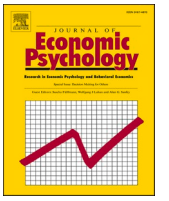
Journal of Economic Psychology 105 (2024) 102770

Contents lists available at ScienceDirect

Journal of Economic Psychology

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



Entering a gender-neutral workplace? College students' expectations and the impact of information provision☆

Francesca Barigozzi [a], José J. Domínguez [b], Natalia Montinari [a,*]

[a] University of Bologna, Department of Economics, Piazza Scaravilli 2 40126, Bologna, Italy
[b] University of Granada, Department of Economic Theory and History, Campus Universitario La Cartuja, 18071, Granada, Spain

ARTICLE INFO

JEL Codes:
D03
C91
J71

Keywords:
Gender Gaps
Expectations
Information provision
Competition
Hiring
Task assignment

ABSTRACT

This paper explores whether college students are aware of gender disparities in academic performance and labor market outcomes, and examines the effect of providing information about these gaps. The study uses a lab experiment that includes (i) a questionnaire eliciting beliefs, (ii) a task assignment game where participants act as employers, and (iii) a game measuring willingness to compete. The experiment features two feedback treatments: one providing information only on gender gaps in labor market outcomes, and the other including information on both academic performance and labor market outcomes. In another treatment, the questionnaire was administered without providing new information to make gender salient. Results indicate that most participants are unaware of gender gaps. Feedback treatments did not significantly affect hiring decisions but, making gender salient, positively influenced women's assignment to the difficult task, particularly among those previously unaware of the gaps, possibly due to social desirability bias. Men with implicit stereotypes were more inclined to compete regardless of treatment, while women with implicit stereotypes competed more after receiving information on the gap in academic performance. Overall, the study suggests that highlighting gender issues and informing women who hold implicit stereotypes can have mild positive effects.

1. Introduction

In most countries, girls outperform boys in school (Salvi del Pero and Bytchkova, 2013). Later in life, women also attain higher levels of education than men (Bertrand, 2020) and, in many countries, outperform men in college (see, among others, Conger and Long, 2010, for the USA; Piazzalunga, 2018, AlmaLaurea, 2022, Bovini et al. 2024, for Italy; Verbree et al., 2023, for the Netherlands; Carroll, 2023, for the UK). Nevertheless, women are still less likely than men to be employed, and when they are employed, they earn less (Bertrand, 2020). An important question is whether prospective workers are aware of these figures or, conversely, they expect to

☆ The data and code used in this study are available at https://researchbox.org/2174&PEER_REVIEW_passcode=ZHTGVQ. This repository contains the complete dataset and the code used for the data processing and statistical analysis. We thank Federico Lenzi and Francesca Madeddu for their excellent research assistance in preparing the Questionnaire and the Feedback treatments. We thank Irene Buso for her precious assistance in the Lab. We are grateful to the participants at the ESA Meeting 2022 for their useful comments and suggestions. Francesca Barigozzi and Natalia Montinari acknowledge funding from the University of Bologna RFO.

* Corresponding author.
E-mail addresses: francesca.barigozzi@unibo.it (F. Barigozzi), jjdominguez@ugr.es (J.J. Domínguez), natalia.montinari2@unibo.it (N. Montinari).

https://doi.org/10.1016/j.joep.2024.102770
Received 29 December 2023; Received in revised form 18 September 2024; Accepted 20 September 2024
Available online 26 September 2024

enter a gender-neutral workplace.

Workers' expectations largely influence labor market outcomes (Kim et al., 2019; Balleer et al., 2023; Mueller and Spinnewijn, 2023). Specifically, success in the workplace is influenced by investments and signals related to productivity, which prospective workers undertake based on their expectations of future job prospects. However, it remains uncertain to what extent the gender gap in labor market outcomes can be attributed to gender differences in beliefs. As an example, women's decision to opt out of competitive job selection procedures might be rational if they believe a gender gap in employment exists. The extensive media debate that ensued after movements like #MeToo# and the extended media coverage regarding the evidence of heightened gender gaps in the labor market due to the Covid-19 pandemic has highlighted the presence of gender inequalities in the last years. Then, it appears unclear whether prospective workers, like college students, hold correct expectations of early gender inequalities in the labor market and whether those expectations affect their labor market decisions.

Suppose that college students' expectations are not accurate. Would providing information be an appropriate intervention? There is little evidence on whether information provision on gender gaps can impact the behavior of men and women in the labor market, nor is it clear if such an intervention would be desirable. On one side, accurate information about *future career prospects* enables job seekers to make more efficient choices (Chen and He, 2021). For instance, having accurate information on the workplace empowers job candidates to decide if and to what extent they should compete with their peers during job selection procedures and throughout their employment. On the other side, enlightening women about persistent gender gaps in labor market outcomes carries the risk of triggering a self-fulfilling prophecy. This could potentially lead women to feel discouraged and ultimately accept the prevailing unequal labor market conditions, inadvertently perpetuating the status quo. Another piece of information that might be relevant for job seekers and employers is the existing *academic performance gap favoring women* in many countries (see, among others, Carroll, 2023, for the U.K.; Piazzalunga, 2018, for Italy; Conger and Long, 2010, for the USA; Verbree et al., 2023, for the Netherlands). Such information could influence women's readiness to compete with peers (see Wozniak et al., 2014). Additionally, knowledge of the gender gap in academic performance can aid employers in making effective decisions during hiring and task allocation. Finally, by jointly learning about women's superior academic performance and their disadvantages in the labor market, prospective workers and employers can infer that differences in labor market outcomes are unlikely to be attributable to academic ability.

Overall, it is important to study how information provision affects labor market behavior and whether it is an effective intervention to reduce the gender gap in labor market outcomes. In this study, we design a laboratory experiment on college students to examine i) expectations of gender gaps in academic performance and labor market outcomes and ii) the impact of information provision on labor demand and supply in experimental labor markets.

Our research aims to address the following questions: Do individuals on the brink of entering the labor market possess accurate beliefs regarding gender gaps in academic performance and job market opportunities? If not, what kinds of biases are present? Does feedback about the gender gap in academic performance and labor market outcomes influence employers and job market candidates in experimental labor markets? If so, how does the impact vary with the accuracy of students' expectations?

We collected data on expectations regarding gender gaps in academic performance and early labor market outcomes from 408 college students. As for early labor market outcomes, we collect field-specific beliefs on gender disparities one year and five years after graduation, including employment rates and wages, and information on wage differences and the likelihood of self-employment five years post-graduation. Then, we provide field-specific feedback to students in Economics and Management, Engineering, and Law using data from students who graduated from the University of Bologna (Italy) a few years before. These fields differ in male/female enrollment ratios, early employment, and wage gaps.

Treatments vary in three key aspects: first, the timing of eliciting participants' beliefs about labor market inequalities; second, the point at which feedback was provided—either immediately following belief elicitation or at the study's conclusion; and third, the content of the feedback—with and without information on academic performance. Hence, according to the feedback content, we distinguish between *Mixed Feedback* and *Outcome Feedback Treatment*, where *Outcome Feedback* indicates that only information on labor market outcomes is provided. Note that eliciting beliefs about gender gaps necessarily makes those gaps salient. Hence, in one treatment, participants play the Demand-side and Supply-side Games before belief elicitation (*Baseline Treatment*) and, in another treatment, beliefs are elicited before the two games, but feedback is provided only after the games are played (*Salience Treatment*). The structured comparison across treatments enables us to derive causal inferences regarding the possible effect of information about gender disparities and to disentangle its impact from the effect of making those gender gaps salient.

As mentioned before, in the *Mixed Feedback Treatment*, we offer information on gender gaps in labor market outcomes *and* academic performance. Notably, in this treatment, information provision offers both positive and negative evidence. Female students receive positive news about their relative college performance, while negative news pertains to early gender gaps in market outcomes. The reverse applies to males. Overall, the information provided in this feedback suggests that early labor market disparities aren't due to academic ability differences.

The role of information provision depends on the accuracy of students' expectations in several ways. First, feedback can be interpreted as information provision only for subjects not previously aware of those gaps. For students already aware or partially aware, feedback may also or mainly function as a "reminder" or a potential "debiasing intervention." Second, students may be aware of some gender gaps but not others. For instance, a student might know about early gender gaps in labor market outcomes but not about the gap in academic performance. If such a student is female, she may interpret this information as "positive and encouraging." By accounting for participants' pre-existing awareness of gender gaps and implicit biases, measured by the Implicit Association Test, we aim to determine whether our intervention's effectiveness arises from making gender gaps salient, from acting as a reminder, a debiasing mechanism, from information transmission, or a combination of these factors.

We examine the effect of information provision on students' behavior in stylized experimental labor markets: specifically on

individuals' decisions to engage in competitive scenarios (addressed in the Supply-side Game) and on their choices related to hiring and task allocation (examined in the Demand-side Game). All participants play both the Demand-side and the Supply-side Game whose order is randomized in each treatment. Note that the gender gap in early-career outcomes results from the interaction of employers and job seekers, i.e., the interaction of demand and supply. For example, employers might be less likely to select women, especially as leaders and women might refrain from competitive environments (Buser et al., 2014). Hence, it is important to study the impact of information on the two sides of the labor market.

Our Demand-side Game exploits the task assignment game designed by Domínguez and Montinari (2021). All participants assume the role of employers and must choose three workers from a mixed-gender pool of candidates. Among these candidates, two must be designated for a simple task (such as summing up three-digit numbers), while the other must be assigned to a challenging task involving mathematical problem-solving (such as quantitative multiple-choice questions). Employers' earnings are contingent on the performance of the workers they select. These employers are provided with information about the candidates' age, gender, and field of study and receive a performance-related signal. Finally, the Supply-side Game follows the approach of Niederle and Vesterlund (2007). Participants engage in the same simple task as earlier (summing up three-digit numbers). They choose between receiving compensation through a non-competitive (low) piece rate or opting for a competitive tournament with a potentially higher payoff. This game is used to determine if information plays a role in influencing gender-related disparities in willingness to compete.

We have derived three key sets of results from our study.

Concerning students' expectations, we have found that most participants lack awareness of gender gaps in academic performance and labor market outcomes. Despite widespread media coverage, it appears that many Italian college students still hold misconceptions about gender gaps in the labor market. A significant portion of participants expect to enter a gender-neutral workplace, with equal prospects and opportunities for both men and women. Participants are more likely to correctly predict labor market outcomes occurring five years after graduation compared to one year after graduation. Overall, *w*omen tend to exhibit a higher awareness of gender gaps in labor market outcomes than men, especially in Engineering.

Exploring the impact of Feedback treatments on the behavior of employers and job market candidates in our stylized experimental labor market, we observe the following. In the Demand-side Game, Feedback treatments do not impact hiring decisions, nor the probability of female candidates being assigned to the challenging task compared to the *Salience Treatment.* However, the *Salience Treatment* shows a positive effect relative to the *Baseline Treatment*, driven primarily by unaware subjects and possibly due to an increased sense of accountability and/or a social desirability bias. Our results thus suggest that salience is the driving mechanism increasing the probability for female workers to be assigned to the challenging task, and that it operates mainly for subjects who are unaware of early gaps in labor market outcomes. In the Supply-side Game, we find that information provision does not affect the gender gap in competitiveness. Men with higher implicit stereotypes are more willing to compete (regardless of treatment), while women with implicit biases respond positively to academic performance information in the *Mixed Feedback Treatment*. Hence, Feedback treatments seem to work as a debiasing mechanism for women.

Moving to awareness, results indicate that treatments increase the competitiveness of aware participants, while no significant effects are found for those unaware of gender gaps. Results from the Supply-side Game suggest that men naturally lean toward competition, especially when they read about women's superior college performance. Importantly, feedback does not reduce women's willingness to compete. Thus, our experiment does not support concerns about the potential negative consequences of providing information about existing gender inequalities in the labor market, particularly when the feedback includes details about women's superior academic achievements.

We conclude this Introduction with two important remarks. First, our experimental labor market is built on two games already employed to study labor market behaviors in previous literature. The Niederle and Vesterlund game we used to mimic the supply side of the labor market has been proven to capture in the lab a trait −i.e., competitiveness- which explains subsequent relevant outcomes in the "real" labor market, such as the gap in earnings among young professionals (Reuben et al., 2019) as well as the choice of more competitive university tracks (Buser et al. 2014). As for the demand side of the labor market, following the results of the baseline treatment in Domínguez and Montinari (2021), we expect that gender stereotypes regarding women's math ability may affect employers' hiring and task assignment decisions. When this is the case, feedback provision might work as a de-biasing mechanism, leading to hiring and assignment choices that are more favorable to female candidates. However, the Implicit Association Test is not significant in our estimations of the Demand-side Game, suggesting that salience is a more plausible mechanism. Secondly, our information treatment does not offer game-specific details about the participants' abilities or past behavior in the two stylized experimental games. Instead, it provides insights into the gender gap in college performance and early market outcomes *among peers* (individuals from the same university and field of study). In this context, our feedback treatment serves as a primer on the existence of field-specific gender gaps, also suggesting that early gender inequalities in the labor market are not likely attributable to academic ability. As mentioned before, the impact of this treatment depends on the participants' awareness of these gaps.

The paper is organized as follows. Section 2 describes the related literature. Section 3 explains the experiment design and execution. Section 4 derives our three sets of results. Section 5 concludes.

2. Literature review

A couple of empirical papers used survey data to elicit students' expectations of labor market outcomes and are thus related to the first part of our lab experiment aimed at eliciting students' awareness about gender gaps. Filippin and Ichino (2005) investigated whether male and female college students hold similar expectations about future wages and whether the gender wage gap is correctly anticipated. They explore the extent to which second-year students at Bocconi University, a prestigious Business and Management

private college in Milan (Italy), anticipate the gender wage gap. In addition, wage expectations of (N=1154) students are matched with the actual wages of a sample of similar Bocconi graduates. The authors find that the gender gap implied by students' expectations one year after graduation is consistent with the gender gap implied by the earnings of their elder counterparts. There is, instead, a misperception of the gender gap later in the career after graduation because students expect the gender gap to be roughly constant, while realizations indicate an increasing gap with experience. Our elicitation of students' awareness about the existence of a gender wage gap is similar to Filippin and Ichino's "explicit expected gap" elicitation because we both ask students whether they think a man or a woman with their same background is more likely to earn a given wage level. While Filippin and Ichino elicit expectations about the gender wage gap also indirectly by comparing personal wage expectations with the actual wage of similar graduates, we elicit students' expectations about other gender gaps such as academic performance and unemployment rate for graduates in their same field of study.

Only partially related to our paper is Sipe et al. (2009), who elicited students' expectations of gender *discrimination* in the labor market instead of students' expectations about gender gaps in labor market outcomes as we do. The authors analyze survey data (N=1,373) collected among undergraduate students enrolled in a Business major in the College of Business Administration in a large public university in the USA. Their findings suggest that students underestimate the extent and consequences of gender discrimination and tend to consider themselves immune to gender discrimination.

Some empirical studies document the emergence of an early gender gap in labor market outcomes. Among them, Combet and Oesch (2019) used two cohort studies of young adults from Switzerland and matched the two sexes based on their intellectual ability and educational attainment before they entered the labor market. They find an unexplained gender wage gap of between 3 and 6 percent in favor of men and discuss anticipation of family roles and (implicit) biases in hiring as possible driving forces.

Several experimental studies have focused on the possible mechanisms that explain the gender gap in employment and wages from both sides of the labor market. Among them, we report below the papers that inspired our lab experiment's design of the Demand-side and Supply-side games.

On the supply side, gender differences in taste for competition have been hypothesized as one of the potential predictors of labor market outcomes (Buser et al., 2014; Dohmen and Falk, 2011; Heinz et al., 2016). If women have a stronger aversion to competitive workplaces than men of similar ability, the low share of women participating in competitive sectors could explain the gender gap in labor market success (Niederle and Vesterlund, 2007; Flory et al., 2014; Samek, 2019). Niederle and Vesterlund (2007) find that twice as many men as women choose to enter a tournament in the laboratory, despite no differences in performance. Using a field experiment with more than 35,000 university students, Samek (2019) finds that competitive, performance-based compensation schemes disproportionately deter women from applying. Previous studies suggested that men's overconfidence is the main driver of the gender gap in competition. However, the very recent study by Lénárd et al. (2024) shows that, even if both genders had the same level of confidence, a persistent gender gap in preference for competition would persist in the group of participants who can guess their relative performance correctly, with "realistic men" being more willing to compete than "realistic women."

The Supply-side game of our experiment is related to those papers, especially Niederle and Vesterlund (2007). Some papers study feedback provision in the Niederle-Vesterlund game. Ertac and Szentes (2011) present evidence that gender gaps in competitiveness decline significantly when individuals are given performance feedback before making their incentive scheme choice. Similarly, Wozniak et al. (2014) show that relative performance feedback moves high-ability females toward more competitive forms of compensation and low-ability men toward less competitive forms, thus eliminating gender differences in choices. Baier et al. (2024) highlight that gender differences in task choice and competitiveness are contingent on feedback, the underlying task, and the task choice set. Our feedback provision differs from those in the above studies because we inform participants of women's superior performance in college (for students from the same university and field of study). Such an ability gap suggests that women should perform at least as well as men in the tasks assigned in the supply-side game. Furthermore, feedback on gender gaps in labor market outcomes and the gap in academic performance together suggest that women's lower success in the labor market is not due to lower academic ability.

Still focusing on the supply side of the labor market, different outcomes can also depend on gender gaps in applicants' preferences. Cortes et al. (2023) investigate gender differences in the job search process in the field and lab. They collect information on initial job offers and acceptances from undergraduate students and find a clear gender difference in the timing of job offer acceptance, with women accepting jobs substantially earlier than men, and a sizable gender earnings gap in accepted offers. Their findings highlight the importance of risk preferences and beliefs for gender differences in job-searching behavior. Save-Soderbergh (2019) explores gender gaps in salary requests in the field and shows that the gender gap in negotiation behavior explains a substantial portion of the gender gap in starting salaries. Specifically, although female applicants are slightly more likely to negotiate their starting salaries, when negotiating, they make lower salary requests than males do, especially among males and females that state high requests. Different outcomes in competing environments can also depend on gender gaps in the ability to perform under pressure. The recent study by Bühren et al. (2024) documents the existence of gender differences in response to expectations and choking under pressure in skill-based tasks, with men failing more often than women.

Career achievements do not, however, depend only on the supply side of the labor market. On the demand side, factors such as employers' stereotypes and discrimination can also contribute to the gender wage gaps. Notwithstanding equal performance by female and male workers, employers' gender biases may result in assigning less profitable activities to women in the organization (De Pater et al., 2010). The experimental evidence on task assignment shows that discrimination against women operates not only via the possibility of promotion in the vertical hierarchy, women being less likely to be selected as team leaders (Reuben et al., 2012; Peterle & Rau, 2017), but also horizontally, having less likelihood of being assigned to high-promotability tasks within the organization, compared to men (Babcock et al., 2017a; von Bieberstein et al., 2020).

In a recent lab experiment, Domínguez and Montinari (2021) investigate how gender quotas in hiring affect the allocation of workers into different tasks within an organization. Participants played the role of employers and were asked to hire a team of six workers from a pool of 15 and assign them to one of two tasks, which differed in complexity and profitability. Employers had information about workers' age, gender, and university major as well as a signal of performance. Workers assigned to the Simple Task had to complete additions, while workers assigned to the Challenging Task had to solve mathematical problems. Their findings suggest that gender quotas to increase the number of women hired may backfire, resulting in their more likely assignment to less-challenging, low-promotability tasks compared to men and impairing their chances of career advancement. The focus and objective of our paper are different from those of Domínguez and Montinari (2021); however, our Demand-side game is inspired by their setting and differentiates between simple and challenging tasks.

Finally, Carlana (2019) and Alesina et al. (2024) document the negative effect of teachers' implicit stereotypes, measured by the Implicit Association Test (IAT), on students' academic achievement. The biases considered in the two papers refer to girls' performance in math and immigrant students' overall academic performance, respectively. Alesina et al. (2024) show that informing teachers about their own biases helps to reduce the negative effects of implicit stereotypes. The authors also examine whether learning about one's own bias (i.e., learning their IAT score) generates different behavior compared to learning about bias in general. The average impact of the two information treatments is similar, but the effects are heterogeneous: teachers with stronger negative stereotypes adjust their behavior when informed about their own bias, but do not react to generic debiasing messages. Unlike Alesina et al., our study uses the IAT score as a control and provides feedback to correct misconceptions about gender gaps in early labor market outcomes and academic performance, without giving individual-specific feedback. In the demand-side game, the IAT score never achieves statistical significance. However, in the supply-side game, we show that male students with higher implicit stereotypes are more willing to compete, regardless of treatment. For female participants, we find a positive effect of the *Mixed Feedback* Treatment for those holding implicit stereotypes, suggesting that our intervention primarily functions through a combination of reminder and de-biasing mechanisms.

3. Experimental design and procedures

We conducted a laboratory experiment comprising three distinct tasks: 1) a Questionnaire (Q), 2) a Demand-side Game (D), and 3) a Supply-side Game (S). While the Questionnaire is either present as a first or last task depending on the treatment, the order of the Demand-side and Supply-side Games is randomized.

In the following paragraphs, 3.1, 3.2, and 3.3, we provide a detailed description of each part. Section 3.4 outlines the treatments used in the experiment, and Section 3.5 elucidates the feedback content provided. Section 3.6 formulates our hypothesis, and Section 3.7 presents the experimental procedures.

3.1. Belief elicitation: Questionnaire (Q)

Participants, who were students from three schools—Economics and Management, Law, and Engineering—were requested to complete a questionnaire. They were encouraged to consider their responses carefully. Each respondent in the questionnaire was asked to provide their expectations regarding performance disparities within their specific school and the labor market for graduates from that school. Specifically, the questionnaire mentioned three hypothetical individuals: "*Giovanni, Angela, and Hassan"* who *"just graduated from the School of Economics and Management (*or *Engineering,* or *Law) of the University of Bologna.*".

Respondents were initially asked to estimate the probability of everyone achieving a graduation grade equal to or higher than a specific threshold, as indicated in Question A:

- **Question A:** "*Based on your own experience and information, what is the probability that Giovanni (*then *Angela,* and then *Hassan) graduates with a grade at least as high as 105 out of 110?"*

The grade mentioned in Question A corresponds to the *average graduation grade* achieved by graduates from one of the Master' programs in the field of Economics and Management offered by the University of Bologna in the years 2010, 2011, and 2012. The average graduation grades for Engineering and Law are 106, and 102, respectively.[1] Importantly, all questions (question A and the following) in the questionnaire were field-dependent, meaning that subjects answered questions specifically concerning peers who graduated in the same field of study as themselves.

We chose the names 'Giovanni' and 'Angela' as representative of typical Italian names, reflecting a man and a woman, respectively. In addition, we introduced the name 'Hassan' to represent a student with an immigrant background. This was done to make gender less salient and reduce the desirability bias in the questionnaire's replies. This approach is similar to the one used by Carlana (2019) and Alesina et al. (2024), who simultaneously elicit teachers' attitudes about the gender gap in math performance and the immigrant-native gap in overall school performance.

[1] The University of Bologna provides a wide range of Master's programs, including eighteen programs in the field of Economics and Management, three programs in the field of Law, and twenty-eight programs in Engineering. At the University of Bologna, the average graduation grade for Engineering is 106/110, while for Law, it is 102/110. These average graduation grades have been computed based on data collected from individuals who graduated in 2010, 2021, and 2012. See Section A.3 of the supplementary section where the exact feedback provided is reproduced.

The objective of Question A was to gauge respondents' beliefs regarding the presence of a gender gap in academic performance within their field of study, operating under the assumption that they may possess better knowledge about peers who attended similar master's programs. So, we compare the probabilities assigned to *Angela* and *Giovanni* in each field of study.

The subsequent section of the questionnaire aims to elicit expectations regarding gaps in early labor market outcomes following graduation for the same three hypothetical individuals. Specifically, six questions are presented, following the same structure as Question A. Each question begins with the following information: "*Giovanni/Angela/Hassan just graduated from the School of Economics and Management/Engineering/Law at the University of Bologna with a grade of 105/106/102 out of 110.*" Participants were then asked to express their expectations by assigning probabilities to specific labor market outcomes relevant to their field of study. As an illustration, we provide the questions with threshold wages that apply to graduates from the School of Economics and Management. These wage figures were extracted from the wage distribution of graduates in the three fields of study who graduated between 2010 and 2012, one year and five years after graduation. We obtained this data from AlmaLaurea (https://www.almalaurea.it/), the statistical office of Italian public universities, and the choice of these specific questions is a consequence of the available administrative information.

In your view, what is the

- *probability that Giovanni (*then *Angela* and then *Hassan) will be employed on a permanent contract?* **Question B1 (permanent contract).**
- *probability that Giovanni (*then *Angela* and then *Hassan) will earn 1170 Euro per month one year after graduation?* **Question B2 (median wage after 1 year).**[2]
- *probability that Giovanni (*then *Angela* and then *Hassan) will earn 1420 Euro per month one year after graduation?* **Question B3 (high-stake wage after 1 year).**
- *probability that Giovanni (*then *Angela* and then *Hassan) will earn 1460 Euro per month five years after graduation?* **Question B4 (median wage after 5 years).**
- *probability that Giovanni (*then *Angela* and then *Hassan) will earn 1760 Euro per month five years after graduation?* **Question B5 (high-stake wage after 5 years).**
- *probability that Giovanni (*then *Angela* and then *Hassan) will be self-employed five years after graduation?* **Question B6 (self-employment after 5 years).**

Participants received a fixed payment of 10 Euros upon completing the questionnaire, regardless of the correctness of their answers. Note that we decided to ask indirect questions rather than the participants' estimate of the gaps to limit the confounding of personal aspirations and characteristics on beliefs.

3.1.1. Robustness of the belief's elicitation strategy

To verify the robustness of our beliefs' elicitation strategy, we conducted an online experiment in May 2024 with 152 participants. In this experiment, participants were only confronted with the beliefs' elicitation task (i.e. the Questionnaire) and we varied (i) the incentive structure of the payment for the beliefs elicitation and (ii) the order of the names of the three fictitious students.

Regarding the incentives, instead of providing a fixed payment, we implemented a mechanism where one of the seven survey questions was randomly selected, and participants were paid based on the accuracy of their guesses for that specific question. If their guess was within $\pm$ 5 % of the actual value, the participant received an additional 1 euro; otherwise, they received no additional money. In each question, participants had to make three guesses, one for each fictitious student, and they were paid for each correct guess. Thus, in a randomly selected question, participants could earn up to 3 euros in addition to the 2 euros show-up fee. Our findings, detailed in Appendix B, indicate that the results from both elicitation methods are qualitatively the same.

Note that, in our main treatments, the order of the names of the three fictitious students was fixed (Giovanni first, then Angela, and finally Hassan). In the online experiment, we also tested whether varying the order of the names affected the beliefs elicitation. Implementing all possible orders (i.e., $3! = 6$) was deemed excessive, so we selected three specific orders. Specifically, we varied the first name in the list: order 2 (Angela, Hassan, Giovanni; N=77, 50.66 %) and order 3 (Hassan, Giovanni, Angela; N=43, 25.66 %). Our findings, detailed in Section B of the Appendix, show that the results are also robust to the order of the names.

3.2. Demand-side game (D)

In the demand-side game, all participants in a session assumed the role of employer for a firm. Employers were tasked with making two consecutive decisions following the approach of Domínguez and Montinari (2021). Firstly, they had to select three candidates from a pool of six to hire for their firm. Subsequently, employers had to assign the chosen candidates to two different tasks that varied in

[2] The wages of questions B2 and B4 correspond to the median wage of graduates from the University of Bologna with a Master's degree in the field of Economics and Management one year and five years after graduation, respectively. For graduates with a Master's degree in Engineering and Law, the corresponding wages are 1280 and 660 Euros after one year, respectively, and 1460 and 1270 Euros after five years. The high-stake wages of questions B3 and B5 represent the upper bound of the third quartile in the wage distribution of graduates with a Master's degree in Economics and Management one year and five years after graduation. For graduates with a Master's degree in Engineering and Law, the corresponding wages are 1450 and 1130 Euros after one year, respectively, and 1920 and 1560 Euros after five years.

difficulty and profitability for the employer. Specifically, each employer had to select two candidates for the Simple Task (ST) and one candidate for the Challenging Task (CT). The ST involved summating as many three-digit numbers as possible within a 6-minute timeframe. On the other hand, the CT required solving as many mathematical problems as possible within a 10-minute timeframe.

An example of both tasks is provided below in Fig. 1:

Employers had access to information such as the candidates' age, gender, and field of study. Additionally, they received a performance signal indicating the productivity of the candidates. This performance signal was a noisy measurement based on the number of correct answers provided by the candidate during the first three minutes of the six-minute Standard Task. Nothing is said about the candidate's productivity in the challenging task in the idea that it is quite easy for employers to retrieve information about workers' productivity in simple and ordinary tasks while it is more difficult to get signals about challenging ones. The candidate profiles used in the game were derived from the candidate pools analyzed by Domínguez and Montinari (2021).[3] The six candidates appearing in each pool are characterized by a performance signal equal to or above the median.[4] In the Appendix, Table A1 reports for each pool of six candidates the information provided to the employers and the ex-post-performance in the Simple Task and the Challenging Task. Candidate pools from which employers must choose three candidates are designed so that, based on productivity signals, it is always advantageous to hire at least one man and one woman. The third candidate to be hired must be selected from a subset of the four remaining candidates, where men and women have similar productivity signals. However, the two tasks that employers must assign to the chosen candidates are gender-stereotyped, therefore, participants' implicit biases may play a role and affect the interpretation of the signal, together with the other pieces of information provided (e.g. gender, field of study and age).

Each employer participated in four rounds, encountering a different group of candidates in each round. The four pools of candidates were randomly rotated across rounds to ensure that all employers evaluated the same pool by the end of the experiment and to control for order effects. No information was provided to employers after each round.

During each round, employers were compensated based on the performance of their assigned candidates. They received 10 cents for each addition correctly solved by candidates assigned to the Standard Task and 1 Euro for each problem correctly solved by the candidate assigned to the Challenging Task. Before the decision stage, employers engaged in a non-rewarded practice session to familiarize themselves with the two tasks before making their decisions in the game. Tables A1a and b in the Appendix detail the candidate pools and their performance in the two tasks. Note that employers only see the performance signal in the Standard Task (column "Signal" in Table A1a), not the data in the last two columns, which explain the payoff for each chosen candidate. Table A1b shows no significant gender differences in the performances of the 24 individuals.

3.3. *Supply-side game (S)*

The supply-side game comprises three stages, following the protocol by Niederle and Vesterlund (2007): experimental sessions were perfectly balanced concerning gender. Subjects were asked to perform an adding task (like the Standard Task) in all stages. They had three minutes to solve as many additions of five two-digit numbers as possible. As it is well known, even if there are no gender differences in the average ability to perform this specific task, men's ability to solve mathematical tasks is generally perceived as relatively higher than women's (see e.g. Balafoutas and Sutter, 2012, and Niederle et al. 2013). The payment scheme differed at each stage. In Stage 1 (*Piece rate*), subjects earned 0.50 Euro per each correct calculation. In Stage 2 (*Tournament*), subjects were allocated into groups of three male and three female students. The gender composition of the groups was common knowledge, but the identity of the group members was not revealed. The payments in Stage 2 depended on the subjects' relative performance. The tournament winners were the two subjects who correctly solved the largest number of additions. Winners received 1.50 Euro per addition correctly solved, while other subjects in the group did not receive any payment. In Stage 3 (*Choice*) of the game, subjects decided on the preferred payment scheme. They were paid 0.50 Euro per correct addition if the piece rate was selected. If the tournament was selected, the individual's performance in Stage 3 was compared to that of other members of the same group in Stage 2. Specifically, participants in the tournament won if they solved correctly at least as many additions as the winner ranking second in Stage 2. Again, winners received 1.50 Euro per calculation correctly submitted. In both Stage 2 and Stage 3, ties were broken randomly. Information on the relative performance of subjects in each group in Stages 2 and 3 was not disclosed until the end of the study. At the end of Stage 3, we elicited individual beliefs about subjects' relative ranking for all subjects and same-gender subjects in their group and gender differences in performance under tournament (Stage 2). Beliefs were incentivized. Subjects received 1 Euro for each of the three guesses if the answer was correct and 0 otherwise. The questions used to elicit beliefs are reported in Section D (Instructions) of the supplementary section.

[3] Dominguez and Montinari (2021) compare two treatments: a baseline treatment where employers are no constrained in hiring, and a Quota treatment where employers are forced to hire at least 50% of women while they are free in the assignment to the challenging task. The authors find that the Quota treatment closes the gap in hiring but fails to close the gap in task assignment. Moreover, employers earn lower profits under the Quota treatment since they are worse at selecting the candidates in the hiring stage.

[4] In Domínguez and Montinari (2021), 60 subjects performed ST and CT, but only 56 provided valid answers. No significant gender differences exist in those individuals' absolute performances in ST and CT. The information provided to employers in the current experiment was derived from a subsample of 24 candidates extracted from their pool by selecting the 6 candidates with a signal equal or above to the median.

Solve the following addition:	**2y +3=3x- 5 passes through the point**:
526 + 414+780=_____	A. (1, 5/2) B. (1, -5/2) C. (0,0) D. (-1, -5/2)

a. Example of Simple Task (ST) **b.** Example of Challenging Task (CT)

Fig. 1. a. Example of Simple Task (ST). b. Example of Challenging Task (CT).

3.4. Treatments

To study the impact of feedback (F) provision on supply and demand side behaviors, we implemented four treatments varying the feedback timing and the feedback content: *Baseline*, *Salience*, *Mixed Feedback (MFT),* and *Outcome Feedback (OFT)*. We considered a between-subjects design where subjects did not participate in more than one treatment.

In the *Baseline Treatment (BT),* either the Demand-side or the Supply-side Game was played first. The Questionnaire was introduced after both games to avoid potential biases in the decisions. The Feedback was provided after the Questionnaire at the end of the lab experiment. To control for the effect of filling out the Questionnaire (without feedback) before the decisions in the two games, we study the *Salience Treatment* (*ST*), which is identical to the *Baseline Treatment*, except that the Questionnaire was presented as the first task before the Supply-side and the Demand-side Games and the feedback was provided only at the end of the study. Playing the questionnaire first in *Salience Treatment* makes the gender dimension salient, even if no feedback is provided. Previous research has indeed evidenced that simply exposing individuals to stimuli that evoke negative stereotypes associated with their identity (e.g. asking their gender or their ethnicity) may impact behavior, and for example may activate a stereotype threat effect (see Shih et al.; 1999).

In the *Mixed Feedback Treatment*, the Questionnaire was administered first, followed by Feedback provided immediately before implementing the Demand-side or Supply-side Game. Here, the questionnaire and the feedback content covered both academic performance and labor market outcomes. The *Outcome Feedback Treatment* was identical to the *Mixed Feedback Treatment*, but the questionnaire and feedback content focused solely on labor market outcomes.[5] In other words, the question and feedback on academic performance were excluded, so the treatment only conveyed negative information for women (corresponding to positive information for men) on labor market outcomes. By comparing the *Outcome Feedback Treatment* and the *Mixed Feedback Treatment* we aim to disentangle the impact of different information content. More details on the feedback provision are provided in Section 3.5.

Note that *Mixed Feedback* and *Output Feedback Treatments* do modify two aspects of *Baseline Treatment:* the timing of the questionnaire (and therefore the salience of the gender dimension) and the timing of the feedback provision. For this reason, we consider as benchmark our *Salience* treatment to estimate the treatment effects.[6]

Within treatments, the order of the Demand-side and Supply-side Games was randomized to ensure the consistency of the behavior in each game. Table 1 summarizes the treatments and specifies the order of their parts (Q=Questionnaire, D=Demand-side Game, S=Supply-side Game, F=Feedback) and the number of observations per treatment, with male and female participants being perfectly balanced. Note that having an equal number of male and female participants is needed to implement the Supply-side Game which relies on the experimental design by Niederle and Vesterlund (2007).

We are aware that our treatments may induce experimental demand effects. In this respect, our treatments can be ranked based on the potential for such effects, with the highest experimental demand likely associated with the Feedback treatments, where the content and layout explicitly compare the two genders. In contrast, we expect a lower potential for experimental demand effects in the Salience treatment for two reasons: no feedback is provided and, to obscure our main research question, we ask about not only a male and a female character but also about a third character with a non-Italian sounding name.

[5] The *Outcome Feedback* treatment was conducted in May 2024 following the suggestions of an anonymous reviewer and the Guest Editor. We used the same laboratory and applied the same criteria for selecting the subject pool. More details are provided in the appendix.

[6] The questionnaire in the *Output Feedback Treatment* differs from the one used in the *Salience Treatment* because it does not include the question about academic performance, and consequently, no feedback is provided on this issue. This choice was motivated by the hypothesis that salience is influenced by highlighting the gender dimension itself rather than the content of a specific question. When comparing both the likelihood of correctly answering questions B1-B6 and the probability of being classified as aware of gaps at 1 year or 5 years, we find that subjects in the *Outcome Feedback Treatment* are significantly more likely to hold correct beliefs than those in the *Salience Treatment*. This difference appears to be more an effect of the participants in the 2024 cohort being more informed compared to those who participated in 2019–2021, while the qualitative trends of the results are confirmed (such as women, on average, holding more correct beliefs than men, particularly women in Engineering). For further details, see the discussion in Section A of the Appendix (Tables A3a and b) and in Section B of the Appendix.

Table 1
Treatments, sequences of games, and participants.

Treatment	*Order*	*% Female*	*N*
Baseline (*BT*)	D-S-Q-MF S-D-Q-MF	50 %	72
Salience (*ST*)	Q-D-S-MF Q-S-D-MF	50 %	132
Mixed Feedback (*MFT*)	Q-MF-D-S Q-MF-S-D	50 %	108
Outcome-Feedback (*OFT*)	Q-OF-D-S Q-OF-S-D	50 %	96
Total	–	**50 %**	**408**

Note: The second column indicates the order in which the parts of each treatment were presented. D=Demand-side Game; S=Supply-side Game; Q=Questionnaire; F=Questionnaire feedback. OF identifies the *Output Feedback Treatment* (where feedback content covered only labor market outcomes) while MF identifies the *Mixed Feedback Treatment* (where feedback content covered both academic performance and labor market outcomes). The difference between MFT and OFT consists therefore in the content of the information provided. Gender balance in each treatment holds within the order variations *and* each treatment.

3.5. Feedback's content

The feedback used data from three cohorts of graduates from the University of Bologna collected from Almalaurea and is reported in detail in section F of the supplementary section. We calculated the exact probabilities requested in each question for both men and women.[7] The overall information is qualitatively similar for the three fields of study. In a nutshell, women like Angela outperform men like Giovanni and Hassan in college, while a gap exists in labor market outcomes in favor of men.

Table 2a reports a Summary of the information contained in the *Mixed Feedback Treatment* by question, school, and gender, while Table 2b reports a summary of the information contained in the *Outcome Feedback Treatment* and in the online experiment by question, school, and gender (which were run 4 years later and for this reason are presented separately).

Specifically, in Table 2a (relevant for the *Mixed Feedback Treatment* where we elicit beliefs on academic performance), we observe that the average graduation grade is higher for women (Question A). In both Table 2a and 2b, women are less likely to be employed in a permanent contract (question B1), to earn a median and high wage one year after graduation (questions B2 and B3, respectively), to earn a median and a high wage five years after graduation (questions B4 and B5, respectively), to be self-employed five years after graduation (question B6). Thus, in the *Mixed Feedback Treatment* the feedback provides one piece of information representing good news for women (feedback on Question A) and another that is bad news (feedback on Questions B1-B6).

Note that the *Mixed Feedback Treatment* presents unbalanced information, highlighting good news for men in several dimensions, with only one positive news for women. This format reflects the ongoing media debate about women typically outperforming men in educational achievements, yet men outperforming women in various labor market outcomes (such as salary, career achievements, and entrepreneurship) throughout their careers. As mentioned before, to isolate the effects of the positive and negative information provided in the *Mixed Feedback Treatment*, we implement the *Outcome Feedback Treatment,* which excludes information about academic performance.

3.6. Experimental hypothesis

We tested three main hypotheses, which, for the two stylized games in the baseline treatment, are essentially derived from the results documented by previous studies employing similar games.

Hypothesis 1. On average, participants are unaware of gender gaps in market outcomes among their peers, particularly of gender gaps that arise one year after graduation. Compared to male participants, females expect a larger gap in these outcomes.

The first sentence of this hypothesis is based on the results reported by Sipe et al. (2009) and Filippin and Ichino (2005) documenting, respectively, a tendency to underestimate the existence of gender discrimination in the labor market as well as a tendency to underestimate the development of gender gap in earnings. Moreover, Filippin and Ichino (2005) also find that females always expect a higher gap than males. Based on this evidence, we will use one-sided tests to detect gender differences in beliefs.

We do not have a prior hypothesis on awareness of the gender gap in academic performance.

Hypothesis 2. In the demand-side game, compared to the Baseline treatment, we expect the Salience and the two Feedback treatments to influence the behavior of male and female participants by increasing their likelihood of hiring women and assigning them to

[7] An example of the feedback provided to students enrolled in a master's program offered by the School of Economics and Management is reproduced in Appendix F. Interestingly, descriptive statistics indicate that those male graduates who, before enrolling in the University of Bologna, obtained a high school or a BA degree outside the European Union perform equally well in college and in the labor market as male native students. This is why, in *Mixed Feedback Treatment*, participants are informed that gender gaps in academic performance are the same for immigrants and native men. Similarly, in both the *Mixed Feedback* and *Outcome Feedback Treatment*, participants are informed that gender gaps in labor market outcomes are the same for immigrants and native men.

Table 2a
Summary of the information contained in the **Mixed Feedback** Treatment by question, school, and gender.

	School of Economics and Management		School of Engineering		School of Law	
	Men	Women	Men	Women	Men	Women
Question A (average graduation grade)	104/11	107/110	105/110	106/110	99/110	103/110
Question B1 (% in permanent contract, 1 year)	17.5 %	17.1 %	18.6 %	13.6 %	18.4 %	16.9 %
Question B2 (% above median wage, 1 year)	65 %	44 %	53 %	42 %	60 %	43 %
Question B3 (% above high wage, 1 year)	37 %	13 %	27 %	20 %	34 %	18 %
Question B4 (% above median wage, 5 years)	58 %	39 %	53 %	40 %	68 %	44 %
Question B5 (% above high wage, 5 years)	43 %	11 %	27 %	18 %	40 %	19 %
Question B6 (% self-employed, 5 years)	19 %	16.6 %	12.1 %	11.6 %	56.4 %	50.3 %

Note: The probabilities reported in this table were extracted from the wage distribution of graduates in 2010, 2011, and 2012 one year and five years after graduation in the three fields of master's programs. AlmaLaurea (https://www.almalaurea.it/), the statistical office of Italian public Universities, provided us with the data.

Table 2b
Summary of the information contained in the **Outcome Feedback** Treatment by question, school, and gender.

	School of Economics and Management		School of Engineering		School of Law	
	Men	Women	Men	Women	Men	Women
Question B1 (% in permanent contract, 1 year)	29.5 %	21.7 %	37.0 %	28.6 %	11.8 %	10.9 %
Question B2 (% above median wage, 1 year)	50.7 %	40.3 %	28.7 %	22.6 %	44.0 %	39.2 %
Question B3 (% above high wage, 1 year)	23.0 %	13.6 %	8.3 %	4.9 %	31.0 %	20.3 %
Question B4 (% above median wage, 5 years)	45.7 %	36.0 %	27.6 %	20.2 %	50.1 %	40.3 %
Question B5 (% above high wage, 5 years)	26.7 %	12.0 %	9.0 %	4.0 %	37.6 %	22.6 %
Question B6 (% self-employed, 5 years)	20.5 %	15.0 %	10.0 %	11.0 %	55.0 %	49.6 %

Note: The probabilities reported in this table were extracted from the wage distribution of graduates in 2015, 2016, and 2017 one year and five years after graduation in the three fields of master's programs. AlmaLaurea (https://www.almalaurea.it/), the statistical office of Italian public Universities, provided us with the data.

the Challenging Task. We expect awareness about gender gaps to mediate the effect of treatments.

In the control treatment presented by Dominguez and Montinari (2021), women have lower chances of being hired and assigned to the Challenging Task than men. Thus, in the Baseline Treatment, we expect to observe the same trend. As mentioned in the introduction, in this game, gender stereotypes regarding women's math ability may affect employers' hiring and task assignment decisions.

Comparing the two *Feedback* treatments with the *Baseline* treatment introduces changes in two things simultaneously: the salience of the gender dimension and the provision of feedback. For this reason, as mentioned before, we will use the *Salience* Treatment as a benchmark.

By making the gender dimension salient, we expect that individuals may alter their hiring and assignment decisions in ways that are more favorable to women. We do not have a prior hypothesis on the relative strength of the effects associated with the *Salience* and Feedback treatments. However, the *Mixed Feedback* treatment could be more effective than the *Outcome Feedback* treatment, because it also provides information about women's superior academic achievements. Overall, the effect of the two *Feedback Treatments* may depend on participants' awareness of the existing gaps.

To conclude, the intervention's effectiveness could arise from various channels such as making gender gaps salient, activating debiasing mechanisms, serving as a reminder, or through the information in the feedback. As mentioned previously, the specific mechanism at work depends on awareness; it can function as a reminder for those who are aware or as a pure information provision for those who are not.

Hypothesis 3. In the supply-side game, we expect that Salience Treatment induces lower competitiveness by female participants (especially if they are affected by biases) by activating a stereotype threat. Feedback treatments may increase or decrease women's willingness to compete. We expect the Mixed Feedback Treatment to discourage women's willingness to compete less than the Outcome Feedback Treatment because it also provides information about women's superior academic achievements.

Our Supply-side Game is based on the experimental protocol of Niederle and Vesterlund (2007), which has been extensively replicated (see the discussion in Niederle 2017). Based on the documented previous results, in the Baseline Treatment, we expect to observe a gender gap in competition in favor of men.

Our Hypothesis 3 refers to the impact of saliency and information about gender gaps on female willingness to compete. Previous literature has shown that making gender salient in a (stereotypical) mathematical task may induce women to compete less (see, e.g. Shih et al.; 1999). Thus, we expect that women's willingness to compete may decrease in the *Salience Treatment* compared to the *Baseline Treatment.*

Regarding feedback provision, while we are uncertain about the effect of Feedback treatments on women's willingness to compete, which may either increase or decrease, we anticipate the relative impact of the two feedback treatments as follows. We expect female participants to feel less discouraged when they learn about women's superior academic performance compared to when they learn

solely about women's poorer labor market outcomes. This awareness may lead to a relatively higher willingness to compete under the *Mixed Feedback Treatment* than under the *Outcome Feedback Treatment*.

Regarding the mechanism driving these expected results, by considering participants' pre-existing awareness of gender gaps and implicit biases, we aim to determine whether our intervention acts as a reminder, a de-biasing mechanism, information transmission, or a combination of these factors.

We do not have a prior hypothesis on the behavior of male participants. Therefore, we refrain from formulating specific predictions about the impact of feedback on men's willingness to compete.

3.7. Experimental procedures

The experiment was conducted at BLESS (the University of Bologna Experimental Laboratory) using zTree (Fischbacher, 2007). The sample consists of N=312 individuals recruited via ORSEE (Greiner, 2015) between October 2019 and April 2021 and of N=96 individuals recruited using the same criteria and procedures, in May 2024, for a total of 408 participants.[8] Participants are students enrolled in Economics and Management, Engineering, or Law programs at the University of Bologna, either in the last year of a BA or a Master's program. The sample was divided into 19 sessions that lasted about 90 min each. Treatments were randomly allocated to different sessions and once arrived at the lab, participants were randomly allocated to seats by randomly drawing a number corresponding to the desk where they were seated. Subjects were informed that one randomly selected part of the experiment would become relevant for payments at the end of the experiment. Average earnings were 17 Euro, including a 5 Euro show-up fee. Subjects were also asked to complete a post-experimental questionnaire and an online version of the Implicit Association Test about Science and Gender (IAT, Greenwald et al., 1998). We used the IAT Gender – Science, which often reveals a relative link between "Liberal arts" and "Females" and between "Science" and "Males." Appendix G reproduces an English version of the instructions, while the original experiment was run in Italian. We focused on the three mentioned schools to have subsamples with different gender segregations: Engineering, with most male students (66.1 %); Economics and Management, whose students' pool is gender balanced (49.1 %); and Law, where most students are female (72 %). As displayed in Table 2a and 2b, female students outperform male students in all schools but are less successful in the labor market. Table A2 in the Appendix reports summary statistics of our participants by field of study.

4. Results

In this section, we present the results of the experiment. First, we focus on the subjects' expectations about ability (Question A) and labor market inequalities (Questions B1-B6). Second, we analyze the decisions in the Demand-side and the Supply-side Games, focusing on the impact of our treatments. We aim to understand i) students' awareness of gender gaps by gender and field of study; and whether Feedback provision iia) increases the probability of success of female candidates in both hiring and task assignment (Demand-side Game) and iib) increases women's willingness to compete in a math task in which men are generally perceived as better performers (Supply-side Game).

In our analysis, we used the *Salience Treatment* as the omitted category. By comparing *Salience* and *Baseline* treatments, we isolate the impact of making gender salient; by comparing *Salience* and *Outcome Feedback* treatments we isolate the effect of information provision on labor market outcomes (i.e., information that is negative for women); by comparing the two *Feedback* treatments we study the impact of adding information on academic achievements (i.e., information that is positive for women).

Moreover, we use the Questionnaire's answers to control whether information provision has a differential impact depending on the subjects' awareness of the existing gender inequalities among peers. Because we are testing simultaneously multiple hypotheses, we followed the Benjamini-Hochberg False Discovery Rate method (Benjamini and Hochberg, 1995) for multiple test adjustment: we sorted all the p-values in ascending rank and multiplied each by the number of separate tests being performed (specified in each case) before dividing each by its rank- thus greater adjustments are made to smaller p-values.

4.1. Awareness of gender gaps in ability and labor market outcomes

To understand whether information provision affects subjects' behaviors in the two games, we first analyze awareness of gender gaps in our sample. To do so, we study replies to questions A and B1-B6 and focus on the *relative magnitude* of the probabilities assigned to women (*Angela*) and men (*Giovanni*). We then check whether the relative magnitude of the probabilities assigned to men and women is correct according to descriptive statistics reproduced in the Feedback (see Table 2a and 2b). For example, in question A (gender gap in academic ability), aware subjects assigned a higher probability to Angela compared to Giovanni. Table A4 in the Appendix summarizes the average probabilities assigned by our participants to Angela and Giovanni in each question.

The proportion of aware subjects for each question is summarized in Table 3, which also reports p-values from a set of Mann-Whitney tests of gender differences in accuracy for each question and school corrected using the Benjamini-Hochberg False Discovery Rate method (Benjamini and Hochberg, 1995) to account for the fact that we are testing gender differences on 7 questions and 4

[8] Between 2019 and 2021 data collection has been interrupted for about one year because of Covid-19 restrictions. Before Covid-19 each session was hosting up to 30 participants, while due to the restrictions on the distance between participants, after Covid-19 we could host a maximum of 12 participants in each session. To account for possible variations in beliefs and behavior around labor market prospects, we checked and observed no systematic changes before and after the pandemic. See Table A2b in the Appendix for details.

Table 3
Answers to the questionnaire: percentage of aware subjects for each question.

Field of Studies	Economics			Engineering			Law			Total		
Gender	*Female*	*Male*	*p*	*Female*	*Male*	*p*	*Female*	*Male*	*p*	*Female*	*Male*	*p*
A: Ability												
Ability	49.1	40.0	0.602	31.7	33.8	1.137	38.3	65.4	0.084	40.4	41.0	1.211
B1-B6: Labor market outcomes												
Permanent contract	24.4	39.8	0.119	66.0	30.4	0.018	43.1	32.1	0.604	40.5	34.5	0.473
Median wage 1	58.1	47.0	0.378	51.1	27.2	0.045	34.7	21.4	0.478	48.3	34.5	0.045
High-stake wage 1	59.3	51.8	0.590	53.2	31.5	0.067	34.7	25.0	0.602	49.3	38.9	0.115
Median wage 5	54.7	48.2	0.658	61.7	32.6	0.018	43.1	32.1	0.604	52.2	38.9	0.050
High-stake wage 5	58.1	53.0	0.787	70.2	30.4	0.000	43.1	32.1	0.604	55.6	39.9	0.018
Self-employed	59.3	39.8	0.066	48.9	52.2	1.034	48.6	35.7	0.521	53.2	44.8	0.276
Aware Individuals (i.e. individuals who correctly answer 2 out of 3 questions in B1-B3 and B4-B6)												
Aware 1y	54.7	45.8	0.500	57.4	30.4	0.024	34.7	28.6	0.839	48.3	36.5	0.068
Aware 5y	62.8	50.6	0.305	63.8	29.3	0.018	43.1	32.1	0.604	56.1	38.4	0.018

Note: The table presents the percentage of aware subjects by question, field of study, and gender. In question A (gender gap in ability), aware subjects assigned a higher probability to Angela (i.e. women). In all questions B (gender gaps in labor market outcomes), aware subjects assigned a higher probability to Giovanni (i.e. men). Columns *p* report one-sided p-values from a set of Mann-Whitney tests on gender differences. All p-values reported are corrected using the Benjamini-Hochberg False Discovery Rate method (Benjamini and Hochberg, 1995). Specifically, in this case, we are executing 36 tests: (7 questions + 2 individual classifications on 4 subgroups corresponding to the 3 fields of study and the total).

subgroups corresponding to the 3 fields of studies and the overall group.

At the aggregate level, Table 3 shows that, coherently with our Hypothesis 1, about 41 % of the students in our sample are unaware of the gender gaps existing among peers in academic performance without significant gender differences. This result is captured by Question A, which is not elicited for the *Outcome Feedback Treatment.*[9] There are no significant gender differences in awareness among students from Economics, Management, and Law. However, gender differences in awareness are evident among Engineering students. In this field, women are more aware than men in assessing the probabilities of various labor market outcomes, except for the likelihood of being self-employed five years after graduation. Furthermore, there are a few differences between fields of study regarding awareness. Overall, women demonstrate greater awareness than men in predicting some of the gender gaps in labor market outcomes. This finding is primarily driven by the high level of awareness observed among women in Engineering. This heightened awareness among female Engineering students may be attributed to greater curiosity and willingness to acquire information about future career prospects. Data from AlmaLaurea, the statistical office of Italian public universities, indicate that female students in Engineering, as well as those in medical schools, possess the strongest backgrounds. They only graduate from high schools, such as scientific and classical lyceums, that prepare students for tertiary education and they achieve the highest graduation grade, by a significant margin.

At the individual level, we distinguish between awareness of gender gaps in labor market outcomes one year after graduation and five years after graduation. We define "Aware 1y" and "Aware 5y" individuals as those who correctly respond to two or more questions among B1-B3 and B4-B6, respectively. In other words, aware subjects correctly respond to at least two out of three questions on labor market outcomes. The percentages of individuals classified according to this definition are reported in the last two rows of Table 3.

When considering the answers to questions B1-B3 we find that overall, 23 % of subjects (N=95/408) correctly replied to one question, while 15 % (N=61/408) correctly replied to all three questions. Furthermore, 34 % (N=140/408) were correct in any of the questions. When considering the answers to questions B4-B6, we find that 25 % of the sample (N=102/408) replied correctly to all three questions, while 30 % (N=122/408) replied correctly to any of the questions. See also Figure A1 in Section C of the Appendix.

Based on our definitions, students classified as Aware 1y represent 42.4 % of the sample (N=173/408). Women are significantly more aware of early gender gaps than men, 48 % compared to 36 % (Mann-Whitney test $z = 2.417$, $p = 0.02$) and the distribution significantly differs depending on the field of studies (χ^2, $p = 0.002$). This finding is driven by women in Engineering, who were more aware of gender gaps in early market outcomes than men in the same field, 57 % compared to 30 % ($z = 3.070$, $p = 0.002$). Moreover, women in Engineering are more aware of gender gaps in market outcomes than women in Law (34 %) ($z = 2.433$, $p = 0.02$) and women in Economics and Management (54 %) ($z = 0.309$, $p = 0.76$).

Students classified as Aware 5y represent 47 % of the sample, with women being more aware than men (56 % vs 38 %, $z = 3.571$, $p = 0.001$). Again, this finding is driven by the gender differences observed among subjects from Engineering (64 % vs 29 %, $z = 3.896$, $p = 0.001$). Similarly, women from Engineering are more aware of long-term gender gaps (69 %) than women from Law (43 %) ($z = 2.207$, $p = 0.03$). We also find that men from Engineering are less aware of outcomes at 5 years than men from Economics and Management (29 % vs 50 %, $z = 2.865$, $p = 0.004$).

According to a set of χ^2 tests, we find that there is a significant association between the number of correct answers in questions B1-

[9] In Section A of the Appendix, we have split Table 3 into two tables (Table A1a and Table A3b): one for the *Outcome Feedback Treatment* and one for the other three treatments. As mentioned before, awareness appears to have increased over time when comparing the *Outcome Feedback Treatment* run in May 2024 to the other treatments conducted between 2019 and 2021. However, the results remain qualitatively the same when comparing genders and fields of study. In Section D of the Appendix, we replicate our analysis by excluding the *Outcome Feedback Treatment* and including the variable accounting for awareness about gender gaps in academic performance.

B3 and the number of correct answers in questions B4-B6, $p < 0.001$ with a Spearman correlation coefficient equal to 0.5579, $p < 0.001$. This result holds when we repeat the χ^2 tests considering males and females separately ($p < 0.001$ in both cases).

Figure C1 in Section C of the Appendix displays the distribution of correct answers to questions B1-B3 and B4-B6, both overall and by gender. Our results remain qualitatively unchanged if, instead of using the dummies Aware 1y and Aware 5y, we employ a different definition of awareness. In addition, Section C of the Appendix provides results from a robustness analysis of our definition of awareness. Specifically, beside Figure C1 which plots two continuous variables ranging from 0 to 3 for each subset of questions and Figure C2 uses the average percentage deviation from the true value in each question.

A group of subjects who deserve special attention is represented by participants who expect gender equality in the labor market. Specifically, these subjects assign equal probabilities to Giovanni and Angela when answering B1-B3 and B4-B6 questions.[10] There is a clear correlation between holding gender-neutral beliefs and being classified as an unaware subject. Specifically, we find that when considering questions B1-B3, 50 % reported equal beliefs in at least two questions and this percentage is equal to 49 % when considering questions B4-B6. These participants hold gender-neutral beliefs, and the impact of information may be different compared to other groups. Figure C.3 in Section C of the Appendix reports the distribution of the number of questions in which individuals predict gender equality. Interestingly, for both sets of questions B1-B3 and B4-B6, there is a U-shape which evidences a bimodal distribution with the two most frequent situations being represented by subjects who hold gender-neutral beliefs in all questions and subjects who hold gender-neutral beliefs in none of the questions.

Our findings are summarized in Result 1, which is consistent with our Hypothesis 1.

Result 1: Students' awareness.

Only 41 % of our participants are aware of gender gaps in academic performance. When considering the gaps in labor market outcomes, we find that 42 % are aware of gaps at 1 year, while 47 % are aware of gaps at 5 years (i.e., they answered correctly to at least two out of three questions among B1-B3 and B4-B6, respectively). Overall, participants are more likely to correctly predict labor market outcomes occurring 5 years after graduation compared to 1 year after graduation. Women demonstrate a higher awareness of gender gaps in labor market outcomes than men. A significant portion of participants expect a gender-neutral workplace, with about 30 % anticipating no gender differences in any questions for both gaps at 1 and 5 years.

Overall, students are better at estimating gender gaps that emerge 5 years after graduation than those that appear 1 year after graduation. Additionally, a significant portion of students believe they will be entering a gender-neutral workplace. While these students may be aware of gender disparities among more established members of the workforce or for example associated with motherhood, they either ignore or underestimate the early emergence of these gaps, which occur as soon as one year after graduation.

Similarly, a general lack of awareness is evident regarding the gender gap in academic performance, a topic that received less media attention. Surprisingly, most students (41 %) incorrectly believe that male students achieve better grades. An interesting exception is observed among male students in Law, where 65 % correctly acknowledge that their female counterparts perform better.

4.2. *Demand-side game: Hiring and task assignment decisions under information*

In this section, we analyze the effect of Feedback provision on the probability of female candidates being hired and assigned to the Challenging Task, i.e., the most complex and profitable task. In other words, we show how information affects the gender gap in employment and leadership. Fig. 2 summarizes the proportion of male and female candidates hired and assigned to CT within each treatment.[11] In Figure A1 of Appendix A, we also report the proportion of women hired and assigned to the Challenging Task at the employers' level.

Inspection of Fig. 2 reveals that the proportion of hired women does not differ significantly across treatments, meaning neither the salience of the gender dimension nor the feedback provision impacts hiring decisions. Table A6 in the Appendix shows regressions on the probability of candidates being hired and reports the same result. Thus, regarding the hiring behavior, our Hypothesis 2 is not supported.

A possible explanation for this null result could be that the information provided in the *Mixed Feedback Treatment* is unbalanced. Specifically, the information about academic performance might be too weak compared to that on labor market outcomes, insufficiently impacting employers' behavior. However, if this were the mechanism, we would expect to see an even lower likelihood of hiring female candidates in the *Outcome Feedback Treatment*, where (positive) information about academic performance is omitted. In addition, we would anticipate a lower likelihood of hiring female candidates in the *Outcome Feedback Treatment* compared to the *Salience Treatment*, which does not provide information on labor market outcomes. Since we do not observe any difference in hiring behaviors between the *Mixed Feedback* and the *Outcome Feedback Treatments*, or between the *Salience* and *Outcome Feedback Treatments*, we conclude that neither the informational content of the feedback nor the salience of the gender dimension is driving this null result.

[10] One might assume that participants exerting low effort in the study would assign equal probabilities to the three characters when stating their beliefs. However, our findings indicate otherwise. The number of subjects who assign the same expected probability to the three characters is extremely low, ranging from 0 out of 408 (for questions about median wages at 1 and 5 years, and the probability of having a permanent contract) to 5 out of 408 (for the question about the probability of being self-employed 5 years after graduation).

[11] In the pool of candidates, 42% are women. Furthermore, in the *Baseline Treatment,* only 25% of the people hired for the Challenging Task are women. However, 50/50 is not a relevant benchmark in our setting. The highest performance signal is either from a man (2 groups) or a tie between one man and one woman (2 groups); see Table A1a. So, any participants who picked the highest signal for the complex task and randomized ties would imply that only 25% of the time women would be hired.

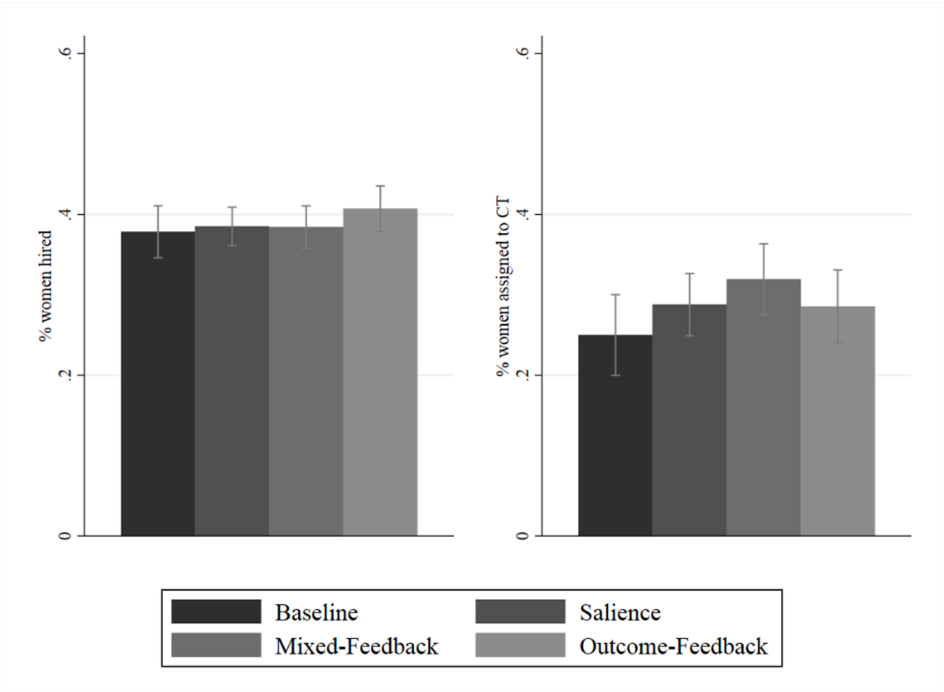


Fig. 2. Employers' hiring and task assignment decisions in the Demand-side Game with CT=Challenging Task. **Note:** The graph shows the proportion of female and male candidates among all candidates hired by employers (N=4896, 408 employers x 3 candidates hired x 4 rounds; 864 in BT; 1584 in ST; 1296 in MFT, 1152 in OFT) and among all candidates assigned to CT (N=1636, 408 employers x 1 candidate x 4 rounds; 288 in BT; 528 in ST; 432 in MFT, 1632 in OFT).

We now move to the assignment decisions. We find that the proportion of female candidates assigned to the CT is significantly higher in *Mixed Feedback Treatment* (32 %) than in *Baseline Treatment* (25 %) (Mann-Whitney $z = -2.007$, $p = 0.04$),[12] while the share of women assigned to the CT in *Salience Treatment* (29 %) and in the *Outcome Feedback Treatment* (29 %) is not significantly different from the *Baseline* one, suggesting that the *Mixed Feedback Treatment* only has an impact on the Demand side game (F v OF: $z = 1.046$, $p = 0.30$; B v OF: $z = -1.031$, $p = 0.30$, S v OF: $z = 0.071$, $p = 0.94$).

However, the significant difference between *Baseline Treatment* and *Mixed Feedback Treatment* needs deeper scrutiny. Decisions in the Demand-side Game are not only contingent on gender; thus, we cannot claim a significant presence or absence of treatment effects on task assignment decisions until controlling for several factors that may affect employers' decisions. To account for such heterogeneity, Table 4 analyzes the probability of candidates being assigned to the CT and presents the marginal effects of different specifications of a probit model. The models regress a dummy variable (*Task*) that equals one if the candidate is assigned to CT and 0 otherwise on the candidates' gender (*Female* candidate), age, school, and the relative ranking of the candidate in her group according to the signal of performance (*Ranking*, continuous). As mentioned before, to evaluate the impact of treatments, we use as the omitted category the Salience treatment and include dummies for the other three treatments. In addition, we control for employers' characteristics such as gender, awareness of gender gaps in market outcomes at 1 and 5 years from graduation,[13] and implicit gender stereotypes in the association of women and science gathered from the implicit association test (captured by the variable IAT_score).[14] The models also consider a set of covariates that account for the effect of the decision environment, such as the order in which the pools are displayed, the number of women in each group, the number of women in top-1 according to the signal, the position in which candidates are presented in the decision screen and the period. A formal specification of the variables used in the analysis of Table 4 is reproduced in Appendix E. We also control the order of the experimental parts, i.e., whether the Demand-side Game was played before

[12] The significant increase in the percentage of women assigned to the Challenging Task from *Baseline Treatment* to *Mixed Feedback Treatment* is also supported by the rank order statistic Somers' D (provided by the `somersd' package in Stata). The test looks at the proportion of women assigned to the Challenging Task accounting for the presence of clusters at the level of employers in the data (D=0.081, p=0.04).

[13] In Appendix C, Table D1 reports results for the same estimations performed on the *Baseline*, *Salience,* and *Mixed Feedback* treatments (excluding the *Outcome Feedback*) and includes controls for awareness about academic performance. Results are unchanged and the coefficients associated with awareness about academic performance do not achieve statistical significance.

[14] 299 out of 408 subjects (73%) obtained a score strictly higher than zero in the Implicit Association Test. A positive score translates into subjects holding implicit gender-stereotyped beliefs on mathematical ability. Almost the same number of men and women (150 women and 149 men) scored positive on the IAT.

Table 4
Treatment effect: Assignment to the Challenging Task in the Demand-side Game.

Dependent variable: Pr (Task = 1) **Estimation technique**: Probit Regression, Marginal Effects				
Population: **Model**:	All (1)	All (2)	All (3)	All (4)
Female	−0.112**	−0.115**	−0.138**	−0.137**
	0.013	0.018	0.016	0.019
MFeedback T	0.001	−0.008	0.001	−0.007
	0.001	0.007	0.001	0.007
MFeedback T x Female	–	0.026	–	0.023
		0.022		0.022
Baseline T	−0.002	0.007	−0.002*	0.006
	0.001	0.007	0.001	0.007
Baseline T x Female	–	−0.027	–	−0.027
		0.023		0.023
OFeedback T	−0.001	−0.001	−0.001	0.004
	0.001	0.007	0.001	0.007
OFeedback T x Female	–	0.002	–	−0.016
		0.022		0.022
Ranking	−0.017**	−0.016**	−0.017**	−0.017**
(Signal, continuous)	0.006	0.006	0.006	0.006
Age (continuous)	−0.012**	−0.012**	−0.012**	−0.012**
	0.002	0.002	0.002	0.002
STEM (0–1)	0.242**	0.242**	0.242**	0.242**
	0.012	0.012	0.012	0.012
Gender of the	−0.001	−0.001	−0.001	−0.001
Employer (0–1)	0.001	0.001	0.001	0.001
IAT score > 0 (0–1)	−0.001	−0.001	−0.001	−0.001
	0.001	0.001	0.001	0.001
Aware 1y (0–1)	0.000	0.000	−0.017**	−0.018**
	0.001	0.001	0.006	0.006
Aware 5y (0–1	0.000	0.001	−0.003	−0.003
	0.001	0.001	0.006	0.006
Supply Game first	−0.002*	−0.002*	−0.002*	−0.002*
	0.001	0.001	0.001	0.001
Aware 1y (0–1) x Females	–	–	0.050**	0.053**
			0.019	0.018
Aware 5y (0–1) x Females	–	–	0.009	0.009
			0.019	0.02
Additional Controls	✓	✓	✓	✓
Observations	9,792	9,792	9,792	9,792
Wald chi2	2623.02***	2023.23***	1470.16***	1298.49***
Linear combination post-estimation tests				
Female + Feedback x Female		−0.089**		−0.114**
Female + Baseline x Female		−0.142**		−0.640**
Female + Outcome Feedback		−0.114**		−0.153**
Baseline x Female − MFeedback x Female		−0.054**		−0.051**
Baseline x Female − OFeedback x Female		−0.023		−0.012
MFeedback x Female − OFeedback x Female		0.025		0.039

Note: The table shows the marginal effects of different probit estimations of the probability of candidates being assigned to the Challenging Task. The overall number of observations corresponds to 9792 (408 evaluators x 6 candidates x 4 rounds). Additional controls include candidates' position in the decision screen, number of women in the pool, number of women among the top 1 performers according to the signal, and period. Standard errors, clustered at the employer level, are in parentheses. All p-values are corrected using the Benjamini-Hochberg False Discovery Rate method (Benjamini and Hochberg, 1995). Specifically, in this case, we are executing 12 comparisons: (3 treatments, 2 genders, and 2 games).
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

or after the Supply-side Game.

In Model 1, we observe that female candidates have a lower probability of being assigned to CT compared to male candidates, while the treatment coefficients do not display significant additional effects. Among the other controls, we find that the signal about the ability as well as the age and the field of study of the candidate have a significant impact on the probability of being hired. In Model 2

we introduce the interactions between Female candidates and Treatments. The coefficients of the interactions do not achieve statistical significance, while the dummy Female candidate remains negative and highly significant.

In Models 3 and 4 we add to Models 1 and 2 an interaction between the dummy accounting for awareness of the decision maker concerning the gaps in outcomes at 1 and 5 years and the dummy Female capturing the gender of the candidate.[15] Including this interaction deserves specific attention since as previously discussed, the impact of the information provided in the different treatments may depend on the accuracy of students' expectations about the labor market outcomes. Results already evidenced in Models 1 and 2 remain unchanged, except for the fact that, in Model 3, being in the *Baseline Treatment* has an additional negative and significant effect compared to the *Salience Treatment*, while the two feedback treatments continue not to display significant additional impact. When considering the interaction effect between the awareness dummies and the gender of the candidate, we find a positive and significant effect for the dummy capturing awareness about the gender gaps at 1 year both in Models 3 and 4 suggesting that −irrespective of treatment- individuals who are aware of gaps in labor market outcomes are also more likely to assign female candidates to the Challenging Task.[16] This finding suggests that more informed participants, acting as employers, may feel greater sympathy and identification with the young job seekers they are selecting. These participants are likely to empathize more with the job candidates listed in the demand-side game, perceiving them as peers—young graduates from similar fields of study.

Post-estimation tests also reveal that, compared to the *Mixed Feedback Treatment* women in the *Baseline Treatment* have a significant lower probability of being assigned to the Challenging Task of about 5 %. Interestingly, in none of the models, the IAT coefficient achieves statistical significance, as well as the gender of the employer who decides about the task assignment.

Results displayed in Table 4 remain unchanged if instead of using the dummy about being aware of the gap we use the number of correct answers (see Table A7 in Section A of the Appendix).

To better analyze the role of awareness evidenced in Table 4, we conduct two additional analyses exploiting the classification of aware subjects. First, we split the sample into four groups based on participants' awareness/unawareness of the labor market gap at 1 and 5 years from graduation and for each group, we replicate the analysis from Table 4. Results of the first exercise are reported in section A of the Appendix, Table A8a for subjects aware and not aware of gaps at 1 year and Table A8b for the subjects aware and not aware of gaps at 5 years. Results show that, for both groups, the treatment effect evidenced in Table 4 is driven by unaware individuals.

Second, we replicate the analysis displayed in Table 4 by comparing individuals who are aware of both gaps to those who are unaware of both gaps. Table C1 in Section C of the Appendix shows the distribution of subjects who are both Aware 1y and Aware 5y and those not aware in either domain. Results of this second exercise are reported in Table 5, where Models 1–4 refer to individuals who are aware of both domains while Models 5–8 refer to individuals who are not aware of either domain.

Inspection of Table 5 reveals that, as already evidenced from the findings in Tables A1a and b, the treatment effect observed in Table 4 is replicated only for participants who are not aware of gender gaps. Other results remain unchanged.

A possible explanation for this result is that, for participants who are unaware of gender gaps, the *Salience Treatment* may activate a sense of accountability. They might not have previously considered that their decisions could be judged based on gender fairness. Making the gender dimension salient may increase their awareness of potential scrutiny, leading them to behave more equitably (Lerner and Tetlock, 1999). Another possible explanation could be social desirability. If unaware subjects have not previously felt the need to conform to gender equality norms, the *Salience Treatment* may make the pressure to align with social norms more apparent, significantly influencing their behavior (Cialdini and Goldstein, 2004). In contrast, aware subjects, who already recognize gender disparities and have likely integrated this understanding into their decision-making processes, are not influenced by making the gender dimension salient. They are already operating with a heightened sensitivity to gender issues as evidenced by the results reported in Models 3 and 4 of Table 4.

The main findings from this section are summarized in Result 2.

Result 2: Demand-side game (hiring and tasks assignment). *Feedback treatments do not impact hiring decisions, nor the probability of female candidates being assigned to the Challenging Task compared to the Salience Treatment. However, the Salience Treatment shows a positive effect relative to the Baseline Treatment, driven primarily by unaware subjects and possibly due to an increased sense of accountability and/or a social desirability bias. When examining awareness of gaps in labor market outcomes, we find that individuals aware of gaps at 1 year are more likely to assign women to the Challenging Task, while awareness of gaps at 5 years does not significantly influence task assignment.*

Taken together, our results in this section suggest that salience is the driving mechanism affecting individual decision-making and that it operates mainly for subjects who are not aware of early gaps in labor market outcomes. For these subjects, making the gender dimension salient seems to activate a behavioral change in the assignment to the Challenging Task, which however is not compatible with a reminder effect.

We complete this subsection with a remark on the impact of the feedback on efficiency. Unlike Reuben et al. (2014), our experiment does not find any impact of treatments on efficiency. We defined efficiency from two perspectives: i) in terms of employers' earnings and ii) as a percentage deviation from the maximum attainable earnings. Using these two measures, we show that efficiency differences are not significant across treatments, except for the *Outcome Feedback Treatment*. Subjects in this treatment deviate from the maximum

[15] Since, in the *Outcome Feedback Treatment*, we eliminated the question about grades, this analysis does not include any variables accounting for awareness about grades. Section D of the Appendix replicates our analysis, excluding the *Outcome Feedback Treatment*, and incorporating a dummy variable for awareness about grades.

[16] In Appendix D, Table D2 replicates the analysis presented in Models 1 and 2 of Table 4 splitting the sample based on whether employers are aware of gaps in academic performance and early gaps in market outcomes. For all these models, the coefficient of Female candidate is negative and significant, indicating a detrimental effect for women in the *Baseline Treatment* while the coefficient of *Feedback* is not statistically significant.

Table 5
Treatment effect: Assignment to Challenging Task in the Demand-side Game depending on the awareness of both gaps.

Dependent variable: Pr (Task = 1) **Estimation technique**: Probit Regression, Marginal Effects								
Population:	**Aware 1y and 5y**				**NOT Aware 1y and 5y**			
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	−0.056**	−0.066*	−0.056*	−0.066*	−0.125**	−0.110**	−0.125**	−0.110**
	0.02	0.031	0.02	0.031	0.019	0.025	0.019	0.025
MFeedback T	−0.002	−0.013	−0.002	−0.013	0.002	0.002	0.002	0.002
	0.002	0.014	0.002	0.014	0.002	0.01	0.002	0.01
MFeedback T x Female	–	0.031	–	0.031	–	−0.001	–	−0.001
		0.038		0.038		0.033		0.033
Baseline T	0.000	−0.011	0.000	−0.011	−0.005**	0.015	−0.005**	0.015
	0.003	0.02	0.003	0.02	0.002	0.009	0.002	0.009
Baseline T x Female	–	0.029	–	0.029	–	−0.063*	–	−0.063*
		0.052		0.052		0.03		0.03
OFeedback T	−0.001	−0.001	−0.001	−0.001	−0.001	0.008	−0.001	0.008
	0.002	0.012	0.002	0.012	0.003	0.01	0.003	0.01
OFeedback T x Female	–	−0.002	–	−0.002	–	−0.026	–	−0.026
		0.035		0.035		0.033		0.033
Ranking (Signal, continuous)	−0.030**	−0.030**	−0.030*	−0.030**	−0.011	−0.011	−0.011	−0.011
	0.01	0.01	0.01	0.01	0.008	0.008	0.008	0.008
Age (continuous)	−0.013**	−0.013**	−0.013**	−0.013**	−0.008**	−0.008**	−0.008**	−0.008**
	0.003	0.003	0.003	0.003	0.002	0.002	0.002	0.002
STEM (0–1)	0.238**	0.238**	0.238**	0.238**	0.224**	0.224**	0.224**	0.224**
	0.02	0.02	0.02	0.02	0.019	0.019	0.019	0.019
Gender of the	0.000	0.000	0.000	0.000	−0.002	−0.002	−0.002	−0.002
Employer (0–1)	0.001	0.001	0.001	0.001	0.002	0.002	0.002	0.002
IAT score > 0 (0–1)	0.003*	0.003*	0.003*	0.003*	−0.001	−0.001	−0.001	−0.001
	0.001	0.001	0.001	0.001	0.002	0.002	0.002	0.002
Supply Game first	0.000	0.000	0.000	0.000	−0.004**	−0.004**	−0.004**	−0.004**
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Additional Controls	✓	✓	✓	✓	✓	✓	✓	✓
Observations	3,480	3,480	3,480	3,480	4,488	4,488	4,488	4,488
Wald chi2	810.72***	641.07***	810.72***	641.07***	1427.46***	939.44***	1427.46***	939.44***
Linear combination post-estimation tests								
Female + Feedback x Female		−0.035	–	−0.035	–	−0.111	–	−0.111
Female + Baseline x Female		−0.037	–	−0.037	–	−0.173	–	−0.173
Female + Outcome Feedback		−0.0675	–	−0.068	–	−0.136	–	−0.136
Baseline x Female − MFeedback x Female		−0.002	–	−0.002	–	−0.062	–	−0.062
Baseline x Female − OFeedback x Female		0.031	–	0.031	–	−0.037	–	−0.037
MFeedback x Female − OFeedback x Female		0.033	–	0.033	–	–	–	–

Note: The table shows the marginal effects of different probit estimations of the probability of candidates being assigned to the Challenging Task. The overall number of observations corresponds to 7488 (312 evaluators x 6 candidates x 4 rounds). Additional controls include the order of the tasks, candidates' position in the decision screen, number of women in the pool, number of women among the top 1 performers according to the signal, and period. Standard errors, clustered at the employer level, are in parentheses. All p-values are corrected using the Benjamini-Hochberg False Discovery Rate method (Benjamini and Hochberg, 1995). Specifically, in this case, we are executing 12 comparisons: (3 treatments, 2 genders, and 2 games). *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

earnings associated with ST (7.2 %) significantly more than subjects in the Baseline ($z = 2.749$, $p = 0.006$), Salience ($z = 3.821$, $p = 0.001$), and *Mixed Feedback Treatment* ($z = 2.366$, $p = 0.02$). However, both the total deviation and the deviation from the maximum earnings in the challenging task are not significantly different. See Table A9 in Section A of the Appendix for a more detailed analysis. The minimal difference across treatments depends on the specific groups' composition (displayed in Tables A1a and b in Section 1 of the Appendix) and on the performance of the groups' members in the two tasks, which do not display substantial gender differences.

4.3. Supply-side game: Willingness to compete under feedback provision

In this section, we focus on the supply side of the market. Following the literature, the willingness to compete is expressed through the dichotomous decision in Stage 3 of the Supply-side Game, where subjects decide whether to select a tournament payment scheme or opt for a non-competitive payment scheme (piece rate) to be applied to their performance in the addition task. In our sample, where the number of men and women is almost perfectly balanced by design (205 vs 203/408), the proportion of men who decided to compete (67 %) doubles that of women (33 %) (Mann-Whitney $z = 7.022$, $p < 0.001$). When considering the performance, we find that men outperformed women under the piece rate payment (we refer to the piece rate of Stage 1): men correctly solved, on average, 5.16 calculations (Std. Dev. = 2.73), and women solved 4.48 calculations (Std. Dev. = 2.16) correctly (Mann-Whitney test $z = 2.744$, $p =$

0.006). When considering Stage 2, the gender differences in performance are again significant in the tournament stage: men correctly solved 5.98 (Std. Dev. = 2.82) calculations, and women correctly solved 5.24 (Std. Dev. = 2.45) calculations ($z = 0.518$, $p = 0.01$). While most studies using the Niederle-Vesterlund math task do not detect gender differences in performance, Große and Riener (2010) and Kessel et al. (2021) found that men significantly outperformed women, a result that is in line with our findings. In our experiment, participants in the Supply-side game only receive feedback about the tournament outcome at the end of the entire experiment, therefore their updating of beliefs about gender differences in performance during the experiment is limited. Nonetheless, all our econometric analyses account for prior experience with the Supply game before participating in the Demand-side game; see Tables 4 and 5.

Consider now Fig. 3. Panel A shows the proportion of subjects who decided to compete in each treatment. It can be noted that the *Mixed Feedback Treatment* significantly increases the share of women choosing to compete compared to the *Baseline*, but this does not help to close the gender gap in willingness to compete because men's willingness to compete significantly increases as well. Eventually, the difference in the competing gap between *Baseline* and *Mixed Feedback* is not significant. Within each treatment, men decided to compete more often than women (*Baseline*: $z = 2.625$, $p = 0.008$; *Salience*: $z = 4.340$, $p < 0.001$; *Mixed Feedback*: $z = 3.271$, $p = 0.001$; *Output Feedback*: $z = 3.697$, $p = 0.001$).

Panel B displays the decision to compete by gender over the pool of same-sex subjects. Results mirror the ones depicted in panel A, with the difference that the *Baseline* vs *Mixed Feedback* comparison does not achieve statistical significance. Other comparisons are not statistically significant as well.

Beyond gender and the treatment variation, we expect that other aspects may influence the decision to compete such as subjects' self-confidence,[17] the decision environment, and the subjects' personal characteristics. We control for these characteristics in different specifications of a probit model that regresses the probability of choosing the competitive payment scheme in Stage 3 (*Choice*). The marginal effects of these models are presented in Table 6.

We consider the following independent variables: the subjects' gender, the number of correct calculations under tournament conditions in Stage 1 and Stage 2, the field of study, age, self-confidence, whether the subject played the Supply-side Game before or after the Demand Game, awareness of the gap in market outcomes, and implicit gender stereotypes gathered from the Implicit Association Test.

In Model 1, which considers the entire sample, the factors significantly predicting the probability of choosing to compete are gender, tournament performance, field of study, and self-confidence. These findings align with previous literature (Niederle and Vesterlund, 2007; Kamas and Preston, 2012). Specifically, higher tournament performance and studying Engineering or Economics and Management (as opposed to studying Law) positively influence the choice to compete, while being female and underconfident negatively influence it. When considering the dummies capturing the treatment effect, we do not find any significant results. However, individuals classified as aware 5y are significantly more likely to compete. Model 2 is identical to Model 1 but introduces an interaction between being female and the treatment dummies. The interactions between Female and the treatment dummies are not significant, while other results remain unchanged. Interestingly, for both models, we find no evidence of an order effect between the games: playing the Supply-side Game before or after the Demand Game does not affect the willingness to compete.

Models 1 and 2 compare the relationship between gender and the three treatments. Models 3 to 6 split the sample by gender to determine the pure within-gender effect of the treatments. In Models 3 and 4, which only consider women, we do not find any impact of the treatments, suggesting that feedback provision about labor market outcomes has a positive impact on women's willingness to compete. We also observe that tournament performance, the field of study, and confidence play a meaningful role in women's decision to compete, while awareness of gender gaps does not achieve significance. When in Model 4 we consider the interactions between the IAT_score > 0 dummy and the treatments, we find that being in the *Mixed Feedback Treatment* for women who hold an IAT_score > 0 is associated with a higher competitiveness. This could suggest that the *Mixed Feedback Treatment* works as a debiasing mechanism for females who hold implicit stereotypes by inducing them to compete. Since the dummy for the *Outcome Feedback Treatment* is not significant, this effect is likely driven by the feedback about academic performance. In Table D3 in section D of the Appendix, we replicate the analysis contained in Model 6 excluding the *Outcome Feedback Treatment* and including the dummy accounting for awareness of academic performance.

In Models 5 and 6, which consider men alone, the field of study is not relevant anymore in explaining the decision to compete as well as self-confidence. The tournament performance and holding an implicit stereotype (IAT_score > 0) are essentially the only factors playing a key role together with the awareness of labor market outcomes at 5 years. Model 6 confirms that performance in previous stages is the only explanator of men's willingness to compete.

Tables A10a and b in Section A of the Appendix repeat the analysis presented in the models displayed above but split the sample into different subgroups based on subjects' awareness of labor market outcomes. In Table A10a, we consider individuals who are aware of both gaps at 1 and 5 years, while in Table A10b, we only consider individuals who are not aware of the gaps. In both tables, irrespective of awareness of gender gaps in market outcomes, women are less likely to choose competition than men. The coefficient for Female is negative and significant in all models, with a larger magnitude for aware subjects. Additionally, we observe that female participants holding implicit stereotypes significantly increase their willingness to compete in both the *Mixed Feedback*, the *Outcome Feedback,* and *Baseline Treatments*, supporting the debiasing mechanism suggested in Model 6. For unaware women holding stereotypes

[17] 77 subjects (25%) are categorized as underconfident individuals (39 are men and 38 are women). This is done by comparing individuals' expected ranking to their actual ranking in the tournament. 50% of the sample were overconfident, and the remaining 25% were correct in their guess.

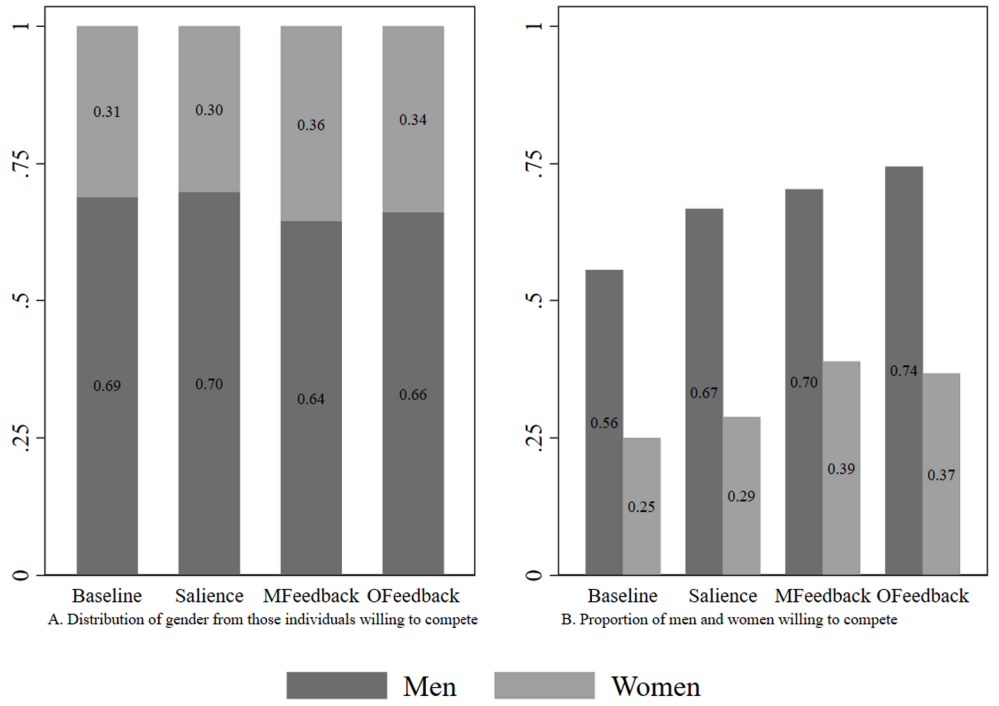


Fig. 3. Proportion of subjects willing to compete in the Supply-side Game. **Note:** Panel A shows the proportion of women and men among those subjects who decided to compete (N=29 in *Baseline*, N=63 in *Salience*, N=59 in *Mixed Feedback*, and N=53 in *Outcome Feedback*). Panel B shows the proportion of men and women who decided to compete over the pool of same-sex subjects (*Baseline*: n = 36, *Salience* n = 66, *Mixed Feedback*: n = 54, *Outcome Feedback*: n = 49 female, 47 male).

(Table A10b), the *Baseline* has a negative effect while the *Outcome Treatment* confirms a positive effect. However, due to the limited sample size, we do not further discuss this finding.

Our findings on willingness to compete are summarized in Result 3.

Result 3: Supply-side game (willingness to compete). *Feedback treatments enhance the willingness to compete of both male and female participants, yet they do not alter the gender gap in competition. The Mixed Feedback Treatment significantly increases the willingness to compete for female participants who hold implicit biases (IAT>0) compared to females in the Salience Treatment. This result suggests the existence of a debiasing mechanism operating in the sample of women. Holding implicit biases and being aware of the gaps at 5 years positively affects men's willingness to compete irrespective of treatment. Under-confidence negatively affects the willingness to compete for female but not for male participants.*

Taken together, our analysis of the Supply-side Game shows that besides the already known determinants for individual willingness to compete (e.g., past performance and underconfidence) specific features of our experiment design deserve attention. Holding implicit biases as captured by the IAT has a differential impact on men and women. While men holding higher implicit stereotypes are more willing to compete irrespective of treatment, women holding implicit biases are likely to respond to the positive piece of information on academic performance contained in the Mixed Feedback Treatment, as discussed above.

Moving to awareness, we find that, in the Supply-side Game, being aware of gaps in labor market outcomes displaying after 5 years from graduation has a positive impact on competitiveness, however, this effect is driven by males. Moreover, we find that, for women who are aware of both gaps at 1 and 5 years, *Mixed Feedback* has a positive impact on competitiveness, but this effect seems to be driven by women who hold implicit stereotypes. For them, also the *Outcome Feedback* and the *Baseline Treatment* have an impact on the willingness to compete compared to the *Salience Treatment*.

The coexistence of negative news regarding gender gaps in market outcomes and positive news about women's higher academic performance might help explain why *Mixed Feedback Treatment* fails to influence the behaviors of women unaware of gender gaps. Our interpretation posits that, for women lacking awareness of gender gaps, the positive and negative pieces of information may have offset each other. Nevertheless, the insightful and policy-relevant takeaway from this lack of reaction to *Mixed Feedback* by these women is that information about gender gaps in labor market outcomes doesn't appear to deter (previously unaware) female students from continuing to compete at their previous levels. In addition, the fact that the coefficient of the *Outcome Feedback Treatment* does not achieve statistical significance suggests that the positive impact of *Mixed Feedback Treatment* is driven by the piece of positive news about academic performance. Regarding the effect of information on women's academic performance, our results can be seen as complementary to those of Neschen and Hügelschäfer (2021). They designed two lab experiments showing that information about quotas harmed the performance evaluation of all women (not only the quota's targets), as it served as a negative signal regarding women's abilities. Our results suggest instead that information on the gender gap in academic ability can be encouraging for women.

Table 6
Treatment effect by gender: Willingness to compete in the Supply-side Game.

Dependent Variable	Pr (Choice = 1)					
Estimation Technique	Probit Regression, Marginal Effects					
Population	All	All	Women	Women	Men	Men
Model	(1)	(2)	(3)	(4)	(5)	(6)
Female (0–1)	−0.339**	−0.348**	–	–	–	–
	0.057	0.100				
MFeedback (0–1)	0.102	0.092	0.132	−0.199	0.037	0.269
	0.07	0.102	0.088	0.164	0.089	0.177
MFeedback (0–1) x Female	–	0.020	–	–	–	–
		0.142				
Baseline (0–1)	−0.038	−0.055	−0.039	−0.025	−0.057	−0.010
	0.084	0.115	0.114	0.178	0.100	0.199
Baseline (0–1) x Female	–	0.035	–	–	–	–
		0.167				
OFeedback (0–1)	0.090	0.095	0.179	0.044	−0.031	0.208
	0.094	0.125	0.111	0.212	0.122	0.268
OFeedback (0–1) × Female	–	−0.011	–	–	–	–
		0.150				
Performance Piece Rate	0.022	0.022	0.015	0.017	0.020	0.015
	0.016	0.016	0.021	0.022	0.018	0.018
Performance Tournament	0.052**	0.051**	0.059**	0.063**	0.036	0.041*
	0.016	0.016	0.021	0.022	0.018	0.019
Engineering (0–1)	0.192*	0.191*	0.296**	0.262**	0.033	0.033
	0.076	0.076	0.097	0.098	0.098	0.101
Economics (0–1)	0.168*	0.167*	0.140	0.118	0.114	0.104
	0.074	0.074	0.086	0.087	0.109	0.113
Age (Continuous)	0.009	0.009	0.015	0.017	0.000	−0.002
	0.011	0.011	0.012	0.012	0.014	0.015
Underconfident (0–1)	−0.217**	−0.217**	−0.263**	−0.548**	−0.135	−0.187
	0.064	0.064	0.085	0.206	0.078	0.151
IAT_score > 0 (0–1)	0.029	0.028	−0.03	−0.128	0.110	0.285**
	0.064	0.064	0.077	0.131	0.077	0.134
Supply Game first (0–1)	0.064	0.065	0.068	0.086	0.050	0.045
	0.055	0.055	0.070	0.071	0.070	0.072
Aware 1 (0–1)	−0.098	−0.099	−0.036	−0.04	−0.176	−0.162
	0.068	0.068	0.081	0.082	0.095	0.095
Aware 5 (0–1)	0.190**	0.189**	0.072	0.043	0.267**	0.242*
	0.073	0.073	0.09	0.093	0.101	0.102
Underconfident (0–1) × Baseline	–	–	–	0.338	–	0.281
				0.283		0.224
Underconfident (0–1) × MFeedback	–	–	–	0.215	–	0.066
				0.253		0.206
Underconfident (0–1) x OFeedback	–	–	–	0.468*	–	0.074
				0.266		0.224
IAT_score > 0 × MFeedback	–	–	–	0.426*	–	−0.308
				0.208		0.202
IAT_score > 0 × Baseline	–	–	–	−0.071	–	−0.184
				0.220		0.220
IAT_score > 0 × OFeedback	–	–	–	0.140	–	−0.346
				0.228		0.275
Observations	408	408	266	266	142	142
Wald chi2	72.73***	30.42***	3.84	7.02	4.56	1036.62***
Linear combination post-estimation tests						
Female + Feedback x Female −		−0.329	–	–	–	–
Female + Baseline x Female	–	−0.313	–	–	–	–
Female + Outcome Feedback	–	−0.359	–	–	–	–
Baseline x Female − Feedback x Female		0.015	–	–	–	–
Baseline x Female − OFeedback x Female		0.046	–	–	–	–
Feedback x Female − OFeedback x Female		0.030	–	–	–	–

Note: The table shows the marginal effects of different probit estimations of candidates' probability of choosing the competitive payment scheme in Stage 3 of the Supply-side Game. Standard errors, clustered at the individual level, are in parentheses. All p-values are corrected using the Benjamini-Hochberg False Discovery Rate method (Benjamini and Hochberg, 1995). Specifically, in this case, we are executing 12 comparisons: (3 treatments, 2 genders, and 2 games).
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

5. Conclusion

In our laboratory experiment, we first assessed Italian college students' awareness of gender gaps in their specific fields of study within the Italian labor market. We then analyzed how providing information about these gaps influenced their behavior in a series of stylized games linked to real labor market scenarios. Specifically, we examined the impact of information on their choices as workers or employers. Our feedback treatments focused on gender gaps relevant to the students' fields of study in the local labor market, providing two key pieces of information crucial for decision-making in a competitive environment.

In the *Outcome Feedback Treatment*, students received information about various gender gaps in labor market outcomes specific to their fields of study—such as employment probabilities, salary levels, and the likelihood of self-employment one and five years after graduation—all of which favored men. In the *Mixed Feedback Treatment*, in addition to this local labor market information, students were informed about the academic performance gap in their college and field of study. In Italy, as in other countries (references are provided in the Introduction), women outperform men academically across all fields of study. The feedback provided was based on publicly available data from AlmaLaurea (https://www.almalaurea.it/), the statistical office of Italian public universities, reflecting the academic performance and the labor market outcomes of graduates from the same programs a few years earlier. Since eliciting beliefs about gender gaps inherently highlights these disparities, we compared the impact of the two feedback treatments with the effect of simply filling out the questionnaire before playing the games (*Salience Treatment*).

We found that Italian college students, regardless of their field of study, lack awareness of gender gaps. First, most participants incorrectly believe that male students achieve better grades in their college and field of study. Second, when asked about local labor market outcomes specific to their fields of study, most respondents are unaware of the gaps present one year and five years after graduation. Specifically, most respondents answered correctly to only one or none of the three questions regarding employment probability, salary levels, and the likelihood of self-employment one year and five years after graduation. Additionally, a significant portion of students (about 30 %) mistakenly believe they will enter a gender-neutral workplace where men and women achieve the same labor market outcomes.

The results suggest that the observed change in employer behavior in the Demand Game, specifically assigning more women to the challenging task, is primarily driven by making gender salient. This effect is most pronounced among participants who were previously unaware of labor market outcome gaps. Highlighting gender seems to prompt a behavioral change that may not be due to a simple reminder effect but could instead arise from a sense of accountability and/or a desire to conform to socially desirable norms. Conversely, results from the Supply Game indicate that women's increased willingness to compete in response to the feedback treatments is likely driven by a debiasing intervention.

From our findings, we can draw several policy implications.

Since college students appear to be largely unaware of the existing early gender gaps in labor market outcomes, our first policy implication underscores the necessity of continuing to provide information, despite the ongoing media debate on this issue.

Secondly, our result about the increased assignment of women to the challenging task points to the relevance of interventions targeting employers. In this respect, note that feedback treatments do not increase the share of hired women. Still, emphasizing women's academic performance helps increase the number of women selected for the challenging task. Hence, our findings suggest that priming gender gaps in academic performance and early gaps in labor market outcomes could help increase gender equity inside organizations by decreasing the glass ceiling.

Finally, a primary concern when presenting information about early gender gaps in the labor market is the potential for this knowledge to negatively impact women's confidence and motivation. However, our result on female students' willingness to compete suggests that providing a positive message about women's academic abilities may counteract this discouragement. Notably, we observed that women did not reduce but instead increased their willingness to compete.

Funding

This study was carried out within the project Gender Stereotypes and Women's Achievements (GENS), CUP J53D23004290001.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.joep.2024.102770.

References

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

AlmaLAurea (2022). Rapporto 2022: LAUREATE E LAUREATI: SCELTE, ESPERIENZE E REALIZZAZIONI PROFESSIONALI. Report available at https://www.almalaurea.it/sites/almalaurea.it/files/convegni/gennaio2022/6_almalaurea_rapportocompleto_laureatelaureati.pdf.
Babcock, L., Recalde, M. P., & Vesterlund, L. (2017a). Gender Differences in the Allocation of Low-Promotability Tasks: The Role of Backlash. *American Economic Review, 107*(5), 131–135.
Baier, A., Davis, B., & Jaber-Lopez, T. (2024). Gender, choice of task, and the effect of feedback on competition: An experiment. *Journal of Economic Psychology, 103*, Article 102743.
Balafoutas, L., & Sutter, M. (2012). Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science, 335*(6068), 579–582.
Balleer, A., Duernecker, G., Forstner, S., & Goensch, J. (2023). Biased expectations and labor market outcomes: Evidence from German survey data and implications for the East-West wage gap. CESifo Working Paper, No. 10336.
Benjamini, Y., & 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.
Bertrand, M. (2020). Gender in the 21st Century. *American Economic Association Papers and Proceedings, 110*(5), 1–24.
Bovini, G., De Philippis, M., & Rizzica, L. (2024). *The Origins of the Gender Pay Gap: Education and Job Characteristics*. Bank of Italy: Temi di Discussione.
Bühren, C., Gschwend, M., & Krumer, A. (2024). Expectations, gender, and choking under pressure: Evidence from alpine skiing. *Journal of Economic Psychology, 100*, Article 102692.
Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics, 129*(3), 1409–1447.
Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics, 134*(3), 1163–1224.
Carroll, M. (2023). *Sex gaps in education*. Cambridge University Press & Assessment.
Chen, Y., & He, Y. (2021). Information acquisition and provision in school choice: An experimental study. *Journal of Economic Theory, 197*, Article 105345.
Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*(1), 591–621.
Combet, B., & Oesch, D. (2019). The gender wage gap opens long before motherhood: Panel evidence on early careers in Switzerland. *European Sociological Review, 35*(3), 332–345.
Conger, D., & Long, M. C. (2010). Why are men falling behind? Gender gaps in college performance and persistence. *The ANNALS of the American Academy of Political and Social Science, 627*(1), 184–214.
Cortés, P., Pan, J., Pilossoph, L., Reuben, E., & Zafar, B. (2023). Gender differences in job search and the earnings gap: Evidence from the field and lab. *The Quarterly Journal of Economics, 138*(4), 2069–2126.
De Pater, I. E., Van Vianen, A. E., & Bechtoldt, M. N. (2010). Gender differences in job challenge: A matter of task allocation. *Gender, Work & Organization, 17*(4), 433–453.
Dohmen, T., & Falk, A. (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review, 101*(2), 556–590.
Domínguez, J. J., & Montinari, N. (2021). Gender Quotas and Task Assignment in Organizations. WP 21/13, Department of Economic Theory and Economic History of the University of Granada.
Ertac, S., & Szentes, B. (2011). The effect of information on gender differences in competitiveness: Experimental evidence. Koç University-TUSIAD Economic Research Forum Working Papers 1104, Koc University-TUSIAD Economic Research Forum.
Filippin, A., & Ichino, A. (2005). Gender wage gap in expectations and realizations. *Labour Economics, 12*(1), 125–145.
Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental economics, 10*(2), 171–178.
Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of personality and social psychology, 74*(6), 1464.
Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association, 1*(1), 114–125.
Große, N. D. ; Riener, G. (2010). Explaining gender differences in competitiveness: Gender-task stereotypes, Jena Economic Research Papers, No. 2010,017, Friedrich Schiller University Jena and Max Planck Institute of Economics, Jena.
Heinz, M., Normann, H. T., & Rau, H. A. (2016). How competitiveness may cause a gender wage gap: Experimental evidence. *European Economic Review, 90*, 336–349.
Kamas, L., & Preston, A. (2012). The importance of being confident; gender, career choice, and willingness to compete. *Journal of Economic Behavior & Organization, 83*(1), 82–97.
Kessel, D., Mollerstrom, J., & Van Veldhuizen, R. (2021). Can simple advice eliminate the gender gap in willingness to compete? *European Economic Review, 138*, Article 103777.
Kim, S., Klager, C., & Schneider, B. (2019). The effects of alignment of educational expectations and occupational aspirations on labor market outcomes: Evidence from NLSY79. *The Journal of Higher Education, 90*(6), 992–1015.
Lénárd, T., Horn, D., & Kiss, H. J. (2024). Competition, confidence and gender: Shifting the focus from the overconfident to the realistic. *Journal of Economic Psychology, 104*, Article 102746.
Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*(2), 255.
Mueller, A. I., & Spinnewijn, J. (2023). Expectations data, labor market, and job search. *Handbook of Economic Expectations*, 677–713.
Neschen A., & Hügelschäfer S. (2021). Gender bias in performance evaluations: The impact of gender quotas. *Journal of Economic Psychology*, 85, Article 102383.
Niederle, M. (2017). A gender agenda: A progress report on competitiveness. *American Economic Review, 107*(5), 115–119.
Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics, 122*(3), 1067–1101.
Piazzalunga, D. (2018). The gender wage gap among college graduates in Italy. *Italian Economic Journal, 4*(1), 33–90.
Peterle, E., & Rau, H. A. (2017). Gender differences in competitive positions: Experimental evidence on job promotion. CEGE Number 303 – February 2017.
Reuben, E., Rey-Biel, P., Sapienza, P., & Zingales, L. (2012). The emergence of male leadership in competitive environments. *Journal of Economic Behavior & Organization, 83*(1), 111–117.
Reuben, E., Sapienza, P., & Zingales, L. (2014). How stereotypes impair women's careers in science. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, 111(12), 4403–4408.
Reuben, E., Sapienza, P., & Zingales, L. (2019). Taste for competition and the gender gap among young business professionals. Columbia Business School Research Paper No. 15-92. Available at SSRN: https://ssrn.com/abstract=2677298 or http://dx.doi.org/10.2139/ssrn.2677298.
Salvi del Pero, A., & Bytchkova, A. (2013). A Bird's Eye View of Gender Differences in Education in OECD Countries. OECD Social, Employment and Migration Working Papers, No. 149, OECD Publishing, Paris, https://doi.org/10.1787/5k40k706tmtb-en.
Samek, A. (2019). Gender differences in job entry decisions: A university-wide field experiment. *Management Science, 65*(7), 3272–3281.
Säve-Söderbergh, J. (2019). Gender gaps in salary negotiations: Salary requests and starting salaries in the field. *Journal of Economic Behavior & Organization, 161*, 35–51.
Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science, 10*, 80–83.
Sipe, S., Johnson, C. D., & Fisher, D. K. (2009). University students' perceptions of gender discrimination in the workplace: Reality versus fiction. *Journal of Education for Business, 84*(6), 339–349.
Verbree, A. R., Hornstra, L., Maas, L., & Wijngaards-de Meij, L. (2023). Conscientiousness as a predictor of the gender gap in academic achievement. *Research in Higher Education, 64*(3), 451–472.
von Bieberstein, F., Jaussi, S., & Vogel, C. (2020). Challenge-seeking and the gender wage gap: A lab-in the-field experiment with cleaning personnel. *Journal of Economic Behavior & Organization, 175*, 251–277.
Wozniak, D., Harbaugh, W. T., & Mayr, U. (2014). The menstrual cycle and performance feedback alter gender differences in competitive choices. *Journal of Labor Economics, 32*(1), 161–198.