Stereotypes and Tournament Self-Selection: A Theoretical and Experimental Approach*

Iñigo Hernandez-Arenaz†

January 31, 2020

Abstract

We present a theoretical model in which agents have imperfect self-knowledge about their abilities and have to self-select into either a high-paying or a low-paying tournament. The model shows that negative (positive) stereotypes generate underrepresentation (overrepresentation) of stereotyped agents in the high-paying tournament even when the stereotype is false. This is because stereotypes affect self-assessment and consequently subsequent behavior. We call this mechanism *self-stereotyping*. We run a lab experiment in which we use subjects' beliefs about the gender nature of a real-effort task to test the predictions of the theoretical model. The results of the experiment are in line with the predictions of the model for men but not for women, which partially validates the model and the self-stereotyping mechanism.

Keywords: Competitive sorting; Segregation; Self-assessment; Self-stereotyping; Stereotypes.
JEL classification numbers: C79, C91, D84, J16, J24

*I especially thank Nagore Iriberri for all her invaluable comments, suggestions and insights. I am also in-debted to Jorge Alcalde-Unzu, Antonio Cabrales, Daniel Cardona, Subhasish M. Chowdhury, Bernardo García-Pola, Oihane Gallo, Javier Hualde, Elena Iñarra, Jaromír Kováříck, Friederike Mengel, Patrick Nolen, Pedro Rey-Biel, Rudi Stracke and seminar participants at various universities for helpful comments. I am also grateful to Jose Miguel Lana, Jorge Nieto and Gloria Sanz from the Public University of Navarre for allowing me to conduct the pre-experimental survey. Financial support from Vicerrectorado de Investigación de la UPV/EHU (PIF//13/015) and Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco (IT869-13) is gratefully acknowledged. All errors are my own.

†University of the Balearic Islands. E-mail: iharenaz@gmail.com.

1 Introduction

Assessing one's own abilities is a key step in decision-making processes regarding education and labor market choices, as people have to forecast their chances of success . However, people are unable to directly assess their true abilities but rather assess them by means of imperfect signals. Previous research in social psychology suggests that stereotypes play an important role in how signals are interpreted by behaving as anchors for their interpretation (Lenney, 1977; Beyer, 1990; Beyer and Bowden, 1997; Pomerantz et al., 2002; Ehrlinger and Dunning, 2003; Chatard et al., 2007). Following this line of research, stereotypes may affect judgments and make subjects behave in ways consistent with stereotypes.

This paper studies how stereotypes, defined as *beliefs* that there are differences between social groups, affect self-selection into a high paying or a low paying tournament. This setting captures the labor market feature of competition being unavoidable in most situations but differing in intensity from one job to another. Market participants can choose the degree of competition that they want to face, so this feature directly relates to sorting in the labor market. To study this issue, we first present a game theoretical model in which different effects of stereotypes can be identified. Then we run a laboratory experiment where we test the main predictions of the theoretical model.

The theoretical model, as in Morgan et al. (2017), considers a mass of agents who must simultaneously self-select into one of two perfectly discriminatory tournaments, such that winners are declared based on their ability at performing a particular task. These tournaments differ in the prizes that they award: One pays more than (and is therefore preferred to) the other. Agents suffer from imperfect self-knowledge as they do not know their own real abilities but merely observe informative signals about them.

We introduce stereotypes into the model by splitting the whole population into two different social groups. One is targeted by a negative (positive) stereotype, which states that, on average, the stereotyped group is worse (better) than the non-stereotyped one.[1] This gives rise to beliefs about how abilities are distributed across the whole population which characterizes agents' behavior and thus, the equilibrium.

The most interesting effect of the existence of stereotypes is the appearance of a mechanism that we call *self-stereotyping*. This term reflects the fact that by including the stereotype into his or her priors an agent "*perceives himself or herself as an interchangeable exemplar of a social group rather than as a unique individual*" (Latrofa et al., 2010). In particular, this self-stereotyping mechanism manifests in the model by leading stereotyped and non-stereotyped agents who observe the *same* signal to differ in their self-assessments and, consequently, to behave differently in ways consistent with the stereotype. If the stereotype is false –i.e. no behavioral differences are justified– the model predicts group segregation between tournaments, with the group assumed ex-ante to be better being overrepresented in the high-paying tournament.

To empirically test the implications of the theory, and in particular the self-stereotyping mechanism outlined above, we run a lab experiment that replicates the setting of the theoretical model. In the experiment, we make use of a novel task that proves to be gender

[1] See Bordalo et al. (2016) for a rationalization of the existence of stereotypes

neutral in performance but highly heterogeneous regarding individual beliefs as to its gender nature. This heterogeneity in beliefs reveals how the perceived gender nature of the task affects self-assessment and the subsequent choices in self-selection into tournaments.

The bulk of the experiment consists of the following. First, subjects perform the following real-effort task 14 times: for three seconds, subjects see a picture of a glass containing a certain quantity of water. Then the picture disappears and a picture of an empty glass appears on which subjects must indicate the level of water shown in the first picture. This task proves to be gender neutral in performance and shows no clear nature as to the abilities required to perform it well, which favors heterogeneity as to which gender performs better on average. Importantly, subjects receive no feedback about their performance during or after this stage. Secondly, subjects are shown their performance in one randomly chosen repetition of the task and the average performance distribution in the session. Based on this information, they are asked to self-assess by estimating their average performance over the 14 times. Thirdly, subjects must choose between two options (tournaments) –A and B– under nine different situations that vary in how many prizes each option offers and in the relative size of those prizes. In all situations, option A offers the highest prize and represent the high-paying tournament from the theoretical model. Within each option, prizes are awarded to those subjects who choose that option and show the best average performances in the real-effort task until the prizes available in that option are exhausted, i.e. if more subjects choose an option than there are prizes on offer, those with the lowest average performance in the real-effort task are left without a prize.

Once the decision as to which tournament to choose is made, we elicit beliefs about which gender performs better at the real-effort task in an incentive-compatible way. We use those beliefs to calculate a measure of how male-biased each subject perceives the task to be and analyze how those beliefs affect self-assessment, behavior, and final outcomes during the experiment.

The experimental results partly support the predictions of the theoretical model. Women do not seem to react to their beliefs about which gender performs better in the real effort-task, but men raise (lower) their self-assessment the more male-biased (female-biased) they think the task is. Accordingly, as the belief about the degree of male-affinity of the task increases, the entry rate and winning rate for men in the high-paying tournament also does. Importantly, the believed gender nature of the task impacts men and women significantly differently in self-assessment and in the probability of entering the high-paying tournament. This is noteworthy because it means that beliefs about the gender nature of the task contribute to the existence of gender gaps in these two variables. In particular, and as predicted by the theoretical model, this gap will favor men if the task is perceived to be male-biased and will favor women if it is perceived to be (sufficiently) female-biased.

To further test these results, we performed two robustness checks. The first one restricts the sample to these subjects whom the theory states are more likely to feel the effects of the self-stereotyping mechanism on their behavior (*marginal* subjects). Results on this subsample are similar to the ones on the full sample but greater in magnitude and significance. Moreover, no results are found for the remaining subjects. This shows that, as predicted by the theory, the self-stereotyping mechanism is everywhere but as regards behavior and outcomes it is only relevant for a very specific subsample of subjects.

The second robustness check controls for potential correlations between self-reported perceptions and unobservable characteristics (e.g. optimism) through a within-subjects treatment in which subjects repeat the self-assessment and the tournament choices but interacting only with others of their own gender. In this *single-sex* environment perceptions should play no role as subjects make decisions in an ex-ante homogeneous group with no room for the self-stereotyping mechanism to appear. Consistently with this, we find that in this setting perceptions have no effects on self-assessment or on tournament choices. This strongly suggests that our early findings were not driven by unobserved characteristics, reverse causality, or spurious correlations. However, the results from this treatment should be taken with a grain of salt, as it could be affected by order effects, saliency of gender, etc. These caveats are discussed further at the end of Section 3.3.2.

Both the theory and the experimental findings reported in paper suggest that in the real world, where many education and labor choices involve choosing between different competitive environments, social groups will sort differently just because of stereotypes. Specifically, the results in this paper highlight the importance of self-stereotyping as a mechanism that favors segregation. Of special interest is the case of gender-based segregation which is at the heart of the gender wage gap (Macpherson and Hirsch, 1995; Blau and Kahn, 1997; Bertrand and Hallock, 2001; Bayard et al., 2003; O'Neill, 2003; Ludsteck, 2014; Card et al., 2016; Blau and Kahn, 2017), as women are negatively stereotyped in at least three areas strongly related to wages: Quantitative skills (Frome and Eccles, 1998; Nosek et al., 2009), leadership (Schein, 2001; Atwater et al., 2004), and general IQ (Furnham and Gasson, 1998; Furnham et al., 2002; Petrides et al., 2004; Bian et al., 2017). Previous research has already shown that gender differences in sorting in the labor market follow stereotypical patterns (Cejka and Eagly, 1999; Thébaud, 2010; Fernandez and Friedrich, 2011; Barbulescu and Bidwell, 2013; Leslie et al., 2015), but their underlying mechanisms are far from understood. According to the model proposed in this paper those stereotypes hurt women's self-assessment even if they are false and therefore undermine their professional goals, leading them to self-select into lower-paying itineraries than men.[2]

In economics, experimental studies have also shown the importance of self-assessment of own abilities in decision-making. The paper most closely related to ours is that by Coffman (2014). She shows that when individuals are asked general knowledge questions they display different degrees of confidence in their answers depending on how gender-congruent the topic is, which leads to inefficiencies in cooperative settings. By contrast, we assess the effects of the self-stereotyping mechanism in a competitive setting in which subjects' incentives are not aligned.[3] In a similar way, Coffman et al. (2019) use a measure similar to ours

[2]In addition to a purely normative motivation concerned with fairness, recent studies have shown labor market segregation has important effects at aggregate level. In particular, a recent study by Hsieh et al. (2019) estimates that around 40% of the growth in the US between 1960 and 2010 can be attributed to a change in the labor market sorting of blacks and women.

[3]There are other noteworthy differences between the experiment in this paper and that in Coffman (2014). To start with, in Coffman (2014) the stereotypes are confirmed by the data –the stereotype is true, on a statistical basis– while in our experiment the task used is completely gender neutral, reinforcing the idea that beliefs are what matters most. Also, our experiment exploits individual perceptions and not average ones, providing direct evidence that *individual* and not necessarily commonly held perceptions are at the heart of self-stereotyping. In addition, her work cannot fully separate the role of normative and descriptive stereotypes while in our experiment there is only room for descriptive ones. Finally it can be argued that, given the incentive alignment

to show in the lab that the more male-oriented a question is perceived to be, the greater the propensity is for a man to be chosen to report the answer on behalf of a group. Another closely related piece of research is Klinowski (2019) where, in line with our findings, holding gender-congruent stereotypes is found to improve men's self-confidence but does not affect women's. Cubel and Sanchez-Pages (2017) also put the focus on the effects of stereotypes and find strong evidence supporting the argument that women who believe that women are better at the beauty contest game engage in a greater depth of reasoning.

The effects of stereotypes have been also studied in the context of the decision whether to compete. It is widely documented that women are less likely to participate in competitive environments (for a review, see Niederle and Vesterlund, 2011), but previous studies have shown that stereotypical perceptions behind tasks and cultural differences are important factors in explaining this gender gap (Gneezy et al., 2009; Günther et al., 2010; Booth and Nolen, 2012; Shurchkov, 2012; Dreber et al., 2014; Grosse et al., 2014; Halladay, 2017; Klinowski, 2019). This paper goes one step further and shows that stereotypical perceptions also matter in *sorting between different competitive environments*. Thus, not only are men and women more likely to compete in gender-congruent fields but even when they decide to compete the sorting within different competitive environments seems to be conditioned by stereotypes.

Finally, this paper is also related in spirit to the literature on statistical discrimination (Phelps, 1972; Arrow, 1973; Coate and Loury, 1993). However, unlike these models, here we apply the notion of stereotypes to the supply side of the labor market rather than to the demand side. This enables us to show theoretically and experimentally that the existence of stereotypes suffices to generate segregation even in the absence of discrimination of any kind or of belief as to its existence.

The rest of the paper is organized as follows. Section 2 develops the theoretical model: Subsection 2.1 characterizes the equilibrium; subsection 2.2 analyzes the equilibrium prediction to identify the effects of stereotypes on segregation and establishes our main hypotheses for experimental testing. Section 3 is devoted to the experiment: Subsections 3.1 and 3.2 detail the design of the experiment and introduce basic descriptive analyses. Subsection 3.3 analyzes the experimental data in depth. Section 4 concludes.

2 A Model of Self-Selection with Stereotypes

Consider a continuum of risk neutral agents of mass 1 (a cohort) which is split into two different social groups: A mass λ of S-agents and a mass $1-\lambda$ of N-agents.[4] Agents know which social group they belong to.

Each agent is endowed with an unobservable ability level at a task. Instead, each agent

within teams in Coffman (2014), social preferences may be playing an important role in magnifying the effect of stereotypes by "washing one's hands of" when a topic is gender incongruent.

[4]The assumption of risk neutrality plays no role in the model's conclusions but it is made for the sake of clarity. In particular all the results in subsections 2.1 and 2.2 remain qualitatively the same under (homogeneous) non-neutral risk preferences.

observes a signal of his/her ability level, r_i, which is given by the following expression:

$$r_i = a_i + \mu_i$$

where a_i is the real ability with which the agent is endowed and μ_i is a realization of a random shock distributed as a zero-mean normal with variance σ_μ^2 which is i.i.d. across agents and abilities. Let $\Phi(.)$ and $\phi(.)$ denote the CDF and PDF of this random shock and let these functions be known by the agents.

Assume that there is a stereotype concerning social group S which holds that the distribution of abilities of social group S is equal to that of social group N except for a shift in the mean. For simplicity, it is further assumed that the stereotype is public information and believed to be true by all agents.[5] Notice that at this stage we allow the stereotype to be true or not. Let agents believe that the distribution of abilities for social group $g \in \{S, N\}$ follows a normal distribution $(\widetilde{a}|g) \sim \mathcal{N}(M^g, \sigma_a^2)$ with associated CDF and PDF $F^g(.)$ and $f^g(.)$, respectively. Throughout the paper we assume that the beliefs about the distribution of abilities of social group N are correct and distributed as $(\widetilde{a}|N) \sim \mathcal{N}(M^N, \sigma_a^2)$. The stereotype can thus be modeled by assuming that agents believe that the mean ability for social group S is $M^S = (1+\alpha)M^N$, where the parameter $\alpha \in \mathbb{R}$ captures the strength and direction of the stereotype. Therefore, an $\alpha < 0$ represents a negative stereotype, while an $\alpha > 0$ a positive one.

Under this setting, $F^g(a)$ represents the prior held by an agent from group $g \in \{S, N\}$ as to the probability of having an ability lower than a and $f^g(a)$ the density of having an ability a. However, after observing the signal, the agent performs a Bayesian updating such that the posterior beliefs about his ability can be summarize through the following distribution:

$$(a|r_i, F^g) \sim \mathcal{N}\left((1-\gamma)M^g + \gamma r_i, (1-\gamma)\sigma_a^2\right) \tag{1}$$

with $\gamma = \sigma_a^2/(\sigma_a^2 + \sigma_\mu^2)$. Denote by $\hat{F}(a|r_i, F^g)$ and $\hat{f}(a|r_i, F^g)$ the CDF and PDF of this updated distribution of beliefs concerning abilities. That is, $\hat{F}(a|r_i, F^g)$ is a function which returns the probability that an agent from social group g observing a signal r_i attaches to the event of having an ability lower than a while $\hat{f}(a|r_i, F^g)$ is the density attached to having an ability level equal to a. In other words, the functions $\hat{F}(a|r_i, F^g)$ and $\hat{f}(a|r_i, F^g)$ define agent's self-assessment. From expression (1) it is clear that, if $\alpha < 0$, we obtain $\hat{F}(a|r_i, F^S) > \hat{F}(a|r_i, F^N)$. This means that for the same observed signal r_i, S-agents put more weight on the belief of having an ability level lower than a than N-agents. If $\alpha > 0$, the opposite happens. This implies a difference in self-assessment consistent with the stereotype and establishes formally what we call the *self-stereotyping mechanism*.

Agents' problem is to choose in which of two perfectly discriminating and mutually exclusive tournaments, $t \in \{A, B\}$, to participate.[6] Tournament t offers a mass $\delta_t < 1$ of prizes

[5] This assumption is required for the common prior assumption to hold without having to deal with higher-order beliefs which would require introducing extra structure and assumptions in the model. Importantly, the channel we identify through which stereotypes manifest into behavior and the object of study of the experimental design, the *self-stereotyping* mechanism, is independent of this assumption since it does not depend on the particular way in which the interaction between agents takes place. In addition, this assumption seems legitimate to the extent that a stereotype is defined as "*a fixed idea or image that many people have of a particular type of person or thing, but which is often not true in reality*" (Oxford Dictionary).

[6] We could alternatively assume that tournaments are not perfectly discriminating but there is noise in mea-

each worth $W_t \geq 0$. Within each tournament, prizes are awarded to those with the highest ability levels: if the mass of entrants in tournament t is lower than δ_t, then all entrants receive the prize W_t. Otherwise, only the mass δ_t of entrants with the highest ability levels in tournament t get the prize W_t and the remaining entrants get 0. The structure of the tournaments (amount of prizes and number of prizes available) is public information and hence known by the agents. Two final assumptions are made. First $\delta_A + \delta_B \geq 1$, which means that there is no scarcity of prizes, i.e. if there is coordination everyone could get a prize. Second, without loss of generality, tournament A is more attractive than B, i.e. $W_A > W_B$.

This setting represents a Bayesian game in which the strategy profiles map types (signal×social group) into actions in an environment that is public information and is characterized by W_A, W_B, δ_A, δ_B, $F^N(.)$, $\Phi(.)$, λ, and α. Formal definitions including the derivation of generic expected utilities are given in Appendix A.1.

2.1 Characterization of the Equilibrium

When choosing which tournament to participate in, an agent computes the expected payoff from participating in each tournament and chooses the one that returns the highest payoff. An agent's expected payoff from participating in tournament $t \in \{A, B\}$ is simply the prize awarded by this tournament to its winners (W_t) times the probability that the agent attaches to getting one of the prizes. Notice that this probability is equivalent to the self-assessed probability of being among the mass δ_t of most able entrants in that tournament. Thus, if an agent believes that the mass of entrants in tournament t is greater than δ_t, by equation (1), the agent will expect to get a prize with a probability (strictly) lower than one for all signal levels, although that probability is increasing on the signal observed. On the other hand, if the expected mass of entrants is lower than δ_t, the self-assessed winning probability will take a value of one.

Proposition 2.1. *In equilibrium, agents expect the mass of entrants in tournament A to be greater than δ_A. Consequently, the self-assessed probability of getting a prize in tournament A is strictly lower than one while that for tournament B is one.*

To prove this proposition, suppose that this is not the case and that agents expect fewer entrants in tournament A than there are prizes available there. In that case, all agents will prefer to choose tournament A because then they expect to get the prize W_A ($> W_B$) with a probability of one. This would be anticipated by all agents, so an equilibrium where agents believe that the mass of entrants in tournament A is lower than the mass of prizes cannot exist. This proposition has two important implications. First, since the mass of prizes ($\delta_A + \delta_B$) is greater than the population size, agents should believe that the number of entrants in tournament B is lower than δ_B, i.e. that by choosing tournament B they can guarantee a payoff of W_B. On the other hand, since tournament A is expected to be overcrowded, agents should also expect that they will get the prize only if their ability exceeds a certain threshold.

Proposition 2.2. *Given the environment defined by the tuple $(W_A, W_B, \delta_A, \delta_B, F^N(.), \Phi(.), \lambda, \alpha)$, the equilibrium of the game is characterized by a belief as to which minimum ability is required to get a prize in tournament A (a^α), a signal level for N-agents*

suring the ability levels of agents. Given the characterization of the equilibrium (proposition 2.2) this would not change the results. This change would merely imply a common, higher degree of uncertainty. In particular, it would imply that the self-assessed probability of winning has greater variance which would not bring any additional insight under the risk neutrality assumption.

(r^α_N), and a signal level for S-agents (r^α_S), such that all agents from social group $g \in \{S, N\}$ who observe a signal higher (lower) than the signal r^α_g will choose to participate in tournament A (B). The tuple $(a^\alpha, r^\alpha_N, r^\alpha_S)$ must hold simultaneously

1. $[1 - \hat{F}(a^\alpha | r^\alpha_g, F^g)] = W_B/W_A, \forall g \in \{S, N\}$
2. $\int_{a^\alpha}^{\infty} \lambda[1 - \Phi(r^\alpha_S - a)] f^S(a) + (1-\lambda)[1 - \Phi(r^\alpha_N - a)] f^N(a) da = \delta_A$

Proposition 2.2 formalizes the conditions that characterize the Bayesian Nash Equilibrium of the game so that beliefs and choices are consistent with each other.[7] Next, we present the intuition behind proposition 2.2 as well as of its proof.

Point 1 of proposition 2.2 shows that if agents believe that the minimum ability needed to get a prize in tournament A is a^α, there is a signal level r^α_g that makes an agent from social group $g \in \{S, N\}$ indifferent between choosing tournament A or B. This signal threshold r^α_g should be such that the self-assessed probability that the agent attaches to having an ability greater than a^α, $[1 - \hat{F}(a^\alpha | r^\alpha_g, F^g)]$, is W_B/W_A. If this holds then the expected payoff from tournament A for an agent who observes signal r^α_g will be W_B which, by proposition 2.1, is what agents expect to get by choosing tournament B. Since by equation (1) the self-assessed probability for exceeding a^α is increasing on the signal observed, the same goes for the expected payoff from tournament A. Thus, any agent who observe a signal greater (lower) than r^α_g will choose tournament A (B).

The left-hand side of the equation in point two of proposition 2.2 accounts for the mass of agents with ability greater than a^α that will enter into tournament A if agents use the signals (r^α_S, r^α_N) as the thresholds to choose tournament A. To see this, notice that the term $[1 - \Phi(r^\alpha_g - a)]$ shows the probability with which an agent from social group g with ability a is expected to receive a signal $r_i > r^\alpha_g$ and thus will choose tournament A.[8] Therefore, point 2 in proposition 2.2 simply says that, given the signal thresholds (r^α_S, r^α_N), agents expect a mass δ_A of agents with ability greater or equal than a^α entering tournament A. In other words, agents believe that prizes from tournament A can only be won by having an ability strictly greater than a^α.[9]

Thus, the tuple $(a^\alpha, r^\alpha_N, r^\alpha_S)$ by holding points 1 and 2 of proposition 2.2 makes beliefs and actions to be consistent with each other, giving rise to an equilibrium. Moreover, this equilibrium always exists and is unique.[10]

[7] Importantly, that W_A, W_B, δ_A, δ_B, $F^N(.)$, $\Phi(.)$, λ, and α are public information guarantees that the common prior assumption holds (Harsanyi, 1967-1968). Thus, all agents should agree in their posteriors beliefs (Aumann, 1976) making common knowledge that all agents, irrespectively of their social group, believe that the minimum ability to get a prize in tournament A is a^α and that agents from social group $g \in \{S, N\}$ will choose to participate in A iff $r_i > r^\alpha_g$.

[8] In particular, since $r_i = a_i + \mu_i$, this probability is equal to $Prob(r > r^\alpha_g | a) = Prob(\mu > r^\alpha_g - a) = 1 - \Phi(r^\alpha_g - a)$. Since we are working with a continuum of agents, this probability can be interpreted as the fraction of agents from social group g with ability a that observes a signal greater than r^α_g.

[9] Notice that this also implies that the total mass of entrants in A is greater than δ_A. In particular, a believed threshold of a^α makes agents to believe that the total mass of entrants in tournament A is $\delta_A + \int_{-\infty}^{a^\alpha} \lambda[1 - \Phi(r^\alpha_S - a)] f^S(a) + (1-\lambda)[1 - \Phi(r^\alpha_N - a)] f^N(a) da$ with the last term being the expected mass of agents with ability lower than a^α but observing by chance a signal greater than r^α_g.

[10] See Appendix A.2 for the formal characterization of the equilibrium and Appendix A.3 for the proof corresponding to its existence and uniqueness. As can be inferred from the two paragraphs above, these proofs rely upon the existence of two functions – one mapping minimum abilities into signals and another mapping signals into minimum abilities– and in the existence of a fixed point in which both actions and beliefs are consistent.

Corollary 2.3. *For any environment* $(W_A, W_B, \delta_A, \delta_B, F^N(.), \Phi(.), \lambda, \alpha)$ *the equilibrium signal thresholds of social groups* S *and* N *are related such that*

$$r_N^\alpha - r_S^\alpha = \alpha M^N \left[\frac{1-\gamma}{\gamma} \right]$$

The above corollary comes directly from point 1 of proposition 2.2 and summarizes the most important effect of the existence of stereotypes and the main implication of the self-stereotyping mechanism generated by these stereotypes on equilibrium behavior: In order to choose tournament A, S-agents and N-agents must observe different signals. In other words, an S-agent and an N-agent who observe the same signal and potentially have the same ability may differ in their choice of which tournament to compete in. Thus, through the self-stereotyping mechanism, the existence of stereotypes creates *behavioral differences* between agents who are exactly comparable except in which social group they belong to.

2.2 Representation

The above section shows that N-agents will enter tournament A under a different set of signals from S-agents. We now address how this behavioral difference affects the *representation* of each social group in each tournament. Notice that the results in the above subsection are independent of whether the stereotype is true or false as they merely describe the optimal behavior provided that the stereotype is *believed to be true*. However, from now on we restrict the analysis to the case in which the stereotype is false but believed to be true by the agents.[11]

If the stereotype is false, the distribution of abilities for both social groups is the same at $F^N(.)$. To avoid confusion, denote the CDF for the common distribution when the stereotype is false as $F(a)$ with associated PDF $f(a)$. All proofs of results in this section are relegated to Appendix A.4. We start by defining what we mean by representation.

Definition *We say that a social group is underrepresented (overrepresented) in tournament* A *if the proportion of agents from that social group in tournament* A *is lower (higher) than would be expected under perfect information.*

To manage the concept of overrepresentation and underrepresentation presented in the above definition we focus solely on the S group. Notice that under the assumptions that the stereotype is false and that agents have perfect information about their real abilities the proportion of agents in tournament A from social group S should be exactly λ.[12] Under the equilibrium conditions displayed in Proposition 2.2, only those agents from social group $g \in \{S, N\}$ who observe a signal exceeding r_g^α will enter tournament A. So when the stereotype is false we should observe a mass $\lambda \int_{-\infty}^{\infty}[1 - \Phi(r_S^\alpha - a)]f(a)da$ of S-agents and a mass $(1 -$

[11] The case in which the stereotype is true is hard to manage, as whether it causes over/underrepresentation depends not only on whether the stereotype is positive or negative but also on the joint values of the magnitude of the stereotype, the spread of prizes, and the actual variance of abilities. For this reason no clear conclusion can be drawn, so we stick to the case in which the stereotype is false. In any event this can be argued to be the most interesting case.

[12] Under perfect self-knowledge (when own ability levels are perfectly known to the agents) the decision rule for both social groups would be the same (to choose A if ability exceeds a certain threshold) independently of whether or not there is a stereotype as the self-stereotyping mechanism cancels out. Therefore the social composition in each tournament would replicate that of the whole population. The role of stereotypes in this setting would be just to decrease/increase the believed minimum ability required for obtaining W_A. This effect does not create any asymmetry between social groups.

$\lambda) \int_{-\infty}^{\infty} [1 - \Phi(r_N^\alpha - a)] f(a) da$ of N-agents entering tournament A. Therefore, in equilibrium the proportion of agents from social group S in tournament A would be

$$\lambda_A^\alpha = \frac{\lambda \int_{-\infty}^{\infty} [1 - \Phi(r_S^\alpha - a)] f(a) da}{\int_{-\infty}^{\infty} \lambda [1 - \Phi(r_S^\alpha - a)] f(a) + (1 - \lambda)[1 - \Phi(r_N^\alpha - a)] f(a) da} \tag{2}$$

Accordingly, social group S will be underrepresented when $\lambda_A^\alpha < \lambda$ and overrepresented when $\lambda_A^\alpha > \lambda$. Thus when the stereotype is false any deviation from $\lambda_A^\alpha = \lambda$ is undesirable as it evidences representation problems.

Proposition 2.4 (Segregation). *If the stereotype is false, agents from social group S are underrepresented (overrepresented) when the stereotype is negative (positive).*

This proposition implies that when the stereotype is false there is always a representation gap in the pool of participants in tournament A independently of the particular environment in which the choice takes place. However, the excess of entries in tournament A means that not all agents will obtain a prize, so this imbalance in participation may not translate into an imbalance in the set of winners. This is of great concern, especially taking into account corollary 2.3, which implies that if the stereotype is false then S-agents who participate in tournament A will have a higher average rate of success when the stereotype is negative because on average they will have higher ability levels than those from social group N, and vice versa when the stereotype is positive. However a look at how the mass δ_A of prizes is split between the two social groups reveals that these differences in winning rates do not cancel out the imbalance in participation.

Proposition 2.5 (Wage gap). *If the stereotype is false, agents from social group S are underrepresented (overrepresented) in the set of winners from tournament A when the stereotype is negative (positive).*

The reason for this is that as the stereotype is false, the distribution of abilities is the same for both groups and, therefore, the distribution of signals is also the same. However, since N-agents have a lower signal threshold to enter tournament A (if the stereotype is negative), more N-agents than S-agents will enter this tournament (as shown in proposition 2.4). In particular, all those high-ability N-agents who by chance observe a signal in the interval (r_S^α, r_N^α) will enter tournament A while those high-ability S-agents observing a signal within the same interval will not. If the stereotype is positive the opposite reasoning applies. Notice that this implies a wage gap at social group level. Among those who win any prize (get a job, for example), S-agents are more likely to get lower-paid position than equally skilled N-agents when a negative stereotype exists and vice versa if the stereotype is positive. It is important to remark that this result is found in the absence of any kind of discrimination and arise naturally from the existence of a stereotype in which agents believe.

3 Experimental Test

We ran a laboratory experiment to test the relevance of the theoretical model on explaining self-selection into tournaments under stereotypes. The experiment explicitly tests for the self-stereotyping mechanism described in the theoretical model as it represents the source for differences in self-selection.

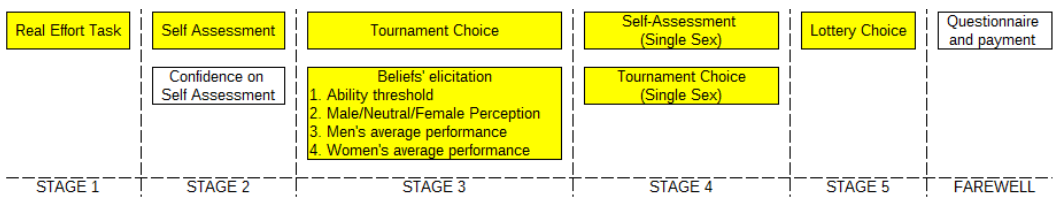
The experiment took place at the Bilbao Laboratory of Experimental Analysis (Bilbao Labean) at the University of the Basque Country. It involved a computer-based form using z-Tree experimental software (Fischbacher, 2007). Subjects were recruited through ORSEE (Greiner, 2015), which resulted in 217 participants –106 (48.85%) men and 111 (51.15%) women– split into six different sessions. Recruiting was carried out in such a way that the gender balance in each session could be assured but that subjects were unaware of this at the time of recruiting.[13]

At the beginning of each session, subjects were provided with written general instructions which were read aloud to guarantee that they were public knowledge. In particular, these general instructions informed them that the experiment consisted of 5 different stages, instructions for which would be displayed on their computer screens immediately before the start of each stage. A translation of all the instructions can be found in Appendix B. All subjects took all decisions simultaneously, so all agents entered all stages at the same time. Each session lasted for around one and a half hours, including payment. Average earnings were 15.90 euro (s.d. 3.51) including a show-up fee of 3 euro, and total earnings ranged from 7.2 euro to 24.4 euro.

3.1 Experimental Design

The experiment was designed to follow closely the theoretical framework presented in Section 2. In particular, in stages 1, 2 and 3 the setting from the theoretical model is transferred to the lab.[14] At the end of stage 3, subjects' beliefs about the gender nature of the task and beliefs about other participants' behavior are collected. Stage 4 repeats stage 2 and 3 but in a single sex setting. Stage 5 consists of a lottery choice intended to elicit risk preferences (Eckel and Grossman, 2002). The experiment ended with a non-incentivized questionnaire. For a graphical summary of the experiment see Figure 1, which is explained in detail below.

FIGURE 1–ROADMAP OF THE EXPERIMENTAL DESIGN



Notes: All incentivized tasks are shown in yellow.

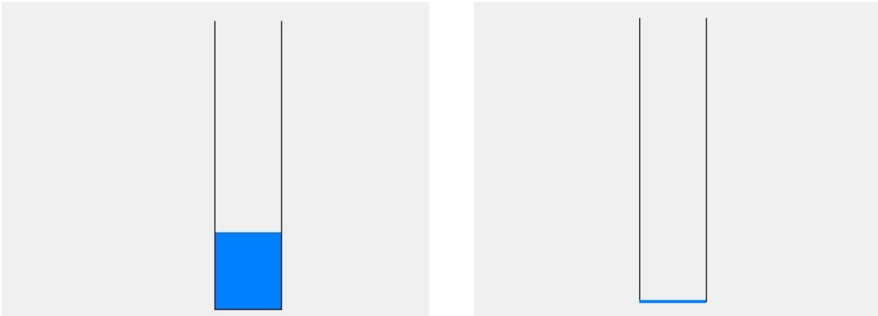
Stage 1: The real-effort task

The first stage of the experiment consisted of a real-effort task. In this part, subjects saw a glass with some water in it for 3 seconds (Figure 2, picture on the left). The picture then disappeared and an empty glass appeared with a blue slider at the bottom (Figure 2, picture

[13] The number of participants in each session and the percentage of women were 36 (47.2%), 37 (54%), 33 (57.6%), 34 (44.1%), 38 (50%) and 39 (53.85%) respectively. All sessions were statistically gender balanced as intended. Subjects could see each other at the beginning and during the experiment, so it is assumed that they were aware of this gender balance.

[14] To avoid framing effects, the experiment referred to choices as "options" rather than "tournaments".

FIGURE 2–EXAMPLE OF THE PICTURES SHOWN TO SUBJECTS DURING THE REAL-EFFORT TASK



on the right). Subjects were asked to drag the slider and locate it as close as possible to the level of water shown in the previous picture. To prevent cheating, the first picture was displayed in the top half of their screens while the picture of the empty glass was in the bottom half.[15] Subjects had to repeat this task fourteen times for different initial pictures. The only thing that changed from one picture to another was the water level shown. All subjects saw the same pictures in the same order. Before starting the task, subjects had a practice run to get used to how the task worked and to settle any potential doubts. Subjects received no feedback on how well they did on this trial and it was made clear that it had no effect on the rest of the experiment.

The score awarded to each subject in each repetition was inversely proportional to the distance between the actual level of water in the first picture and their guess. Subjects were told that for payment just one of the fourteen repetitions would be randomly selected.[16]

This particular task, which is novel in the literature, was selected for several reasons. First, participants were very unlikely to have any prior experience with this or any related task. This means that they had no expectations about their own performance or that of others prior to participating in the experiment. Second, the nature of the actual ability required to perform well is not clear. This helps to ensure heterogeneity in the beliefs of subjects as to what abilities are important to perform well and therefore as to the gender nature of the task. To further check this point, a pre-experiment survey was run on a similar subject pool which confirmed that the task was perceived on average as gender neutral but that there was a significant degree of heterogeneity in beliefs as to its gender nature.[17] Third, it is very complicated to figure out how well one has performed in the task without feedback.

[15]In addition, the layout of the screens in the lab (in boxes inside the desks, covered by a glass and set at a 45^o angle) prevented them from using their hands or a pen to measure the initial water level accurately.

[16]Specifically, the score in each run was computed as $100 - 5 \times Distance|\textit{Water Level}, \textit{Slider Location}|$ such that the maximum score attainable was 100. Notice that the minimum score varies from picture to picture. Negative scores were extremely unusual and account for only 8 of the 3038 times that the task was performed in total (0.26%). Only 1 of the 217 subjects received a negative signal. The rule for payment was the following: $Payment_{stage\_1} = max(5 \times score\_selected/100, 0)$.

[17]The survey was conducted at the Public University of Navarre. We gathered information about the beliefs held by 195 undergraduate students concerning the task. Once the task had been described in the same way as in the experiment, participants were asked which gender they thought would perform better at it. Immediately afterwards they were asked to explain their choice in an open form question. Most subjects reasoned their choice by linking the task to a particular ability. These abilities were extremely heterogeneous, which indicates that unanimity in this regard is far from being the case. The required abilities reported include intuition, precision, logic, visual memory, spatial vision, and others.

Fourth, the task is extremely simple to understand and no gender differences in performance were expected. This was indeed corroborated in the data (t-statistic=-0.72, two-sided p-value=0.47).

Stage 2: Self-assessment stage

In stage 2, participants were asked to estimate their average performance over the fourteen runs at the stage 1 task. Instructions emphasized that this average performance represented their *ability* at that particular task. We will refer to it as the average performance in the task throughout the text.[18] Participants were asked to select where in a ranking of 11 possible levels they thought their ability lay. These ranks ranged from "<70" to "97-100" in intervals of 3. Subjects were informed that if their estimation was correct they would be paid an additional 1.5 euro. At this stage we used two different sets of instructions.

In three of the six sessions, before they enter the estimation subjects were shown their scores in one randomly chosen run out of the fourteen times that they performed the real effort-task. Thus, similarly to the theoretical model, subjects received an informative but not perfectly correlated *signal* from which they had to estimate their average performance.[19] They were also shown the distribution of the average performances of all participants in the session. This information was presented as a chart and as a histogram (see Figure 3).[20]

In the remaining three sessions, before subjects were shown their signal a brief description of the task was provided in which the task was framed as being related to spatial vision and manipulation of visual information, and it was stated that such skills were essential for engineering. Note that the worst case scenario for this experiment is the case in which the general perception is gender neutrality or the absence of gender-based perceptions as the model would predict no behavioral differences. This additional information sought to increase the proportion of subjects who reported that the task was male-biased. This manipulation did indeed increase the perception of the task as being male by 12 percentage points (p-value<0.05) but it had no other noticeable effects, so we pooled the data in all sessions.[21]

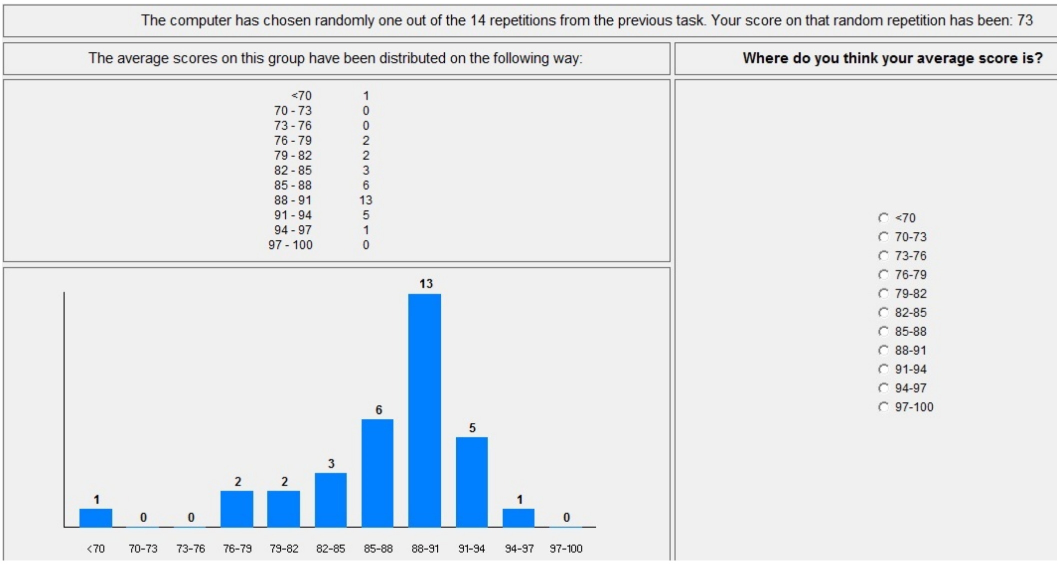
[18]We chose the wording ability for the experimental instructions to emphasize the existence of an inherent ability to perform well. Although we believe the average performance during these 14 runs is a sufficiently good proxy of ability, as in the theoretical model what we are really interested in is knowing subjects' self-assessment of its inherent ability and not so about the performance during the experiment. This wording, without incurring into experimental deception, makes subjects more likely to think in terms of inherent ability than on their particular performance during the experiment, making more meaningful the perceived gender nature of the task. Notice that, to prevent subjects from getting confused, in the instruction set we always clarify that ability is the average performance in the real effort task.

[19]The actual correlation between signals and average performances for the full sample was 0.43 (p-value<0.001).

[20]This feature of the experimental design makes the distribution of average performance common knowledge so that the self-stereotyping mechanism could be properly identified. Failing to provide this information could lead to subjects in the same session to differ in how the average performance is believed to be distributed. Differences in beliefs would generate differences in self-assessments and behaviors not related to the self-stereotyping mechanism.

[21]Results go in the same direction regardless of whether the analyses are conducted on just the manipulated or the non-manipulated group. See tables C7 and C11 in Appendix C. Moreover, all the analyses performed in this paper include session fixed effects and thus control for this manipulation. Excluding session fixed effects and including instead a dummy variable with a value of 1 if the session was manipulated and 0 otherwise and the interaction effect of this variable with the *Female* dummy show that the manipulation itself has no noteworthy effect on the variables studied.

FIGURE 3–SELF-ASSESSMENT TASK



The rest of the instructions were exactly the same as in the two previous sessions.

Stage 3: Tournament choice and elicitation of beliefs

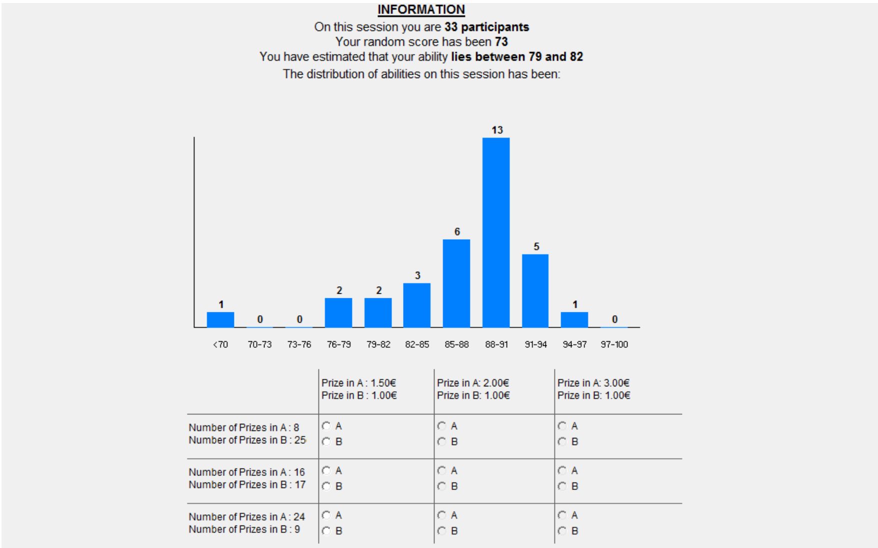
This stage involved three different parts. In the first part, subjects were given a decision matrix that showed 9 different situations. In each situation they then had to choose between option A and option B, each of which offered different prizes. Subjects were informed that if they chose an option in which the number of participants choosing that same option was lower than the number of prizes available under it, they would be sure to receive the option's prize. When the number of participants opting for that option exceeded the number of prizes available for it in that situation, only those with the highest average performance in the real-effort task would get prizes. When choosing which tournament to enter, subjects thus first had to interpret the signal observed in stage 2 to estimate their average performance in the real-effort task so as to determine their chances of winning. This resembles the theoretical framework set out in Section 2.

The 9 situations differed in the number of prizes offered by option A and option B and in the value of the prize in option A. Importantly for this part of stage 3, the sum of the number of slots in option A and option B was always the same as the number of participants in that session of the experiment. This means that the only way in which everyone could get a prize was by perfectly coordinating on choices. Option B always offered prizes worth one euro while option A's prizes could be 1.5 euro, 2 euro or 3 euro. Similarly, the number of prizes offered in option A (B) were 1/4 (3/4), 1/2 (1/2) or 3/4 (1/4) of the number of subjects participating in each session. The combination of these 3 spreads and 3 possible allocations of prizes gives the 9 situations observed by subjects (see Figure 4).[22] Subjects were told that at the end of the experiment one of the 9 situations would be taken randomly and the amount resulting from their choice and that of the other participants in that situation would

[22]After hearing the instructions for this stage, subjects had to answer a test with control questions to ensure that they had understood correctly how winners were chosen.

be paid. When choosing between option A and option B subjects could see at all times the number of participants in the session, their signals obtained from their performance in stage 1, their estimated ranking of their average performance from stage 2, and the distribution of average performance in the session.

FIGURE 4–TOURNAMENT SELF-SELECTION



For the sake of completeness, it is worth saying that the second part of stage 3 was identical to the first except that in all 9 situations option B now offered the same number of prizes as there were subjects in the session. Decisions in this setting, however, are not really about *where to compete* but about *whether to compete or not*. Given that we use gender as the basis for social groups this distinction seems to be important.[23]

Stage 3 concluded with 5 questions –which appeared one by one on different screens– related to the session. At the end of the experiment the computer selected one of these questions and subjects received 1.5 euro if they had answer it correctly. A translation of the exact wording can be found in Appendix B. Questions 1 and 2 were intended to measure potential heterogeneity in forecasting the behavior of others by asking subjects to predict the minimum average performance required to obtain a prize in tournament A in the first and in second part of stage 3 respectively. In terms of the theoretical model, we were asking subjects to forecast a^{α}. Subjects' answers to these questions showed that the vast majority of them correctly expected tournament A to be overcrowded in both parts of stage 3.[24]

[23]For example, Filippin and Crosetto (2016) and Crosetto and Filippin (2017) show that a gender gap in attitudes toward risk *only* arises when the elicitation method includes a safe option. In line with this, under this second setting we find a gender gap in entry into tournament A significantly larger than in the setting analyzed in this paper. However, for this setting perceptions do not appear to play a significant role in explaining behavior, although coefficients always go in the direction predicted by the theory. The consistency of coefficients with the theory suggests that the aforementioned gender difference in behavior may be hindering our analyses when we look at perception-related behavior. Results from this setting are analyzed in Hernandez-Arenaz (2018).

[24]In the instructions for this question, subjects were informed that if they believed tournament A was going to be uncrowded –so that there was no minimum ability–, they should choose ability "<70". Only 8 subjects (4.3%) made this choice. As Table C1 in Appendix C shows, tournament A was effectively overcrowded in all situations and all sessions. These two facts support the idea of treating the theoretical and experimental frameworks as strategically close.

Questions 3, 4 and 5 were intended to measure subjects' perception of the gender nature of the task performed in stage 1. Question 3 sought to directly elicit beliefs about stereotypes by asking subjects to guess who had performed better on average in the real effort task: men/no differences/women. Questions 4 and 5 aimed to gauge subjects' perceptions as to the average performance of men and women respectively by using an eleven-option item ranging from "<70" to "97-100" in intervals of 3.

Stage 4: Single-sex environment

Stage 4 resembles stages 2 and 3 except that *subjects only interacted with those of their same sex*.[25] We refer to this stage as the *single-sex* setting. In this setting, subjects again had to estimate their ability levels based on the same signal observed in stage 2. The only difference with respect to stage 2 was that this time we showed them the distribution of abilities within their own gender instead of that for the whole session. They were told that if their estimation was correct, they would be paid an additional 1 euro. Next they again had to choose between tournament A and B. The rules were the same as in stage 3, but this time subjects only interacted with their own gender. This meant that at all levels the session was split into two gender-homogeneous groups and subjects played stages 2 and 3 within their corresponding group.

Stage 5: Risk attitudes and final questionnaire

Stage 5 consisted of a lottery choice to elicit risk attitudes. The elicitation method used resembles the one in Eckel and Grossman (2002), enabling 8 different degrees of risk attitude to be differentiated.

Finally, subjects also completed a non-incentivized questionnaire which collected data on some sociodemographic variables, social risk attitude (Weber et al., 2002), and the big-five personality traits (Gosling et al., 2003). Additionally, in two sessions subjects were asked to choose from a closed list at the beginning of the questionnaire what ability they thought was most important to perform well in the real effort task: *Attention to details*, *Observational capacity*, *Intuition*, *Logic*, *Visual memory*, *Precision* or *Spatial vision*.[26] Immediately afterwards, subjects were asked to rate each ability as male-biased, gender-neutral, or female-biased.

3.2 Perceptions and Descriptive Statistics

We gather information about subjects' perception as to the gender nature of the task using the answers provided at the end of stage 3 (questions 3, 4, and 5). In particular, question 3 provides us with a discrete variable on whether subjects perceived the task in stage 1 as male-biased, female-biased or gender-neutral. We refer to this variable as *Perception* of the

[25] There is no data from this stage for session 5 as there was a technical problem and data did not store correctly. Thus, for stage 5 we have data from only 179 subjects.

[26] Before running the experiment, we ran a pre-experiment survey in which a subject pool of undergraduate students was given with the instructions used in the experiment for the real-effort task. The instructions were read aloud and subjects then had to rate the task as male or female and to answer an open form question eliciting the reason for their choice. Their answers suggest that, as the task is novel, they relate the ability to perform well with general skills such as precision, intuition, spatial vision, etc. We used these answers to construct the closed list used in the experimental sessions.

task. Questions 4 and 5 ask about the perceived average performance for men and women, respectively. Based on subjects' answers to these two questions we can compute the difference between the reported averages in abilities for men and for women. We call this variable *Maleness*, and it is our main variable of interest in the main analysis. A *Maleness* score of zero implies that the subject believes there are no gender differences in average performance in the real-effort task. A positive score implies that the subject believes that women perform in average worse than men. The contrary is implied if the score is negative. Moreover, higher absolute values indicate a greater bias in favor of men (if the score is positive) or in favor of women (if it is negative). We call the absolute value of *Maleness* the *strength* or *intensity* of the bias. We favor this measure of *Maleness* over the alternative of *Perception* as it enables us not only to look at the direction of the bias but also at its strength, thus enriching the analysis. An alternative analysis of the main results making use of the *Perception* variable is presented in Table C2 in Appendix C. It shows that all results remain qualitatively the same, although slightly weakened.

Notice that *Maleness* and *Perception* should be consistent with each other so that if a subject claims that men have a higher ability level than women (*Maleness*>0) his/her associated *Perception* of the task should be male-biased. Otherwise we cannot identify subjects' beliefs correctly, because the direction of gender bias will change depending on which measure is considered. Thus, for the sake of consistency, in our main analysis we consider only those subjects who were consistent in these measures. This leaves us with 188 out of the total sample of 217 subjects (87% of the original sample).[27] Table C3 in Appendix C shows how the original sample differs from the sample used in the main analysis.

We take this *Maleness* variable as our variable for assessing the impact of the gender-based perception of the task, so it deserves a closer look. First, one may ask where the heterogeneity in *Maleness* comes from. Answering this question is always tricky, but we collected some data that could shed light on the issue. Remember that at the beginning of the questionnaire in the last two sessions (77 subjects in total, 68 consistent), subjects had to indicate which of a closed list of seven skills they thought was the most important to perform well and then, rate all seven on a discrete scale where 1 meant female-biased, 2 gender-neutral, and 3 male-biased. From this last question, similarly to Coffman et al. (2019), we computed the perceived maleness of each skill on the list by averaging subjects' answers. With this, we constructed a *Skill Maleness* variable such that for each subject who stated that one particular skill was the most important to perform well, we impute the corresponding average score. That is, this *Skill Maleness* tells us for each subject how male is perceived (on average) to be the skill he/she believes is the most important to perform well in the real-effort task. Regressing the *Maleness* variable on *Skill Maleness*, we find a positive and significant (p-value<0.01) coefficient, which suggests that the more male-biased the most important skill to perform well in the task is believed to be, the more male-biased the task is believed to be. See Tables C4 and C5 in Appendix C. This exercise thus suggests that the heterogeneity of the *Maleness* variable comes from heterogeneity in beliefs as to which skill is the most necessary to perform well and that it is beliefs concerning that skill that drive

[27] Including all 217 subjects in the analysis does not change the results but, on the contrary, strengthens them (see Tables C7 and C11 in Appendix C). We consider as consistent those subjects who hold a non-neutral *Perception* but claim *Maleness*=0 (24 subjects). Because the believed average performances for men and for women are reported in intervals of three, these two statements are not mutually exclusive (for example, if the gender difference is believed to be very narrow).

the perception of the task. Importantly, notice that at this point we are indeed talking about stereotypes as this *Skill Maleness* variable reflects the average perception as to the skill and not just individual beliefs.

Although the *Maleness* variable is constructed based on incentivized elicitation, it is crucial to rule out potential endogeneity and self-selection issues.[28] Given that beliefs about the gender nature of the task are not randomly allocated, our main concern is to address whether there is any systematic bias in the way in which these beliefs are set. In particular, we look at whether subjects used perceptions to self-justify their performances. It is plausible that men (women) observing high (low) signals might justify their own performances by reporting the task as male-biased and vice versa if the signals observed are low (high). That is, the reported *Maleness* of the task might be correlated with the signal observed. If that were the case then our analysis would suffer from a strong identification problem as we could not properly identify the effect of perceptions as to the gender nature of the task on the outcome variables of interest. The regression analysis shown in Table 1 checks for and rules out this possibility.

TABLE 1–OLS REGRESSION FOR MALENESS

Sample:	*Men* (1)	*Women* (2)	*All* (3)
Female			-0.294
			(0.985)
Signal	1.093	0.753	1.086
	(0.833)	(0.808)	(0.801)
Signal∗Female			-0.357
			(1.115)
Constant	-0.153	-0.338	-0.0741
	(0.789)	(0.766)	(0.733)
Session FE	YES	YES	YES
Observations	95	93	188
R-squared	0.061	0.148	0.136

Notes: Female takes a value of 1 when the subject is a woman and 0 otherwise. *Signal* is a continuous variable accounting for the signal observed by the subject during the experiment. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Columns (1) and (2) of Table 1 reports analyses run separately for men and women. Regressions show no significant effects in either gender. The analysis in column (3) checks for whether the size of the effect of the signal in determining *Maleness* differs for men and women. The analysis further confirms that it does not. These analyses enable us to reject the idea that there is a self-selection problem driven by the signals observed which affects the reported *Maleness* of the real-effort task.

Related with the previous point, we further check what the optimal decision of each subject should have been during the experiment if all subjects had also chosen their optimal behavior. In other words, we can see whether a subject would have won the prize in A if all other subjects had participated in tournament A. Table C6 in Appendix C shows that

[28] When comparing the distribution of beliefs across genders as regards *Maleness* (see Figure C1 in Appendix C), we find an own-gender bias such that both men and women are more likely, on average, to state that their gender performs better. This translates into a small gender difference in beliefs (p-value<0.1) but should not be problematic for analyzing the data.

this optimal behavior, i.e. the behavior that would maximize the subject's payoff, does not depend on the perceived *Maleness* for either gender.

TABLE 2–CORRELATIONS FOR SUBJECTS' CHARACTERISTICS AND OUTCOME VARIABLES WITH MALENESS

PANEL A: SUBJECTS' CHARACTERISTICS

Obs.	*Men* 95 (1)	*Women* 93 (2)	Obs.	*Men* 95 (1)	*Women* 93 (2)
Signal	0.1094	0.0711	Risk Pref.	0.1733*	-0.172*
	(0.291)	(0.498)		(0.093)	(0.099)
Taste Comp.	0.0870	0.301***	Age	0.1175	0.2311**
	(0.402)	(0.003)		(0.257)	(0.026)
Difficulty	-0.0486	-0.0056	Min.Ab.Win	0.2265**	0.0735
	(0.640)	(0.957)		(0.027)	(0.484)
Extraversion	-0.0194	-0.0602	Agreeableness	-0.1002	0.1697
	(0.852)	(0.566)		(0.334)	(0.104)
Conscientiousness	0.0071	-0.0327	Neuroticism	0.0249	0.0409
	(0.946)	(0.756)		(0.811)	(0.697)
Openness	0.0245	0.0012	Social Risk	0.0227	-0.1602
	(0.813)	(0.991)		(0.827)	(0.125)

PANEL B: OUTCOME VARIABLES

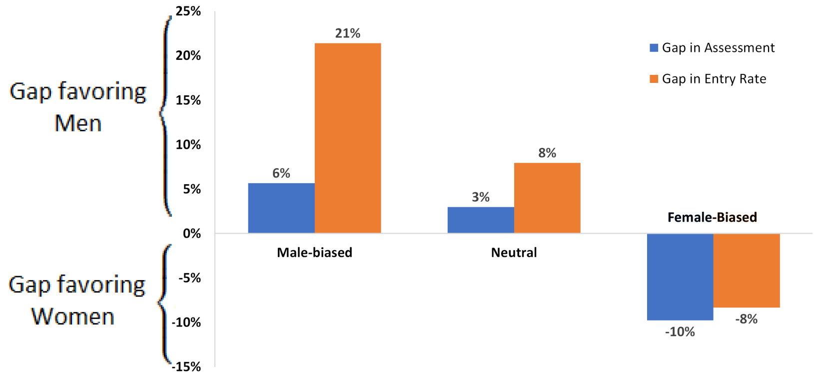
Obs.	*Men* 95 (1)	*Women* 93 (2)	Obs.	*Men* 95 (1)	*Women* 93 (2)
Assessment	0.2138**	-0.0259	Prob(W_A)	0.2693***	0.0008
	(0.037)	(0.805)		(0.008)	(0.994)
Prob(A)	0.2994***	-0.0117	Mean Earnings	0.2019**	-0.0110
	(0.003)	(0.912)		(0.049)	(0.916)

Notes: Signal is the random draw observed by the subject after the real effort task. *Taste Comp* is a 5-level discrete variable accounting for self-reported taste for competition with 1 meaning "Not at all" and 5 "A lot". *Difficulty* is a 5-level item measuring subjects' perceptions on how difficult is to perform well at the real-effort task with 1 meaning "Extremely easy" and 5 "Extremely difficult". *Min.Ab.Win* is an 11-level item measuring subjects' beliefs about the minimum average performance required to win the prize in tournament A in intervals of three with 1 meaning "<70" and 11 ">90". *Risk Pref* is an 8-level variable measuring subjects' risk attitude. *Age* is subjects' age in years. *Extraversion*, *Agreeableness* , *Conscientiousness*, *Neuroticism*, and *Openness* are 7-level items corresponding to the big-five personality traits (Gosling et al., 2003). *Social Risk* is subjects' score in the social risk attitude scale (Weber et al., 2002). *Assessment* is subjects' self-assessment. *Prob(A)* is subjects' average entry rate into tournament A across the 9 situations. *Prob*(W_A) is the average probability with which subjects get the prize at tournament A across the 9 situations. *Mean Earnings* is the mean of the earnings across the 9 situations. p-values in parentheses.

Panel A of Table 2 looks at potential self-selection patterns based on other control variables obtained during the experiment by showing the correlations within each gender with the *Maleness* variable. Only a few control variables are correlated. To check whether these correlations are the drivers of our results, Tables C7 and C11 in Appendix C include a specification in which all these variables are controlled for. This analysis returns consistent results.

Panel B of Table 2 presents an overview of the results by looking at the raw data for the main variables of interest: Subjects' self-assessment (*Assessment*), whether subjects chose tournament A (*Prob(A)*), whether subjects (independently of choosing to participate in tournament A or in tournament B) won the prize in A (*Prob*(W_A)), and the mean earnings obtained across all 9 choices of tournaments (*Mean earnings*). For men, it shows a positive, significant correlation between the perceived maleness of the task and subjects' self-assessment, entry rate into tournament A, probability of obtaining the high-prize W_A, and mean earnings derived from the decisions as to which tournament to participate in. This

FIGURE 5–GENDER GAP IN SELF-ASSESSMENT AND ENTRY RATE FOR TOURNAMENT A



Notes: Average gap in self-assessment (blue bars) and in entry rate for tournament A (red bars) relative to women's average values for each *Perception* category (x-axis). A positive (negative) score shows a gender gap in favor of men (women).

set of correlations is consistent with the self-stereotyping hypothesis and its implications on behavior and outcomes as shown in the theoretical model. However, although the direction is as predicted by the theoretical model, for women the perceived maleness of the task is not significantly correlated with self-assessment, or behavior or outcomes. Panel B of Table 2 also suggests that gender gaps in self-assessment and in behavior depend, through men's reactions, on particular beliefs as to the gender nature of the real-effort task. Figure 5 shows these gaps for subjects who hold beliefs as to the task being male-biased, gender-neutral, and female-biased. The picture seems compatible with the predictions of the theoretical model in this regard: a gender gap in self-assessment and in entry rate in tournament A favoring men appears among subjects who believe the task is male-bias, a gender gap favoring women appears among subjects who believe the task is female-biased, and gender gaps in absolute terms are the lowest among subjects who believe that the task is gender-neutral.

Below, we conduct formal regression analyses to establish the relationship between beliefs as to the gender nature of the task (*Maleness*) and the three main outcome variables: self-assessment (*Assessment*), entry rate into the high-paying tournament (*Prob(A)*), probability of winning the high-paying tournament prize (*Prob*(W_A)), and earnings across tournaments (*Earnings*).[29] All analyses include as control variables the signal observed by subjects (*Signal*) and session fixed effects. In addition, when looking at *Prob(A)* and *Prob*(W_A) we also include tournament fixed effects. The independent variables of interest are *Female*, *Maleness*, and the *Maleness*∗*Female* interaction effect. This interaction effect captures how the gender gap in the dependent variable changes when the perceived maleness of the task increases.[30] This specification thus enables us to separate pure gender effects (coefficient of *Female*) from gender differences arising from perceptions (coefficient of *Maleness*∗*Female*). It also enables us to compare men and women with the same perception *and* perception strength, and to check whether the impact of perceptions comes from men's reactions, women's reactions or

[29]Table C2 in Appendix C shows the main regressions using *Perception* as the independent variable. Analyses using this variable both as continuous and as dummy return similar conclusions to the ones using *Maleness*, although results become slightly weaker.

[30]When looking at *Prob(A)* and *Prob*(W_A), we use a probit model and report marginal effects for the sake of interpretation. Marginal effects of interacted terms cannot be computed straightforwardly, so all regressions present the coefficient for the *Maleness*∗*Female* interaction effect corrected as suggested in Ai and Norton (2003).

a combination of the two. This last analysis is relegated to Appendix C but discussed in the main text.

3.3 Results

Table 3 shows the results for the main variables of analysis: *Assessment*, *Prob(A)*, *Prob(W_A)*, and *Mean Earnings*. The analysis presented in this table tests the main implications of the theoretical model and, in particular, the self-stereotyping mechanism.

TABLE 3– REGRESSION ANALYSES

Dep. Variable:	*Assessment* (1)	*Prob(A)* (2)	*Prob(W_A)* (3)	*Mean Earnings* (4)	*Prob(A)* (5)	*Prob(W_A)* (6)	*Mean Earnings* (7)
Signal	6.416***	0.401***	1.364***	2.052***	-0.0648	0.947***	1.667***
	(1.060)	(0.117)	(0.160)	(0.327)	(0.116)	(0.172)	(0.395)
Female	-0.0263	-0.0632**	0.0253	0.106	-0.0587**	0.0284	0.108
	(0.175)	(0.0292)	(0.0384)	(0.0914)	(0.0269)	(0.0375)	(0.0913)
Maleness	0.245***	0.0580***	0.0530**	0.0906	0.0389**	0.0414*	0.0759
	(0.0874)	(0.0197)	(0.0239)	(0.0549)	(0.0183)	(0.0241)	(0.0553)
Maleness*Female	-0.325**	-0.0595**	-0.0594	-0.125	-0.0355	-0.0457	-0.105
	(0.142)	(0.0282)	(0.0355)	(0.0784)	(0.0252)	(0.0354)	(0.0806)
Assessment					0.0690***	0.0545***	0.0600*
					(0.0118)	(0.0153)	(0.0316)
Constant	1.813*			-0.554*			-0.684**
	(0.993)			(0.311)			(0.327)
Tournament Controls	–	YES	YES	–	YES	YES	YES
Session FE	YES	YES	YES	YES	YES	YES	YES
Number of Clusters	–	188	188	–	188	188	–
Observations	188	1,692	1,692	188	1,692	1,692	188
R-squared	0.445			0.226			0.237

Notes: Column (1) shows the OLS estimates for the subjects' self-assessments. Columns (4) and (7) shows the OLS estimates for the subjects' mean earnings across all 9 tournaments. Columns (2)–(3) and (5)–(6) show the marginal effects of the probit model for the probability of choosing tournament A and the probability of getting W_A. Interaction corrected using Norton et al. (2004) for columns (2), (3), (5), and (6). Robust standard errors for columns (1), (4), and (7) and clustered standard errors at subject level for columns (2), (3), (5), and (6) are displayed in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Column (1) presents the results from an OLS regression that tests for the self-stereotyping mechanism. First, notice that this analysis returns a non-significant coefficient for the *Female* variable, which indicates that under neutral perceptions (*Maleness*=0) there are no gender differences in how men and women self-assess. However, the coefficient of *Maleness*$*$*Female* shows that how male the task is perceived to be determines the size and direction of the gender gap in self-assessment (p-value=0.023). These two results taken together imply that when the task is perceived as male-biased (*Maleness*>0) self-assessment by men is significantly higher than that by comparable women, but when the task is perceived as female-biased (*Maleness*<0) the opposite is true. This result is robust to the inclusion of other covariates and the use of different samples (see Table C7 in Appendix C).

Table 3 shows that men react significantly to their beliefs as to the gender-nature of the real effort task (p-value<0.01) but also suggests that women are not affected by those beliefs. Effectively, the corresponding hypothesis test (*Maleness*+*Maleness*$*$*Female*=0) reveals that women hardly react to their beliefs (p-value=0.44). To further check this result, separate regressions for men (column (2)) and women (column (3)) are displayed in Table C8 in Appendix C. This analysis confirms the previous insight: The average effect of *Maleness* is positive and significant on men (β=0.248, p-value<0.01), but for women it is negative but not significant (β=-0.089, p-value=0.388). This result is in line with the finding in Klinowski (2019) that the perception held by men affects their self-assessment but does not affect women's. Thus, the change in the gender gap shown in Table 3 is due to the sensitivity of men to the believed gender nature of the real-effort task.

Result 1: *The more male-biased the task is perceived to be, the higher the self-assessment of men is in comparison with that of women. Moreover, when the task is perceived as gender neutral, men and women's self-assessments do not differ.*

Result 1': *The self-stereotyping mechanism is only validated for men: Men react positively to the believed maleness of the real-effort task by raising their self-assessment while women do not react significantly to these beliefs.*

According to the theoretical model formalized in section 2, the strategy on whether to choose tournament A or tournament B takes the form of a signal threshold such that if the signal is above the threshold the agent chooses to participate in tournament A while otherwise the agent chooses tournament B. Column (2) in Table 3 shows that subjects in the experiment are to some extent following such a strategy as the *Signal* variable has a positive, significant impact on the probability of choosing tournament A. However, the negative, significant coefficient of the *Female* variable says that when men and women perceived the task as gender-neutral (*Maleness*=0), men have to observe lower values of the signal in order to participate in tournament A.[31] This result is in line with previous research in different settings (Niederle and Vesterlund, 2011), but it seems to be driven by one exceptional session in which there is a substantial gender gap in entry rate. In the remaining 5 sessions the coefficient is far from significant and is not even consistent in its sign. More importantly, as implied by corollary 2.3, column (2) shows that this difference in signal thresholds is further shaped by the perceived gender-nature of the real-effort task (α in the theoretical model). If the perceived maleness of the task increases the gap in signal thresholds increases, while if the perceived maleness of the task decreases the gap in thresholds also decreases and even can change signs.

As in the case of self-assessment, a look at the effects of the perceived maleness of the task on the probability of choosing to participate in tournament A, shows that this variable affects men's behavior positively (β=0.055, p-value<0.01) but does not affect women's (β=-0.011, p-value=0.53). This can be seen in columns (2) and (3) of Table C9 in Appendix C.

Result 2: *The more male-biased the task is perceived to be, the higher the entry rate of men in tournament A is in comparison with that of women. Moreover, when the task is perceived as gender neutral, men self-select into tournament A at a higher rate than women.*

Result 2': *Beliefs about the gender nature of the real-effort task affect only men's propensity to enter in tournament A: Men react positively to the believed maleness of the real-effort task by increasing their entry rate in tournament A while women do not react significantly to those beliefs.*

[31] Notice that the *Signal* variable positively affects the probability of entering tournament A. Given the negative estimate for *Female*, a man and a woman who are identical in everything except their gender will have the same probability of competing in tournament A if and only if the woman observes a higher signal than the man.

As a final check for the effect of perceptions, column (3) of Table 3 shows how perceptions as to the real-effort task affect the probability of ultimately obtaining a prize in tournament A. Although the sign is as predicted by the theory, the coefficient for *Maleness∗Female* does not reach conventional levels of significance after we correct for the interaction term as proposed by Ai and Norton (2003) (p-value=0.103). However, the effect is still substantial, suggesting that there is a lack of power and that the perceived gender nature of the real-effort task plays some role in determining whether there is a gender gap in the probability of finally getting the prize in tournament A and if so how big it is. Similar results hold in regard to the probability of getting the prize in tournament B but in the opposite direction (see columns (7) and (8) in Table C10 in Appendix C).[32] Nonetheless, this lack of significance was expected since it follows from the theoretical model that a gender gap in the entry rate translates in a less meaningful gender gap in winning rate (see subsection 2.2).[33]

A separate look at each gender in Table C10 in Appendix C and the corresponding hypothesis testing on table 3 reveal that beliefs as to the real-effort task affect the likelihood of getting the prize in tournament A for men (β=0.173, p-value=0.041) but not for women. This means that the model correctly predicts the effects of beliefs about the gender nature of the task on outcomes for men but not for women. In particular, if we compare two men who differ ex-ante –i.e. before deciding in which tournament to participate– only in their beliefs as to the gender-bias of the task, the one who believes that the task is more biased towards men is more likely to get the high-prize.

Result 3: *Our results provide suggestive evidence in favor of the perceived gender nature of the real-effort task modifying the gender gap in the probability of winning the high-prize.*

Result 3': *Beliefs as to the gender nature of the real effort task affect only men's likelihood of getting the prize in tournament A: The more male-biased a man believes the task to be, the more likely he is to get the prize in tournament A.*

Result 3' establishes that the more male-biased the task is perceived to be, the more likely men are to get the prize in tournament A, but at the same time column (7) in Table C10 in Appendix C shows that this also makes them less likely to get the prize in tournament B. This raises the question of whether the net result is positive, negative or null. In column (4) of Table 3, we check this by estimating the effect of the gender bias on subjects' mean earnings across all nine tournaments. None of the estimates of interest reach conventional levels of significance but they are close enough to consider this analysis as suggestive evidence that gender perceptions as to the task have some effect on final earnings. In particular, the more male-biased the task is perceived to be, the higher the earnings obtained by men are in comparison with those of comparable women (p-value=0.113). When we look at both genders separately (see Table C12 in Appendix C), we find that the more male-biased the task is

[32]Notice that, in contrast to *Prob(A)*, where not participating in tournament A implies participating in tournament B (so that analyzing *Prob(B)* returns the same estimates but with the opposite sign), not getting the high-prize does not imply getting the low-prize as the subject may end up with no prize at all. Table C10 in Appendix C shows the results for the probability of getting the prize in tournament B (*Prob(W_B)*) and the probability of ending without a prize (*Prob(∅)*).

[33]Interestingly, even if men enter tournament A more often than women when the task is perceived as gender-neutral, this gap does not translate to the probability of winning the prize in tournament A but. More strikingly, the estimate for *Female* becomes positive. However, this last can be explained by the fact that among those who claim the task to be gender-neutral, although not significantly (p-value=0.126), women have a slightly higher average performance in the real-effort task. In line with this argument, when we control for average performance in the real-effort task the estimate for *Female* becomes negative (though not significant).

perceived to be, the more men earn (β=0.0946, p-value=0.103), while women's earnings are hardly affected (β=-0.0138, p-value=0.818)

Result 4: *Our results do not provide sufficiently strong evidence to claim that the perceived gender nature of the real-effort task can modify the gender gap in earnings.*

Result 4': *Our results provide suggestive evidence in favor of men earning more the more male-biased the task is perceived to be.*

In columns (5), (6) and (7), we perform the same analyses as in columns (2), (3), and (4) but include subjects' self-assessments as an extra control to check whether it is possible to explain the results on the effect of perceptions on behavior and outcomes through the effects of perceptions in self-assessment. In other words, we check whether the results 2-4 are due to the self-stereotyping mechanism.

As predicted by the theoretical model, controlling for individual self-assessment decreases the effect of the *Maleness* variable on determining gender gaps in all columns compared to the specification in which it is not controlled for.[34] In addition, Table 3 as well as Tables C11 and C9 in Appendix C show that including self-assessment as an extra control brings the role of perceptions as to the gender nature of the task closer to 0 for both genders. However, merely adding this variable does not suffice to fully explain the effect of maleness for men, which remain significant in some instances. This could be due in part to the fact that the self-assessment variable is discrete and does not completely capture all the variation that would be required to offset the effect of maleness.

Result 5: *Controlling for self-assessment reduces the effect of perceptions but does not fully offset them.*

Finally, notice that the *Assessment* variable is a good predictor of subjects' behavior as it always returns a significant estimate in the expected direction. Importantly, incorporating this variable as a predictor of the probability of entering tournament A, causes the *Signal* variable to become non-significant, which indicates that the information that the signal conveys is transferred in full to subjects' self-assessments. This confirms that subjects rely heavily on their reported self-assessments to make payoff-relevant decisions. Together with the fact that column (1) in Table 3 shows that perception as to the gender nature of the task affects self-assessment, this leads naturally to the conclusion that self-assessment actually conveys information about the gender nature of the real effort task and that by controlling for self-assessment we are thus also controlling for the gender perception of the real-effort task. This link between *Assessment* and *Maleness* may be the result of a spurious correlation or even reverse causality, but the model developed in Section 2 predicts it through the self-stereotyping mechanism, i.e., subjects' beliefs as to the gender nature of the real-effort task affect self-assessment and not the other way around.

3.3.1 Robustness 1: Marginal Subjects

The above results support the theoretical predictions of Section 2 for men but not for women. However, the model makes a stronger prediction: Although self-stereotyping affects self-assessment for the full sample, its behavioral consequences should show up more strikingly

[34] The actual changes in the significance levels from columns (2)–(4) to columns (5)–(7) are from 0.043 to 0.183 for *Prob(A)*, from 0.103 to 0.208 for *Prob*(W_A), and from 0.101 to 0.194 for *Mean Earnings*.

in those subjects who are borderlines as regards being winners in tournament A. This excludes those with very high (low) signals, who are almost sure that they will (will not) obtain a prize in tournament A, so their decision on whether to participate or not should not be affected by their particular perception as to the gender bias of the task. To address this issue, we construct a dummy variable called *Marginal* which takes a value of 1 when the subject in any particular situation can be considered to be reasonably uncertain of his/her chances of winning in tournament A and 0 otherwise. In particular, we set 50% of the sample as marginal in order to guarantee that all agents are considered as *Marginal* for some situations and *Nonmarginal* for others.[35] This analysis is thus close to a within-subjects comparison. Notice that whether a subject is marginal or not is defined at the tournament-level, so this analysis can only be performed for *Prob(A)* and *Prob(W_A)* as they are the only outcomes that are defined at the tournament-level.

TABLE 4–PROBIT MODEL FOR *Prob(A)* AND *Prob(W_A)*: MARGINAL *vs* NONMARGINAL

Sample:	*Marginal*				*Nonmarginal*			
Dep. Variable:	*Prob(A)* (1)	*Prob(W_A)* (2)	*Prob(A)* (3)	*Prob(W_A)* (4)	*Prob(A)* (5)	*Prob(W_A)* (6)	*Prob(A)* (7)	*Prob(W_A)* (8)
Signal	0.298*	1.688***	-0.231	1.232***	0.225*	1.226***	-0.107	0.930***
	(0.158)	(0.288)	(0.187)	(0.282)	(0.117)	(0.193)	(0.121)	(0.216)
Female	-0.0672*	0.0932*	-0.0639*	0.0930*	-0.0531	-0.0284	-0.0446	-0.0230
	(0.0350)	(0.0514)	(0.0339)	(0.0501)	(0.0373)	(0.0369)	(0.0352)	(0.0359)
Maleness	0.0890***	0.0735**	0.0714***	0.0578	0.0270	0.0335	0.00833	0.0255
	(0.0260)	(0.0360)	(0.0258)	(0.0367)	(0.0213)	(0.0211)	(0.0200)	(0.0215)
Maleness∗Female	-0.0833**	-0.1108**	-0.0153	-0.0927*	-0.0282	-0.0649	-0.0047	-0.0070
	(0.0387)	(0.0477)	(0.0366)	(0.0487)	(0.0308)	(0.0352)	(0.0284)	(0.0350)
Assessment			0.0792***	0.0740***			0.0581***	0.0373**
			(0.0150)	(0.0223)			(0.0142)	(0.0177)
Tournament Controls	YES	YES	YES	YES	YES	YES	YES	YES
Session FE	YES	YES	YES	YES	YES	YES	YES	YES
Number of Clusters	188	188	188	188	188	188	188	188
Observations	846	846	846	846	846	846	846	846

Notes: Marginal effects are reported. Interaction corrected using Norton et al. (2004). Clustered standard errors at subject level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 4 shows that performing this classification and replicating the analyses in columns (2), (3), (5), and (6) of Table 3 separately for subjects considered as marginal and for those who are not results in the effect of maleness on behavior (*Prob(A)*) and outcomes (*Prob(W_A)*) showing up as significant only for marginal individuals. For both variables of interest, checks show that the coefficient in this subsample of marginal subjects is greater and more significant than for the whole sample. Importantly, a significant effect of the perceived maleness of the task on the gender gap in the probability of getting the prize in tournament A arises for this subsample. The fact that it is precisely in this subsample, where the effect was predicted to be stronger, that we find a significant result supports the assertion made for the full sample that this effect of beliefs on the gender gap shows as not significant because its effect is weakened (see *Result 3* above). Finally, as with the full sample, the effects of percep-

[35] In particular, the subsample of marginal subjects was identified as follows: We sorted all subjects within each session by the signal that they observed. Then, for a situation in which tournament A offers a fraction of prizes δ_A, we considered that the behavior of those who observed a signal above (below) the percentile $[1-\delta_A]+25$ ($[1-\delta_A]-25$) of the signal distribution was unlikely to be affected by the perceptions held. The results are similar when different interval amplitudes are used.

tions affect men's behavior and outcomes but not women's (see Table C13 in Appendix C) and controlling for the reported self-assessment diminishes but does not totally remove the effects of perceptions.[36]

3.3.2 Robustness 2: Tournament Self-Selection in a Single-Sex Environment

In stage four of the experiment, subjects go through the exact same choices (self-assessment and choice of tournament) but in a single-sex environment. The difference with respect to the previous situation is that subjects now interact with and see the information for their own gender only. Importantly, except for the fact that subjects now observe the distribution of abilities of those of their own gender, the information that they hold in terms of their own and others' abilities is the same. More precisely, they observe the same signal provided in stage 2 of the experiment and no information is provided about their payoffs at stage 3.[37]

According to the theory, in this environment perceptions should not play any role in affecting subjects' decisions: Subjects observe the ability distribution *for their own gender only*, so their beliefs about whether the task is more or less male-biased should be completely irrelevant given that in all cases they see themselves in an *ex-ante homogeneous* group in which such perceptions cannot trigger the self-stereotyping mechanism.

This treatment thus enables us to further reject other potential selection biases regarding perceptions or, by contrast, to admit that there are important unobserved characteristics which play a role in agents' decisions. For example, men who claim that the task is male-biased and men who claim that it is female-biased could differ in key aspects such as general optimism or self-esteem. If so, it should still be possible to observe an effect of perceptions on the new estimates of ability and on signal-based behavior. On the other hand, the loss of the effect of perceptions in this *single-sex* environment would be further evidence in support of the initial premise that the channel driving the results in Table 3 is actually the self-stereotyping mechanism rather than unobserved differences in subjects' characteristics across perceptions. Notice that the latter case also rules out a potential problem of *reverse causality* of the "I believe I am good (bad) at the task so my gender should be better (worse)" kind as if this were the case we should still expect a relationship between perceptions and the new reported self-assessment.

Table 5 shows the same analyses performed in Table 3 for the data from this *single-sex* environment. The first noteworthy finding is that having information about the distribution of average performances of their own gender brings about changes in the coefficient *Maleness* and in the *Maleness*∗*Female* interaction term important enough to cancel out all the effects of subjects' perceptions on self-assessment (column (1)).

Given the above, the natural expectation would be to find that the beliefs as to the gender bias of the task to have no effect in behavior or outcomes. Regression analysis confirms that this is the case. Column (2) confirms that with the same signal observed men and women

[36] Notice that, as with the full sample, the perception of subject as to the gender-bias of the task is not related to the payoff-maximizing decision that the subject should make. See Table C6 in Appendix C.

[37] Unfortunately, a technical issue resulted in data for one of the sessions not being correctly stored. Thus, for this stage of the experiment we have a final sample of 150 subjects instead of 188 so the coefficients are not fully comparable. However, the results from the main analysis remain qualitatively the same and excluding this session from the dataset actually strengthens them. See Table C14 in Appendix C.

TABLE 5–RESULTS FOR THE SINGLE-SEX ENVIRONMENT

Dep. Variable:	*Assessment$_{ss}$*	*Prob(A)$_{ss}$*	*Prob(A)$_{ss}$*	*Prob(W_A)$_{ss}$*	*Prob(W_A)$_{ss}$*	*Mean Earnings*
Sample:	*All*	*All*	*Marginal*	*All*	*Marginal*	*All*
	(1)	(2)	(3)	(4)	(5)	(6)
Signal	7.920***	0.559***	0.672***	1.167***	1.244***	1.634***
	(0.998)	(0.129)	(0.2041)	(0.251)	(0.3748)	(0.4417)
Female	0.0975	0.0573	0.0092	-0.0286	-0.0077	0.116
	(0.204)	(0.0844)	(0.123)	(0.0997)	(0.136)	(0.2592)
Maleness	0.123	0.0262	0.0454	0.0315	0.0428	0.0544
	(0.0993)	(0.0236)	(0.0353)	(0.0309)	(0.0431)	(0.0665)
Maleness∗Female	-0.0378	-0.0503	-0.0793	-0.0284	-0.0799	-0.0479
	(0.126)	(0.0326)	(0.0533)	(0.0397)	(0.0533)	(0.0886)
Tournament Controls	–	YES	YES	YES	YES	–
Session FE	YES	YES	YES	YES	YES	YES
Number of Clusters	–	150	150	150	150	–
Observations	150	1,350	675	1,350	675	150
R-squared	0.643					0.135

Notes: Column (1) shows the OLS estimates for the subject's self-assessment. Column (6) shows the OLS estimates for the subjects' mean earnings across all 9 tournaments. Columns (2)–(3) and (4)–(5) show the marginal effects of the probit model for the probability of choosing tournament A and the probability of getting W_A for the full sample and the subsample of agents classified as marginal, respectively. Interaction corrected using Norton et al. (2004). Robust standard errors for column (1) and (6) and clustered standard errors at subject level for columns (2)–(5) are displayed in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

have the same probability of entering tournament A in the *single-sex* environment and that this behavior is not affected by their beliefs as to the gender nature of the task. Column (3) shows that this is also the case for those agents classified as marginal. Columns (4) and (5) show that perceptions do not affect subjects' probabilities of winning the high-prize either for the full sample or for the sample of marginal subjects and, as before, that for the same signal observed men's and women's likelihood of winning the prize from tournament A is the same. Column (7) shows that perceptions also have no effect on mean earnings.

This robustness provides further support for the self-stereotyping mechanism by bringing down the concerns that the results presented in Table 3 are due to unobserved characteristics, inverse causality or spurious correlations.

However, several caveats regarding the design of the experiment lead us to take data from this stage with a grain of salt. Firstly, subjects may be subject to an order effect and/or influenced by past decisions. The most likely effect of this would be for subjects to show a degree of inertia in behavior which pushes them to replicate past decisions at this stage. Notice that this point is contrary to our findings as it implies that perception-related behavior must be still present through this inertia. The second caveat is that the aim of the experiment is partially disclosed to subjects, given that in stage 3 they are asked about perceptions. They could react to that knowledge by changing their behavior. The last concern is that subjects may learn more accurately what their average performance is as we provide them with new information by displaying the distribution of average performances of their own gender only. This may make the signal more informative, which would reduce the role played by perceptions. However, the data does not support this learning argument. First, a Kolmogorov-Smirnov test rejects on all sessions that the distributions of average performances of men and women are different. Therefore the picture of the distribution of average performances in the first part of the experimental session and that observed at this stage should be similar. This reasoning is supported by the fact that subjects' guess rates for *Assessment* and $Assessment_{ss}$ are 29.7% and 32.6%, figures that are not statistically different. The drawback is that since the *mixed-sex* and *single-sex* distributions of average performances shown to subjects are very similar, subjects who hold non-neutral perceptions may learn that there are no gender differences and thus behave in the same way as agents who hold neutral perceptions. We cannot rule out this last possibility, but this alternative interpretation of the results of this section still confirms that the findings in the *mixed-sex* environment are driven by perceptions and not by unobserved characteristics.

4 Concluding Remarks

Overall, this paper presents a theoretical model and an experiment to test it, the results of which lead us to believe that interaction between stereotypes and self-assessment plays an important role in shaping major decisions of agents both previously and in the labor market. This paper thus helps understand education and labor segregation. Needless to say, we do not posit that the gender segregation observed in the market is generated exclusively through this self-stereotyping mechanisms. Obviously other factors that are highly likely to play a role in shaping it, such as the prevalence of gender roles and social norms (Akerlof and Kranton, 2000), discrimination (Phelps, 1972) or, simply, gender differences in preferences (Croson and Gneezy, 2009). However, the self-stereotyping mechanism brought to light in this paper is worth considering, especially in view of the fact that there is already documentation showing that, in many cases, the segregation observed in the data is consistent with existing gender stereotypes (Cejka and Eagly, 1999; Haveman and Beresford, 2012; Barbulescu and Bidwell, 2013).

Understanding whether gender-based segregation in the labor market can be explained through existing stereotypes is an extremely important issue because women are the targets of negative stereotypes in at least three areas strongly related to wage levels: Quantitative skills (Frome and Eccles, 1998; Nosek et al., 2009), leadership (Schein, 2001; Atwater et al., 2004), and general IQ (Furnham and Gasson, 1998; Furnham et al., 2002; Petrides et al., 2004; Bian et al., 2017).

According to the theoretical model proposed in this paper those stereotypes hurt women's self-assessment and therefore undermine their career goals, leading them to self-select into lower paying itineraries than men. The results of the experiment presented in this paper are very much in line with the theoretical predictions, but only for men. Despite the model's failure to predict women's behavior, this partial validation of the theoretical model is encouraging as it also provides interesting insights into how stereotypes affect choices and suggests directions for policies to deal with them.

In particular, corollary 2.3 sheds light on the factors that influence asymmetry in the behavior of the two social groups. According to the theory, a crucial aspect that moderates the impact of stereotypes is the informativeness of the signal (γ in the theoretical model): the less informative the signal is, the more important the self-stereotyping mechanism becomes and the greater its implications will be in behavior and outcomes. This teaches us that the effects of stereotypes are likely to vanish as agents become more experienced in the task itself. Note, however, that acquiring more experience might not be trivial. If an agent persists in an activity in which he/she is believed to be at a disadvantage, he/she could realize that he/she is really not, but given that he/she believes from the very beginning that he/she is at a disadvantage, his/her incentives to persist in that activity might be low and he/she may well drop out before collecting enough observations. In short, stereotypes could be a vicious circle. This has clear policy implications, as any intervention that provides more and better information on agents' abilities would lower the impact of stereotypes. Previous research indirectly shows that such interventions can be successful (Brandts et al., 2014; Joensen and Nielsen, 2015; Alan and Ertac, 2019). These potential interventions are not tested in the current experiment but they indicate a promising direction for further research into this topic.

References

Ai, C. and Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics letters*, 80(1), 123–129.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, (pp. 715–753).

Alan, S. and Ertac, S. (2019). Mitigating the gender gap in the willingness to compete: Evidence from a randomized field experiment. *Journal of the European Economic Association*, 17(4), 1147–1185.

Arrow, K. (1973). The theory of discrimination. *Discrimination in labor markets*, 3(10).

Atwater, L. E., Brett, J. F., Waldman, D., DiMare, L., and Hayden, M. V. (2004). Men's and women's perceptions of the gender typing of management subroles. *Sex Roles*, 50(3-4), 191–199.

Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics*, 1236–1239.

Barbulescu, R. and Bidwell, M. (2013). Do women choose different jobs from men? mechanisms of application segregation in the market for managerial workers. *Organization Science*, 24(3), 737–756.

Bayard, K., Hellerstein, J., Neumark, D., and Troske, K. (2003). New evidence on sex segregation and sex differences in wages from matched employee–employer data. *Journal of Labor Economics*, 21(4), 887–922.

Bertrand, M. and Hallock, K. F. (2001). The gender gap in top corporate jobs. *Industrial & Labor Relations Review*, 55(1), 3–21.

Beyer, S. (1990). Gender differences in the accuracy of self-evaluations of performance. *Journal of Personality and Social Psychology*, 59(5), 960–970.

Beyer, S. and Bowden, E. M. (1997). Gender differences in self-perceptions: Convergent evidence from three measures of accuracy and bias. *Personality and Social Psychology Bulletin*, 23(2), 157–172.

Bian, L., Leslie, S.-J., and Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, 355(6323), 389–391.

Blau, F. D. and Kahn, L. M. (1997). Swimming upstream: Trends in the gender wage differential in the 1980s. *Journal of Labor Economics*, (pp. 1–42).

Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3), 789–865.

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

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131(4), 1753–1794.

Brandts, J., Groenert, V., and Rott, C. (2014). The impact of advice on women's and men's selection into competition. *Management Science*, 61(5), 1018–1035.

Card, D., Cardoso, A. R., and Kline, P. (2016). Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women. *The Quarterly Journal of Economics*, 131(2), 633–686.

Cejka, M. A. and Eagly, A. H. (1999). Gender–stereotypic images of occupations correspond to the sex segregation of employment. *Personality and Social Psychology Bulletin*, 25(4), 413–423.

Chatard, A., Guimond, S., and Selimbegovic, L. (2007). "How good are you in math?" the effect of gender stereotypes on students' recollection of their school marks. *Journal of Experimental Social Psychology*, 43(6), 1017–1024.

Coate, S. and Loury, G. C. (1993). Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, 83(5), 1220–1240.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, 129(4), 1625–1660.

Coffman, K. B., Flikkema, C. B., and Shurchkov, O. (2019). Gender stereotypes in deliberation and team decisions. *Harvard Business School*.

Crosetto, P. and Filippin, A. (2017). Safe options induce gender differences in risk attitudes. *IZA Discussion Papers*.

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

Cubel, M. and Sanchez-Pages, S. (2017). Gender differences and stereotypes in strategic reasoning. *The Economic Journal*, 127(601), 728–756.

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

Eckel, C. C. and Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, 23(4), 281–295.

Ehrlinger, J. and Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology*, 84(1), 5–17.

Fernandez, R. M. and Friedrich, C. (2011). Gender sorting at the application interface. *Industrial Relations: A Journal of Economy and Society*, 50(4), 591–609.

Filippin, A. and Crosetto, P. (2016). A reconsideration of gender differences in risk attitudes. *Management Science*.

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

Frome, P. M. and Eccles, J. S. (1998). Parents' influence on children's achievement-related perceptions. *Journal of Personality and Social Psychology*, 74(2), 435–452.

Furnham, A. and Gasson, L. (1998). Sex differences in parental estimates of their children's intelligence. *Sex Roles*, 38(1-2), 151–162.

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

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

Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6), 504–528.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1), 114–125.

Grosse, N., R., G., and Dertwinkel-Kalt, M. (2014). Explaining gender differences in competitiveness: Testing a theory on gender-task stereotypes. *Working paper, University of Mannheim*.

Günther, C., Ekinci, N. A., Schwieren, C., and Strobel, M. (2010). Women can't jump?–an experiment on competitive attitudes and stereotype threat. *Journal of Economic Behavior & Organization*, 75(3), 395–401.

Halladay, B. (2017). Perception matters: The role of task gender stereotype on confidence and tournament selection. *Mimeo*.

Harsanyi, J. C. (1967-1968). Games with incomplete information played by "bayesian" players, i–iii. *Management Science*, 14(3), 159–182, 320–334, 486–502.

Haveman, H. A. and Beresford, L. S. (2012). If you're so smart, why aren't you the boss? explaining the persistent vertical gender gap in management. *The ANNALS of the American Academy of Political and Social Science*, 639(1), 114–130.

Hernandez-Arenaz, I. (2018). Exploring gender gaps in competitive-sorting. *Mimeo*.

Hsieh, C.-T., Hurst, E., Jones, C. I., and Klenow, P. J. (2019). The allocation of talent and us economic growth. *Econometrica*, 87(5), 1439–1474.

Joensen, J. S. and Nielsen, H. S. (2015). Mathematics and gender: Heterogeneity in causes and consequences. *The Economic Journal*.

Klinowski, D. (2019). Selection into self-improvement and competition pay: Gender, stereotypes, and earnings volatility. *Journal of Economic Behavior & Organization*, 158, 128–146.

Latrofa, M., Vaes, J., Cadinu, M., and Carnaghi, A. (2010). The cognitive representation of self-stereotyping. *Personality and Social Psychology Bulletin*, 36(7), 911–922.

Lenney, E. (1977). Women's self-confidence in achievement settings. *Psychological Bulletin*, 84(1), 1–13.

Leslie, S.-J., Cimpian, A., Meyer, M., and Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, 347(6219), 262–265.

Ludsteck, J. (2014). The impact of segregation and sorting on the gender wage gap: Evidence from German linked longitudinal employer-employee data. *Industrial & Labor Relations Review*, 67(2), 362–394.

Macpherson, D. A. and Hirsch, B. T. (1995). Wages and gender composition: why do women's jobs pay less? *Journal of Labor Economics*, 13(3), 426–471.

Morgan, J., Sisak, D., and Várdy, F. (2017). The ponds dilemma. *The Economic Journal*, 128(611), 1634–1682.

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

Norton, E. C., Wang, H., Ai, C., et al. (2004). Computing interaction effects and standard errors in logit and probit models. *Stata Journal*, 4, 154–167.

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, and others (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, 106(26), 10593–10597.

O'Neill, J. (2003). The gender gap in wages, circa 2000. *The American Economic Review*, 93(2), 309–314.

Petrides, K. V., Furnham, A., and Martin, G. N. (2004). Estimates of emotional and psychometric intelligence: Evidence for gender-based stereotypes. *The Journal of Social Psychology*, 144(2), 149–162.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The American Economic Review*, 62(4), 659–661.

Pomerantz, E. M., Altermatt, E. R., and Saxon, J. L. (2002). Making the grade but feeling distressed: Gender differences in academic performance and internal distress. *Journal of Educational Psychology*, 94(2), 396–404.

Schein, V. E. (2001). A global look at psychological barriers to women's progress in management. *Journal of Social Issues*, 57(4), 675–688.

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

Thébaud, S. (2010). Gender and entrepreneurship as a career choice: do self-assessments of ability matter? *Social Psychology Quarterly*, 73(3), 288–304.

Weber, E. U., Blais, A.-R., and Betz, N. E. (2002). A domain-specific risk-attitude scale: measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, 15(4), 263–290.

Appendices

A Technical Appendix (For Online Publication)

A.1 Formal derivation of expected utilities

Let $q_g : r \to [1,0]$ be a generic strategy profile for agents from social group $g \in \{S,N\}$ mapping each signal observed into the probability of entering tournament A and let $q_g^{\alpha}(r)$ be the corresponding one at equilibrium.

Given any arbitrary pair of strategy profiles $(q_s(r), q_n(r))$, the expected entry rate into tournament A of those agents whose ability is exactly a given the distributional assumptions $F^S(a)$ and $F^N(a)$ would be

$$E\left[\lambda_a q_s(r) + (1-\lambda_a) q_n(r)|a\right] = \int_{-\infty}^{\infty} \left[\lambda_a q_s(a+\mu) + (1-\lambda_a) q_n(a+\mu)\right] \phi(\mu) d\mu$$

where λ_a is the believed proportion of S-agents among those with ability a, which equals $\lambda_a = \lambda f^S(a)/[\lambda f^S(a) + (1-\lambda) f^N(a)]$. Thus, $E\left[\lambda_a q_s(r) + (1-\lambda_a) q_n(r)|a\right]$ represents, provided that the stereotype is believed true and that the behavioral rule $(q_s(r), q_n(r))$ is common knowledge, the proportion of agents with ability a who will enter tournament A.[38] Consequently, among the participants in tournament A, only those whose real ability exceeds a certain threshold a_A holding[39]

$$\int_{a_A}^{\infty} E\left[\lambda_a q_s(r) + (1-\lambda_a) q_n(r)|a\right] dF^{\alpha}(a) \leq \delta_A \quad (= \text{ if } a_A > -\infty)$$

will expect to obtain a prize, where $F^{\alpha}(a) \equiv \lambda F^S(a) + (1-\lambda) F^N(a)$ is the CDF of the (commonly) believed distribution of abilities in the whole population. Therefore, given the pair of strategy profiles $(q_s(r), q_n(r))$, the expected return (utility) from participating in tournament A for an agent who observes signal r_i from social group $g \in \{S,N\}$ is

$$U[A|q_s(r), q_n(r), r_i, F^g] = W_A[1 - \hat{F}(a_A|r_i, F^g)] \tag{3}$$

where W_A is the prize that can be obtained in tournament A and $1 - \hat{F}(a_A|r_i, F^g)$ is just the self-assessed probability of having a real ability level higher than a_A.

Similarly, the expected entry rate in tournament B of those agents whose ability is exactly a is $1 - E\left[\lambda_a q_s(r) + (1-\lambda_a) q_n(r)|a\right]$ so only those with a real ability higher than a certain threshold a_B which holds

$$\int_{a_B}^{\infty} 1 - E\left[\lambda_a q_s(r) + (1-\lambda_a) q_n(r)|a\right] dF^{\alpha}(a) \leq \delta_B \quad (= \text{ if } a_B > -\infty)$$

will expect to obtain a prize. Therefore, given the pair of strategy profiles $(q_s(r), q_n(r))$, the expected return from participating in tournament B for an agent who observes signal r_i from social group $g \in \{S,N\}$ is

$$U[B|q_s(r), q_n(r), r_i, F^g] = W_B[1 - \hat{F}(a_B|r_i, F^g)]$$

where W_B is the prize that can be obtained in tournament B and $1 - \hat{F}(a_B|r_i, F^g)$ is the self-assessed probability of having a real ability level higher than a_B.

---

[38] In equilibrium, the behavioral rule $(q_s(r), q_n(r))$ should be common knowledge as the common prior assumption holds (Aumann, 1976). To see that the common prior assumption holds, notice that the fraction of stereotyped agents in the society (λ), the prior about how abilities are distributed within groups $(F^N(.), F^S(.))$, the stereotype (α), and the structure of the tournaments $(\delta_A, \delta_B, W_A, W_B)$ are assumed to be public information, by rationality this is also common knowledge. Thus, we can extend the game one step backwards and assume, with the same information as before being public knowledge, that in a first stage nature chooses with probability λ each agent to be an S-agent and with probability $1-\lambda$ to be an N-agent. As implied by Harsanyi (1967-1968), this game is identical to the one presented in the main text and thus the equilibrium would be exactly the same.

[39] By replacing and doing the maths this expression can be transformed into

$$\lambda \int_{a_A}^{\infty} \left[\int_{-\infty}^{\infty} q_s(a+\mu)\phi(\mu) d\mu\right] dF^S(a) + (1-\lambda) \int_{a_A}^{\infty} \left[\int_{-\infty}^{\infty} q_n(a+\mu)\phi(\mu) d\mu\right] dF^N(a) \leq \delta$$

A.2 Characterization of the equilibrium

First I demonstrate formally that the equilibrium strategy profiles, $(q_s^\alpha(r), q_n^\alpha(r))$, should induce a situation in which tournament A is necessarily overcrowded. Consider any pair of strategy profiles $(q_s'(r), q_n'(r))$ such that the induced standard needed to beat for tournament A, a_A, is $-\infty$. Then it is easy to see that $\forall g \in \{S, N\}$

$$U[A|q_s'(r), q_n'(r), r_i, F^g] = W_A > W_B \geq U[B|q_s'(r), q_n'(r), r_i, F^g] \quad \forall r_i$$

which clearly contradicts the fact that the pair $(q_s'(r), q_n'(r))$ represent an equilibrium as all agents have incentives to choose tournament A. This has two immediate consequences. First, since tournament A is overcrowded the standard for winning in tournament A at equilibrium should be given by a value $a_A \equiv a^\alpha > -\infty$ holding

$$\int_{a^\alpha}^{\infty} E\left[\lambda_a q_s^\alpha(r) + (1-\lambda_a) q_n^\alpha(r)|a\right] dF^\alpha(a) = \delta_A \tag{4}$$

which furthermore implies that $a^\alpha < \infty$. Second, since $\delta_A + \delta_B \geq 1$ in equilibrium tournament B should end up with vacancies, i.e. $\int_{a_B}^{\infty} 1 - E\left[\lambda_a q_s^\alpha(r) + (1-\lambda_a) q_n^\alpha(r)|a\right] dF^\alpha(a) < \delta_B$, $\forall a_B$ which implies the existence of a corner solution at $a_B = -\infty$, so $U[B|q_s^\alpha(r), q_n^\alpha(r), r_i, F^g] = W_B$, $\forall r_i, g$. This means that in equilibrium agents' decision can be simplified to two options: getting W_B for sure or getting W_A if his real ability happens to be greater than a^α and 0 otherwise.

Since the believed ability threshold needed to beat at tournament A to get the prize in equilibrium, a^α, is an interior point in $(-\infty, \infty)$, it can be checked from equation (3) that the boundary conditions for agents' expected utility in tournament A at the equilibrium are $\lim_{r_i \to -\infty} U[A|q_s^\alpha(r), q_n^\alpha(r), r_i, F^g] = 0$ and $\lim_{r_i \to \infty} U[A|q_s^\alpha(r), q_n^\alpha(r), r_i, F^g] = W_A$.

Moreover, from equation (3) follows that $U[A|q_s(r), q_n(r), r_i, F^g]$ is differentiable and strictly increasing on r_i for any pair $(q_s(r), q_n(r))$ that induce an ability threshold $a^\alpha \in (-\infty, \infty)$. This strictly monotone increasing property of $U[A|q_s(r), q_n(r), r_i, F^g]$ together with its boundary conditions allows us to claim that the group-specific strategy profile in equilibrium must be also monotone increasing. In particular, for any pair of strategy profiles $(q_s^\alpha(r), q_n^\alpha(r))$ that constitutes an equilibrium there must be a signal value r_g^α for each $g \in \{S, N\}$ such that

$$U[A|q_s^\alpha(r), q_n^\alpha(r), r_g^\alpha, F^g] \equiv W_A[1 - \hat{F}(a^\alpha|r_g^\alpha, F^g)] = W_B \tag{5}$$

so that, by the strictly monotone property, $U[A|q_s^\alpha(r), q_n^\alpha(r), r_i, F^g] \gtreqless U[B|q_s^\alpha(r), q_n^\alpha(r), r_i, F^g] = W_B$ if $r_i \gtreqless r_g^\alpha$. Therefore it can be concluded that, provided a^α, the equilibrium strategy profiles for social group g is characterized by a signal contingent strategy profile such that $q_g^\alpha(r_i) = 1$ if $r_i > r_g^\alpha$ and $q_g^\alpha(r_i) = 0$ if $r_i < r_g^\alpha$. Substituting this strategy profiles into equation (4) follows that the believed real ability threshold at the equilibrium should hold[40]

$$\int_{a^\alpha}^{\infty} \lambda[1 - \Phi(r_S^\alpha - a)] f^S(a) + (1-\lambda)[1 - \Phi(r_N^\alpha - a)] f^N(a) da = \delta_A \tag{6}$$

that together with equation (5) fully characterizes the equilibrium of this game.

A.3 Existence and uniqueness of the equilibrium

By Appendix A.2 above, we know that the equilibrium is characterized by a threshold strategy summarized by a signal value r_g^α for each social group. Assume an arbitrary pair signal threshold $\bar{r}_S, \bar{r}_N$. Given imperfect self-knowledge, there are some agents from social group $g \in \{S, N\}$ who have very low (high) ability levels who received a signal $r > (<) \bar{r}_g$, and thus mistakenly opted for tournament

[40] See footnote 39 for the intermediate step going from equation (4) to equation (6).

A (B). Specifically, the probability of an agent from social group g with real ability a participating in tournament A is conditional on he/she receiving a signal $r > \bar{r}_g$. This probability can be computed as

$$Prob(r > \bar{r}_g | a) = Prob(a + \mu > \bar{r}_g) = 1 - \Phi(\bar{r}_g - a)$$

Therefore, given an arbitrary pair of threshold signals $(\bar{r}_S, \bar{r}_N)$, the mass of agents participating in tournament A in equilibrium can be computed as $\int_{-\infty}^{\infty} \lambda[1 - \Phi(\bar{r}_S - a)]f^s(a) + (1-\lambda)[1 - \Phi(\bar{r}_N - a)]f^N(a)da$. Consequently agents believe that the CDF of abilities *within* those participating in tournament A is summarized by the function

$$\mathcal{A}^{\bar{r}_S,\bar{r}_N}(a') \equiv \frac{\int_{-\infty}^{a'} \lambda[1 - \Phi(\bar{r}_S - a)]f^s(a) + (1-\lambda)[1 - \Phi(\bar{r}_N - a)]f^N(a)da}{\int_{-\infty}^{\infty} \lambda[1 - \Phi(\bar{r}_S - a)]f^s(a) + (1-\lambda)[1 - \Phi(\bar{r}_N - a)]f^N(a)da}$$

Thus, agents believe that within those agents participating in tournament A there is a proportion of agents $\mathcal{A}^{\bar{r}_S,\bar{r}_N}(a')$ whose ability levels are lower than a', while the remaining proportion $1 - \mathcal{A}^{\bar{r}_S,\bar{r}_N}(a')$ has ability levels strictly higher than a'.

On the other hand, notice that there is a mass of prizes of exactly δ_A for participants in tournament A. As the mass of agents in that tournament is $\int_{-\infty}^{\infty} \lambda[1-\Phi(\bar{r}_S - a)]f^s(a) + (1-\lambda)[1-\Phi(\bar{r}_N - a)]f^N(a)dy$, the proportion of agents participating in tournament A who obtain W_A is

$$\Delta_A^{\bar{r}_S,\bar{r}_N} = \frac{\delta_A}{\int_{-\infty}^{\infty} \lambda[1 - \Phi(\bar{r}_S - a)]f^s(a) + (1-\lambda)[1 - \Phi(\bar{r}_N - a)]f^N(a)da} > \delta_A$$

Thus the probability of obtaining a prize in tournament A is equal to the probability of being in the top $\Delta_A^{\bar{r}_S,\bar{r}_N}$ of the distribution $\mathcal{A}^{\bar{r}_S,\bar{r}_N}(a)$. As we are working with a continuum of agents, for any pair $(\bar{r}_S, \bar{r}_N)$ there is an ability level $\bar{a}$ such that

$$1 - \mathcal{A}^{\bar{r}_S,\bar{r}_N}(\bar{a}) = \Delta_A^{\bar{r}_S,\bar{r}_N}$$

In other words, given $(\bar{r}_S, \bar{r}_N)$, agents believe that the prize in A will only be won by those subjects participating in A and with ability level higher than $\bar{a}$. This gives rise to the following lemma:

Lemma A.1. *For any arbitrary pair $(\bar{r}_S, \bar{r}_N)$ there is a uniquely determined ability level $\bar{a}$ such that $1 - \mathcal{A}^{\bar{r}_S,\bar{r}_N}(\bar{a}) = \Delta_A^{\bar{r}_S,\bar{r}_N}$. This implies the existence of a function mapping pairs of signal thresholds (r_S, r_N) into ability levels. This function is denoted by $\mathbf{a} : (r_S, r_N) \to a$ and $\forall (r_S, r_N)$ solves*

$$\int_{\mathbf{a}(r_S,r_N)}^{\infty} \lambda[1 - \Phi(r_S - a)]f^S(a) + (1-\lambda)[1 - \Phi(r_N - a)]f^N(a)da \leq \delta_A \quad (= \textit{ if } \mathbf{a}(r_S, r_N) > -\infty) \quad (7)$$

Moreover, $\frac{\partial \mathbf{a}(r_S,r_N)}{\partial r_g} < 0 \; \forall r_g$*,* $g \in \{S, N\}$*,* $\lim_{(r_S,r_N)\to(-\infty,-\infty)} \mathbf{a}(r_S, r_N) \equiv a^{Max} \in (-\infty, \infty)$*, and* $\lim_{(r_S,r_N)\to(\infty,\infty)} \mathbf{a}(r_S, r_N) = -\infty$.

Proof: *The implicit solution for $\mathbf{a}(r_S, r_N)$ in equation (7) is immediately apparent if we substitute $\mathcal{A}^{r_S,r_N}(\mathbf{a}(r_S, r_N))$ and $\Delta_A^{r_S,r_N}$ by their extended expressions and do the math. On the other hand, uniqueness is easily proven once it is noted that equation (7) is monotonically decreasing in $\bar{a}$.*

If $(r_S, r_N) \to (-\infty, -\infty)$ means that all agents will enter tournament A and equation (7) yields that $\mathbf{a}(r_S, r_N)$ implicitly solves

$$\int_{\lim_{(r_S,r_N)\to(-\infty,-\infty)} \mathbf{a}(r_S,r_N)}^{\infty} \lambda f^s(a) + (1-\lambda)f^s(a)da = 1 - F^{\alpha}\left(\lim_{(r_S,r_N)\to(-\infty,-\infty)} \mathbf{a}(r_S, r_N)\right) = \delta_A$$

from what it emerges that the upper bound for the ability threshold is

$$\lim_{(r_S,r_N)\to(-\infty,-\infty)} \mathbf{a}(r_S, r_N) = (F^{\alpha})^{-1}(1 - \delta_A) = a^{Max} \in (-\infty, \infty)$$

Similarly, if $(r_S, r_N) \to (\infty, \infty)$ equation (7) shows that there is no interior solution for $\mathbf{a}(r_S, r_N)$ as

$$\lim_{(r_S,r_N)\to(\infty,\infty)} \int_{a'}^{\infty} \lambda[1-\Phi(r_S-a)]f^S(a)+(1-\lambda)[1-\Phi(r_N-a)]f^N(a)da = \int_{a'}^{\infty} 0dy = 0 < \delta_A, \quad \forall a' \in (-\infty, \infty)$$

so the corner solution applies, leading to $\lim_{(r_S,r_N)\to(\infty,\infty)} \mathbf{a}(r_S, r_N) = -\infty$.

The last part of the lemma is proven straightforwardly by noting that for any change in the pair (r_S, r_N) the value of $\mathbf{a}(r_S, r_N)$ should also change such that equation (7) holds. Thus, the total derivative with respect to r_g, $g \in \{S, N\}$, in the foregoing expression should be zero, i.e.

$$-[1 - \Phi(r_g - \mathbf{a}(r_S, r_N))]f(\mathbf{a}(r_S, r_N))\frac{\partial \mathbf{a}(r_S, r_N)}{\partial r_g} - \int_{\mathbf{a}(r_S,r_N)}^{\infty} \phi(r_g - y)f(y)dy \equiv 0 \Rightarrow \frac{\partial \mathbf{a}(r_S, r_N)}{\partial r_g} < 0$$

Thus, the mapping $\mathbf{a} : (r_S, r_N) \to a$ can be represented by means of a monotonically decreasing, continuous function in both arguments whose codomain is the set $(-\infty, a^{Max})$. □

The above lemma simply states that $\mathbf{a}(r_S, r_N)$ can be considered as a function mapping each pair of potential signal levels behaving as a threshold to exactly one compatible real ability threshold level. Moreover, it implies that this function is upper bounded at a^{Max} and that it is decreasing on r_g $\forall g \in \{S, N\}$ such that $\mathbf{a} : (r_S, r_N) \in \mathbb{R}^2 \to a \in (-\infty, a^{Max})$.[41]

Now fix any arbitrary ability threshold $\bar{a}$. Given $\bar{a}$, the assigned probability of an agent from social group $g \in \{S, N\}$ observing signal r_i for winning the prize in tournament A is just the probability that he/she assigns –given distributional assumptions– to having a real ability greater than $\bar{a}$ conditional on the signal he/she has observed, i.e. $1 - \hat{F}(\bar{a}|r_i, F^g)$.

Lemma A.2. *There exists a function for each social group $g \in \{S, N\}$ mapping abilities thresholds (a) into signal thresholds. This function is denoted by $\mathbf{r_g} : a \to r_g$ and $\forall a \in \mathbb{R}$ solves*

$$W_A[1 - \hat{F}(a|\mathbf{r_g}(a), F^g)] = W_B \tag{8}$$

Moreover, $\frac{\partial \mathbf{r_g}(a)}{\partial a} > 0 \; \forall a \in \mathbb{R}$, $\lim_{a\to-\infty} \mathbf{r_g}(a) = -\infty$, *and* $\lim_{a\to\infty} \mathbf{r_g}(a) = \infty$.

Proof: *The proof is straightforward once it is noted that for any a, $\mathbf{r_g}(a)$ should be such that equation (8) holds for any value of $a \in \mathbb{R}$. Thus, the total derivative with respect to a of equation (8) should be zero, i.e.*

$$W_A \left[-\hat{f}(a|\mathbf{r_g}(a), F^g) - \frac{\partial \hat{F}(a|\mathbf{r_g}(a), F^g)}{\partial \mathbf{r_g}(a)} \frac{\partial \mathbf{r_g}(a)}{\partial a} \right] \equiv 0 \Rightarrow \frac{\partial \mathbf{r_g}(a)}{\partial a} > 0$$

The boundary conditions are straightforward and can be checked by applying the limits to equation (8). Thus, the mapping $\mathbf{r_g} : a \in \mathbb{R} \to r_g \in \mathbb{R}$ can be represented by means of a monotonically increasing, continuous function. □

Having established Lemma A.1 and A.2, we can state the main proposition of this subsection which establishes the existence and uniqueness of the equilibrium:

Proposition A.3. *There is a real ability threshold, a^α, such that its uniquely induced behavioral rules, $r_S(a^\alpha), r_N(a^\alpha)$, induce simultaneously a^α as a real ability threshold, i.e.*

$$a(r_N(a^\alpha), r_S(a^\alpha)) = a^\alpha$$

Moreover, given the properties of the mappings $a : (r_N, r_S) \to a$ and $r_g : a \to r$ stated in lemmas A.1 and A.2, this element a^α exists and is uniquely determined.

[41]Notice that this value a^{Max} is just the ability threshold needed to beat if everyone enters tournament A. Therefore, it can be thought as to represent the minimum ability an agent should have in order to consider tournament A a strictly dominant strategy. This value a^{Max} also represents the minimum ability required to participate in tournament A under perfect self-knowledge.

Proof: *From lemmas A.1 and A.2 we find that any real ability value established as standard induces one signal threshold level for each social group that at the same time induce a new standard. Therefore, in equilibrium we should find that both the signal threshold levels (induced by the real ability standard) and the real ability level (induced by the signal threshold levels) should be compatible with the beliefs held by the agents. Summarizing, we are looking for a value* a^α *that represents a fixed point for the system* $\mathbf{a}(\mathbf{r_N}(a^\alpha), \mathbf{r_S}(a^\alpha)) = a^\alpha$*. Now we can prove A.3 very easily.*

The first thing to notice is that, given the properties on lemmas A.1 and A.2

$$\frac{d\mathbf{a}(\mathbf{r_N}(a), \mathbf{r_S}(a))}{da} = \frac{\partial \mathbf{a}(r_N, r_S)}{\partial r_N}\frac{\partial \mathbf{r_N}(a)}{\partial a} + \frac{\partial \mathbf{a}(r_N, r_S)}{\partial r_S}\frac{\partial \mathbf{r_S}(a)}{\partial a} < 0$$

and that $\mathbf{a}(\mathbf{r_N}(a), \mathbf{r_S}(a))$ *is continuous on* a*. Thus, if there is* a^α *holding* $\mathbf{a}(\mathbf{r_N}(a^\alpha), \mathbf{r_S}(a^\alpha)) = a^\alpha$*, it must be unique.*

Next it can be observed that the limit of $\mathbf{a}(\mathbf{r_N}(a), \mathbf{r_S}(a))$ *when* $a \to -\infty$ *is*

$$\lim_{a\to-\infty} \mathbf{a}(\mathbf{r_N}(a), \mathbf{r_S}(a)) = \lim_{(r_N, r_S)\to(-\infty,-\infty)} \mathbf{a}(r_N, r_S) = a^{Max}$$

and that the limit of $\mathbf{a}(\mathbf{r_N}(a), \mathbf{r_S}(a))$ *when* $a \to \infty$ *is*

$$\lim_{a\to\infty} \mathbf{a}(\mathbf{r_N}(a), \mathbf{r_S}(a)) = \lim_{(r_N, r_S)\to(\infty,\infty)} \mathbf{a}(r_N, r_S) = -\infty$$

Therefore, we find that the compound function $\mathbf{a}(\mathbf{r_N}(a), \mathbf{r_S}(a)) : a \in \mathbb{R} \to (-\infty, a^{Max})$ *is continuous on* a *and it is strictly decreasing. Thus, exists a unique value* $a^\alpha \in (-\infty, a^{Max})$ *holding* $\mathbf{a}(\mathbf{r_N}(a^\alpha), \mathbf{r_S}(a^\alpha)) = a^\alpha$*.* □

A.4 Proof of propositions: Representation

Proof of Proposition 2.4: Assume that the stereotype is negative so $\alpha < 0$. By corollary 2.3, $r_S^\alpha > r_N^\alpha$ so $1 - \Phi(r_S^\alpha - a) < 1 - \Phi(r_N^\alpha - a)$ $\forall a$ which actually means that the probability of an N-agent participating in A is greater than for an S-agent with the same real ability. With this it is easy to check that the denominator of the representation index shown in equation (2) is

$$\int_{-\infty}^{\infty} \lambda[1 - \Phi(r_S^\alpha - a)]f(a) + (1-\lambda)[1 - \Phi(r_N^\alpha - a)]f(a)da >$$
$$\int_{-\infty}^{\infty} \lambda[1 - \Phi(r_S^\alpha - a)]f(a) + (1-\lambda)[1 - \Phi(r_S^\alpha - a)]f(a)da = \int_{-\infty}^{\infty} [1 - \Phi(r_S^\alpha - a)]f(a)dy$$

which proves that $1 > \frac{\int_{-\infty}^{\infty}[1-\Phi(r_S^\alpha - a)]f(a)da}{\int_{-\infty}^{\infty}\lambda[1-\Phi(r_S^\alpha - a)]f(a)+(1-\lambda)[1-\Phi(r_N^\alpha - a)]f(a)da} = \frac{\lambda_A^\alpha}{\lambda}$. The analogous analysis for $\alpha > 0$ proves that $\frac{\lambda_A^\alpha}{\lambda} > 1$ □

Proof of Proposition 2.5: Assume that the stereotype is negative. Since the stereotype is by assumption false, a^α does not represent the real minimum ability necessary to win a prize in tournament A. Instead, as S-agents are ex-ante more skillful than predicted by the stereotype the required minimum ability for obtaining W_A given the behavioral rules (r_S^α, r_N^α) is some $\hat{a} > a^\alpha$. Thus, the proportion of S-agents who will win a prize W_A from social group S can be expressed as

$$\omega_A^\alpha = \frac{\lambda\int_{\hat{a}}^{\infty}[1 - \Phi(r_S^\alpha - a)]f(a)da}{\int_{\hat{a}}^{\infty}\lambda[1 - \Phi(r_S^\alpha - a)]f(a) + (1-\lambda)[1 - \Phi(r_N^\alpha - a)]f(a)da}$$

so the equivalent index to the one shown in equation (2) is $\frac{\omega_A^\alpha}{\lambda}$. Like the index for participation, this index reflects under/overrepresentation when $\frac{\omega_A^\alpha}{\lambda} \neq 1$. As in the proof of proposition 2.4 it is straightforward to see that

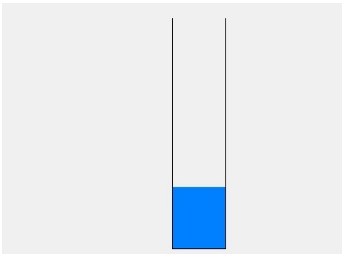
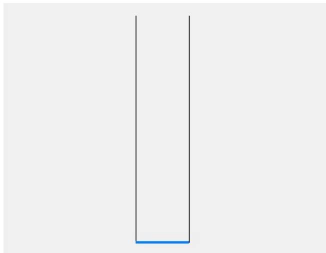
$$\int_{\hat{a}}^{\infty}[1 - \Phi(r_S^\alpha - a)]f(a)da < \int_{\hat{a}}^{\infty}\lambda[1 - \Phi(r_S^\alpha - a)]f^S(a) + (1-\lambda)[1 - \Phi(r_N^\alpha - a)]f^N(a)da$$

which comes directly from the fact that as $r_S^\alpha > r_N^\alpha$ so $1 - \Phi(r_S^\alpha - a) < 1 - \Phi(r_N^\alpha - a)\ \forall a$. This implies that $\frac{\omega_A^\alpha}{\lambda} < 1$. If the stereotype is positive an analogous analysis takes place, but taking into account that $\hat{a} < a^\alpha$ and $r_S^\alpha < r_N^\alpha$. $\square$

B Instructions for the experiment (For Online Publication)

B.1 Instructions for Stage 1

In this stage you will be asked to perform a task that involves two pictures. The first picture will show a glass containing some water (like the one in the left picture) and will be located at the top of the screen. This first picture will be displayed for 3 seconds. After that time, the picture of the full glass will disappear and, a second picture of an empty glass will appear at the bottom of your screen with a blue bar at its base (like the picture on the right).

Your task is to indicate on the second picture the water level from the first one. To do this, left click your mouse on the blue bar in the second picture and drag it until it matches the level of water shown in the first picture. Once you are satisfied with the location of the bar, press the “OK” button that will be displayed at the bottom right of your screen. This will take you automatically to a new picture of glass with water. The place where the bar is when you press “OK” determines the location of the bar for the purpose of computing your score.

You will perform this task 14 consecutive times with different initial pictures. The score you awarded for each repetition will be computed according to the following formula:

$$Score = 100 - 5 * distance(Real\_Level\_Water - Bar\_Position)$$

where 100 is the height of the glass and distance(Real_Level_Water-Bar_Position) is the absolute value of the distance between the water level in the first picture and the level where you place the bar in the second picture. If you do not understand the formula exactly, don’t panic. The important thing is that the closer you leave the bar to the actual water level, the higher your score will be. NOTE: If the distance is greater than 20 your score will be negative.

PAYMENTS:
The computer will choose at random one of the 14 times that you perform the task and you will be paid according to the following rule

$$Payment = max(5 * Score/100, 0)$$

For example, if the distance is 0 in the randomly chosen repetition (i.e. you placed the bar at exactly the right water level), your score will be 100 (100=100-0) and you will get €5 (€5=€5*1). If the distance is 4 in the randomly chosen repetition your score will be 80 (80=100-5*4) and you will get €4 (€4=€5*0.8). If the distance between the water level and the place where you locate the bar in the randomly chosen repetition is greater than 20 your score will be negative and you will get €0.

B.2 Instructions for Stage 2 (Not Manipulated)

In this part of the experiment you will be asked to estimate your average score over the 14 repetitions of the task you performed previously.

This average represents your real ability in the task. If your estimation matches your actual average score you will be paid €1.5. Otherwise you will get €0.

Before you enter your estimation, we will provide you with certain information that may help you. Please press “OK” to continue.

(INFORMATION_SIGNAL) Now you can see on your screen the score chosen randomly by the computer from among the 14 times you have performed the task, i.e. the score that you see is your real score from one of the 14 times that you have completed the “glass task”.

Please, press “OK” to continue.

(INFORMATION_OTHERS) Now you can see the distribution of the average scores have been distributed for the session in intervals of three. As you can see, this information is available in both chart and plot form.

Please, press “OK” to continue.

(SELF-ASSESSMENT) Now you can see at the bottom right of your screen a list of options. All you have to do is choose the group to which you believe you belong from among the 11 possibilities available. If you guess correctly you will get an extra €1.5. Once you have selected an option press “OK” to confirm your choice. You will then be asked another question that you must also answer. Once everyone has answered this additional question you will go on to stage 3 of the experiment.

B.3 Instructions for Stage 2 (Manipulated)

In this part of the experiment you will be asked to estimate your average score over the 14 repetitions of the task you performed previously.

This average represents your real ability in the task. If your estimation matches your actual average score you will be paid €1.5. Otherwise you will get €0.

Before you enter your estimation, we will provide you with certain information that may help you. Please press "OK" to continue.

(INFORMATION_FRAMING)
DESCRIPTION OF THE TASK: The task that you have just performed is closely linked to your ability to perceive the proportions of one object and transfer them to another. In particular, the previous task is based in your spatial vision capability and your subsequent handling of visual information. These abilities are essential for performing tasks in the field of engineering.

According to the description of the BACHELOR'S DEGREE IN ENGINEERING provided by the UPV/EHU one of the BASIC SKILLS required for engineers is:
"The capability for spatial vision and knowledge of graphic representation techniques, including both conventional methods of metric geometry and descriptive geometry, such as computer-aided design applications."

Please, press "OK" to continue.

(INFORMATION_SIGNAL) Now you can see on your screen the score chosen randomly by the computer from among the 14 times you have performed the task, i.e. the score that you see is your real score from one of the 14 times that you have completed the "glass task".

Please, press "OK" to continue.

(INFORMATION_OTHERS) Now you can see the distribution of the average scores have been distributed for the session in intervals of three. As you can see, this information is available in both chart and plot form.

Please, press "OK" to continue.

(SELF-ASSESSMENT) Now you can see at the bottom right of your screen a list of options. All you have to do is choose the group to which you believe you belong from among the 11 possibilities available. If you guess correctly you will get an extra €1.5. Once you have selected an option press "OK" to confirm your choice. You will then be asked another question that you must also answer. Once everyone has answered this additional question you will go on to stage 3 of the experiment.

B.4 Instructions for Stage 3 (Coordination)

In this part of the experiment, you will be presented with two options (option A and option B) under 9 different situations. Your task will be to choose between option A and option B for each of the 9 situations. Each of these two options will give you the chance to obtain different prizes. However, each option has a maximum number of prizes for those who choose that option. The maximum number of prizes awarded under each option varies from one situation to another but in all the situations the sum of the prizes for A and B is equal to the number of people in this room. That is, in all the situations the number of prizes from A and the number of prizes from B add up to —.

If more people choose one option than there are prizes for that option in the situation in question, prizes will be awarded only to those subjects who have chosen the option who showed the highest ability levels in the task on which you worked in the stage 1 (measured as the mean score over the 14 repetitions). For example, if an option offers 5 prizes and 7 people choose it, only the 5 with the highest ability levels will obtain the prize.

The prize in option B is always be €1.00. However, the prize available in option A varies in each of the 9 situations.

At the end of the experiment, the computer will choose randomly one of the 9 situations and you will be paid what you have won in that situation given your choice and the choice of the rest of the participants in that situation.

An example is shown below. Please press“OK” to go through the example.

EXAMPLE: Assume that there are 10 people participating (so the total number of prizes between options A and B is always 10). In this case, the next screen will show you a matrix similar to the one below where the 9 different situations are presented:

	Prize in A: 1.50€ Prize in B: 1.00€	Prize in A: 2.00€ Prize in B: 1.00€	Prize in A: 3.00€ Prize in B: 1.00€
Number of Prizes in A : 2 Number of Prizes in B : 8	A B	A B	A B
Number of Prizes in A : 5 Number of Prizes in B : 5	A B	A B	A B
Number of Prizes in A : 8 Number of Prizes in B : 2	A B	A B	A B

The first row of the table presents 3 different situations in which there are always 2 prizes in A and 8 in B. However, in each of the three situations in the first row the prize in option A is different.

Similarly, the first column shows three situations in which always the prize of option A is €1.50 and that in option B €1.00. However, the maximum number of prizes in options A and B change in each of the situations in column 1.

(CONTROL_QUESTION_COOR) To check whether you have understood how your decisions and those of others determine your payments in this stage, we will now run through an example, at the end of which you will be asked to answer some questions.

Once everyone has answered these questions correctly you will move on to the screen where you have to take the decisions.

Example: Assume that there are 4 participants in this session (yourself and 3 others). Your ability (average score) in the task was 65 and in the randomly chosen repetition you obtained a score of 90. The average scores of the other 3 participants were 60, 70, and 80 respectively. Assume that you are making your choice in a situation in which option A offers 1 prize and option B offers 3. Based on this information, answer the following questions.

The 4 questions that you must answer are shown below. Once you have answered them, press "OK". If your answers are correct you will be moved on automatically to the decision-making screen as soon as all your partners have also answered the questions correctly.

1. If one of your partners had choose option A and the other two option B, will you win a prize if you choose option A?
   - Yes
   - No
   - **Depends**
2. If two of your partners choose option A and the other option B, will you win a prize if you choose option A?
   - Yes
   - **No**
   - Depends
3. If you all choose option A, who will win the prize? (choose as many as you think are correct)
   - Me
   - The one whose ability level is 60
   - The one whose ability level is 70
   - **The one whose ability level is 80**
4. If you all choose option B, who will win the prize? (choose as many as you think are correct)
   - **Me**
   - The one whose ability level is 60
   - **The one whose ability level is 70**
   - **The one whose ability level is 80**

(CONTROL_QUESTION_OUT) To check whether you have understood how your decisions and those of others determine your payments in this stage, we will now run through an example, at the end of which you will be asked to answer some questions.

Once everyone has answered these questions correctly you will move on to the screen where you have to take the decisions.

Example: Assume that there are 4 participants in this session (yourself and 3 others). Your ability (average score) in the task was 65 and in the randomly chosen repetition you obtained a score of 90. The average scores of the other 3 participants were 60, 70, and 80 respectively. Assume that you are making your choice in a situation in which option A offers 1 prize. Based on this information, answer the following questions.

The 4 questions that you must answer are shown below. Once you have answered them, press "OK". If your answers are correct you will be moved on automatically to the decision-making screen as soon as all your partners have also answered the questions correctly.

1. If one of your partners had choose option A and the other two option B, will you win a prize if you choose option A?
   - Yes
   - No
   - **Depends**
2. If two of your partners choose option A and the other option B, will you win a prize if you choose option A?
   - Yes
   - **No**
   - Depends
3. If you all choose option A, who will win the prize? (choose as many as you think are correct)
   - Me
   - The one whose ability level is 60
   - The one whose ability level is 70
   - **The one whose ability level is 80**
4. If you all choose option B, who will win the prize? (choose as many as you think are correct)
   - **Me**
   - **The one whose ability level is 60**
   - **The one whose ability level is 70**
   - **The one whose ability level is 80**

B.5 Instructions for Stage 3 (Belief elicitation)

Next you will be asked 5 questions regarding this session. At the end of the experiment the computer will chose one of these questions at random and you will be paid €1.5 if your answer is correct according to the data collected during the session and €0 otherwise.

Question 1: Consider the situation above in which option A offered 16 prizes of €2.00 each and option B 17 prizes of €1.00 each. In this situation you have chosen option —.
Given your choice and your partners' choices in this situation, in what range do you think the minimum ability required to win tournament A's prize lies?
Note: If you think that in this situation there are fewer participants than prizes in option A, you should answer "<70"

Available options: <70, 70-73, 73-76, 76-79, 79-82, 82-85, 85-88, 88-91, 91-94, 94-97 and 97-100

Question 2: Consider the situation above in which option A offered 16 prizes of €2.00 each and option B — prizes of €1.00 each. In this situation you have chosen option —.
Given your choice and your partners' choices in this situation, in what range do you think the minimum ability required to win tournament A's prize lies?
Note: If you think that in this situation there are fewer participants than prizes in option A, you should answer "<70"

Available options: <70, 70-73, 73-76, 76-79, 79-82, 82-85, 85-88, 88-91, 91-94, 94-97 and 97-100

Question 3: In this session, who do you think has performed better on average in the initial task?

Available options: Men/No Differences/Women

Question 4: In this session the average ability across all the — participants has been —. In what range do you think men's average ability level in the initial task lies?

Available options: <70, 70-73, 73-76, 76-79, 79-82, 82-85, 85-88, 88-91, 91-94, 94-97 and 97-100

Question 5: In this session the average ability across all the — participants has been —. In what range do you think women's average ability level in the initial task lies?

Available options: <70, 70-73, 73-76, 76-79, 79-82, 82-85, 85-88, 88-91, 91-94, 94-97 and 97-100

B.6 Instructions for Stage 4 (Ability estimation, single sex)

This part of the experiment is very similar to stage 3. The difference is that now you will interact only with persons of your own sex

On the next screen you will be asked again to estimate your ability level in the initial task (average over the 14 repetitions). To that end, you will be shown the distribution of abilities of all the participants of your own gender. That is, if you are a man all the other participants shown will also be men and if you are a woman they will all be women.

If your estimation matches your real ability you will get €1.00. Otherwise you will get €0.

B.7 Instructions for Stage 4 (Coordination, single sex)

In this part of the experiment you will again be presented with two options (A and B) under 9 different situations. The decision matrix that you will see will be similar to the one in stage 3A.

The difference from stage 3A is that now you will only interact with participants of your own gender, i.e. if you are a man all the other participants will also be men and if you are a woman they will be women.

In this session there are — participants of your own gender. Therefore, the total number of prizes between option A and option B will in all cases be —.

As in part 3A, the computer will randomly pick one of the situations and you will be paid according to your choice and the choices of those with whom you interact.

B.8 Instructions Stage 5 (Risk Elicitation)

On the next screen you will be presented with 8 different options, each of which offers two different quantities that you can win by choosing that option. In all the options, each outcome has a probability of 50%, i.e., the result of choosing an option depends exclusively on luck. At the end of the experiment the computer will randomly pick one result from the option you have chosen and you will be paid accordingly.

Below this text you will find the 8 available options. To see in more detail how to read this table, consider option 5. In this option the possible results are €0.7 and €2.7. Both are equally likely, which means that the computer will choose €0.7 as the payment on one of every 2 occasions and €2.7 the other.

You must choose one of the 8 possible options. To that end, an empty box will appear where you must enter the number of the option (from 1 to 8) that you want to choose.

	Probability 50%	Probability 50%
1	€1.5	€1.5
2	€1.3	€1.8
3	€1.1	€2.1
4	€0.9	€2.4
5	€0.7	€2.7
6	€0.6	€2.8
7	€0.4	€2.9
8	€0	€3

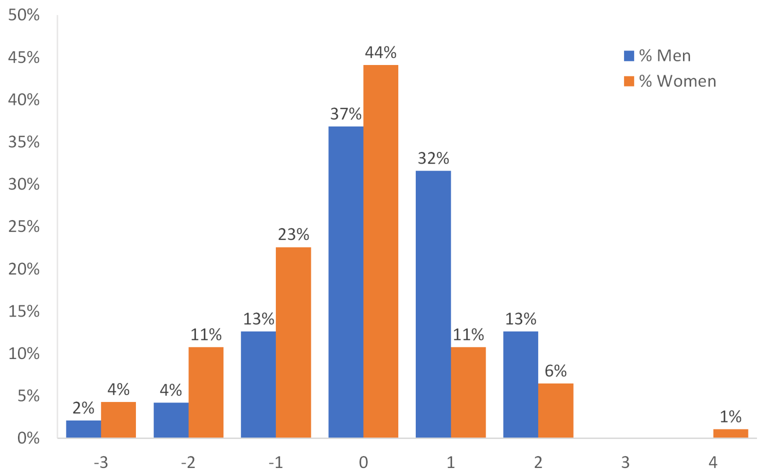
C Supplementary Analyses (For Online Publication)

TABLE C1–PROPORTION OF ENTRANTS IN TOURNAMENT A

	(δ_A, W_A) (0.25,1.5)	(0.5,1.5)	(0.75,1.5)	(0.25,2)	(0.5,2)	(0.75,2)	(0.25,3)	(0.5,3)	(0.75,3)
Session 1 (n=33)	0.33	0.63	0.85	0.56	0.78	0.85	0.59	0.85	0.89
Session 2 (n=34)	0.39	0.65	0.90	0.52	0.81	0.87	0.68	0.84	0.84
Session 3 (n=36)	0.33	0.64	0.88	0.42	0.82	0.91	0.58	0.94	0.97
Session 4 (n=37)	0.38	0.69	0.76	0.55	0.72	0.86	0.48	0.83	0.86
Session 5 (n=38)	0.37	0.76	0.95	0.42	0.84	0.84	0.50	0.89	1.00
Session 6 (n=39)	0.37	0.70	0.93	0.40	0.90	0.93	0.57	0.90	1.00
Total	0.36	0.68	0.88	0.48	0.81	0.88	0.57	0.88	0.93

Notes: Proportion of entrants in tournament A at each of the nine situations determined by the number of slots available (δ_A) and the prize offered (W_A) by tournament A. Notice that in each session there is a different number of subjects. The parameter δ_A should be interpreted as the number of prizes in tournament A per subject in the session such that coordination occur only when the entry rate in tournament A is δ_A.

FIGURE C1–DISTRIBUTION OF THE MALENESS VARIABLE WITHIN GENDERS



Note: The variable *Maleness* shows the difference between the reported averages in abilities for men and for women.

TABLE C2–ALTERNATIVE ANALYSES USING *Perception* AS INDEPENDENT VARIABLE

Panel A: *Perception* as continuous variable

Dep. Variable:	*Assessment*	*Prob(A)*	*Prob(W_A)*	*Mean Earnings*	*Prob(A)*	*Prob(W_A)*	*Mean Earnings*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Signal	6.163***	0.402***	1.339***	2.021***	-0.0436	0.955***	1.689***
	(0.974)	(0.109)	(0.158)	(0.314)	(0.111)	(0.169)	(0.379)
Female	0.399	0.0425	0.0985	0.266*	0.00701	0.0840	0.238
	(0.311)	(0.0488)	(0.0623)	(0.149)	(0.0444)	(0.0620)	(0.152)
Perception	0.460***	0.0983***	0.0895***	0.174**	0.0640**	0.0711**	0.150*
	(0.163)	(0.0268)	(0.0321)	(0.0764)	(0.0253)	(0.0330)	(0.0800)
Perception*Female	-0.389*	-0.0925**	-0.0531	-0.110	-0.0581*	-0.0387	-0.0864
	(0.228)	(0.0379)	(0.0483)	(0.112)	(0.0350)	(0.0482)	(0.115)
Assessment					0.0661***	0.0514***	0.0522
					(0.0115)	(0.0154)	(0.0318)
Constant	1.780**			-0.779**			-0.861***
	(0.851)			(0.307)			(0.318
Tournament Controls	–	YES	YES	–	YES	YES	–
Session FE	YES	YES	YES	YES	YES	YES	YES
Number of Clusters	–	188	188	188	188	188	188
Observations	188	1,692	1,692	188	1,692	1,692	188
R-squared	0.451			0.238			0.245

Panel B: *Perception* as Dummy Variable (Base Level: *Perception=Neutral*)

Dep. Variable:	*Assessment*	*Prob(A)*	*Prob(W_A)*	*Mean Earnings*	*Prob(A)*	*Prob(W_A)*	*Mean Earnings*
	(1)	(2)	(3)		(4)	(5)	
Signal	6.290***	0.394***	1.299***	1.935***	-0.0474	0.927***	1.627***
	(1.056)	(0.110)	(0.165)	(0.307)	(0.110)	(0.171)	(0.364)
Female	-0.0739	-0.0407	0.0690	0.207	-0.0336	0.0744	0.210
	(0.267)	(0.0492)	(0.0729)	(0.178)	(0.0446)	(0.0705)	(0.176)
Perception=Male	0.253	0.0871*	0.0356	0.0328	0.0661	0.0265	0.0205
	(0.224)	(0.0510)	(0.0723)	(0.173)	(0.0467)	(0.0707)	(0.171)
(Perception=Male)*Female	-0.375	-0.152***	-0.0477	-0.345*	-0.120**	-0.0343	-0.309
	(0.318)	(0.0529)	(0.0666)	(0.190)	(0.0504)	(0.0662)	(0.192)
Perception=Female	-0.740**	-0.110*	-0.155*	-0.0455	-0.0619	-0.127	-0.0271
	(0.353)	(0.0567)	(0.0796)	(0.153)	(0.0529)	(0.0786)	(0.156)
(Perception=Female)*Female	0.600	0.0397	0.0967	0.260	0.00133	0.0797	0.231
	(0.379)	(0.0529)	(0.0672)	(0.158)	(0.0486)	(0.0667)	(0.161)
Assessment					0.0661***	0.0500***	0.0489
					(0.0116)	(0.0153)	(0.0322)
Constant	2.305**			-0.405			-0.518
	(0.921)			(0.343)			(0.363)
Tournament Controls	–	YES	YES	–	YES	YES	–
Session FE	YES	YES	YES	YES	YES	YES	YES
Number of Clusters	–	188	188	–	188	188	–
Observations	188	1,692	1,692	188	1,692	1,692	188
R-squared	0.462			0.257			0.263

Notes: Perception is a variable taking value 0, 1, or 2 if the the subject perceived the task as Female affine, Neutral, or Male affine, respectively. Column (1) shows the OLS estimates for subject's self-assessment. Columns (2)–(5) show the marginal effects of the probit model for the probability of choosing tournament A and the probability of getting W_A. Robust standard errors for column (1) and clustered standard errors at subject level for columns (2)–(5) displayed in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE C3–CONSISTENCY BETWEEN MALENESS AND PERCEPTION

		Perception			
		Male	Neutral	Female	Total
Maleness	-4	0	0	0	0
	-3	0	0	6	6
	-2	0	3	14	17
	-1	2	10	33	45
	0	13	52	11	76
	1	40	9	1	50
	2	18	3	1	22
	3	0	0	0	0
	4	1	0	0	1
	Total	74	77	66	**217**

		Perception			
		Male	Neutral	Female	Total
Maleness	-4	0	0	0	0
	-3	0	0	6	6
	-2	0	0	14	14
	-1	0	0	33	33
	0	13	52	11	76
	1	40	0	0	40
	2	18	0	0	18
	3	0	0	0	0
	4	1	0	0	1
	Total	72	52	64	**188**

Notes: Relationship between *Maleness* (rows) and *Perception* (columns) of the task for the original sample (left) and for the final sample (right). Inconsistencies are displayed in red.

TABLE C4–PERCEPTIONS ABOUT SKILLS

Ability	Freq.	Percent	Average Maleness	95% Conf. Interval	
Attention to Details	8	10.39%	0.32	0.22	0.43
Observational Capacity	14	18.18%	0.48	0.37	0.59
Intuition	6	7.79%	0.92	0.77	1.07
Logic	2	2.60%	1.13	1.00	1.26
Visual Memory	22	28.57%	0.97	0.84	1.11
Precision	4	5.19%	0.84	0.69	1.00
Spatial Vision	21	27.27%	1.47	1.33	1.60
Total	77	100%			

Notes: Freq. (*Percent*) refers to the number (percentage) of subjects indicating that skill as the most important to perform well in the real effort task. *Average Maleness* shows the average perception of the skill with 0 being female biased, 1 gender neutral, and 2 male biased. The *95% Confidence Interval* refers to the *Average Maleness* variable.

TABLE C5–OLS FOR MALENESS ON SKILL MALENESS

	(1) Maleness
Skill Maleness	0.995***
	(0.339)
Constant	-1.126***
	(0.351)
Observations	68
R-squared	0.125

Notes: Skill Maleness is a continuous variable that refers to how male in average is perceived the ability stated by the subject as the most important to perform well in the real effort task. The lower (upper) bound of this variable is 0 (2) meaning that all subjects perceived the task as female (male) biased. *** p<0.01, ** p<0.05, * p<0.1.

TABLE C6–DETERMINANTS OF THE OPTIMAL BEHAVIOR

Dep. Variable:	*Prob(W_A)*	*Prob(W_A)*	*Prob(W_A)*	*Prob(W_A)*	*Prob(W_A)*	*Prob(W_A)*
Sample:	All	Men	Women	All	Men	Women
				Marginal	*Marginal*	*Marginal*
	(1)	(2)	(3)	(4)	(5)	(6)
Signal	1.352***	1.623***	1.150***	0.3398	0.6980*	0.0256
	(0.197)	(0.367)	(0.237)	(0.2297)	(0.356)	(1.003)
Female	0.0654			0.1041		
	(0.0489)			(0.0647)		
Maleness	0.0141	0.00888	0.00503	-0.0007	0.0005	0.0020
	(0.0322)	(0.0337)	(0.0308)	(0.0396)	(0.1115)	(0.0394)
Maleness∗Female	-0.0241			-0.0246		
	(0.0435)			(0.0548084)		
Tournament Controls	YES	YES	YES	YES	YES	YES
Session FE	YES	YES	YES	YES	YES	YES
Number of Clusters	188	95	93	188	95	93
Observations	1,692	855	837	846	423	423

Notes: Marginal effects for the probability of obtaining W_A unconditional on the behavior of others (probability that tournament A represents a dominant strategy). Interactions corrected using Norton et al. (2004). Clustered standard errors at subject level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE C7–ALTERNATIVE ANALYSES FOR ASSESSMENT

Sample:	*All*	*All+Inconsistent*	*Robust to Outliers*	*Not Manipulated*	*Manipulated*
	(1)	(2)	(3)	(4)	(5)
Signal	5.967***	6.230***	7.110***	3.614**	7.859***
	(1.136)	(0.973)	(0.499)	(1.740)	(1.050)
Female	0.0277	-0.0667	-0.0728	-0.0866	-0.0979
	(0.192)	(0.157)	(0.152)	(0.225)	(0.295)
Maleness	0.170*	0.240***	0.197**	0.307***	0.259
	(0.0870)	(0.0811)	(0.0937)	(0.115)	(0.196)
Maleness$*$Female	-0.268*	-0.295**	-0.227*	-0.426**	-0.243
	(0.143)	(0.124)	(0.127)	(0.201)	(0.216)
Risk Pref.	0.0473				
	(0.0508)				
Min.Ab.Win	0.108*				
	(0.0635)				
Taste Comp.	0.137				
	(0.0855)				
Age	0.0407				
	(0.0291)				
Difficulty	-0.194				
	(0.167)				
Extraversion	-0.0597				
	(0.0752)				
Agreeableness	0.175				
	(0.122)				
Conscientiousness	0.0188				
	(0.0840)				
Neuroticism	-0.0464				
	(0.0692)				
Openness	0.0892				
	(0.104)				
Social risk	0.203				
	(0.214)				
Constant	-0.799	2.316***	1.667***	5.419***	0.967
	(1.786)	(0.846)	(0.464)	(1.589)	(0.944)
Session FE	YES	YES	YES	YES	YES
Observations	188	217	188	100	88
R-squared	0.494	0.439	0.571	0.284	0.556

Notes: Inconsistent subjects are those whose state perception and maleness are not compatible. *Not Manipulated* subjects are those participating in sessions 1, 2, and 5. *Manipulated* are those participating in sessions 3, 4, and 6. Robust standard errors, except for column 3 that performs a regression analysis robust to outliers, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE C8–OLS FOR ASSESSMENT

Sample:	*All* (1)	*Men* (2)	*Women* (3)
Signal	6.416***	6.391***	6.451***
	(1.060)	(1.041)	(2.159)
Female	-0.0263		
	(0.175)		
Maleness	0.245***	0.248***	-0.0890
	(0.0874)	(0.0867)	(0.103)
Maleness∗Female	-0.325**		
	(0.142)		
Constant	2.160**	2.036**	2.203
	(0.912)	(0.864)	(1.938)
Session FE	YES	YES	YES
Observations	188	95	93
R-squared	0.445	0.521	0.363

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE C9–PROBIT MODEL FOR THE PROBABILITY OF ENTERING TOURNAMENT A

Sample:	*All* (1)	*Men* (2)	*Women* (3)	*All* (4)	*Men* (5)	*Women* (6)
Signal	0.401***	0.334**	0.488***	-0.0648	-0.232	0.0792
	(0.117)	(0.143)	(0.175)	(0.116)	(0.153)	(0.133)
Female	-0.0632**			-0.0587**		
	(0.0292)			(0.0269)		
Maleness	0.0580***	0.0555***	-0.0115	0.0389**	0.0307*	-0.00586
	(0.0197)	(0.0177)	(0.0183)	(0.0183)	(0.0164)	(0.0167)
Maleness∗Female	-0.0595**			-0.0355		
	(0.0282)			(0.0252)		
Assessment				0.0690***	0.0860***	0.0588***
				(0.0118)	(0.0140)	(0.0151)
Tournament FE	YES	YES	YES	YES	YES	YES
Session FE	YES	YES	YES	YES	YES	YES
Number of Clusters	188	95	93	188	95	93
Observations	1,692	855	837	1,692	855	837

Notes: Marginal effects are reported. Interactions corrected using Norton et al. (2004). Clustered standard errors at subject level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE C10–PROBIT MODEL FOR THE PROBABILITY OF WINNING A PRIZE

Dep. variable	Prob(W_A)	Prob(W_A)	Prob(W_A)	Prob(W_A)	Prob(W_A)	Prob(W_A)	Prob(W_B)	Prob(W_B)	Prob($\varnothing$)	Prob($\varnothing$)
Sample:	All	Men	Women	All	Men	Women	All	All	All	All
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Signal	1.364***	1.525***	1.259***	0.947***	0.983***	0.896***	-0.131	0.229**	-0.912***	-0.925***
	(0.160)	(0.283)	(0.196)	(0.172)	(0.328)	(0.200)	(0.0981)	(0.105)	(0.132)	(0.161)
Female	0.0253			0.0284			0.0611**	0.0573**	-0.0802*	-0.0800*
	(0.0384)			(0.0375)			(0.0262)	(0.0246)	(0.0429)	(0.0428)
Maleness	0.0530**	0.0496**	0.00109	0.0414*	0.0351	0.00358	-0.0456***	-0.0317**	-0.0123	-0.0128
	(0.0239)	(0.0234)	(0.0249)	(0.0241)	(0.0246)	(0.0245)	(0.0168)	(0.0162)	(0.0259)	(0.0262)
Maleness∗Female	-0.0594			-0.0457			0.0408	0.0242	0.0231	0.0237
	(0.0355)			(0.0262)			(0.0238)	(0.0235)	(0.0380)	(0.0386)
Assessment				0.0545***	0.0678***	0.0473**		-0.0503***		0.00180
				(0.0153)	(0.0254)	(0.0193)		(0.00849)		(0.0146)
Tournament Controls	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Session FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Number of Clusters	188	95	93	188	95	93	188	188	188	188
Observations	1,692	855	837	1,692	855	837	1,692	1,692	1,692	1,692

Notes: Columns (1)-(6) show the marginal effects for winning the prize in tournament A (*Prob(W_A)*). Columns (7) and (8) show the marginal effects for winning the prize in tournament B (*Prob(W_B)*). Columns (9) and (10) show the marginal effects for not winning any prizes (*Prob($\varnothing$)*). Interactions corrected using Norton et al. (2004). Clustered standard errors at subject level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE C11–ALTERNATIVE ANALYSES FOR THE PROBABILITY OF ENTERING TOURNAMENT A

Sample:	*All*	*Marginal*	*Nonmarginal*	*All+Inconsistent*	*Not Manipulated*	*Manipulated*	*All*
Method:	*Probit*	*Probit*	*Probit*	*Probit*	*Probit*	*Probit*	*OLS*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Signal	0.340***	0.196	0.222*	0.402***	0.373**	0.404***	0.442***
	(0.120)	(0.162)	(0.128)	(0.110)	(0.170)	(0.156)	(0.138)
Female	-0.0443	-0.0588*	-0.0216	-0.0833***	-0.0956**	-0.0156	-0.0610**
	(0.0296)	(0.0356)	(0.0384)	(0.0273)	(0.0387)	(0.0435)	(0.0305)
Maleness	0.0435**	0.0801***	0.0103	0.0552***	0.0468*	0.0886***	0.0543***
	(0.0193)	(0.0235)	(0.0213)	(0.0190)	(0.0262)	(0.0306)	(0.0202)
Maleness∗Female	-0.0397	-0.0711**	-0.00482	-0.0571**	-0.0494	-0.0981**	-0.0607**
	(0.0282)	(0.0319)	(0.0306)	(0.0259)	(0.0376)	(0.0422)	(0.0273)
Risk Pref	0.0173**	0.0145	0.0202*				
	(0.00814)	(0.00944)	(0.0105)				
Min.Ab.Win	0.00814	0.00888	0.00833				
	(0.0104)	(0.0111)	(0.0125)				
Taste Comp.	0.0225	0.00747	0.0377**				
	(0.0157)	(0.0204)	(0.0191)				
Age	0.00771*	0.00381	0.00872				
	(0.00434)	(0.00561)	(0.00536)				
Difficulty	-0.0380*	-0.0653**	-0.0104				
	(0.0227)	(0.0286)	(0.0270)				
Extroversion	-0.00715	0.00524	-0.0150				
	(0.0118)	(0.0156)	(0.0149)				
Agreeableness	0.0219	-0.00136	0.0501**				
	(0.0170)	(0.0203)	(0.0233)				
Conscientiousness	0.0216*	0.0308**	0.00690				
	(0.0125)	(0.0153)	(0.0151)				
Neuroticism	-0.00589	-0.00973	-0.00740				
	(0.0129)	(0.0155)	(0.0187)				
Openness	-0.0154	-0.0410**	0.0108				
	(0.0161)	(0.0200)	(0.0189)				
Social Risk	0.0765**	0.0952**	0.0601				
	(0.0331)	(0.0395)	(0.0394)				
Constant							-0.00310
							(0.133)
Tournament Controls	YES	YES	YES	YES	YES	YES	YES
Session FE	YES	YES	YES	YES	YES	YES	YES
Number of Clusters	188	188	188	217	100	88	188
Observations	1,692	846	846	1,953	900	792	1,692
R-squared							0.230

Notes: Inconsistent subjects are those whose state perception and maleness are not compatible. *Not Manipulated* subjects are those participating in sessions 1, 2, and 5. *Manipulated* are those participating in sessions 3, 4, and 6. Marginal effects are reported. Interactions corrected using Norton et al. (2004). Clustered standard errors at subject level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE C12–OLS FOR MEAN EARNINGS

Sample:	*All* (1)	*Men* (2)	*Women* (3)
Signal	2.052***	1.910***	2.205***
	(0.327)	(0.402)	(0.403)
Female	0.106		
	(0.0914)		
Maleness	0.0906	0.0946	-0.0138
	(0.0549)	(0.0574)	(0.0599)
Maleness∗Female	-0.125		
	(0.0784)		
Constant	-0.554*	-0.344	-0.655*
	(0.311)	(0.382)	(0.362)
Session FE	YES	YES	YES
Observations	188	95	93
R-squared	0.226	0.251	0.235

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE C13–PROBIT MODEL FOR *Prob(A)* AND *Prob*(W_A): MARGINAL *vs* NONMARGINAL

Dep. Variable: *Sample:*	*Prob*(W_A) All (1)	*Prob*(W_A) Men (2)	*Prob*(W_A) Women (3)	*Prob*(W_A) All *Marginal* (4)	*Prob*(W_A) Men *Marginal* (5)	*Prob*(W_A) Women *Marginal* (6)
Signal	0.298*	0.123	0.576*	1.688***	2.036***	1.441***
	(0.158)	(0.149)	(0.314)	(0.288)	(0.484)	(0.335)
Female	-0.0672*			0.0932*		
	(0.0350)			(0.0514)		
Maleness	0.0890***	0.0769***	-0.0103	0.0735**	0.0772**	-0.0274
	(0.0260)	(0.0213)	(0.0222)	(0.0360)	(0.0347)	(0.0297)
Maleness∗Female	-0.0833**			-0.1108**		
	(0.0387)			(0.0477)		
Tournament Controls	YES	YES	YES	YES	YES	YES
Session FE	YES	YES	YES	YES	YES	YES
Number of Clusters	188	95	93	188	95	93
Observations	846	423	423	846	423	423

Notes: Marginal effects are reported. Interactions corrected using Norton et al. (2004). Clustered standard errors at subject level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE C14– MAIN ANALYSIS RESTRICTING THE SAMPLE TO THOSE SUBJECTS FOR WHICH WE HAVE DATA ON THE *single-sex* SETTING

Dep. Variable:	*Assessment* (1)	*Prob(A)* (2)	*Prob(W_A)* (3)	*Mean Earnings* (4)	*Prob(A)* (5)	*Prob(W_A)* (6)	*Mean Earnings* (7)
Signal	7.311***	0.423***	1.290***	1.879***	-0.0989	0.936***	1.611***
	(0.975)	(0.141)	(0.176)	(0.341)	(0.146)	(0.228)	(0.487)
Female	-0.0424	-0.0419	0.0119	0.0562	-0.0378	0.0138	0.0578
	(0.184)	(0.0336)	(0.0429)	(0.101)	(0.0317)	(0.0423)	(0.102)
Maleness	0.234**	0.0608***	0.0519**	0.0793	0.0430**	0.0425	0.0708
	(0.0934)	(0.0213)	(0.0256)	(0.0579)	(0.0198)	(0.0262)	(0.0587)
Maleness*Female	-0.393***	-0.0768**	-0.0645*	-0.121	-0.0485	-0.0499	-0.106
	(0.147)	(0.0311)	(0.0371)	(0.0849)	(0.0288)	(0.0381)	(0.0884)
Assessment					0.0682***	0.0437**	0.0367
					(0.0149)	(0.0186)	(0.0398)
Constant	1.434*			-0.375			-0.428
	(0.849)			(0.315)			(0.308)
Tournament Controls	–	YES	YES	–	YES	YES	YES
Session FE	YES	YES	YES	YES	YES	YES	YES
Number of Clusters	–	150	150	–	150	150	–
Observations	150	1,350	1,350	150	1,350	1,350	150
R-squared	0.539			0.194			0.198

Notes: This table replicates Table 3 but restricting the sample to those subjects for which we have data on the *single-sex* setting (n=150). Column (1) shows the OLS estimates for the subjects' self-assessments. Columns (4) and (7) shows the OLS estimates for the subjects' mean earnings across all 9 tournaments. Columns (2)–(3) and (5)–(6) show the marginal effects of the probit model for the probability of choosing tournament A and the probability of getting W_A. Interaction corrected using Norton et al. (2004) for columns (2), (3), (5), and (6). Robust standard errors for columns (1), (4), and (7) and clustered standard errors at subject level for columns (2), (3), (5), and (6) are displayed in parentheses. *** p<0.01, ** p<0.05, * p<0.1.