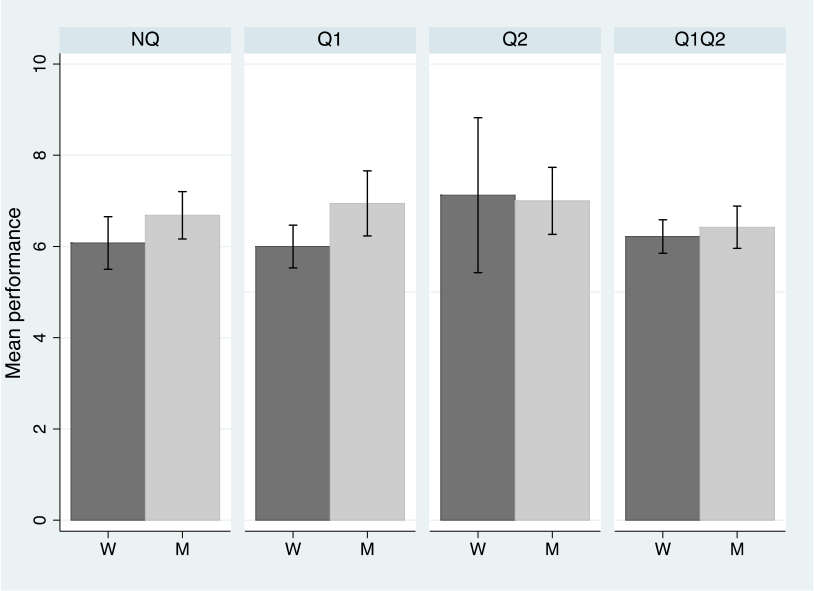


Figure 7. Performance of competition winners in stage 4.
Note. Average performance of winners of competition in stage 4, by gender and across all treatments (N=86 participants; NQT: 22; Q1T: 13; Q2T: 23 and Q1Q2T: 29. The bars show, for each treatment, the average stage 4 performance (number of correct calculations) of those participants who won the tournament in stage 4. Error bars, mean ± SEM.

In Figure 7 we plot the average performance of winning participants in each treatment by gender. Within each treatment, differences between the performances of men and women are not significant (men give, on average, one more correct answer than women do (MWT: z=1.78, p=0.074). Similarly, neither women's nor men's performances differ among treatments (pairwise comparisons, MWT; women: all p>0.267; men: all p>0.074). Overall, we do not find any distortion in the performance of winning participants associated with the introduction of quotas.[28]

[28] When considering the effects of gender quotas on welfare, we observe that the average earnings in stage 3 are not negatively affected by the treatments, except than in Q1. Gender quotas do not have any effect on the average earnings of individuals in stage 4. Support for these results is provided in section A.7 of the Appendix, Table A9.

4. Conclusion

In most developed countries, women's participation in the labor force now matches that of men. A substantial gap remains, however, when both the sector of employment, for instance, the STEM field, and the percentage of women in upper-level positions are taken into account.

Currently, one of the main aspirations of gender policy is to promote women's achievement in different sectors of society without violating meritocratic principles. To accomplish this aim, it is of upmost importance to understand the effect of policy interventions made at different phases of a career path. In this paper, we ran a laboratory experiment to investigate this largely unexplored question.

Our study sheds light on how gender quotas could affect males and females' decision to self-select into a competitive career path. We show that gender quotas are effective in favoring women's career progression within organizations or to encourage their accessing selective educational institutions, but suggests that policies which limit their implementation to the career entry level might be less effective. Our results point in the direction indicating that entry-level quotas might discourage women from competing at higher levels, undermining the necessary self-confidence in their ability to successfully scale the career ladder. More specifically, a gender quota at the career entry level might activate an unintended negative effect, similar to the stereotype threat. Our findings suggest that introducing a gender quota at the highest level successfully prevents best-performing women getting off the career ladder too early, without negatively affecting men`s decisions. Moreover, none of the interventions analyzed have entailed any efficiency loss, measured in terms of both performance of the winning participants and overall earnings. A policy of introducing a gender quota only at the top turns out to be as effective as a policy introducing quotas at all stages of a career.

To the best of our knowledge, we are the first to investigate the role of gender quotas in a career path framework, showing that the stage at which a gender quota is introduced differently impacts women`s willingness to get to the top. We acknowledge, however, that our results have been obtained in a stylized and abstract environment, and that further research complementing our findings is needed to more comprehensively understand the effect of gender policies in determining women's paths to the top of the career ladder. More specifically, our study abstracts away from other relevant factors that may foster quota's success or failure, such as strategic concerns about others' decision to compete; their impact on task assignment and teamwork within organizations (Kölle, 2016) and leadership's attitude or whether negative spillovers would emerge as a result of their implementation (Leibbrandts et al., 2018).

Acknowledgements

The authors gratefully acknowledge the financial support of the Jan Wallander and Tom Hedelius Foundation, Handelsbanken, Sweden, (Research Grant P2016-0126):1 for funding the experiment and the French National Research Agency (ANR, FELIS Grant, ANR-14-CE28-0010-01) for support in the implementation of the research. Valeria Maggian's contribution was made while she was a researcher at the Groupe d`Analyse et de Théorie Economique (GATE) in Lyon, and performed within the framework of the LABEX CORTEX (ANR-11-LABX-0042) of Université de Lyon, within the program Investissements d`Avenir (ANR-11-IDEX-007) operated by the French National Research Agency (ANR). Natalia Montinari's contribution was performed while she was a researcher at the Department of Economics of (the) Lund University, Sweden. We thank Yuki Takahashi for excellent research assistance. We also thank participants at the 4th Annual Behavioral and Experimental Economics Workshop in Lyon, the Behavioural and Experimental Economics Research Seminar (BEERS) in Lyon, the 7th Annual Meeting of the French Association of Experimental Economics (ASFEE) in Cergy-Paris, the 2016 ESA World meeting in Jerusalem, the 2016 ESA European meeting in Bergen, and the XI Nordic Behavioral and Experimental Economics Conference in Oslo, for their useful comments.

Declaration of interests

Declarations of interest: none

5. References

Almås, I., Cappelen, A., Salvanes, K., Sørensen, E., and Tungodden, B. (2016). What Explains the Gender Gap in College Track Dropout? Experimental and Administrative Evidence. *American Economic Review*, 106, 296-302.

Altonji, J. G., and Blank, R. M. (1999). Race and gender in the labor market. *Handbook of Labor Economics*, *3*, 3143-3259.

Azmat, G., Calsamiglia, C., and Iriberri, N. (2016). Gender differences in response to big stakes. *Journal of the European Economic Association*, *14*(6), 1372-1400.

Balafoutas, L., and Sutter, M. (2012). Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science*, *335*(6068), 579-582.

Balafoutas, L., Kerschbamer, R., and Sutter, M. (2012). Distributional preferences and competitive behavior. *Journal of Economic Behavior & Organization*, *83*(1), 125-135.

Barbulescu, R., and Bidwell, M. J. (2013). Do Women Choose Different Jobs From Men? Mechanisms of Application Segregation in the Market for Managerial Workers. *Organization Science*, 24 (3), 737-756.

Bartling, B., Fehr, E., Maréchal, M. A., and Schunk, D. (2009). Egalitarianism and competitiveness. *American Economic Review*, *99*(2), 93-98.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289-300.

Berge, L., Oppedal, L.I., Bjorvatn K., Pires A.J., and Tungodden, B. (2015). Competitive in the lab, successful in the field? *Journal of Economic Behavior and Organization*, 118: 303–17.

Bertrand, M., Black, S. E., Jensen, S., and Lleras-Muney, A. (2014). *Breaking the glass ceiling? The effect of board quotas on female labor market outcomes in Norway* (No. w20256). National Bureau of Economic Research.

Bock, O., Baetge, I., and Nicklisch, A., 2014. hroot: Hamburg registration and organization online tool. *European Economic Review*, 71: 117-120.

Borghans, L., Heckman, J. J., Golsteyn, B. H., and Meijers, H. (2009). Gender differences in risk aversion and ambiguity aversion. *Journal of the European Economic Association*, *7*(2-3), 649-658.

Born, A., Ranehill, E., and Sandberg, A. (2018). A Man's World? The Impact of a Male Dominated Environment on Female Leadership. Working paper, Stockholm University.

Bracha, A., Cohen, A., and Conell-Price, L. (2019). The heterogeneous effect of affirmative action on performance. *Journal of Economic Behavior & Organization*, *158*, 173-218.

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

Buser, T., Peter, N., and Wolter, S.. (2017). Gender, Competitiveness, and Study Choices in High School: Evidence from Switzerland. *American Economic Review*, 107, 125-130.

Bylsma, W. H., and Major, B. (1992). Two routes to eliminating gender differences in personal entitlement: Social comparisons and performance evaluations. *Psychology of Women Quarterly*, *16*(2), 193-200.

Cavalcanti, T., and Tavares, J. (2016). The output cost of gender discrimination: a model-based macroeconomics estimate. *The Economic Journal*, *126*(590), 109-134.

Crosetto, P., and Filippin, A. (2013). The "bomb" risk elicitation task. *Journal of Risk and Uncertainty*, *47*(1), 31-65.

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

Czibor, E. and Dominguez-Martinez, S., (2019). Never too Late: Gender Quotas in the Final Round of a Multistage Tournament. *The Journal of Law, Economics, and Organization*, 35(2), 319–363.

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).

European Commission (2009): The Gender Challenge in Research Funding: Assessing the European National Scenes. Luxembourg: OPOCE.

Eckel, C. C., and Grossman, P. J. (2008). Men, women and risk aversion: experimental evidence. *Handbook of Experimental Economics Results*, *1*, 1061-1073.

Erosa, A., Fuster, L., and Restuccia, D. (2002). Fertility decisions and gender differences in labor turnover, employment, and wages. *Review of Economic Dynamics*, *5*(4), 856-891.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171-178.

Flabbi, L. and Tejada, M. (2012), "Fields of Study Choices, Occupational Choices and Gender Differentials", background paper for the OECD (2012) Gender Gap: Act Now.

Fouad, N. A., Chang, W. H., Wan, M., and Singh, R. (2017). Women's reasons for leaving the engineering field. *Frontiers in Psychology*, *8*, 875.

Galor, O., and Weil, D. N. (1993). *The gender gap, fertility, and growth* (No 4550). National Bureau of Economic Research, Inc.

Gelman, A. (2019). Don't calculate post-hoc power using observed estimate of effect size. *Annals of Surgery*, *269*(1), e9-e10.

Gelman, A., and Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641-651.

Goldin, C., and Rouse, C. (2000). Orchestrating impartiality: the impact of" blind" auditions on female musicians. *American Economic Review*, *90*(4), 715-741.

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

Hoenig, J. M., and Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55*(1), 19-24.

Houston, J., Harris, P., McIntire, S., and Francis, D. (2002). Revising the competitiveness index using factor analysis. *Psychological Reports*, *90*(1), 31-34.

Ibañez, M., Rai, A., and Riener, G. (2015). *Sorting through affirmative action: three field experiments in Colombia* (No. 183). DICE Discussion Paper.

Kölle, F. (2016). Affirmative Action and Team Performance. Discussion Papers 2016-07. The CEDEX, School of Economics, University of Nottingham.

Kulis, S., Sicotte, D., and Collins, S. (2002). More than a pipeline problem: Labor supply constraints and gender stratification across academic science disciplines. Research in Higher Education, 43(6), 657–91.

Lawless, J. L., and Fox, R. L. (2005). *It takes a candidate: why women don't run for office*. New York: Cambridge University Press.

Leibbrandt, A., Wang, L. C., and Foo, C. (2018). Gender Quotas, Competitions, and Peer Review: Experimental Evidence on the Backlash Against Women. *Management Science*, 64(8), 3501-3516.
Ley, T. J., and Hamilton, B. H. (2008). The gender gap in NIH grant applications. *Science*, *322*(5907), 1472-1474.

Lyons, S. T., Ng, E. S., and Schweitzer, L. (2014). Changing demographics and the shifting nature of careers: implications for research and human resource development. *Human Resource Development Review*, *13*(2), 181-206.

Lu, J., Qiu, Y., & Deng, A. (2019). A note on Type S/M errors in hypothesis testing. *British Journal of Mathematical and Statistical Psychology*, *72*(1), 1-17.

Maggian, V., and Montinari, N. (2017). The spillover effects of gender quotas on dishonesty, *Economic Letters*, 159C. Pages. 33-36.

Marsh, H. W., Jayasinghe, U. W., and Bond, N. W. (2008). Improving the peer-review process for grant applications: reliability, validity, bias, and generalizability. *The American Psychologist,* 63 (3), 160-168.

Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation (2017). Higher Education and Research in France, Facts and Figures. Chapter 13: Gender equality in higher education. Editor: Emmanuel Weisenburger

National Research Council. (2009). Gender differences at critical transitions in the careers of science, engineering and mathematics faculty. Washington, DC: National Academies Press.

National Science Foundation. Division of Science Resources Statistics. (2009). Women, minorities, and persons with disabilities in science and engineering: 2009 (NSF 09-305). Arlington, VA.

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

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), 1-16.

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.

Reuben, E., and Timko, K. (2017). On the Effectiveness of Elected Male and Female Leaders and Team Coordination (No. 10497). IZA Discussion Papers.

Shalvi, S., Dana, J., Handgraaf, M.J.J., and De Dreu, C.K.W. (2011). Justified ethicality: observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, 115 (2): 181-190.

UNESCO (2017). Cracking the code: girls' and women's education in science, technology, engineering and mathematics (STEM); report of the UNESCO International Symposium and Policy Forum. ISBN 978-92-3-100233-5.

Wallon, G., Bendiscioli, S., and Garnkel, M. S. (2015). Exploring Quotas in Academia, Policy report, EMBO. Retrieved online at http://www.embo.org/documents/science_policy/exploring_quotas.pdf.

Waisbren, S. E., Bowles, H., Hasan, T., Zou, K. H., Emans, S.J., Goldberg, C., Gould, S., Levine, D., Lieberman, E., Loeken, M., Longtine, J., Nadelson, C., Patenaude, A. F., Quinn, D., Randolph, A. G., Solet, J. M., Ullrich, N., Walensky, R., Weitzman, P., and Christou, H. (2008). Gender differences in research grant applications and funding outcomes for medical school faculty. *Journal of Womens Health*, 17, 207-214.

Williams, W. M., Mahajan, A., Thoemmes, F., Barnett, S. M., Vermeylen, F., Cash, B. M., and Ceci, S. J. (2017). Does gender of administrator matter? National study explores US university administrators' attitudes about retaining women professors in STEM. *Frontiers in Psychology*, *8*, 700.

Zhang, Y. J. (2013). Can Experimental Economics Explain Competitive Behavior Outside the Lab? SSRN: https://ssrn.com/abstract=2292929 or http://dx.doi.org/10.2139/*ssrn.2292929*

6. Appendix A

A.1 Summary Statistics

Summary statistics are reported in Table A1 and in Table A2. In Table A1, we report information about the field of studies of our participants by gender and treatment. All comparisons are based on Pearson χ^2 tests. In the bottom part of Table A1, when making pairwise comparisons of the fields of study by gender and between treatments, we observe no significant difference in the women's fields of study, except between the NQ treatment with the Q1Q2 treatment (χ^2, p-value=0.06). When comparing males, we observe that the distribution of their fields of study is not perfectly balanced. Finally, when analyzing the fields of study of males and females within the same treatment, we observe that in all treatments but Q1Q2 there is evidence of women's fields of study having a different distribution than men's.

In the main text (section 3.1), however, we show that males' performance in the task does not differ depending on the treatment and that males and females performed equally in the compulsory tournament of stage 2 in each treatment, but in Q2 (Mann Whitney test, p=0.016). This evidence suggests that the participants' field of study does not affect their ability to solve additions. Nevertheless, we include dummies for the field of study in all our regression models.

Field of study	**Treatments**								**All**
	NQ		**Q1**		**Q2**		**Q1Q2**		
	Female	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	
1= Management	29.17 N=14	47.92 N=23	45.83 N=22	33.33 N=16	31.25 N=15	41.67 N=20	47.92 N=23	37.50 N=18	39.32 N=151
2= Engineering	29.17 N=14	29.17 N=14	18.75 N=9	60.42 N=29	33.33 N=16	52.08 N=25	31.25 N=15	50 N=24	38.02 N=146
3=Other (Psychology, Economics, Sociology, Medical Science, Languages)	41.67 N=20	22.92 N=11	35.42 N=17	6.25 N=3	35.42 N=17	6.25 N=3	20.83 N=10	12.50 N=6	22.66 N=87
Female vs. Males Pearson χ^2 test	$\chi^2(3)$=4.802 p=0.091		$\chi^2(3)$=21.274 p=0.000		$\chi^2(3)$=12.490 p=0.002		$\chi^2(3)$=3.687 p=0.158		$\chi^2(3)$=29.272 p=0.000
Males	All pairwise treatment comparisons: all p-values> 0.452, except: NQ vs Q1: p-= 0.004; NQ vs Q2: p-value= 0.02; NQ vs Q1Q2: p-value= 0.095.								
Females	All pairwise treatment comparisons: all p-values> 0.171, except: NQ vs Q1Q2: p-value= 0.06.								

Table A1. Summary statistics for field of study by gender and treatment.

Table A2 reports summary statistics on the overall and women's average performance, participants' beliefs about own relative ranking in stage 2 and participants' risk aversion and competitive attitude in each treatment. First, we observe that the number of best-performing women (defined as those participants who gave 5 or more correct answers in stage 4, consistently with the categorization used in the main text as well) is different depending on the treatment, while the number of best performing men remains almost constant across treatments. It is important to note that this difference is not driven by differences in women's fields of study across treatments, (overall: $\chi^2(2)$= 3.039, p=0.219, N=384; females $\chi^2(2)$=1.0521, p=0.591, N=192; males $\chi^2(2)$= 0.904, p=0.636). Interestingly, while in Q1 a significantly lower proportion of women chooses to compete at the top than in Q2 (see Result 2 in the main text), the number of best performing women is higher in Q1 than in Q2. In the same vein, while the number of best performing women in NQ and Q2 is identical, it is only when a gender quota is present at the second stage of competition that a significant number of them is actually willing to compete at the top (see Result 2 in the main text).

	NQ	Q1	Q2	Q1Q2	All Treatments	Test
Sessions	4	4	4	4	16	-
Participants	96	96	96	96	384	-
Women	48	48	48	48	192	-
# Best Performers	43	54	44	55	196	Over treatments: $\chi^2(3)=5.086$, p=0.166 All pairwise comparisons all p-values> 0.538
# Women Best Performers	16	26	16	30	88	Over treatments: $\chi^2(3)=12.756$, p=0.005 Pairwise comparisons: NQ (Q2) vs Q1: p-value= 0.04; NQ (Q2) vs Q1Q2: p-value= 0.004;
Average Performance in stage 1	3.292 (.187)	3.156 (.168)	3.156 (.189)	3.25 (.172)	3.214 (.089)	Mann Whitney tests: Pairwise comparisons all p-values> 0.5266
Average Women's Performance in stage 1	2.917 (.228)	2.938 (.218)	2.979 (.260)	3.021 (.226)	2.964 (.116)	Mann Whitney tests: All pairwise comparisons all p-values> 0.5163
Average Performance in stage 2	3.844 (.183)	4.052 (.199)	3.875 (.203)	3.979 (.160)	3.938 (.093)	Mann Whitney tests: All pairwise comparisons all p-values> 0.3471
Average Women's Performance in stage 2	3.75 (.226)	3.813 (.264)	3.438 (.284)	3.813 (.224)	3.703 (.125)	Mann Whitney tests: All pairwise comparisons all p-values> 0.1430
Average Beliefs (it takes value between 1 and 12, where 1 identifies the best performance and 12 identify the worst performance).	6.125 (.276)	6.094 (.297)	6.281 (.315)	5.510 (.262)	6.003 (.144)	Mann Whitney tests: All pairwise comparisons all p-values> 0.1209; Q2 vs Q1Q2: p-value=0.088
Average Women's Beliefs (it takes value between 1 and 12, where 1 identifies the best performance and 12 identify the worst performance).	6.292 (.380)	7.146 (.378)	7.021 (.428)	5.813 (.332)	6.568 (.193)	Mann Whitney tests: All pairwise comparisons all p-values> 0.1018; Q1 vs Q1Q2: p-value= 0.0198; Q2 vs Q1Q2: p-value= 0.0703.
Average Competitiveness (it takes value between 11 and 55, where higher values identify the psychological trait of competitiveness)	38.698 (.942)	38.229 (.953)	38.083 (.923)	39.781 (.899)	38.698 (.464)	Mann Whitney tests: All pairwise comparisons all p-values> 0.2046
Average Women's Competitiveness Index (it takes value between 11 and 55, where higher values identify the psychological trait of competitiveness)	35.438 (1.258)	35.688 (1.504)	34.542 (1.391)	39.646 (1.325)	36.328 (.696)	Mann Whitney tests: All pairwise comparisons all p-values> 0.4952 NQ vs Q1Q2: p-value= 0.0187; Q1 vs Q1Q2: p-value= 0.0909; Q2 vs Q1Q2: p-value= 0.0120.
Average Willingness to take risk (BRET) It takes value from 0 to 100, where higher number indicates willingness to take risk	43.896 (1.549)	43.323 (1.412)	42.708 (1.466)	46.260 (1.469)	44.047 (.738)	Mann Whitney tests: All pairwise comparisons all p-values> 0.1488 Q2 vs Q1Q2: p-value=0.0853
Average Women's Willingness to take risk (BRET) It takes value from 0 to 100, where higher number indicates willingness to take risk	42.896 (2.303)	43.938 (2.207)	38.833 (2.037)	44.813 (2.180)	42.620 (1.096)	Mann Whitney tests: All pairwise comparisons all p-values> 0.1431 Q2 vs Q1Q2: p-value=0.0451

Table A2. Summary Statistics.

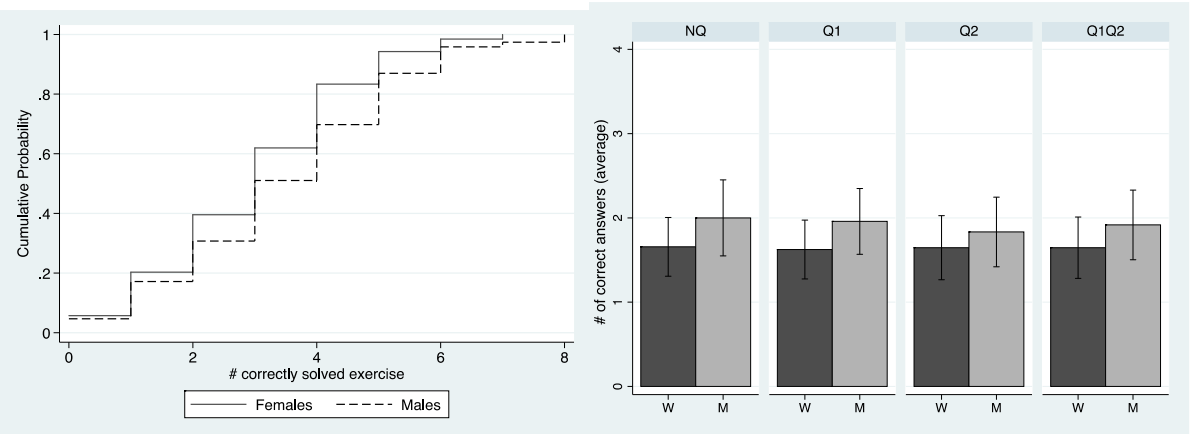
Finally, we observe that women's risk aversion is significantly higher in the Q1Q2 treatment than in the Q2 treatment and that they also have more positive beliefs about their own ranking in the Q1Q2 treatment.
In the post-experimental questionnaire, we also measured the participants' attitudes toward competition. More specifically, participants were asked to grade on a 5-point scale (where 1 = Strongly Disagree and 5 = Strongly Agree) whether they disagree or agree with each of the following statements, taken from the Revised Competitive Index (Houston et al., 2002; Harris and Houston 2010):

- I like competition;
- I am a competitive individual;
- I enjoy competing against an opponent;
- I don't like competing against other people;
- I get satisfaction from competing with others;
- I find competitive situations unpleasant;
- I dread competing against other people;
- I try to avoid competing with other;
- I often try to outperform others;
- I do not like games that are winner-takes-all;
- Competition destroys friendship.

It is interesting to note that females' scores in the Competitive Index are significantly lower than males' in all treatments, except in Q1Q2 (Mann-Whitney test. NQ: p=0.000; Q1: p=0.022; Q2: p=0.000; Q1Q2: p=0.854).

A.2 Performance in stage 1 (piece rate) and stage 2 (compulsory tournament)

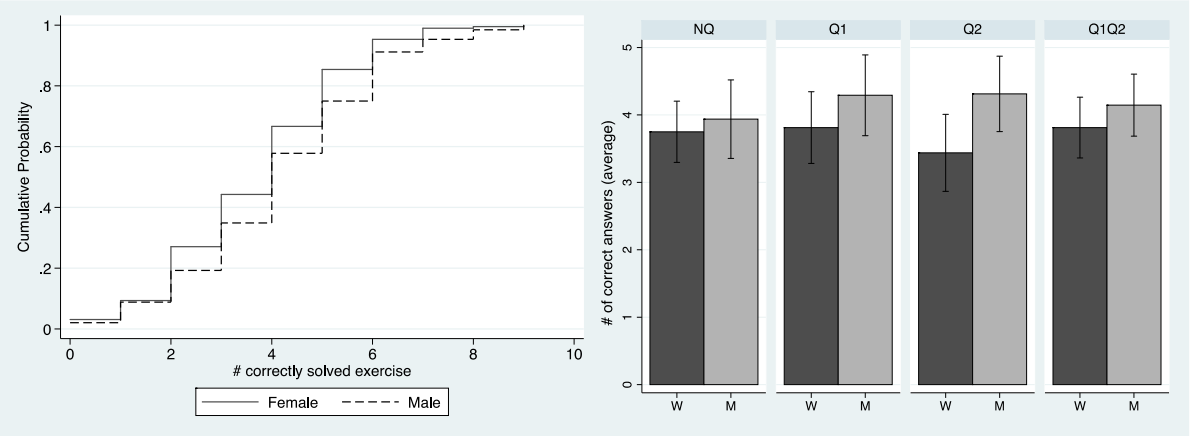
In this section, we report additional results supporting the findings reported in section 3.1. Specifically, we provide details about performances in stage 1 and stage 2 of the experiment, where participants have been exposed to a piece rate and a tournament without the option of a preferred compensation scheme. Figure A1 reports details of the participants' performance in stage 1, where they perform the mathematical task for 4 minutes and are paid according to a piece rate scheme. Panel (a) reports the cumulative distribution of the performance by gender, while panel (b) reports the average performance by gender and treatment. It can be noted that in all treatment men perform better than women do. However, the difference is only significant in the NQ treatment (MW test, z=1.961, p=0.050) and overall (z=2.649, p=0.008).



Panel (a) **Panel (b)**

Figure A1. Performance in stage 1 (piece rate) by gender.
Note. Panel (a) reports the overall cumulative distribution by gender. Panel (b) reports the average performance by gender in each treatment. (N=384 participants; 96 in each treatment); the bars show, for each treatment and gender, the average number of correct answers. Error bars, mean ± SEM.

Figure A2 reports details on the participants' performance in stage 2, where they perform the mathematical task for 4 minutes and are paid according to a tournament scheme. Panel (a) reports the cumulative distribution of the performance by gender, while panel (b) reports the average performance by gender and treatment.
In each matching group, the 4 participants with the best performance (i.e. highest number of correct calculations) get 1.5 euro per correct calculation, while the others get 0. It can be noted that in all treatments men perform better than women do. However, the difference is only significant in the Q2 treatment (MW test, z=2.403, p=0.016) and overall (z=2.403, p=0.016).



Panel (a) **Panel (b)**

Figure A2. Performance in the stage 2 (compulsory tournament) by gender.
Note. Panel (a) reports the overall cumulative distribution by gender. Panel (b) reports the average performance by gender in each treatment. Panel (b) reports the average performance by gender in each treatment. (N=384 participants; 96 in each treatment); the bars show, for each treatment and gender, the average number of correct answers. Error bars, mean ± SEM.

Comparing Figure A1 and Figure A2, it can be noted that the average performance in stage 2 increases compared to stage 1. This finding is in line with that reported by previous works (i.e. Balafoutas and Sutter, 2012 and Niederle et al., 2013). As can be seen from Table A3, which reports the results of a set of Mann - Whitney tests, learning occurs for both men and women in all treatments except for men in the NQ treatment.

Mann-Whitney test stage 1 vs stage 2

	Women	**Men**	**Overall (Women + Men)**
NQ	z=3.427	z=0.986	z=3.013
	p=0.000	p=0.324	p=0.003
Q1	z=2.972	z=3.140	z=4.326
	p=0.003	p=0.002	p=0.000
Q2	z=2.020	z=3.714	z=4.024
	p=0.043	p=0.000	p=0.002
Q1Q2	z=2.571	z=3.133	z=4.016
	p=0.010	p=0.017	p=0.001
Overall	z=5.488	z=5.379	z=7.681
	p=0.000	p=0.000	p=0.000

Table A3. Comparison of the performance in the Piece Rate and the Tournament

A.3 Participants' performance and choice to compete

To provide an initial insight into how men and women differ in their willingness to compete in stage 4, conditional on having chosen to compete in stage 3, depending on their performance level, we split our pool of 384 participants in quartiles based on their performance in stage 2. The first and fourth quartiles respectively identify the worst and best performing participants. According to a Pearson's chi-squared test, the assignment to quartiles is not significantly associated to the treatment ($\chi^2(9) = 10.953$, $p=0.279$), nor to the participant's gender ($\chi^2(3)=7.537$, $p=0.057$).

Since when slicing data in performance the power of our statistical analysis is restricted (see also section A.4), in both Table A4 and Figure A3 we report, respectively, only a numerical and a graphical representation of men and women who decide to compete in stage 4, conditional on having competed in stage 3, divided into quartiles based on their performance in stage 2. A more formal investigation of the treatment effects conditional on participants' performance is provided in the implementation of the econometric analysis in Table 3 in the main text, where we also account for other relevant factors measured at the individual level (such us the field of studies, etc.).

By looking at Figure A3, it can be noted how, irrespective of their performance in stage 2, compared to the NQ treatment, the percentage of women who choose to continue to compete in stage 4 only increases in the Q2 and Q1Q2 treatments. When comparing the Q1 and the NQ treatments, it seems that there is no variation in women's choices to compete between the lowest and with the highest performance (i.e. in the first and fourth quartile, respectively), while women of intermediate performance appear to be less likely to choose competition in stage 4. For men, it can be noted that, irrespective of their performance, treatments Q1 and Q1Q2 always increase their willingness to continue competing compared to the NQ treatment, while the effect of the Q2 treatment is more ambiguous.

Treatment	**NQ**		**Q1**		**Q2**		**Q1Q2**	
	Women	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**
First Quartile	50.00 N=1/2	50.00 N=7/7	66.67 N=4/6	91.67 N=11/12	90.91 N=10/11	75.00 N=6/8	75.00 N=6/8	71.43 N=10/14
Second Quartile	50.00 N=5/10	85.71 N=6/7	38.89 N=7/10	87.50 N=7/8	72.73 N=8/11	61.54 N=8/13	83.33 N=15/18	88.89 N=8/9
Third Quartile	45.45 N=5/11	45.45 N=5/11	28.57 N=2/7	75.00 N=3/4	66.67 N=4/6	57.14 N=4/7	80.00 N=8/10)	85.71 N=6/7
Fourth Quartile	60.00 N=3/5	87.50 N=7/8	57.14 N=4/7	80.00 N=8/15	71.43 N=5/7	72.73 N=8/11	71.43 N=5/7	90.91 N=10/11
Total	50.00 N=14/28	62.50 N=25/40	44.74 N=17/38	84.62 N=33/39	77.14 N=27/35	66.67 N=26/39	79.07 N=34/43	82.93 N=34/41

Table A4. Proportion of women and men who choose to continue to compete in stage 4, conditional on having competed in stage 3, divided into quartiles based on their performance in stage 2.

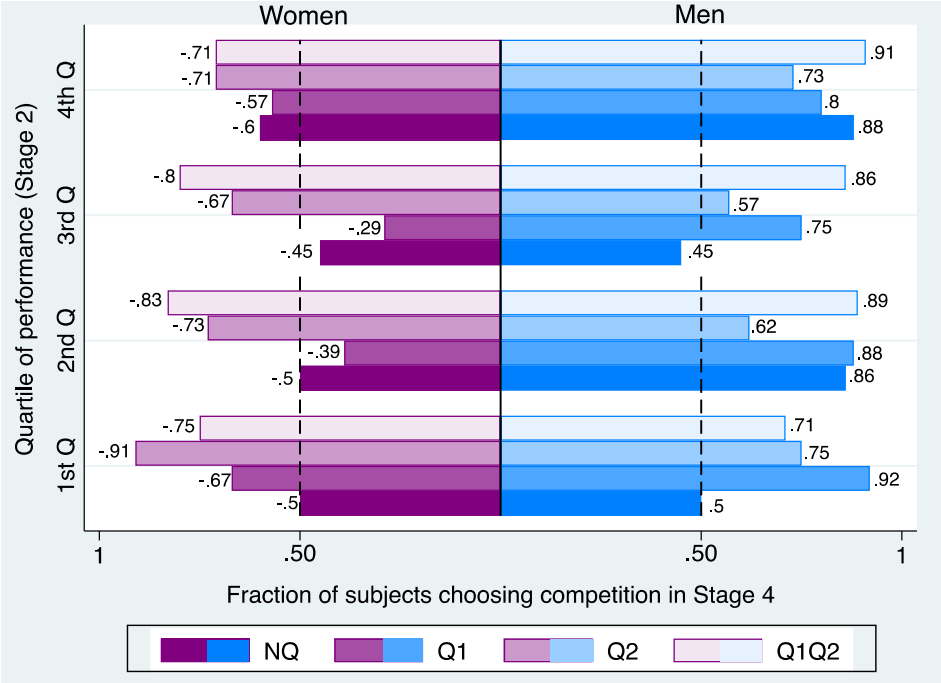


Figure A3. Proportion of women and men who choose to continue to compete in stage 4, conditional on having competed in stage 3, divided into quartiles based on their performance in stage 2.
Note. Fraction of men and women who chose to compete in stage 4, across all treatments, conditional on having chosen the competition in stage 3 and depending on their performance in stage 2 (N=303 participants, 1st Q=152; 2nd Q=87; 3rd Q=69, 4th Q=76). The bars show, for each treatment and quartile, the proportion of participants (between 0 and 1) who chose the tournament in stage 4.

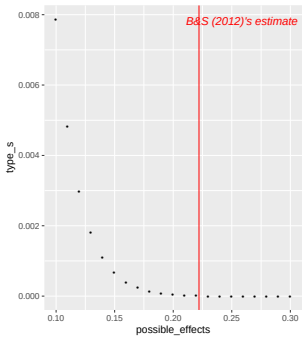
A.4 Retrospective power analysis and multiple comparison p-value adjustment

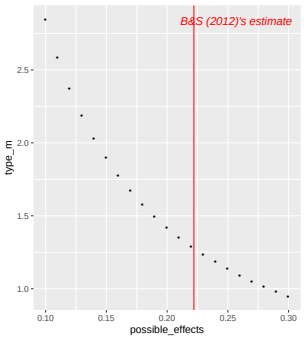
The argument that interpretation of a statistically significant estimate from a small sample requires caution is that when the true effect is also small, a statistically significant estimate is likely to have occurred by chance (type I error) because estimates have to be very large in order to be statistically significant (Gelman and Carlin, 2014). As noted by several studies, confidence intervals – or equivalently standard errors – tend to be very wide in such cases (Amrhein et al., 2019 and Owen, 2016). Since we are dealing with a small sample size (especially when we divide our original sample to consider treatments and the final choice of competition in stage four), we performed two robustness checks to address this concern: (i) retrospective power calculation suggested by Gelman and Carlin (2014) and further developed by Lu et al. (2019) in Section A.4.1, and (ii) multiple comparison p-value adjustment in Section A.4.2. Otherwise, using the standard power formula to make ex-post power calculations would give a much noisier and much larger result than for ex-ante power calculations (Gelman, 2019; Hoening and Heisey, 2001).

A.4.1 Retrospective Power calculation

Below, we perform retrospective power calculations for Tables 2 and 3 using our standard error estimates. For Table 2, we use the t-test to obtain standard errors (the t-test yields almost equivalent statistical significances as the chi-square test). We use these standard errors and theoretically plausible effects we might have expected as outcome of our treatments to calculate the probability of falsely finding results of the opposite sign (type S error) and to compute the expected magnitude of overestimation (type M error). For theoretically plausible possible effects, we follow Balafoutas and Sutter (2012)'s effect size of gender quota introduction for tournament entry, which is 22.2% (see Figure 2 in the main text) and consider a range of about +/− 10% as a possible outcome's variation, with 22.2% being the most plausible effect size. For all power calculations, we implement a two-sided test and consider 5% to be the minimum significance level.

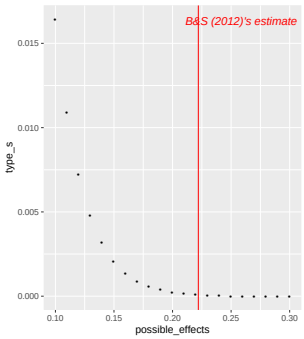
As shown in the different panels of Figure A4, for an effect size of 22.2%, the type S error probability is almost 0 for all statistically significant estimates in Table 2 (Figures A4, panels A4.1, A4.3, A4.5, A4.7), and the type M error rate is at most 1.65 (Figures A4, panels A4.2, A4.4, A4.6, A4.8). This suggests that our estimates are at most twice as large as the true effect size and usually less than 1.5 times as large as the true effect size.

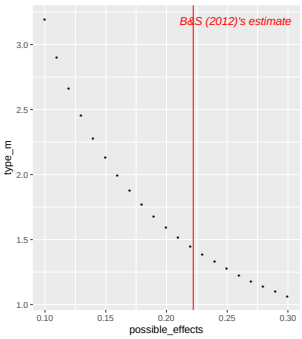




Panel (a) Type S Error Panel (b) Type M Error

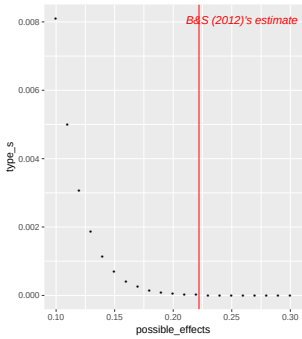
Figure A4.1 Table2 coefficient: NQvsQ1Male

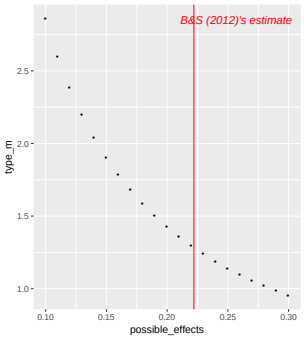




Panel (a) Type S Error Panel (b) Type M Error

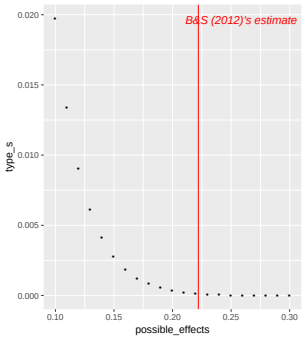
Figure A4.2 Table2 coefficient: NQvsQ1Q2Female

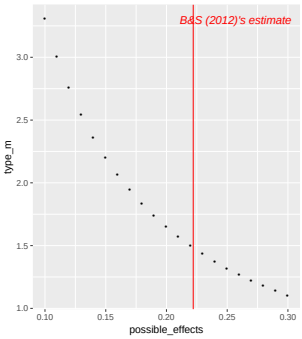




Panel (a) Type S Error Panel (b) Type M Error

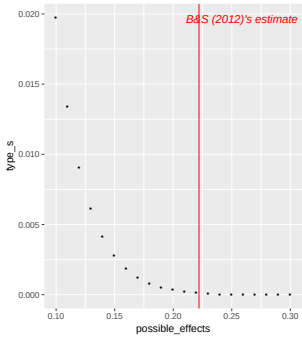
Figure A4.3 Table2 coefficient: NQvsQ1Q2Male

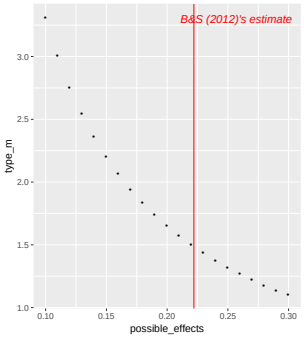




Panel (a) Type S Error Panel (b) Type M Error

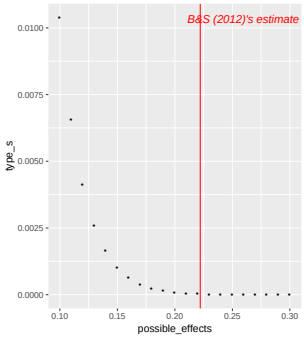
Figure A4.4 Table2 coefficient: NQvsQ2Female

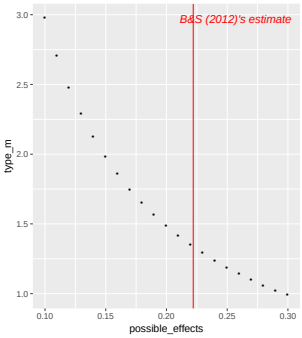




Panel (a) Type S Error Panel (b) Type M Error

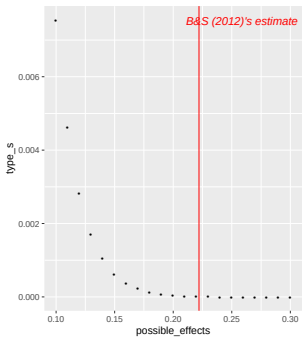
Figure A4.5 Table2 coefficient: NQvsQ2Female

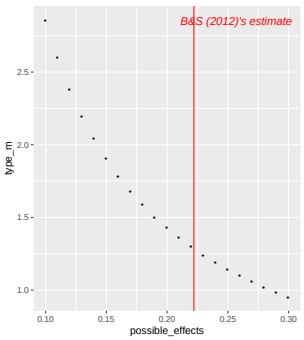




Panel (a) Type S Error Panel (b) Type M Error

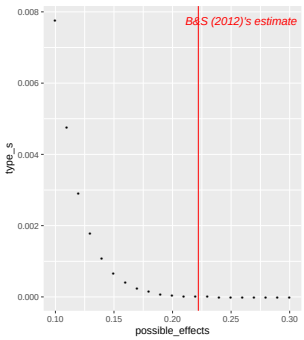
Figure A4.6 Table2 coefficient: Q1vsQ1Q2Female

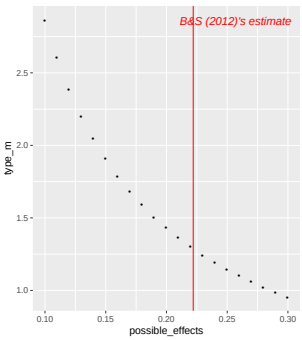




Panel (a) Type S Error Panel (b) Type M Error

Figure A4.7 Table2 coefficient: Q1vsQ2Male



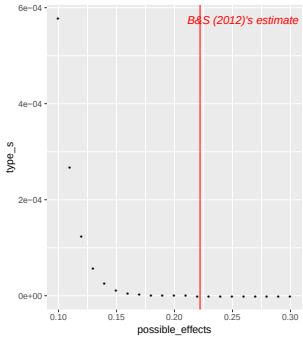


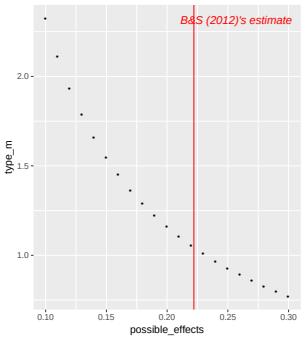
Panel (a) Type S Error Panel (b) Type M Error

Figure A4.8 Table2 coefficient: Q2vsQ12Q2Male

Figure A4. Graphical representation of the type S error probability (panels on the left) and type M error probability (panels on the right) for all statistically significant estimates in Table 2.

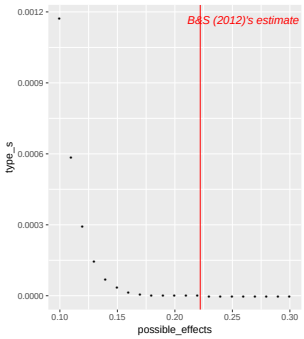
A similar trend is displayed in Figure A5, which refers to the significant estimates of Table 3.

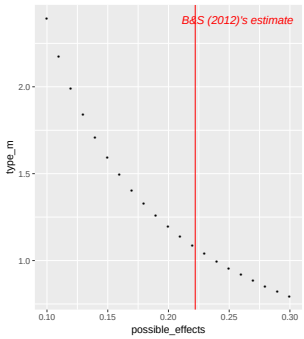




Panel (a) Type S Error Panel (b) Type M Error

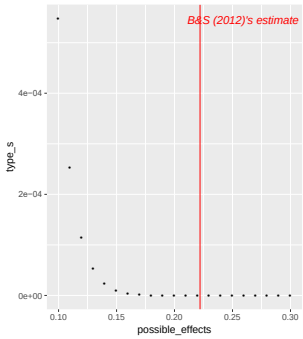
Figure A5.1 Table3, Model 1, coefficient: Q1Q2

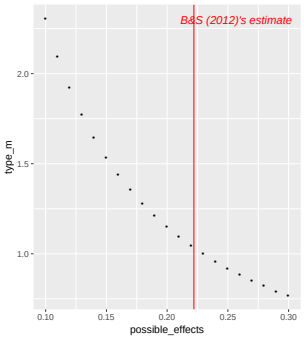




Panel (a) Type S Error Panel (b) Type M Error

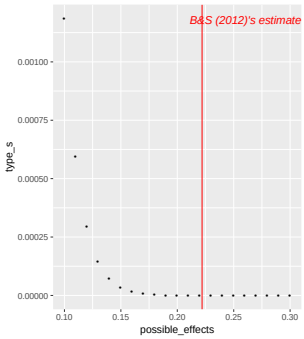
Figure A5.2 Table3, Model 1, coefficient: Q2

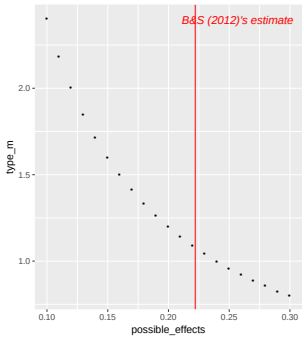




Panel (a) Type S Error Panel (b) Type M Error

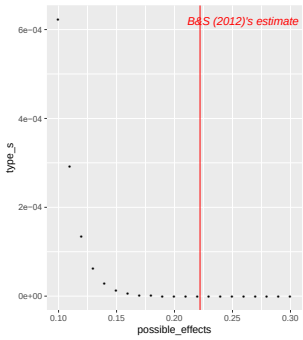
Figure A5.3 Table3, Model 2, coefficient: Q1Q2

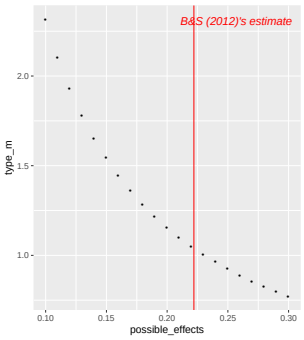




Panel (a) Type S Error Panel (b) Type M Error

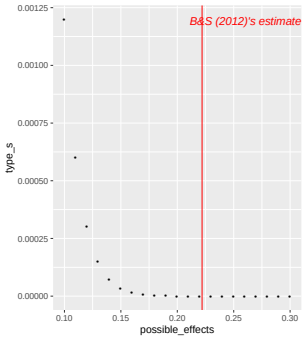
Figure A5.4 Table3, Model 2, coefficient: Q2

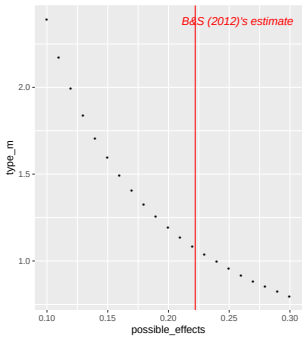




Panel (a) Type S Error Panel (b) Type M Error

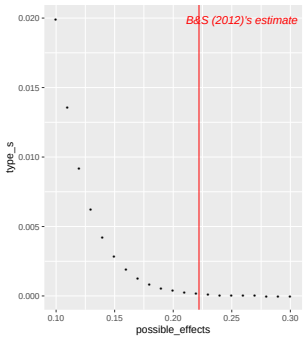
Figure A5.5 Table3, Model 3, coefficient: Q1Q2

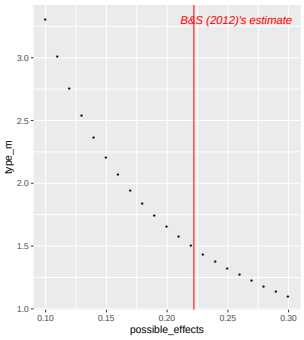




Panel (a) Type S Error Panel (b) Type M Error

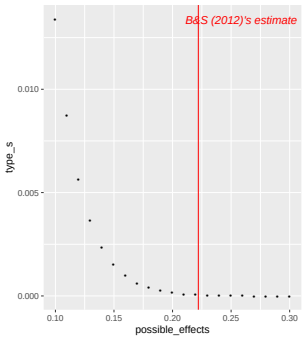
Figure A5.6 Table3, Model 3, coefficient: Q2

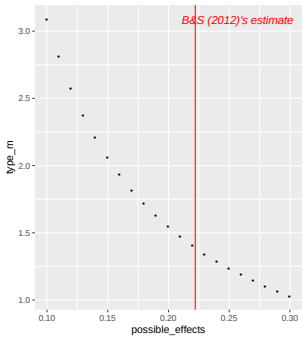




Panel (a) Type S Error Panel (b) Type M Error

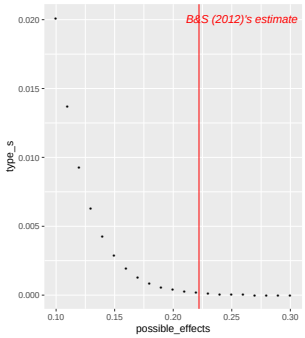
Figure A5.7 Table3, Model 4, coefficient: Q1Q2 x Female



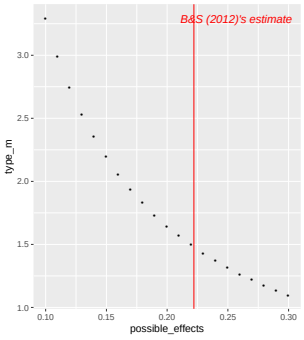


Panel (a) Type S Error Panel (b) Type M Error

Figure A5.8 Table3, Model 4, coefficient: Q2 x Female

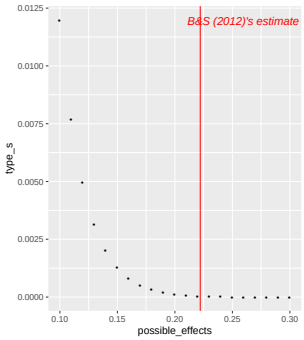


Panel (a) Type S Error

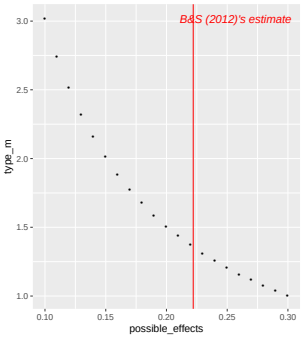


Panel (b) Type M Error

Figure A5.9 Table3, Model 5, coefficient: Q1Q2 x Female

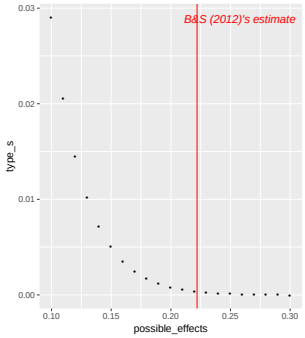


Panel (a) Type S Error

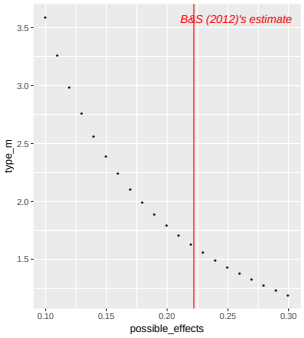


Panel (b) Type M Error

Figure A5.10 Table3, Model 5, coefficient: Q2 x Female

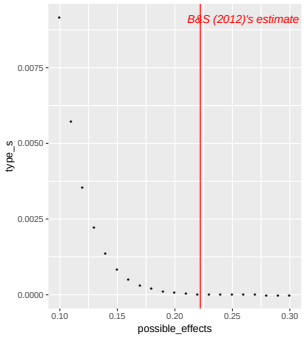


Panel (a) Type S Error

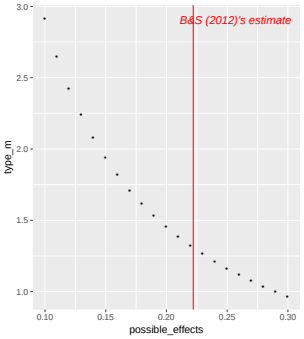


Panel (b) Type M Error

Figure A5.11 Table3, Model 6, coefficient: Q1Q2 x Female



Panel (a) Type S Error



Panel (b) Type M Error

Figure A5.12 Table3, Model 6, coefficient: Q2 x Female

Figure A5. Graphical representation of the type S error probability (panels on the left) and type M error probability (panels on the right) for statistically significant coefficients estimated in the different models of Table 3.

A.4.2 Multiple comparison adjustment

As a further robustness check, we performed multiple comparison p-value adjustment. Since we were initially interested in the effect of gender quota introduction and did not know ex-ante which quotas were more effective in positively affecting women's willingness to compete at the top, the appropriate adjustment is the false discovery rate (Benjamini and Hochberg, 1995). In Table A5, we present the results of the p-value adjustment.

	Treatment comparisons					
	NQvsQ1	**NQvsQ2**	**NQvsQ1Q2**	**Q1vsQ2**	**Q1vsQ1Q2**	**Q2vsQ1Q2**
Females						
Females unadjusted	0.6720	0.0247	0.0105	0.0047	0.0014	0.8376
Females - adjusted vs NQ	0.6720	0.0371	0.0316	-	-	-
Female- unadjusted All Permutation	0.8064	0.0371	0.0211	0.0141	0.0084	0.8376
Males						
Males unadjusted	0,02611	0,6987	0,03878	0,0648	0,8378	0,0932
Males adjusted vs NQ	0,0582	0,6987	0,0582			
Males adjusted All Permutation	0,1163	0,8378	0,1163	0,1296	0,8378	0,1398

Table A5. P-values reported in Table 2 of the main text before and after the false discovery rate adjustment.

First, when comparing the unadjusted and adjusted p-values in rows 1 and 2 of table A5 it can be noted that statistical significances are mostly unchanged. When considering the comparison between the treatments with a quota rather than their relative effect with respect to the NQ treatment, statistical significances are still almost unchanged (row 3). The same is true for the effect of the gender quotas on men, represented in the last three rows of table A5.

Table A6 replicates the same robustness check presented in Table A5, but adjusting the p-values contained in Table 3 in the main text.

It can be noted that, for models 1-3, where have not distinguished the effect of the quota by gender, statistical significances are mostly unaffected by multiple comparison correction. Similarly, for models 4-6, when we adjust the p-values for females only, statistical significances are again mostly unchanged. Last, since in models 4-6 of Table 3 we aim to investigate the effect of imposing a gender quota in different stages of a multi-stage tournament on female and male participants, we should adjust only for female interactions. However, to further test the robustness of our outcomes, we instead adjust all the treatments together for models 4-6. Statistical significances remain mostly unchanged. These measures confirm that our estimates were not obtained by chance.

	P-values of coefficients in the regressions					
	Q1	**Q2**	**Q1Q2**	**femaleXQ1**	**femaleXQ2**	**femaleXQ1Q2**
Model 1 Unadjusted	0.120	0.055	0.000	-	-	-
Model 1Adjusted	0.120	0.083	0.000	-	-	-
Model 2 Unadjusted	0.142	0.046	0.000	-	-	-
Model 2 Adjusted	0.142	0.069	0.000	-	-	-
Model 3 Unadjusted	0.120	0.033	0.000	-	-	-
Model 3 Adjusted	0.120	0.049	0.000	-	-	-
Model 4 Unadjusted	0.207	0.852	0.104	0.769	0.026	0.062
Model 4 Adjusted for female Only	-	-	-	0.769	0.021	0.046
Model 4 Adjusted All	0.311	0.852	0.180	0.852	0.042	0.093
Model 5 Unadjusted	0.282	0.845	0.123	0.989	0.018	0.060
Model 5 Adjusted for female Only	-	-	-	0.989	0.011	0.043
Model 5 Adjusted All	0.423	0.989	0.246	0.989	0.022	0.087
Model 6 Unadjusted	0.214	0.838	0.084	0.945	0.009	0.109
Model 6 Adjusted for female Only	-	-	-	0.945	0.002	0.163
Model 6 Adjusted All	0.320	0.945	0.208	0.945	0.005	0.217

Table A6. P-values reported in Table 3 of the main text before and after the false discovery rate adjustment.

A.5 Frequencies of winning and ex-ante win probabilities

In section 3.2, we define as best-performing participants those subjects who gave 5 or more correct answers in stage 4. This threshold makes the choice of competing payoff-maximizing in all treatments, as shown in Table A7, which reports the frequency of winning the first (Panel a) and second (Panel b) round of the multi-stage tournament.

In order to define the threshold outlined above for the NQ treatment, we follow the procedure described by Niederle et al. (2013). Specifically, we calculate the women (men)'s frequency of winning for each level of performance, when drawing 500,000 groups consisting of 6(5) men and 5(6) women, using the performance

distribution in stage 2 of the entire sample (i.e. 384 participants) with replacement. We compare the frequencies of winning obtained in this way with the ex-ante stage 3 and 4 tournaments' win probabilities, that is 4/12=0.33 and 2/12=0.17. When, for a specific level of performance, the frequencies of winning obtained via simulations are higher than the ex-ante win probabilities, we conclude that choosing to compete would be a payoff-maximizing choice for the participants with that specific level of performance. As a result of this procedure, panels a) and b) of Table A7 show the probability of winning stage 3 and stage 4 competition conditional on stage 2's performance for both men and women in the different treatments.

Panel a) Stage 3: First round of the multistage tournament

	3 correct	**4 correct**	**5 correct**	**6 correct**
Men in all treatments	2.65%	35.62%	85.69%	99.65%
Women in NQ and Q2	2.31%	33.59%	83.67%	99.57%
Women in Q1 and Q1Q2	11.81%	45.54%	84.35%	98.16%

Panel b) Stage 4: Second round of the multistage tournament

	3 correct	**4 correct**	**5 correct**	**6 correct**
Men in all treatments	0.06%	4.08%	33.48%	84.05%
Women in NQ and Q1	0.05%	3.64%	30.72%	82.60%
Women in Q2 and Q1Q2	1.54%	12.67%	44.87%	78.37%

Table A7. Frequency of winning the first (Panel a) and second (Panel b) round of the multi-stage tournament (i.e. stage 3, corresponding to being classified among the top 4 in the group of 12 participants of stage 2 and stage 4, corresponding to being classified among the top 2 in the group of 12 participants of stage 2) for each gender and treatment, conditional on stage 2 performance.

Note that the frequencies of winning in stage 3 and stage 4 for men are not affected by treatments, while varying for women depending on the treatment. In particular, in the treatments where the gender quota is introduced in stage 3 (Q1 and Q1Q2), in order to be a winner of the tournament in stage 3, for a woman it is enough to be among the two best performers out of the six women in stage 2. Similarly, when the quota is introduced in stage 4 (Q2 and Q1Q2) to be the winner of the tournament in stage 4, it is sufficient to be the best women in the group of 6 women.

To calculate the threshold for women in the treatments with quota we proceed as follows: for any given performance level, say, 3, we draw 500,000 groups consisting of 5 women, using the performance distribution of the 192 women with replacement. We then calculate the women's frequency of wins in this set of simulated groups of women, applying the rules of the quota in the different stages. Note that the introduction of a quota increases the frequency of winning for women, but not so much that it exceeds the ex-ante win probabilities. For this reason, the thresholds at which the choice is made to compete payoff-maximizing do not change across genders and treatments.

From Table A7 it can be noted that giving 4 (5) or more correct answers represents the threshold that makes the choice of competing in stage 3 (4) payoff-maximizing in all treatments and for both genders. What changes for women, after the introduction of the quotas, is that the frequencies of winning increase for each level of performance (i.e. for a performance of 3 it increases from 2.31 to 11.81 in stage 3), but not so much to exceed the ex-ante probabilities of winning, altering the threshold. Finally, this threshold identifies as best performers those participants belonging to the third and fourth quartiles, as defined in Table A4 and Figure A3, section A.3.

A.5.1 Alternative definition of Best-Performing participants

Since we focus on the intervention which encourages women to reach the top-positions, we identify as best-performing participants those who, in stage 4, met the threshold previously identified (i.e. who gave 5 or more correct answers). Irrespective from the gender and the treatments, for these participants, competing in stage 4 would have been the payoff-maximizing choice.

An alternative method to identify best-performing participants would have been to consider those subjects who gave 5 or more correct answers in stage 2. This classification would leave our results qualitatively unchanged, as displayed in Figure A6 and by the supporting tests. As shown in Figure A6, in treatment Q1 the proportion of best-performing women entering competition in stage 4 is lower than the proportion of best-performing men ($\chi^2(1)= 3.556$, p=0.059, N=36). Taking treatment Q1 as the benchmark, the fraction of best-performing women who choose to compete increases in treatment Q1Q2 (Fisher's exact test, p=0.070) as well as in treatment Q2, despite such a difference failing to achieving significance ($\chi^2(1)= 1.710$, p = 0.191). The choice of the best performing men is not affected by the treatments ($\chi^2(3)= 1.551$, p=0.671). Moreover, 57.14% (8/14) of the best-performing women choose not to continue competing after stage 3 in treatment Q1, while only 30.77% (4/13) take the same decision in treatment Q2 and 23.53% (4/17) in treatment Q1Q2.

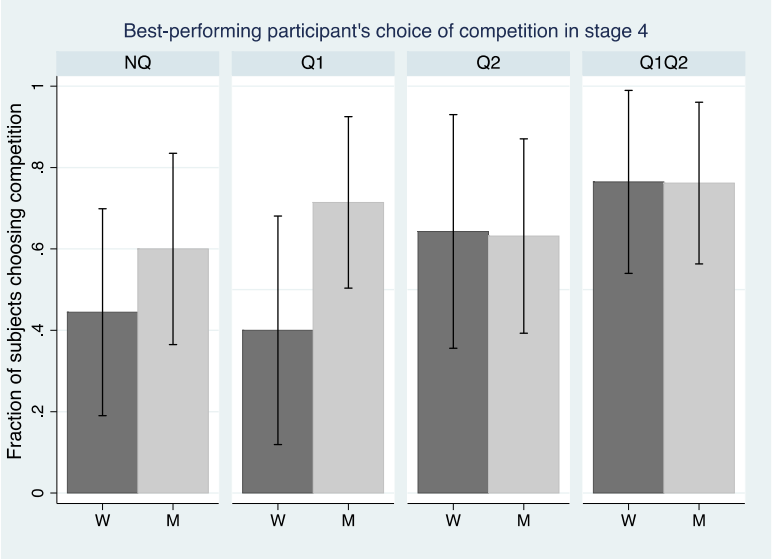


Figure A6. Competition in stage 4, best-performing participants.
Note. Fraction of best-performing female and male participants (identified using the performance in stage 2) who choose to compete in stage 4, across all treatments (N=145 participants; NQT: 38 participants; Q1T: 36 participants; Q2T: 33 participants and Q1Q2T: 38 participants). The bars show, for each treatment and gender, the proportion of participants (between 0 and 1) who choose the tournament compensation scheme in each of both stages of the multi-stage tournament. Error bars, mean ± SEM.

However, we decided not to use the subjects' performance in stage 2 for the identification of best-performing participants in our analysis because subjects' performance in stage 2 is not a good predictor of their performance in the following stages. In particular, the 31.5% (121/384) of those subjects who would have been defined as best-performing in stage 2 would have not been included in the same category based on their performance in stage 4. 71.07% (N=86/121) would have been classified as best-performing participants in stage 4 but not in stage 2, and 28.93% (N=35/121) would have been classified as best-performing participants in stage 2 but not in stage 4. The misclassification does not depend on subjects` choosing to compete in stage 3 and in stage 4, both overall and separately for gender (stage 3: $\chi^2(1)= 2.212$, p=0.137; men: $\chi^2(1)=0.3700$, p=0.543, women $\chi^2(1)= 1.683$, p=0.194; stage 4: $\chi^2(1)= 0.230$, p=0.632; women $\chi^2(1)= 1.550$, p=0.213, men $\chi^2(1)= 0.066$, p=0.798). Rather, it seems to be affected by treatments with the difference driven by women, who are less likely to be misclassified in treatment Q2 compared to the other treatments ($\chi^2(3)=7.083$, p=0.069; women $\chi^2(1)= 8.348$, p=0.039, men $\chi^2(1)= 2.636$, p=0.451).
For this reason, we define the best-performing participants as those subjects who gave 5 or more correct answers in stage 4.

A.6 Beliefs

In the following, we check the robustness of investigation made into the role of beliefs in the main text. More specifically, we i) provide a robustness check on the role of treatments in affecting best performers' beliefs, when categorization depends on participants' performance in stage 2 and ii) replicate the analysis contained in Figure 4 and in Figure 6 in the main text (for the best performing participants) when considering our sample as a whole, in Section A.6.1; Finally, in section A.6.2, we report the results from a set of Tobit regressions on the whole set of our participants..

A.6.1 The impact of affirmative action on participants' beliefs.

In our experiment, beliefs regarding the relative rank achieved in stage 2 are elicited at the end of stage 3, after participants have submitted their choice of the payment scheme for stage 4 (but before being informed whether they are among the winners of the stage 3 competition, and before performing the task in stage 4). Subjects had to indicate their expected rank within the whole group of twelve members, but also within the sub-group of six members composed only of their own gender. Correct guesses were rewarded €1 each, and feedback was not given until after the end of the experiment.

We observe, in Figure A7, that adopting a different definition for best performers' participants (i.e. considering their performance in stage 2, instead of stage 4) than the one used in the main text does not substantially affect our results. In particular, similarly to Figure 5 in the main text, while in the NQ treatment there is no difference in the average beliefs about own ranking between best performing men and women (MWT, z=0.03, p=0.976), in Q1 women are significantly more pessimistic than men about their own ranking (MWT, z=3.85, p=0.000). Moreover, as in the main text, best-performing women rank themselves on average 1.5 positions lower compared to both the NQ treatment, while comparison with other treatments lead to no significant differences (6/12 in Q1 vs. 4.61/12 in NQ; MWT: z=2.20, p=0.028).

Interestingly, the Q1 treatment has the opposite effect on best-performing men, increasing their average beliefs about own ranking by about 1 position (3.57/12 in Q1 vs. 4.55/12 in NQ; MWT: z=1.81, p=0.071).

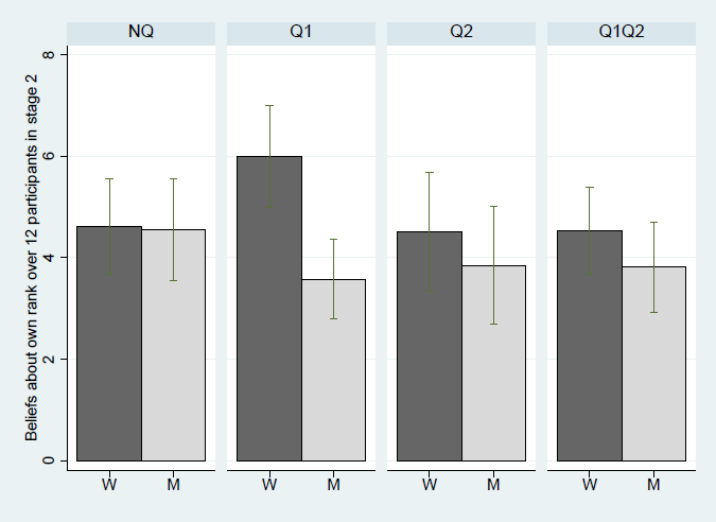


Figure A7. Beliefs about own ranking over 12 participants in stage 2.

Figure A7 reports the average beliefs of best-performing female and male participants (defined according to their performance in stage 2) about their relative ranking over a group of 12 participants in stage 2 (compulsory tournament). The bars show, for each treatment and gender, the average rank guessed by participants (between 1 and 12, where 1 identifies the best performance and 12 identifies the worst performance). Error bars, mean ± SEM. N=145 participants ; NQT: 38; Q1T: 36; Q2T: 33 and Q1Q2T: 38.

Considering all participants, we find that, on average, men believe that they ranked one position above women (5.438 (2.868) vs. 6.568 (2.675), MWT: z=4.239, p=0.000). Figure A8 reports the average beliefs of the whole

pool of our participants by gender and treatment. It can be noted that differences in beliefs depending on gender are significant only for the Q1 and Q2 treatments (Q1: z=4.022, p=0.000; Q2: z=2.344, p=0.019), but not for the NQ and Q1Q2 treatments (NQ: z=0.528, p=0.597; Q1Q2: z=1.643, p=0.100).

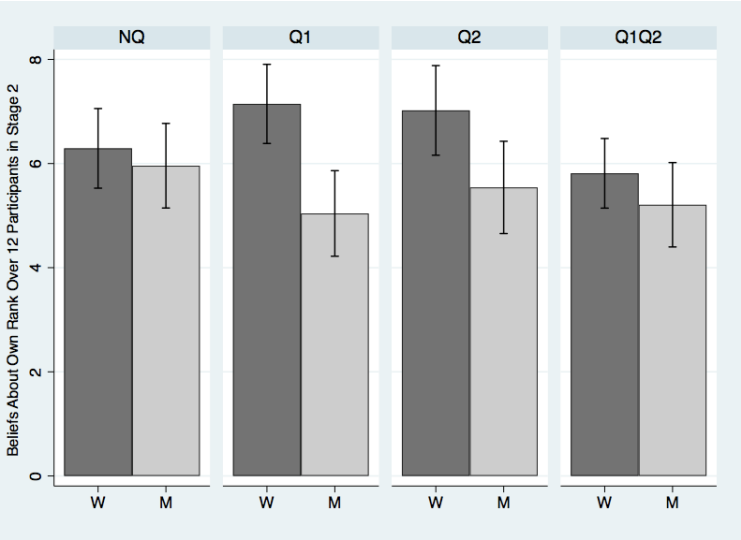
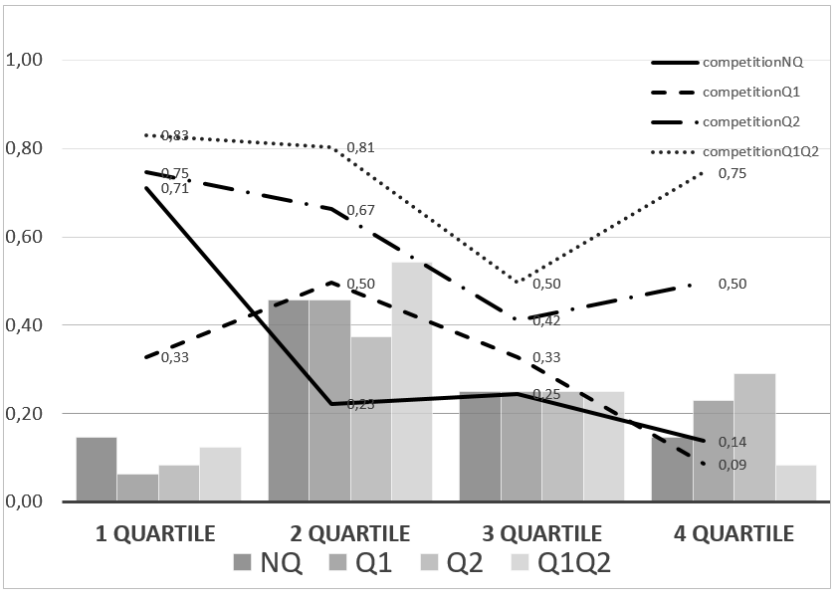
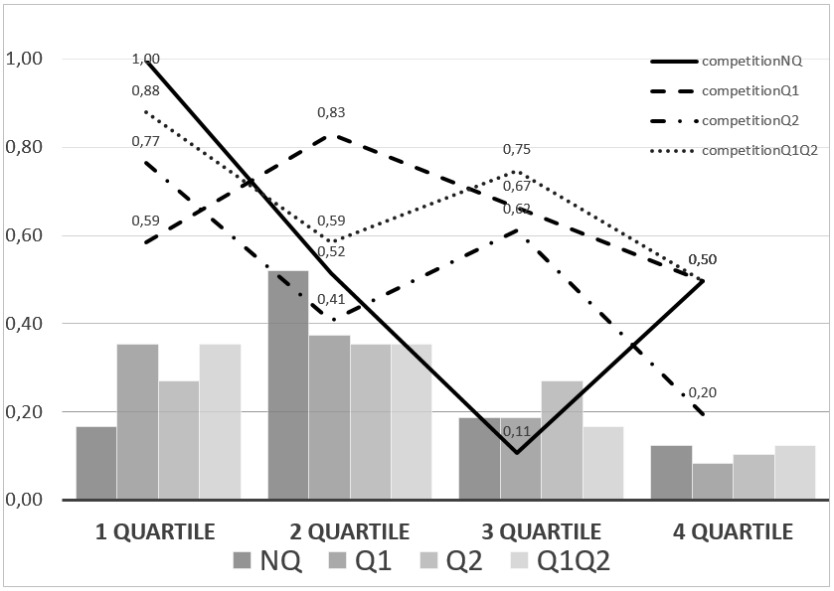


Figure A8. Participants' beliefs.
Note. Average beliefs of female and male participants about their relative ranking over a group of 12 participants in stage 2 (compulsory tournament) (N=384 participants, 96 in each treatment). The bars show, for each treatment and gender, the average rank guessed by participants (between 1 and 12, where 1 indicates the best performance and 12 the worst performance). Error bars, mean ± SEM.

In Figure A9, we replicate the analysis made in the main text with respect to the distribution of beliefs among all participants in each treatment by quartile, also considering the proportion of them deciding to compete in stage 4 depending on the quartile of their beliefs. In particular, in panel (a) we refer to women, while in panel (b) we refer to men.



Panel (a)



Panel (b)

Figure A9. Distribution of participants' beliefs and decisions to compete.

Note. Panel a (b) reports the distribution of beliefs of all female (male) participants about their relative ranking over a group of 12 participants in stage 2 (compulsory tournament) and the fraction of female (male) participants who decide to compete in stage 4. For each treatment, beliefs are grouped in quartiles (expected ranks 1-3 correspond to the first quartile; expected ranks 4-6 correspond to the second quartile; expected ranks 7-9 correspond to the third quartile; expected ranks 10-12 correspond to the fourth quartile).

In the beliefs distribution, when considering women, note how treatment Q1 induces a shift from the first to the fourth quartile, compared to the NQ and Q2 treatments. Regarding the choice to compete depending on beliefs, it can be seen how differences between treatments are more pronounced for the female participants who expect to be ranked in the fourth quartile: in Q1Q2 significantly fewer women decide to compete compared to the NQ and Q1 treatments, according to a set of Fisher's exact test (Q1 vs Q1Q2, p= 0.033; NQ vs Q1Q2, p=0.088). Similarly to what observed when considering only best-performers females, women who expect to be ranked in the first quartile seem also less likely to compete in the fourth stage of competition in the Q1 treatment than in other treatments but this difference fails to reach statistical significance due to the low number of observations.

When also consider panel (b) of Figure A9 we observe the same pattern highlighted in Figure 6: male participants are more likely to believe themselves to be ranked in the first and second quartile, while most female participants believe themselves to be ranked in the second or third quartiles. The decision to compete in stage 4 also significantly differ between males and females both in the NQ and Q1 treatments for participants who expect to be ranked in the second quartile (NQ: p=0.070; Q1: p=0.046), while we observe no significant gender differences when considering other quartiles of beliefs.

A.6.2 Tobit regressions

In Table A8, we report the results of a series of Tobit regressions on individuals' beliefs about their own ranking in the compulsory tournament of stage 2, in which everyone was assigned to a group size of 12. Reporting a higher value indicates an expected worse rank.

In all models, the independent variables include: a dummy for the choice to compete in stage 4, a dummy identifying gender, a dummy identifying best performing participants, both the number of correct answers and the actual ranking in stage 2, a set of dummies accounting for treatments Q1, Q2, Q1Q2, and controls for fields of study.

We observe that, in model 1, females expect to achieve a significantly worse rank than males, the number of correct answers in stage 2 is a good predictor of the participant's expected rank and participants who chose to compete in stage four expect to be better ranked than others.

In specification 2 of table A8, we include the interaction between the dummy identifying treatments and gender. Compared to model 1, the coefficient associated with female loses its statistical significance, while the other results are unchanged. In addition, the interaction between female and the treatment dummy for the Q1 treatment is positive and significant, highlighting that when comparing the NQ and the Q1 treatment, women in the Q1 treatment on average expect to be ranked lower than males. We do not find similar effects for the other treatments.

Finally, in model 3 we also control for individuals' risk preferences and attitudes toward competition: our results are basically unchanged.

	Model 1	Model 2	Model 3
Dependent Variable	Belief on Performance in stage 2 (higher value=lower rank)		
Independent Variables			
Choice of Competing is Stage 4	-.810***	-.779***	-.725***
	(.252)	(.254)	(.261)
Female	.539**	.028	-.033
	(.247)	(.477)	(.483)
Q1 treatment	.307	-.363	-.389
	(.339)	(.477)	(.477)
Q2 treatment	.332	-.008	.002
	(.338)	(.474)	(.474)
Q1Q2 treatment	-.254	-.359	-.368
	(.344)	(.473)	(.474)
Best performers	.241	.217	.209
	(.292)	(.277)	(.277)
Number of correct answers in stage 2	-.917***	-.911***	-.926***
	(.171)	(.170)	(.171)
Ranking in T2	.037	.038	.028
	(089)	(.089)	(.090)
Female x Q1 treatment	-	1.314*	1.349**
		(.673)	(.673)
Female x Q2 treatment	-	.646	.591
		(.670)	(.671)
Female x Q1Q2 treatment	-	.177	.2181
		(.674)	(.677)
Competitiveness	-	-	-.008
			(.014)
Willingness to Take Risk	-	-	-.006
			(.008)
Study: Management	.281	.214	.233
	(.272)	(.273)	(.274)
Study: Others	.528	.443	.460
	(.331)	(.331)	(.331)
Constant	2.289***	9.413***	10.141***
	(.088)	(1.215)	(1.417)
N. of observation	384	384	384
Uncensored	354	354	354
Left-censored(=1)	12	12	12
Right-censored(=12)	18	18	18
Log -likelihood	-830.0878	-827.801	-827.28168
LR	212.10***	216.68***	217.72***
Pseudo R2	0.1133	0.1157	0.1163

Standard errors appear in parentheses. ***, ** and * indicate significance at the 1 % level, at the 5 % level and at the 10 % level, respectively.

Table A8. Tobit regressions: Beliefs about own relative ranking in the compulsory tournament (stage 2, in a group of 12 participants) where a higher value indicates a worse rank.

A.7 Earnings

In section 3.4, we investigate the effects of gender quotas on efficiency, determined as the average performance of the winners in stage 4. In this section, we analyze the effects of gender quotas on the average individuals' earnings in stages 3 and 4.

Individuals' earnings in stages 3 and 4 are determined depending on whether each participant chooses to be paid under a Piece Rate or a Tournament payment scheme. In both stage 3 and 4, if Piece Rate is chosen, the individual gets 0.50 euro per correct answer. In stage 3, if Tournament is chosen, the individual gets €1.50 per correct answer if among the four best performers, zero otherwise. In stage 4, if Tournament is chosen and the individual was among the winners of stage 3's Tournament, the individual gets €3 per correct answer if among the two best performers, zero otherwise; however, even if the Tournament is chosen in stage 4, the Piece Rate payment scheme is applied to stage 4's performance when the individual is not among the winners of stage 3's tournament.

In Table A9, we report the results of a series of OLS regressions on individual earnings in stage 3, in model 1, and on individual earnings in stage 4, in model 2. The independent variables are the treatments Q1, Q2, Q1Q2 and the number of correct answers in previous stages.

We observe that, in model 1, the earnings in stage 3 are negatively affected by the Quota-at-initial-stage treatment (Q1) with respect to the No-Quota treatment. Differently, the Quota-in-final-stage treatment (Q2) and the Quota-in-both-stage treatment (Q1Q2) have no effects with respect to the average earnings of participants in stage 3. In model 2, neither treatment is significant in affecting the amount of an individual's earnings in stage 4 with respect to the No-Quota treatment. Moreover, as shown by the coefficient of *Choice to compete in stage 4*, individuals' earnings are on average positively affected when deciding to compete at the top. Our results confirm that gender quotas, while increasing the number of female best performers at the top, do not negatively affect the overall welfare of the society.

In models 3 and 4 we investigate whether our treatment interventions affect the distribution of earnings between males and females. When adding interactions between the participant`s gender and each treatment to the regression, in model 3, results are almost unchanged with the exception of *Female x Q1Q2*, which is positive and significant: the earnings of the female participants in stage 3 increase with respect to the males' when comparing the NQ and the Q1Q2 treatments. Our results suggest that gender quotas do not have a great impact on the earnings distribution between males and females, when considering stage 3.

Differently, in model 4, we observe that gender quotas affect how earnings are distributed between males and females, in stage 4. First, females earn significantly less than males in the NQ treatment, as shown by the significant and negative coefficient of *Female*. Second, the coefficients of the interaction terms represent the extent to which the difference between the average earnings of males and females changes with respect to the NQ treatments vs the Q1, Q2 and Q1Q2 treatments: the gender quotas also allow a reduction in the gender gap in terms of earnings. Finally, the negative and significant coefficient of *Q1, Q2* and *Q1Q2* provides evidence that these policy interventions marginally reduce males' earnings, with respect to the NQ treatment.

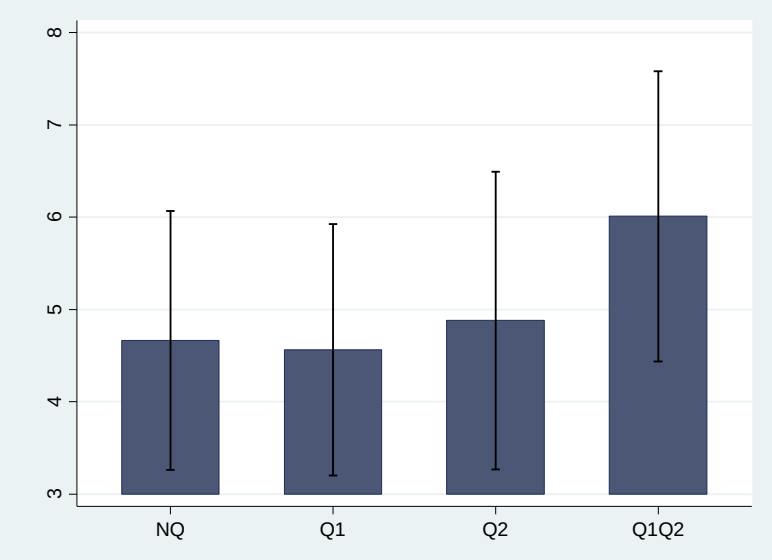
	Model 1	Model 2	Model 3	Model 4
Estimation Method:	OLS Regression			
Dependent Variable:	Individual`s earnings in stage 3	Individual`s earnings in stage 4	Individual`s earnings in stage 3	Individual`s earnings in stage 4
Independent Variables				
Q1 treatment	-1.053** (.540)	-.489 (.686)	-1.381* (.683)	-2.172** (1.048)
Q2 treatment	.291 (.645)	-.734 (.659)	.855 (.735)	-2.021* (1.101)
Q1Q2 treatment	.437 (.468)	-.097 (.700)	-.251 (.647)	-2.321** (.929)
Choice of competition in stage 3	.442 (.332)		.371 (.352)	
Choice of competition in stage 4		3.610*** (.476)		3.493*** (-488)
Female	-.058 (.309)	.303 (.526)	-.305 (.556)	-2.328*** (.812)
Performance in stage 1	.372** (.151)	-.058 (.206)	.381** (.146)	-.172 (.231)
Performance in stage 2	.928*** (.126)	.644*** (.219)	.912*** (.125)	.768*** (.254)
Performance in stage 3	-	1.681*** (.208)	-	1.921*** (.219)
Female x Q1 treatment			.676 (.681)	3.380*** (1.195)
Female x Q2 treatment			-1.117 (.680)	2.588* (1.389)
Female x Q1Q2 treatment			1.403* (.792)	4.503*** (1.185)
Constant	-2.099** (.827)	-6.127*** (.970)	-1.833** (.863)	-4.810*** (-997)
N. of observations	384	384	384	384
Root MSE	3.6843	5.9476	3.6693	5.9123
F	21.53***	25.67***	20.21***	20.12***
R2	0.2697	0.4015	0.2814	0.4133

Note: Table A5 presents the coefficients from a series of OLS regressions. The dependent variable is the Individual`s earnings in stage 3 or in stage 4. Errors appear in parentheses and are clustered at the group level (each group is composed by 12 individuals and, totally, there are 32 groups). ***, ** and * indicate significance at the 1% level, the 5% level and the 10% level, respectively.

Table A9. OLS regressions: Earnings in stage 3 (model 1 and 2) or in stage 4 (model 2 and 4).

In Figure A10, panel (a), we observe that, when considering the average individual`s earnings in stage 4, they are not affected by the introduction of a quota (t-test, allowing for unequal variances in our samples. All pairwise comparisons with NQ, p>0.203). Differently, when considering the distribution of earnings between males and females, in panel (b), we observe that both Q1 and Q2 allow reduction in the gender gap in earnings: the difference between males' and females' earnings in stage 4 is indeed only significant in NQ (t-test, allowing for unequal variances in our samples. NQ: p=0.021, Q1: p=0.212. Q2: p=0.205, Q1Q2: p=0.269).

Panel (a)



Panel (b)

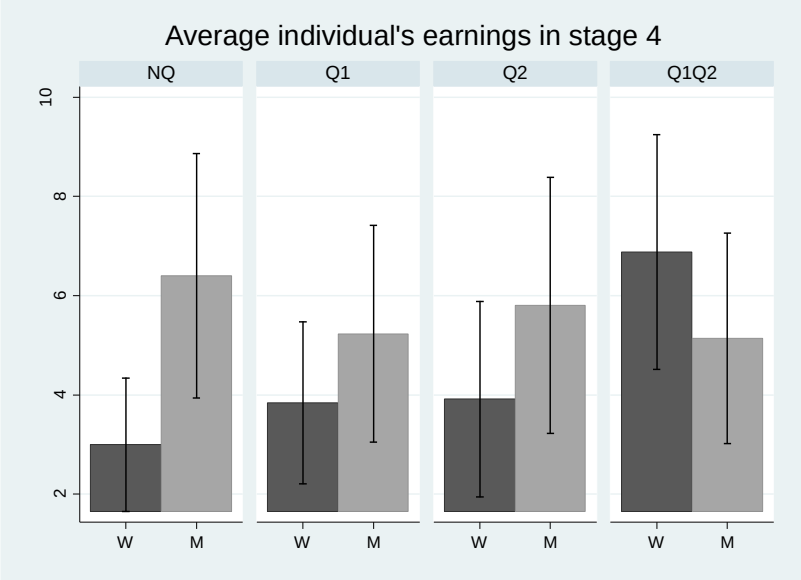


Figure A10. Average individual's earnings in stage 4
Note. Average individual`s earnings in stage 4 by treatment (panel a), and by treatment and gender (panel b). The bars show the average earnings in stage 4, independently of the payment scheme chosen by the individual. N=384 participants; 96 in each treatment, 48 females and 48 males. Error bars, mean ± SEM.

References for the Appendix

Amrhein, V., Greenland, S., and McShane, B. (2019). Scientists rise up against statistical significance, *Nature,* 567, 305-307.

Owen, A. B. (2016). Confidence intervals with control of the sign error in low power settings. *arXiv preprint arXiv:1610.10028*.