



**UNIVERSITÀ
DEGLI STUDI
DI PADOVA**

PhD in Economics and Management
XXXII Cycle
(2016-2019)

Department of Economics
University of Padova

**Essays on Affirmative Action Policies and
Women's Chances of Success:
An Experimental Approach**

Presented by: José Javier Domínguez Ramírez

Supervised by: Prof. Antonio Nicolò

A mi familia

This thesis includes three chapters. The first chapter relies on the continuation of my master dissertation “Positive Discrimination in Childhood and Adolescence: Culture as a Determinant Factor of Individual Behavior”. In this project, we implemented an artefactual experiment in four different countries to observe i) how social preferences develop with age and ii) how the scheme of social preferences changes depending on the ethnicity of the individuals with whom the individuals are interacting. This is done using a range of mini-dictator games from which we classify 665 subjects into a variety of behavioural types. We expand on previous developmental studies of pro-sociality and parochialism by analysing individuals aged 9–67, and by employing a cross country study where participants from Spain interact with participants from different ethnic groups (Arab, East Asian, Black and White) belonging to different countries (Morocco, China, Senegal and Spain). We identify a ‘U-shaped’ relationship between age and egalitarianism that had previously gone unnoticed, and appeared linear. An inverse “U-shaped” relationship is found to be true for altruism. A gender differential is found to emerge in teenage years, with females becoming less altruistic but more egalitarian than males. In contrast to the majority of previous economic studies of the development of social preferences, we report evidence of increased altruism, and decreased egalitarianism and spite expressed towards black individuals from Senegal. Our findings highlight the importance of studying social preferences from both an early age and in later life. As social preferences can enhance efficiency in many workplace interactions, understanding how they develop over the life cycle is important for understanding how socialisation can impact preferences over outcomes. With the working population growing older, and workplaces becoming more diverse, understanding the interaction between social preferences, age and identity is therefore important for the design of institutions and their associated incentives in many societies.

The second and third chapters study the impact of Affirmative Action Policies on female candidates’ chances of success in the labor market. Affirmative action describes policies that support members of a disadvantaged group that has previously suffered discrimination in such areas as education, employment, or housing. Some countries use a quota system, whereby a certain percentage of government jobs, political positions, and school vacancies must be reserved for members of a certain group; an example of this is the reservation system in India. In some other regions where quotas are not used, minority group members are given preference or special consideration in selection processes. In the United States, affirmative action in employment and education has been the subject of legal and political controversy. In 2003, the Supreme Court of the United States maintained the prohibition on the use of quotas, considering race as a plus-factor when evaluating applicants. In other countries, such as the UK, affirmative action is rendered illegal because it does not treat all races equally. However, support for affirmative action has sought to achieve goals such as bridging inequalities and protect

minorities from the discrimination in the labor market and education. In the case of the labor market, despite the considerably growth of female employment derived from the use of Affirmative Action Policies (especially through gender quotas), the gender inequality in the labor market still persists. According to the “2019 Report on Equality between women and men in Europe” (European Commission, 2019), women earn, on average, 14,7% less per hour than men. The gender wage gap is especially higher for managerial positions (23%). Moreover, only the 26.7% of the boards and the 6.5% of the CEOs are female (EU, 2019). The gender gap in labor market outcomes suggests that Affirmative Action Policies could be ineffective in fostering career advancement. Therefore, the effectiveness of these policies should be revised.

In the second chapter, we aim to investigate how gender quotas affect the allocation of workers into different tasks within the organizations. We run a laboratory experiment in which we recruited 128 employers and 120 workers. Employers were asked to hire workers to conform a team of six workers and assign them to two different tasks, in terms of complexity and profitability. We found that gender quotas are useful in increasing the number of high-ability women assigned to the simpler and less profitable task in organizations but ineffective in improving women assignment to more complex and profitable tasks. In fact, gender quotas have a negative effect on the probability that high-ability women are assigned to the most complex and profitable tasks. We contribute to the existing literature studying the effectiveness of gender quotas in addressing the gender gap in organizations by including an analysis of task assignment decisions and highlighting how gender quotas may not be effective in breaking this mechanism at the basis of the documented gender gap in career progression and wages.

Finally, the third chapter provides experimental evidence of the effect of committees’ gender composition on female candidates’ probabilities of being recruited. The underrepresentation of women in the labor market has been attributed, among other factors, to the lack of women in recruitment committees. Therefore, committee quotas are becoming more widespread. To address this question, I designed a laboratory experiment in which groups of three subjects have to jointly select two candidates in a pool of six to perform a task. The results contradict the implicit assumption that more women in committees would automatically benefit female candidates. In fact, male-majority committees increased the probabilities of female candidates to be recruited compared to other committees’ compositions, while committees in which women are majority are the most detrimental for female candidates. The latter can be explained by the fact that men are more influential than women in female-majority groups and they disproportionately proposed to recruit two male candidates.

I would like to acknowledge
Loukas Balafoutas, Nagore Iriberry, Natalia Montinari, Antonio Nicolò, Ramón Cobo-Reyes,
Francisco Lagos, Juan Antonio Lacomba and Fernando García-Quero
for helpful comments and advices.

“Dejaré mi mitad oscura en duermevela
Y a mi otra mitad la haré dueña y señora de mis fiestas
Amaneceré como una nueva versión de humano
Para compensar a este cuerpo poco y mal usado”

Autoterapia, Autoterapia. IZAL (2018)

“Ser valiente
no es sólo cuestión de suerte”

Valiente, Un día en el mundo. Vetusta Morla (2008)

Contents

Chapter 1. The Development of Social Preferences.....	14
1. Introduction	14
2. Experimental design and procedure	17
2.1 Design	17
2.2 Procedure	18
3. Results	20
3.1. Social Preferences.....	20
3.2. Group Biases.....	25
4. Conclusion.....	28
References	30
Appendix A. Experimental Material	34
Appendix B. Statistical appendix	38
Appendix C. Robustness Checks	39
Chapter 2. Gender Quotas and Task Assignment in Organizations	45
1. Introduction	45
2. Experimental Design	48
2.1. General Overview.....	48
2.2. Experiment E: Part 3- Labor Market Game.....	49
2.3. Treatments	50
2.4. Procedure and Samples.....	51
3. Results	51
3.1. Is there a need for the introduction of a gender quota?	51
3.2. The effect of gender quotas on task assignment decisions	59
3.3. Employers' Performance.	61
References	64
Appendix A. Hard Task's problems, Screenshots and Instructions	68
Appendix B. Experiment W.	78
Appendix C. Additional results.....	82

Chapter 3. Committee Quotas and Gender Gap in Recruitment.....	93
1. Introduction	93
2. Experimental Design	96
2.1. Decision 1: Individual Decision and Beliefs.	97
2.2. Decision 2: Group Decision.	97
2.3. Rounds	98
2.4. Candidates	98
2.6. Procedure	99
3. Results	99
3.1. Individual Decisions: The role of evaluators' gender.	100
3.2. Group Decisions: The effect of the gender composition of the committees.	101
3.3. Mechanisms	104
3.4. Ex-post analysis	108
4. Conclusions	110
References	111
Appendix A. Additional Results	116
Appendix B. Instructions and The Implicit Association Test.	123

Chapter 1.

The Development of Social Preferences

Ramón Cobo–Reyes (*American University of Sharjah*), José J. Domínguez (*University of Padova*),
Fernando García–Quero (*University of Granada*), Brit Grosskopf (*University of Exeter*),
Juan A. Lacomba (*University of Granada*), Francisco Lagos (*Zayed University*), Tracy Xiao Liu
(*Tsinghua University*), Graeme Pearce (*University of Exeter*).

Keywords:

Social preferences; Children; Cross–country comparisons; Artefactual field experiment.

Published by Journal of Economic Behavior and Organization.

Printed Version: <https://www.sciencedirect.com/science/article/pii/S0167268119300289>

Article history: Received 25 April 2018; Revised 11 January 2019; Accepted 28 January 2019.

©2019 Published by Elsevier B.V.

1. Introduction

Human social interactions are strongly shaped by social preferences, with evidence from both the laboratory and the field suggesting that such preferences have implications in a range of settings. Previous research shows that prosocial preferences influence outcomes in social dilemmas (Fischbacher & Gächter, 2010), charitable giving (DellaVigna et al., 2012; Falk, 2007) and could even play a role in labour markets and natural competitive market places (Bellemare & Shearer, 2009; Grosskopf & Pearce, 2016; Kube et al., 2012, 2013), affecting welfare distributions and market efficiency (Dufwenberg et al., 2011)¹².

There is also evidence that individuals' concern for others depends on the identity of the person with whom they are interacting (Akerlof & Kranton, 2000; Chen & Li, 2009). For example, there is evidence that subjects behave more charitably (Chen & Li, 2009), cooperatively (Brañas–Garza et al., 2006; Chen et al., 2014; Drouvelis & Nosenzo, 2013) and coordinate more efficiently (Chen & Chen, 2011)

¹ See Camerer & Fehr (2004) and Cooper & Kagel (2009) for comprehensive reviews of the laboratory literature.

² We note that there is mixed evidence on reciprocity in the field. See for example Gneezy & List (2006) and List (2006).

when interacting with the ‘in–group’, i.e. someone they identify with, in comparison to the ‘out–group’. Findings from natural field experiments corroborate these results, with evidence showing that individuals condition their other–regard on the ethnicity of the person they are interacting with (Grosskopf & Pearce, 2016; Mujcic & Frijters, 2013). Bernhard et al. (2006) refer to these types of group biases as parochialism. As it has been argued that social preferences are a ‘fundamental cornerstone’ of humans’ ability to cooperate with genetic strangers (Fehr et al., 2013), understanding the extent to which they are contingent on the ethnicity of others and how this dependency develops, is crucial for the design of institutions and their associated incentives in increasingly diverse societies. Using a unified framework of mini–dictator games, Fehr et al. (2008) examine how altruism, egalitarianism and spite emerge alongside parochialism in children aged 3 to 8 years old. Fehr et al. (2013) expand on this by investigating these behaviours in 8 to 17 year olds using the same experimental design. Both Fehr et al. (2008) and Fehr et al. (2013) examine parochialism using small, ‘interpersonal social groups’ (Brewer & Gardner, 1996) by varying the school from which the person receiving the money in the dictator games is selected. The receiver is either from the same school as the dictator (the ‘in–group’), or a different school (the ‘out–group’). Both studies report evidence of in–group favouritism that increases with age, with subjects behaving more altruistic and less spiteful towards in–group members in comparison to out–group members. However, defining the in–group in this way introduces a potential confound stemming from repeated interactions that could be present when people interact with those that they may be able recognise (List, 2006). Therefore, it may not be surprising that in–group favouritism is found to increase with age. It may be that each additional year of schooling increases a child’s experience with the same peers, potentially resulting in long–term relationships with peers from the same school. In addition, relatively little is known about how these behaviours develop with age later in the life cycle, especially in adulthood. As perceived senses of identity can act as mechanisms that promote, or impede, coordination and cooperation in the workplace, understanding the interaction between age, social preferences and identity, is increasingly important, especially as the working population grows older and becomes more diverse. Our study addresses both of these concerns.

The purpose of this paper is twofold. First, we study how social preferences develop with age, with an extension to include the relatively understudied subject pool of adults. Second, we examine whether and how group biases are manifested in the behaviour of different age groups. This is done using an artefactual field experiment conducted in Granada, Spain, with 665 subjects aged 9 to 67. Utilising mini–dictator games we exploit the unifying framework of Fehr et al. (2008) and Fehr et al. (2013) and categorise subjects from four distinct age groups, Children (aged 9 to 11), Teenagers (aged 15 to 18), Students (aged 18 to 28) and Adults (aged 31 to 67), into one of three behavioural types: altruistic,

egalitarian and spiteful. Following Fehr et al. (2008) and Fehr et al. (2013), dictators were shown a photo of a group of receivers, one of whom they would be matched with at random. We examine group biases by varying the ethnicity of the individuals in the photos: Arab (Morocco), Black (Senegal), East-Asian (China) and White (Spain). We refer to interactions between dictators and White receivers as in-group interactions, and interactions between dictators and receivers from other ethnic groups as out-group interactions. From the perspective of Fehr et al. (2008, 2013), all our interactions could be viewed as out-group, with the out-groups differing in geographical, cultural and economic distance. However, we refer to in-group interactions as those in which individuals share ethnic appearance characteristics, but in which individuals are still strangers. This follows what Brewer & Gardner (1996) call collective rather than interpersonal identity, and addresses the problem of potential repeated interaction effects.

We report a number of observations. First, we observe that for Children, Teenagers and Students, egalitarianism diminishes as they grow older, while altruism increases. In contrast, we find that for Adults, egalitarianism becomes more prominent with age whilst altruism diminishes. Second, following the analysis of Fehr et al. (2008), we report gender differences in egalitarianism emerging in Children and persisting through to adulthood, whilst this differential emerges later for altruistic types, being found in Teenagers. Finally, we report no evidence of favouritism towards the White in-group receivers. Instead, we find that all age groups are more likely to be altruistic when the receiver is Black, except the Adults who do not differentiate based on the receivers' ethnicity. For this age group, we report no evidence of group-contingent behavioural types (Chen & Li, 2009). This paper contributes to the literature in several ways. First, through the inclusion of Adults, we are able to identify two relationships that had previously gone unnoticed, and appeared linear: a 'U-shaped' relationship between age and egalitarianism; and an inverse 'U-shaped' relationship between age and altruism. This complements previous work by List (2004), who finds that contributions in a one shot public goods game are increasing with age. It also provides an informative comparison with House et al. (2013), who report a U-shaped relationship between age and egalitarian choices appearing at a much younger age, i.e. in children aged 3 to 14. While these results seem different to ours, they are obtained in experimental settings where children mostly knew one another, and were incentivised with candy. Second, while previous findings mostly stemming from ultimatum and dictator games have shown that females are more generous than men, we observe that females become more egalitarian with age. Third, our use of a broader sense of identity that considers a larger geographical area to be regarded as the in-group, complements the previous research that studies a narrower sense of identity. In doing so we overcome a potential repeated interaction confound that may be present in previous work.

The remainder of this paper is organised as follows. Section 2 gives details of the experimental design and procedure. Section 3 outlines and discusses the results and Section 4 concludes.

2. Experimental design and procedure

The experiment is designed to examine how social preferences and group biases develop with age. This is done using an artefactual field experiment in which we examine behaviour in a range of mini–dictator games. To examine group biases we use a between–subject design in which we exogenously vary the ethnicity of the receivers with whom the dictators are matched.

2.1 Design

The experiment draws dictators from four distinct age groups: Children aged 9 to 11, Teenagers aged 15 to 18, Students aged 18 to 28 and Adults aged 31 to 67³. The Children and Teenagers were recruited from private, coeducational schools in Granada, Spain⁴. Students were recruited from the experimental subject pool of the University of Granada, and Adults were recruited from the professional staff at the University of Granada⁵.

Table 1. Experimental Design Summary

Subjects' Age	Range	Receiver's ethnicity				Total
		White	Arab	E.Asian	Black	
Children	9-11	51	47	47	33	178
Teenagers	15-18	52	54	48	49	203
Students	18-28	39	45	50	50	184
Adults	31-67	29	26	23	22	100
Total		171	172	168	154	665

Variation in the ethnic identity of the receiver was achieved by showing the dictators a photo of a group of people, one of whom they would be matched with and would act as a receiver to their choices for the duration of the experiment. The ethnicity of the people in the photo was varied by their country of residence: Arab (Morocco), Black (Senegal), East–Asian (China) or White (Spain). Dictators were

³ Although our results are robust to the inclusion of additional control variables, we acknowledge that age has not been randomly assigned in our experiment.

⁴ Consent was obtained from the children, the children's parents and the participating schools.

⁵ To ensure the comparability of subjects of different age groups, we endeavoured to recruit from populations that had been educated in similar institutions. Of our sample, 66% of the Students and 74% of Adults attended a similar primary school to our sample of Children. Further, 81% of Students and 81% of Adults attended a secondary school similar to our Teenagers. 57% of Adults has obtained a university degree.

not informed about the particular country in which the photo was taken, but were told that the recipients were from a foreign country⁶. The receivers in the photos were always strangers and from the same age group as the dictators⁷. The photos contained both males and females.

We selected receivers from these particular countries for a number of reasons. As all our dictators were recruited from the University of Granada, receivers from Spain were selected to serve as a natural ‘in-group’ comparison. The other countries were selected in order to vary the ethnicity of the receivers, and thus the extent to which the dictators may perceive them as out-group. Appearance differences were apparent from the photos, with the receivers from Spain looking most similar to the dictators. Receivers from the other countries differed in appearance to the dictators in their skin tone, hair colour, and facial features. Table 1 presents the number of observations obtained from each treatment for each age group. All experimental materials are given in Appendix A.

2.2 Procedure

Subjects played as dictator in three mini-dictator games taken from Fehr et al. (2008): The Pro-social Game, the Envy Game and the Sharing Game. As the experiments were conducted in a similar manner to the majority of studies that conduct dictator games, our methodology is comparable to the previous literature (see Engel (2011) for a recent meta-analysis of previous dictator game studies). Figure 1 displays the three experimental games graphically. In each game, subjects had to choose between two possible actions: left and right. The action left always resulted in an egalitarian allocation of (5,5) - 5 points for the dictator and 5 points for the receiver. The allocation resulting from right is systematically varied between games and the order in which the games were completed was randomised.

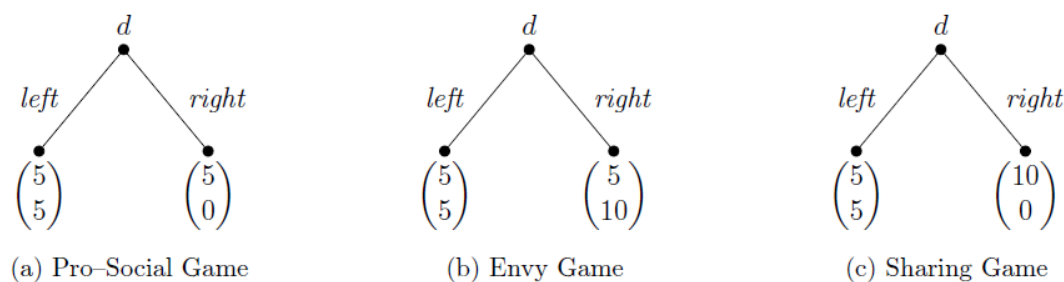


Figure 1. Dictator Games

⁶ Dictators were not told this when the receiver was from Spain.

⁷ To ensure that no dictator potentially knew a receiver, Children and Teenagers interacted with their counterparts from different schools, Students interacted with other students from different subject areas and year groups, and Adults interacted with staff from different colleges.

In the Pro-social Game, the action right results in an allocation of (5,0) - 5 points for the dictator and 0 points for the receiver. This game allows the dictator to avoid advantageous inequality without incurring a cost, and serves to measure the dictator's willingness to avoid it. Choosing left could stem from a preference to avoid inequalities (Fehr & Schmidt, 1999), from efficiency concerns (Charness & Rabin, 2002) or from the desire to maximize the minimum payoff. A self-interested individual is indifferent between either choice. In the Envy Game the action right produces an allocation of (5,10), and serves to provide a measure of the dictator's willingness to costlessly lower the receiver's payoff, reducing disadvantageous inequality. In the Sharing Game, a choice of right produces an allocation of (10,0). Choosing the egalitarian choice in the Sharing Game is costly for the dictator, in contrast to the Pro-social Game, which would show a strong form of inequality aversion.

These games were chosen because the actions taken in each game, when considered collectively, can be used to determine the motives underpinning each dictator's decisions. Following the classifications of Fehr et al. (2008, 2013), each dictator can be categorised as an altruistic, egalitarian or spiteful behavioural type, with strong and weak sub-types, depending on the dictators' choice pattern. Table 2 outlines these classifications in detail. We acknowledge that, as with the studies of Fehr et al. (2008, 2013), a perfectly selfish individual would randomise between left and right in the Pro-social and Envy Games, but select right in the Sharing game, and thus may appear as Weakly Altruistic, Weakly Egalitarian or Spiteful. To address this, we have examined the data to see if we observe similar proportions of Weakly Altruistic, Weakly Egalitarian and Spiteful types. Formally testing this, we can reject the null hypothesis that these proportions are equal ($p < 0.001$, χ^2 Test, $d.f=2$).

In each game, 1 point corresponds to €1, an exchange rate that was employed for all age groups.

We did not want to introduce a potential confound by varying the incentives by age (toys, stickers or sweets as incentives for the Children, and money for the other age groups) as there is evidence that non-monetary incentives result in significantly more pro-social behaviours (see Fehr et al. (2008) versus Fehr et al. (2013)). A similar result is also observed by Moore (2009). This is likely a consequence of sharing norms associated with food and sweets, or different responses to different reward types (House & Tomasello, 2018)⁸.

⁸ Although asking subjects to make multiple decisions may induce moral balancing or licensing, this should be present across all age groups and in all treatments, and as such is independent of age and the in/out-group manipulation.

Table 2. Behavioural Types

Behavioural type	Pro-social	Envy	Sharing
Strongly Egalitarian	(5,5)	(5,5)	(5,5)
Weakly egalitarian	(5,5)	(5,5)	(10,0)
Strongly Altruistic	(5,5)	(5,10)	(5,5)
Weakly Altruistic	(5,5)	(5,10)	(10,0)
Spiteful	(5,0)	(5,5)	(10,0)

Note: Behavioural types are taken from Fehr et al. (2008)

3. Results

In this section, we outline the experimental results. A number of common features are present throughout. Where non-parametric tests are utilised, both the p-value and test used are presented in parentheses. All tests are two sided, unless otherwise stated. All parametric support is obtained from marginal effects estimated from Probit regressions. Tables presenting full regressions are given in Appendix B. We present the results relating to social preferences in Section 3.1 and analyse group biases in Section 3.2.

3.1. Social Preferences

The analysis focuses on each subject's choice pattern across the three games, rather than considering the subjects' choices from each game separately. This enables us to interpret each subject's behaviour within the Fehr et al. (2013) framework, and keeps the analysis concise. We first categorise subjects into each of the behavioural types, as specified in Table 2. Figure 2a presents the distribution of these types, showing the percentage of subjects categorised for each age group. Pooling the weak and strong subtypes, Figure 2b plots the percentage of subjects categorised into three broad categories.

Table 3 presents the estimates of marginal effects from Probit regressions, where in each regression the dependent variable is a dummy that takes a value of 1 (and 0 otherwise) if the subject has been classified into one of the behavioural types - egalitarian in regression (i), altruistic in (ii) and spiteful in (iii). In each regression we include the following variables: the subjects' age in years, the subjects' age in years squared, a dummy variable that takes a value of 1 if the subject is female, the interaction between the subjects' gender and their age, and a dummy that takes a value of 1 if the interaction is in-group (i.e. the receiver is from Spain). We include age squared in the regressions in order to capture any non-linear effects associated with age. The interaction between the subjects' gender and their age is included to account for gender differences that might emerge over time. Finally, we include an in-

group dummy in order to account for any in–group/out–group effects that the literature has previously found to be important⁹.

Table 3. Marginal Effect - Determinants of Behavioural Type

<i>Dependent Variable:</i>	<i>Egalitarian Type</i> (i)	<i>Altruistic Type</i> (ii)	<i>Spiteful Type</i> (iii)
<i>Marginal Effect of Age:</i>			
<i>Children</i>	-0.037*** (0.003)	0.033*** (0.002)	-0.001 (0.003)
<i>Teenagers</i>	-0.034*** (0.004)	0.037*** (0.004)	-0.001 (0.002)
<i>Students</i>	-0.025*** (0.003)	0.029*** (0.003)	-0.001 (0.001)
<i>Adults</i>	0.025*** (0.005)	-0.027*** (0.005)	-0.001 (0.001)
<i>Marginal Effect of Female:</i>			
<i>Children</i>	0.096** (0.044)	-0.045 (0.037)	-0.054* (0.032)
<i>Teenagers</i>	0.132*** (0.043)	-0.104** (0.044)	-0.017 (0.022)
<i>Students</i>	0.136*** (0.039)	-0.132*** (0.042)	0.005 (0.022)
<i>Adults</i>	0.186* (0.096)	-0.285*** (0.097)	0.077** (0.039)

Note: ***, **, * denote significance at the 1%, 5% and 10% levels. Standard errors in parentheses. All reported estimates are from Probit regressions. The marginal effects of *Age* and *Female* are calculated for each regression, and are evaluated at the mean of each age group. *Age* is the reported age of the subject, and treated as a continuous variable. The results remain quantitatively similar if additional control variables are included. The number of observations differ to those reported in Table 1 due to missing entries. The observations from sixteen subjects are dropped, as we were unable to categorise them into either one of the three behavioural types. Full regressions given in Table 6, Appendix B.

Following Fehr et al. (2008), we focus the analysis on the marginal effect of the variable *Age*, which is the subjects' age in years, and then on the marginal effect of *Female*, the dummy variable that takes a value of 1 if the subject is female. We estimate the marginal effect of both these variables for each age group. The age brackets themselves are not included in the regressions. The coefficient estimates are therefore not the level effect of being in one of these age brackets, but the estimated effect of increasing age by one year (or being female), at the empirical mean of the respective age group. Figure 3a and Figure 3b plot the estimated marginal effects of age and female graphically.

⁹ All our estimates are robust to the inclusion of additional control variables. These estimates can be found in Table VII, Appendix B.

In addition to following the methodology of Fehr et al. (2008), we also estimate a multinomial logit model to examine the robustness of our results. We estimate the model using the same explanatory variables as those used in the Probit models, but allow for three unordered outcomes rather than using a binary dependent variable. The estimates of log-odds of age, age squared and the gender dummy are presented in Table 4.

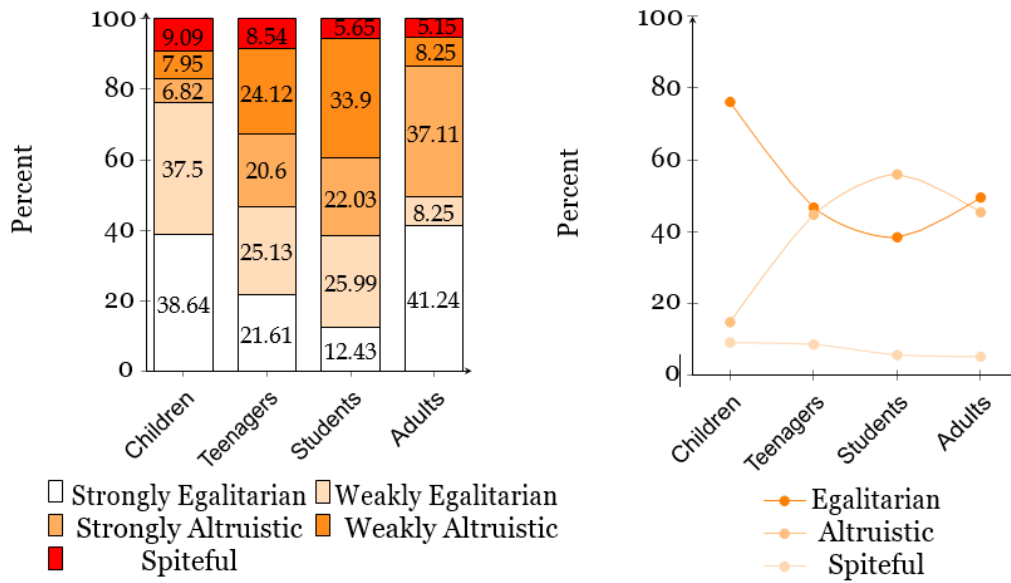
Observation 1. (Development of Behavioural Types) There is a non-linear relationship between egalitarianism and age. Age reduces egalitarianism in Children, Teenagers and Students, but increases egalitarianism in Adults. The inverse holds for altruism.

Support. Combining the Weak and Strong subtypes from Figure 2a, the percentage of egalitarian types is highest for Children (76.1%). Initially, this percentage falls with the subjects' age, being smaller in Teenagers and Students. However, it then increases in Adults. This indicates a 'U-Shaped' relationship, as observed in Figure 2b. In contrast, an inverse 'U-Shape' holds for altruistic types, increasing from Children to Teenagers, peaking for Students, before falling in Adults. Figure 2a also reveals that the majority of Adults' preference types (78.3%) can be categorised as Strong. This compares to just 45.4% of Children, 42.2% of Teenagers and 34.4% of Students.

Table 4. Multinomial Logit Estimates - Determinants of Behavioural Type

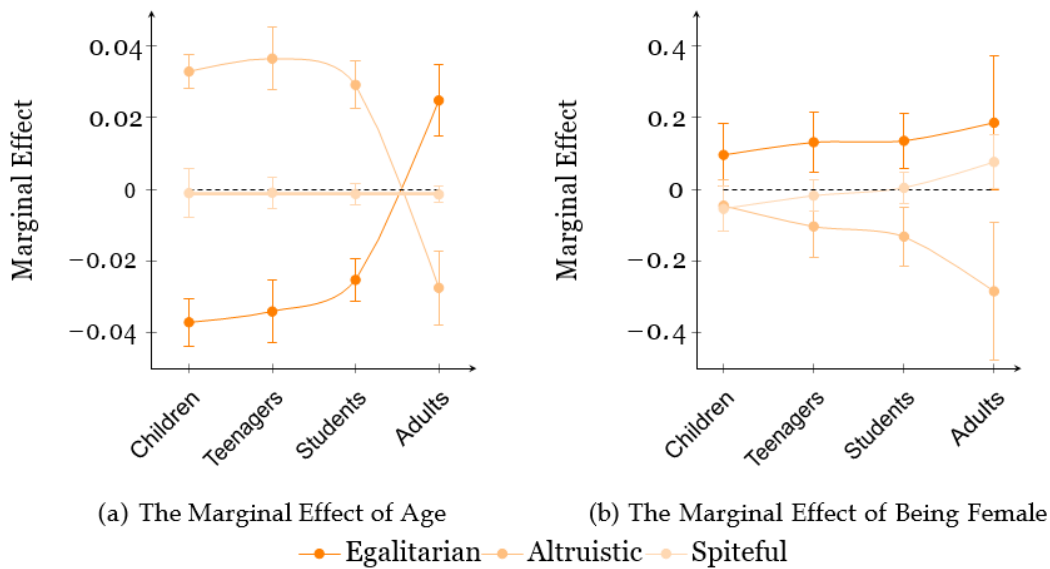
<i>Behavioural Type</i>	<i>Explanatory Variable</i>	<i>Coefficient Estimate</i>	<i>Standard Error</i>
Altruistic	<i>Age</i>	0.329***	(0.042)
	<i>Age2</i>	0.07	(0.064)
	<i>Female</i>	-0.005***	(0.001)
Spiteful	<i>Age</i>	-0.002**	(0.001)
	<i>Age2</i>	-0.19	(0.373)
	<i>Female</i>	-1.548**	(0.805)

Note: ***, **, * denote significance at the 1%, 5% and 10% levels. Standard errors in parentheses. The results remain quantitatively similar if additional control variables are included. The observations from sixteen subjects are dropped, as we were unable to categorise them into either one of the three behavioural types. Egalitarian Types are taken as the baseline. Included variables are identical to those in Table 6.



(a) Distributions of Behavioural Types (b) Percentage of Behavioural Types
Note: For Figure 2a, we were unable to classify sixteen subjects into one of the five behavioural types. Figure 2b plots the combined percentage of weak and strong classifications for each behavioural type for each age group.

Figure 2: Behavioural Types by Age



(a) The Marginal Effect of Age (b) The Marginal Effect of Being Female
 —●— Egalitarian —●— Altruistic —●— Spiteful
Note: The x axis plots the age group at which the marginal effect is evaluated at. In Figure 3a, the y axis plots the marginal effect of age on the probability of being classified into each of the behavioural types. In Figure 3b, the y axis plots the effect of *Female* on the probability of being classified into each of the behavioural types. Marginal effects estimates are given in Table 3. Vertical bars represent 95% confidence intervals.

Figure 3: Behavioural Types - Marginal Effects

Examining the relationship parametrically, Table 3 shows the estimated marginal effect of age on the probability of being classified as one of the behavioural type for each age group. It outlines how age has a negative effect on the probability that a dictator is categorised as an egalitarian type for Children, Teenagers and Students, but has a positive effect for Adults. This is shown graphically in Figure 3. The log-odds estimates in Table 4 corroborate the marginal effect estimates from the Probit models.

Observation 2. (Gender Differences) Females are more likely to be classified as egalitarian, and are less likely to be classified as altruistic, than males. Adult females are more likely to be classified as a spiteful type than Adult males.

Support. Table 3 highlights the significant positive marginal effect of being female on being classified as an egalitarian type for all age groups ($p < 0.05$ for Children, $p < 0.01$ for Teenagers and Students, and $p < 0.1$ for Adults). Table 3 also shows that a negative female effect on altruism emerges in Teenagers and persists into the Adults ($p < 0.01$ for all age groups except Children). There is a weak and small negative female effect in spiteful types for Children ($p < 0.1$) and a positive effect in Adults ($p < 0.05$).

Observation 1 highlights how age negatively impacts egalitarianism, but positively impacts altruism, in Children, Teenagers and Students. This replicates the previous work of Fehr et al. (2013), who report identical results in children aged 8–17. However, the inverse is true for Adults, and through the inclusion of this age group, we are able to identify both a ‘U-shaped’ relationship between age and the proportion of egalitarian types, and an inverted ‘U-shaped’ relationship for the proportion of altruistic types, that had previously gone unnoticed and appeared linear.

We find no evidence that age reduces spitefulness, as found by Fehr et al. (2013). This is likely due to the fact that we observe a significantly smaller proportion of Children to be spiteful types (9.1% compared to 30%) and in line with the findings of Fehr et al. (2013), we find no gender differential in these types. In contrast to Fehr et al. (2013), and highlighted by Observation 2, we observe a later onset of gender differences in altruistic types, as we report the differential emerging in Teenagers rather than in Children. One potential explanation for the observed behaviour is that expectations about what one ought to do (i.e., the injunctive norm) differ across age groups (see House (2018) for a recent discussion of how social norms affect prosocial behaviour). For example, it may be that the 50:50 split is the taught norm in very young children which weakens with age. However, by the time individuals reach adulthood, both egalitarian as well as altruistic behaviour could be seen as normative. An interesting

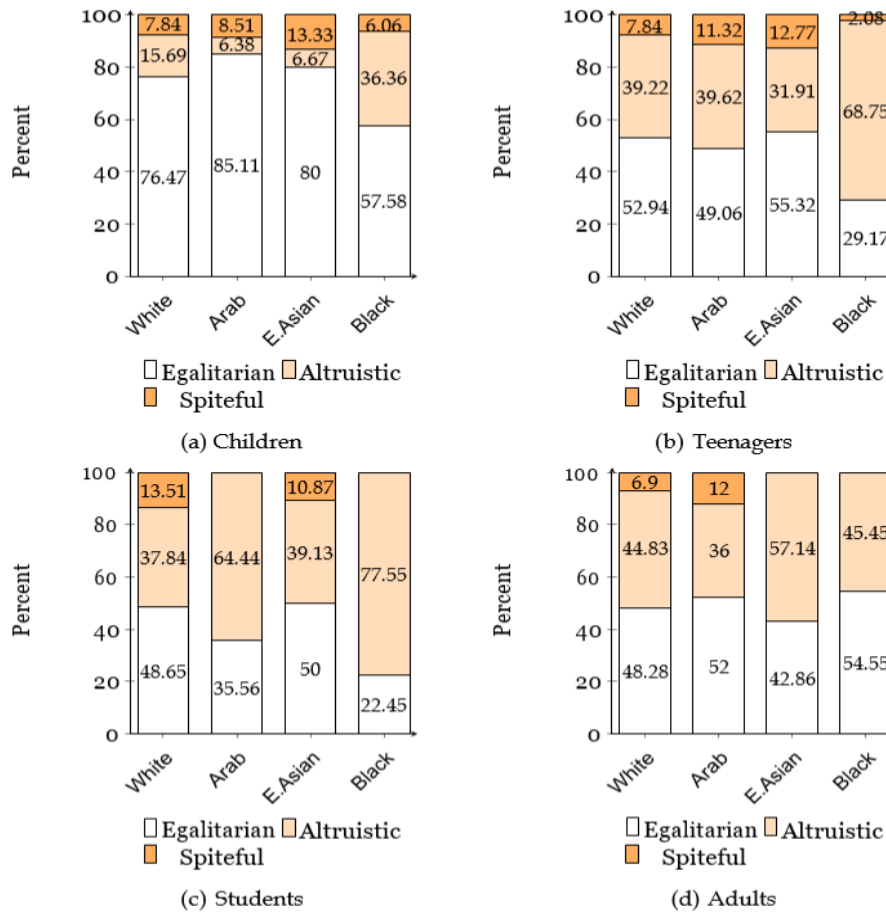
avenue for further research that would solidify the U–shape relationship between age and egalitarianism is to investigate the pro–social behaviour of the elderly.

It is interesting to note that, through the use of the particular constellation of games that we study, i.e. the inclusion of a test of disadvantageous inequality (Envy Game) in addition to a test of advantageous inequality (Sharing Game), females are found to be increasingly egalitarian with age. This contrasts with the conventional finding from standard dictator games, where women are found to be more generous, i.e. they give more than men (see Croson & Gneezy (2009) for an overview of the gender differences literature, and Engel (2011) for a meta-analysis of dictator game results).

3.2. Group Biases

To examine if dictators condition their behaviour on the receivers' ethnicity, we conduct pairwise comparisons of the behavioural patterns of dictators matched to each of the four ethnicities: White, Arab, Black and East–Asian. As in the previous section, we focus on dictators' choice patterns across the three games, rather than considering each game individually. Figure 4 presents the distributions of the three broad behavioural types for each of the four ethnicities we study, by age group. This allows for simple within group comparisons.

Table 5 presents the estimates of marginal effects from Probit regressions, where in each regression the dependent variable is a dummy that takes a value of 1 if the subject has been classified into one of the behavioural types. In each regression, presented in Table 9, Appendix B, we include the same variables as those outlined in Section 3.1, along with three additional dummies, Arab, Black, East Asian, that take values of 1 and 0 otherwise for each of the three ethnicities we examine; White is taken as the baseline. We further include the interaction of these dummies with age. These are included in order to identify any ethnicity effects, and how these effects might develop with age. From each regression we estimate the marginal effect of the ethnicity variables, Arab, East–Asian and Black, on the probability of being classified into each behavioural type, for each age group. As outlined above, White observations are taken as the baseline. The estimates are presented graphically in Figure 5.



Note: We were unable to classify sixteen subjects into one of the behavioural types.
 Figure 4: Behavioural Types by Age Group and the Receivers' Ethnicity

Observation 3. (Group Dependent Behavioural Types) Children, Teenagers and Students are least likely to be an egalitarian and spiteful type, but most likely to be an altruistic type, when the receiver is Black. The Adults' behavioural type is unaffected by the receivers' ethnicity.

Support. Figure 4 highlights how, for all age groups the distribution of types is relatively stable across all ethnicities. The only notable exception is when the receiver is Black: for all age groups, except the Adults, the percentage of subjects classified as being egalitarian is smallest when the receiver is Black. The inverse is true for altruism, with altruists being the most prevalent type when the receiver is Black for all age groups except the Adults.

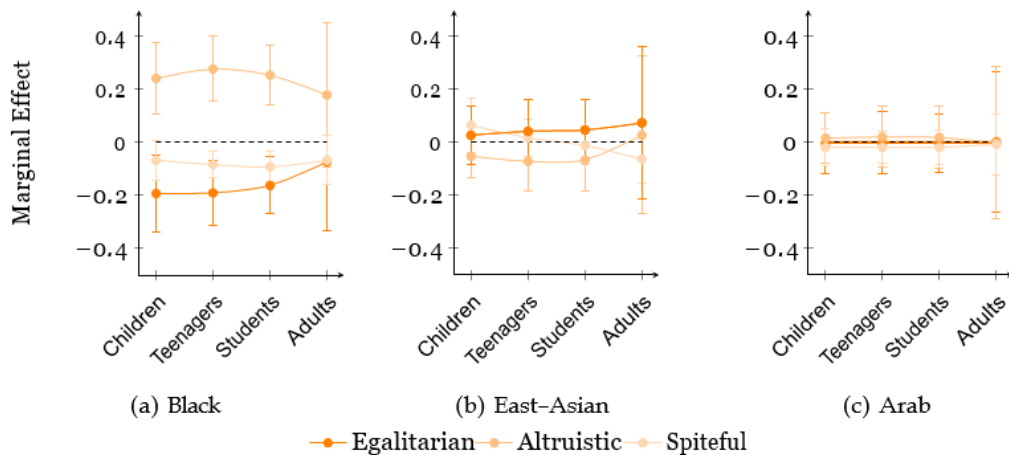
The estimates in Table 5 support this observation formally: when the receiver is Black, the Children, Teenagers and Students are less likely to be egalitarian, more likely to be altruistic in comparison to when the receiver is White. The receivers' ethnicity has no impact on the any of the marginal effects

for the Adults. The marginal effects of all other ethnicities are estimated not to be significant for all age groups. The marginal effects are shown graphically in Figure 5.

Table 5. Marginal Effects - Identity and Behavioural Type

<i>Dependent Variable:</i>	<i>Egalitarian Type</i>			<i>Altruistic Type</i>			<i>Spiteful Type</i>		
<i>Marginal Effect:</i>	<i>Arab</i>	<i>E.Asian</i>	<i>Black</i>	<i>Arab</i>	<i>E.Asian</i>	<i>Black</i>	<i>Arab</i>	<i>E.Asian</i>	<i>Black</i>
<i>Children</i>	-0.003 (0.058)	0.027 (0.057)	-0.192*** (0.073)	0.016 (0.048)	-0.052 (0.042)	0.242*** (0.07)	-0.019 (0.035)	0.065 (0.052)	-0.066* (0.038)
<i>Teenagers</i>	-0.003 (0.06)	0.042 (0.06)	-0.189*** (0.063)	0.021 (0.059)	-0.071 (0.057)	0.278*** (0.063)	-0.019 (0.032)	0.018 (0.035)	-0.082*** (0.026)
<i>Students</i>	-0.002 (0.056)	0.047 (0.057)	-0.161*** (0.055)	0.019 (0.059)	-0.066 (0.06)	0.255*** (0.056)	-0.019 (0.032)	-0.011 (0.038)	-0.091*** (0.029)
<i>Adults</i>	0.003 (0.135)	0.074 (0.135)	-0.073 (0.132)	-0.002 (0.146)	0.027 (0.146)	0.18 (0.139)	-0.008 (0.059)	-0.063 (0.059)	-0.066 (0.047)

Note: ***, **, * denote significance at the 1%, 5% and 10% levels. Standard errors in parentheses. The reported marginal effects are estimated from the regressions given in Table IX, Appendix B, evaluated at the mean for each age group. The results remain quantitatively similar if additional control variables are included. *White* receivers are taken as the baseline.



Note: The x axis plots the age group at which the marginal effects are evaluated. The y axis plots the marginal effect of the receivers' ethnicity on the probability of being classified into each of the behavioural types. The estimated marginal effects are given in Table 5, and are evaluated at the mean. Vertical bars represent 95% confidence intervals. The *Spain* treatment is taken as the baseline.

Figure 5: Behavioural Type and the Receivers' Ethnicity - Marginal Effects

Observation 3 provides evidence of positive discrimination in Children, Teenagers and Students expressed uniquely towards Black receivers. This finding seemingly contrasts with a prevalent result in this literature, in which we typically observe in-group favoritism and out-group discrimination¹⁰. One explanation for why we do not observe in-group favouritism is that simple physical cues and references to a foreign country may not have been enough to induce a sense of identity. This explanation is consistent with the results of Brewer & Silver (2000). Further, as has been previously discussed, all our ethnicity manipulations would be considered to be an out-group by Fehr et al. (2008, 2013)¹¹.

Alternatively, the observed behaviour could be a result of a social desirability bias: subject's may want to be perceived as behaving in a social desirable manner, and thus behave more altruistically towards those they perceive as being the most in need. This, however, would not explain why the Adults do not respond altruistically to Black receivers in the same manner that the Children do. A social desirability bias is likely to be most prevalent in Adults, who are more likely to be sensitive to normative pressures, and least prevalent in Children, who are likely to be unaware of such norms. As speculated by (Baker, 2015), negative prejudices towards the out-group may not necessarily produce animosity. For example, he finds that white Americans are more positive about giving aid when the recipient is of African descent in comparison to those of Eastern-European descent, despite the individuals having similar material needs. Paternalistic behaviour, in the form of altruism towards out-group members, can emerge when subjects feel warmly toward groups they assume to be lacking in a capacity to act. However, as with the other potential explanations, this doesn't seem to be the case for Adults¹².

4. Conclusion

We report evidence of a 'U-shaped' relationship between social preferences and age, with egalitarianism found first to diminish with age, but then to increase as individuals grow older. The inverse U-shaped' relationship is true for altruism. These observations contribute to the literature on the development of social preferences, as previous findings that do not include adults in the analysis had suggested egalitarianism decreases with age, whilst altruism becomes more prevalent. This is

¹⁰ In a meta-analysis of 77 lab studies published in economics, Lane (2016) reports that 93% of these studies report evidence of in-group favouritism.

¹¹ It is possible that the parochialism manipulation may not have been strong enough to induce the group effects observed in some of the previous literature, and other ethnic group cues such as language might have been more effective (see for example Esseily et al. (2016)).

¹² Appendix B shows that our results are robust to the potential issue of multiple hypothesis testing.

important, as altruism has been argued to be a prerequisite for ‘smooth’ workplace interactions (Fehr et al., 2013), being required for individuals to accept inequalities in the workplace. Therefore, our finding that altruism becomes less prevalent in adulthood may have implications for understanding what motivates different age groups in the workplace, particularly in relation to salary disparities.

The differences in behaviour across age groups could be attributed to income differences. For example, Adults earn more money than all other age groups and, as a consequence, payoffs earned in the experiment constitute a relatively smaller amount of the income they have available. However, such an income effect might predict that Adults would behave in a more altruistic way than other age groups, as the choices in all the games that make the recipient better off are relatively less costly. This is not what we observe.

By varying the receivers’ ethnicity, our paper also addresses recent behavioural theories of discrimination that indicate that social preferences are group-contingent. In contrast to previous studies that examine an interpersonal sense of identity, we report evidence of paternalism towards the out-group, rather than preferential treatment of the in-group, when utilising a broader, more collective sense of identity. This is particularly strong in Children, Teenagers and Students. However, out-group favouritism is not ubiquitous. It is only observed in interactions with Black foreign receivers, but not in interactions with East-Asian or Arab foreign receivers.

That Children favour Black foreign receivers, but not the other ethnic groups, is particularly striking given that they could only infer differences through appearance. Although not biased in favour of the in-group as is typically found, the finding that children are both aware and sensitive to the ethnic appearance characteristics of others is in line with previous findings in the developmental psychology literature (Lam et al., 2011).

Our findings highlight the importance of studying social preferences from both an early age and in later life. As social preferences can enhance efficiency in many workplace interactions, understanding how they develop over the life cycle is important for understanding how socialisation can impact preferences over outcomes. With the working population growing older, and workplaces becoming more diverse, understanding the interaction between social preferences, age and identity is therefore important for the design of institutions and their associated incentives in many societies.

References

- Afridi, F., Li, S. X. & Ren, Y. (2015), 'Social Identity and Inequality: The Impact of China's Hukou System', *Journal of Public Economics* 123, 17–29.
- Akerlof, G. A. & Kranton, R. E. (2000), 'Economics and Identity', *Quarterly Journal of Economics* 115(3), 715–753.
- Baker, A. (2015), 'Race, Paternalism, and Foreign Aid: Evidence from US Public Opinion', *American Political Science Review* 109(01), 93–109.
- Bellemare, C. & Shearer, B. (2009), 'Gift Giving and Worker Productivity: Evidence from a Firm–Level Experiment', *Games and Economic Behavior* 67(1), 233 – 244.
- Bernhard, H., Fischbacher, U. & Fehr, E. (2006), 'Parochial Altruism in Humans', *Nature* 442(7105), 912–915.
- Brañas–Garza, P., Cobo–Reyes, R. & Dominguez, A. (2006), "'Si él lo necesita": Gypsy Fairness in Vallecas', *Experimental Economics* 9(3), 253–264.
- Brewer, M. B. & Gardner, W. (1996), 'Who is this "We"? Levels of Collective Identity and Self Representations', *Journal of Personality and Social Psychology* 71(1), 83.
- Brewer, M. B. & Silver, M. D. (2000), 'Group Distinctiveness, Social Identification, and Collective Mobilization', *Self, Identity, and Social Movements* 13, 153–171.
- Camerer, C. & Fehr, E. (2004), *Measuring Social Norms and Preferences using Experimental Games: A Guide for Social Scientists*, in J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr & H. Gintis, eds, 'Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small–Scale Societies', Oxford University Press.
- Charness, G. & Rabin, M. (2002), 'Understanding Social Preferences with Simple Tests', *The Quarterly Journal of Economics* 117(3), 817.

Chen, R. & Chen, Y. (2011), 'The Potential of Social Identity for Equilibrium Selection', *American Economic Review* 101(6), 2562–89.

Chen, Y. & Li, S. X. (2009), 'Group Identity and Social Preferences', *American Economic Review* 99(1), 431–57.

Chen, Y., Li, S. X., Liu, T. X. & Shih, M. (2014), 'Which Hat to Wear? Impact of Natural Identities on Coordination and Cooperation', *Games and Economic Behavior* 84, 58–86.

Cooper, D. & Kagel, J. H. (2009), 'Other Regarding Preferences: A Selective Survey of Experimental Results', *Handbook of Experimental Economics* 2.

Croson, R. & Gneezy, U. (2009), 'Gender differences in preferences', *Journal of Economic literature* 47(2), 448–74.

DellaVigna, S., John, A. L. & Malmendier, U. (2012), 'Testing for Altruism and Social Pressure in Charitable Giving', *Quarterly Journal of Economics* 127(1), 1–56.

Drouvelis, M. & Nosenzo, D. (2013), 'Group Identity and Leading-by-Example', *Journal of Economic Psychology* 39, 414–425.

Dufwenberg, M., Heidhues, P., Kirchsteiger, G., Riedel, F. & Sobel, J. (2011), 'Other-Regarding Preferences in General Equilibrium', *The Review of Economic Studies* 78(2), 613–639.

Engel, C. (2011), 'Dictator Games: A Meta Study', *Experimental Economics* 14(4), 583–610.

Esseily, R., Somogyi, E. & Guellai, B. (2016), 'The relative importance of language in guiding social preferences through development', *Frontiers in Psychology* 7, 1645.

Falk, A. (2007), 'Gift Exchange in the Field', *Econometrica* 75(5), 1501–1511.

Fehr, E., Bernhard, H. & Rockenbach, B. (2008), 'Egalitarianism in Young Children', *Nature* 454(7208), 1079–1083.

- Fehr, E., Glätzle-Rützler, D. & Sutter, M. (2013), 'The Development of Egalitarianism, Altruism, Spite and Parochialism in Childhood and Adolescence', *European Economic Review* 64, 369–383.
- Fehr, E. & Schmidt, K. M. (1999), 'A Theory of Fairness, Competition, and Cooperation', *Quarterly Journal of Economics* 114(3), 817–868.
- Fischbacher, U. & Gächter, S. (2010), 'Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments', *American Economic Review* 100(1), 541–556.
- Gneezy, U. & List, J. A. (2006), 'Putting Behavioral Economics to Work: Testing for Gift–Exchange in Labor Markets using Field Experiments', *Econometrica* 74(1), 1365–1384.
- Grosskopf, B. & Pearce, G. (2016), Do you mind me paying less? Measuring Other–Regarding Preferences in the Market for Taxis, mimeo.
- Holm, S. (1979), 'A simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics* 6(2), 65–70.
- House, B. R. (2018), 'How do social norms influence prosocial development?', *Current Opinion in Psychology* 20, 87 – 91. Early Development of prosocial behavior. URL: <http://www.sciencedirect.com/science/article/pii/S2352250X17301616>
- House, B. R., Silk, J. B., Henrich, J., Barrett, H. C., Scelza, B. A., Boyette, A. H., Hewlett, B. S., McElreath, R. & Laurence, S. (2013), 'Ontogeny of prosocial behavior across diverse societies', *Proceedings of the National Academy of Sciences* 110(36), 14586–14591.
- House, B. R. & Tomasello, M. (2018), 'Modeling social norms increasingly influences costly sharing in middle childhood', *Journal of experimental child psychology* 171, 84–98.
- Kube, S., Maréchal, M. A. & Puppe, C. (2012), 'The Currency of Reciprocity: Gift Exchange in the Workplace', *American Economic Review* 102(4), 1644–1662.
- Kube, S., Maréchal, M. A. & Puppe, C. (2013), 'Do Wage Cuts Damage Work Morale? Evidence from a Natural Field Experiment', *Journal of the European Economic Association* 11(4), 853–870.

Lam, V., Guerrero, S., Damree, N. & Enesco, I. (2011), 'Young Children's Racial Awareness and Affect and their Perceptions about Mothers Racial Affect in a Multiracial Context', *British Journal of Developmental Psychology* 29(4), 842–864.

Lane, T. (2016), 'Discrimination in the Laboratory: A Meta–Analysis of Economics Experiments', *European Economic Review* 90, 375 – 402.

List, J. A. (2004), 'Young, selfish and male: Field evidence of social preferences', *The Economic Journal* 114(492), 121–149.

List, J. A. (2006), 'The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions', *Journal of Political Economy* 114(1), 1–37.

Moore, C. (2009), 'Fairness in Children's Resource Allocation Depends on the Recipient', *Psychological Science* 20(8), 944–948.

Mujcic, R. & Frijters, P. (2013), *Still Not Allowed on the Bus: It Matters If You're Black or White!*, IZA Discussion Paper .

Appendix A. Experimental Material

A.1. Recruitment Procedures

A.1.1. Spanish Receivers

The main part of the experiment was conducted in the city of Granada, Spain. Granada is a southern city of Spain composed of predominantly people who are Caucasian (white) and catholic in their religious beliefs. There is a growing minority of Arabs and Muslims. Granada's economy is dedicated mainly to tourism, (through the attractions of the Alhambra Palace and the Sierra Nevada) and agriculture. It has a population of 236,000. We recruited Children and Teenagers from six different schools in the city. All classes within the schools had between 20 and 35 students, both in the elementary and high schools. The Students were recruited from the University of Granada, which has a student body of 68,000. In our experiment, participants came from different faculties of the University: The Faculty of Economics and Business, the Faculty of Education Sciences and the Faculty of Humanities. Half of the Adults were recruited from the parent body of the Children, with the other half being employees of the University of Granada. The average age was 45 years and 57 percent of the adult

sample holds a university degree.

A.1.2. Chinese Receivers

Receivers from China were recruited in Changde, Hunan province, located in South Central China. The population composition of Changde is more homogeneous than that of east coast cities with similar sized populations and economies because few residents are immigrants. We chose institutions with individuals from urban as well as rural backgrounds, to control for potential Hukou differences between groups. See Afridi et al. (2015) for a discussion of the differences between Hukou groups. Children were recruited from an elementary school located at an intersection between urban and rural districts of Changde. The school consisted of 18 classes with over 1,000 students, with 70% of the students being from a rural area. To match with Spanish participants, we recruited 50 students from a fourth grade class with an equal gender composition, and an age range from 9 to 11. Similar to our recruiting protocol for the Children, we recruited 48 Teenagers aged 13 to 15 from a ninth grade high school class. This school included 30 senior high school classes and 24 junior high school classes with a total of over 3,000 students. Students were recruited from a large undergraduate class in a local college: Hunan University of Arts and Sciences. It has over 23,000 undergraduates, and most of them are local residents or from nearby areas within the Hunan province. In total, 50 Students participated in our study aged between 18 and 24 years old. Lastly, we recruited 23 Adults (teachers) aged 35 to

59 from another elementary school. This school is considered to be the most competitive elementary school in Changde, where 86% teachers have at least a junior college degree.

A.1.3. Moroccan and Senegalese Receivers

In Morocco, the experiment took place in Tangier, the biggest city in North–Western Africa, and located on the Mediterranean coast. In Senegal the experiment took place in Dakar, the capital and largest city of the country. In the Human Development Ranking these countries are, respectively, in positions 126 (Morocco) and 170 (Senegal) of 188 countries. Residents are predominantly Muslim in both countries (Morocco 99.5%; Senegal 90%). With respect to the field work, in both countries we had the full support of an international NGO (Alliance for Solidarity) both before and after the experiment. The NGO requested the permits required for the experiment and completed the translation of the documents (instructions, survey) into the local languages (Arabic and French). They were also in charge of recruitment (following our instructions). Subjects of equal gender composition were recruited from a homogeneous population in low income schools that require their students to wear school uniforms. School uniforms were required as we didn't want clothing to be indicative of potential income differences. The NGO also provided staff members to run the experiment. In Tangier we recruited 47 Children and 54 Teenagers from two elementary schools situated in suburban areas. In Dakar, 33 Children and 49 Teenagers were recruited from the outskirts of Dakar. Students in both countries were recruited from undergraduate classes in two local colleges. Finally, we recruited 48 Adults. In Tangier the adults were part of the staff of a school, as well as parents of the primary school. In Dakar, they belonged to a Neighbourhood Association in the same area of the Children's school. The majority of them were women, given that in both cities they represent the majority of members at this type of association.

A.2. Experimental Instructions

A.2.1. General Comments

Welcome to this experiment. Here you will find the instructions for the tasks you have to fulfill. There are no right or wrong answers, your identity will not be known at any time and we will use only the information you provide. The goal of this experiment is to study how people make decisions. The instructions are very easy and if you follow them carefully you can receive some money. It is very important that you understand the instructions. If you have any questions, do not hesitate to raise your hand and ask the experimenter. Besides these questions, any kind of communication is completely forbidden and you could even be expelled from the experiment.

A.2.2. Specific Instructions – Dictators

This experiment consists of one period. You will be matched with a person from the following group (see picture). There are two types of participants: Type A and Type B. You will participate as Type A and your counterpart (somebody from the picture) will be Type B. You have to make three decisions. For each decision you have to choose between two allocations of money (Payoff A, Payoff B) with the first number indicating the payoff to you and the second number indicating the payoff to Type B. The decisions that you face are shown in the following table.

	Left	Right	Decision
	(Payoff A, Payoff B)	(Payoff A, Payoff B)	Left- Right
Decision 1	(€5,€5)	(€5,€0)	
Decision 2	(€5,€5)	(€5,€10)	
Decision 3	(€5,€5)	(€10,€10)	

Type B will not make any decisions in this task. They will only be informed about what you have chosen. Only one of the three decisions will be selected for payment. Earnings will therefore only depend on your decisions.

A.2.3. Specific Instructions – Receivers

This experiment consists of one period. You will be matched with a person from the following group (see picture). There are two types of participants: Type A and Type B. You will participate as Type B and your counterpart (somebody from the picture) will be Type A. Type A has to make three decisions. For each decision Type A has to choose between two allocations of money (Payoff A, Payoff B) with the first number indicating the payoff to Type A and the second number indicating the payoff to you. The decisions that Type A faces are shown in the following table.

	Left	Right	Decision
	(Payoff A, Payoff B)	(Payoff A, Payoff B)	Left- Right
Decision 1	(€5,€5)	(€5,€0)	
Decision 2	(€5,€5)	(€5,€10)	
Decision 3	(€5,€5)	(€10,€10)	

You as Type B will not make any decisions in this task. You will only be informed about what Type A has chosen. Only one of the three decisions will be selected for payment. Earnings will therefore only depend on Type As decisions.

Appendix B. Statistical appendix

This section presents tables of the complete Probit regression from Section 3.

B.1. Parametric Analysis

This section presents a number of tables of estimates obtained from Probit regressions, along with their corresponding marginal effects.

<i>Dependent Variable:</i>	<i>Egalitarian Type (i)</i>	<i>Altruistic Type (ii)</i>	<i>Spiteful Type (iii)</i>
<i>Age</i>	-0.1645*** (0.0216)	0.187*** (0.0229)	-0.0213 (0.0302)
<i>Age*Age</i>	0.0025*** (0.0004)	-0.0027*** (0.0004)	-0.0002 (0.0006)
<i>Female*Age</i>	0.0052 (0.009)	-0.0153* (0.0093)	0.039* (0.0204)
<i>Female</i>	0.252 (0.2049)	-0.0325 (0.2138)	-0.7289* (0.3729)
<i>In-group</i>	0.0869 (0.1184)	-0.1677 (0.1222)	0.1807 (0.1661)
<i>Constant</i>	1.8214*** (0.2664)	-2.3301*** (0.2853)	-1.0575*** (0.3621)
<i>Obs.</i>	633	633	633

Note: ***, **, * denote significance at the 1%, 5% and 10% levels. Standard errors in parentheses. All reported estimates are from Probit regressions. The results remain quantitatively similar if additional control variables are included. The number of observations differ to those re-ported in Table 1 due to missing entries. The observations from sixteen subjects are dropped, as we were unable to categorise them into either one of the three behavioural types.

Table 6: Probit Estimates - Determinants of Behavioural Type

Appendix C. Robustness Checks

As we examine the data for a number of treatment effects, with 60 hypothesis tests in total (24 in Table 3 and 36 in Table 5) some of the statistical significance that we observe may be an artefact of multiple hypothesis testing (MHT). To account for this, we adjust the calculated p-values used to support Observations 1–3 using the Holm–Bonferonni (HB) correction procedure. We treat all 60 tested hypotheses as being part of the same ‘family’ of tests, and therefore apply the strictest possible correction. This is one of the most standard procedures used to correct for multiplicity in the sciences, and we use this procedure over the more conservative Bonferroni procedure because of its reduced false negative rate, and thus, increased power (Holm, 1979). Table 10 presents the p-values that remain significant once the correction has been applied. The first column provides information on the Table the original p-value is taken from, the second column outlines which Observation the p-value is used to support. The third column shows the dependent variable and the fourth column gives information on the estimated marginal effect. The fifth column details the age group the p-value relates to. The sixth column gives the original p-value, and the final column the Holm-Bonferonni corrected p-value. As an example, consider the first row of Table X. The p-value is taken from Table III, relates to Observation 1, the dependent variable is the egalitarian type and the marginal effect the p-value relates to is the marginal effect of age. It was calculated for Children, had a value of $p < 0.001$ and once corrected is still less than 0.001, and remains highly significant.

As can be seen from the corrected p-values presented in Table 10, Observation 1 can be clearly distinguished from Type 1 error, with the estimated marginal effect of age remaining statistically significant at the $p < 0.001$ level for both the egalitarian and altruistic types. However, Observation 2 is not as robust, with only a gender difference in Students remaining once all p-values have been corrected for. Observation 3 is also found to be robust to criticisms of multiplicity, with all age groups except Adults being more altruistic when the receiver is Black, as originally observed. Thus, we conclude that our main observations are unlikely to be the result of MHT and appear to be robust to such criticisms.

<i>Dependent Variable:</i>	<i>Egalitarian Type</i>	<i>Altruistic Type</i>	<i>Spiteful Type</i>
	<i>(i)</i>	<i>(ii)</i>	<i>(iii)</i>
<i>Marginal Effect of Age:</i>			
<i>Children</i>	-0.028*** (0.005)	0.025*** (0.004)	0.000 (0.004)
<i>Teenagers</i>	-0.029*** (0.006)	0.031*** (0.006)	-0.000 (0.003)
<i>Students</i>	-0.021*** (0.005)	0.025*** (0.005)	-0.001 (0.002)
<i>Adults</i>	0.02*** (0.006)	-0.022*** (0.007)	-0.001 (0.002)
<i>Marginal Effect of Female:</i>			
<i>Children</i>	0.103** (0.051)	-0.052 (0.043)	-0.057 (0.036)
<i>Teenagers</i>	0.137*** (0.053)	-0.103** (0.053)	-0.03 (0.026)
<i>Students</i>	0.129*** (0.047)	-0.119** (0.05)	-0.009 (0.026)
<i>Adults</i>	0.118 (0.121)	-0.203 (0.125)	0.09 (0.059)

Note: ***, **, * denote significance at the 1%, 5% and 10% levels. Standard errors in parentheses. All reported estimates are from Probit regressions. The marginal effects of *Age* and *Female* are calculated for each regression, and are evaluated at the mean for each age group. Age is the reported age of the subject, and treated as a continuous variable. The observations from sixteen subjects are dropped, as we were unable to categorise them into either one of the three behavioural types.

Table 7: Marginal Effects with Additional Controls - Determinants of Behavioural Type

<i>Dependent Variable:</i>	<i>Egalitarian Type</i>	<i>Altruistic Type</i>	<i>Spiteful Type</i>
	<i>(i)</i>	<i>(ii)</i>	<i>(iii)</i>
<i>Marginal Effect of Age:</i>			
<i>Children</i>	-0.037*** (0.003)	0.033*** (0.002)	-0.002 (0.003)
<i>Teenagers</i>	-0.033*** (0.004)	0.036*** (0.004)	-0.001 (0.002)
<i>Students</i>	-0.025*** (0.003)	0.029*** (0.003)	-0.001 (0.001)
<i>Adults</i>	0.02*** (0.003)	-0.02*** (0.003)	-0.001 (0.002)
<i>Marginal Effect of Female:</i>			
<i>Children</i>	0.096** (0.044)	-0.045 (0.037)	-0.054* (0.032)
<i>Teenagers</i>	0.131*** (0.043)	-0.103** (0.043)	-0.018 (0.022)
<i>Students</i>	0.135*** (0.039)	-0.131*** (0.041)	0.004 (0.022)
<i>Adults</i>	0.167* (0.086)	-0.245*** (0.085)	0.073** (0.037)

Note: ***, **, * denote significance at the 1%, 5% and 10% levels. Standard errors in parentheses. All reported estimates are from Probit regressions. The marginal effects of *Age* and *Female* are calculated for each regression for each age group. Age is the reported age of the subject, and treated as a continuous variable. The results remain quantitatively similar if additional control variables are included. The observations from sixteen subjects are dropped, as we were unable to categorise them into either one of the three behavioural types.

Table 8: Average Marginal Effects - Determinants of Behavioural Type

<i>Dependent Variable:</i>	<i>Egalitarian Type</i> (i)	<i>Altruistic Type</i> (ii)	<i>Spiteful Type</i> (iii)
<i>Age</i>	-0.164*** (0.024)	0.182*** (0.025)	0.009 (0.037)
<i>Age</i> × <i>Age</i>	0.002 (0.00)	-0.003 (0.00)	-0.001 (0.00)
<i>Female</i> ×	0.004 (0.00)	-0.013 (0.00)	0.041*** (0.008)
<i>Female</i>	0.28*** (0.017)	-0.104*** (0.021)	-0.767 (0.00)
<i>Arab</i>	-0.014 (0.04)	0.086** (0.042)	-0.156*** (0.054)
<i>E.Asian</i>	0.067* (0.045)	-0.341*** (0.045)	0.737*** (0.092)
<i>Black</i>	-0.627*** (0.028)	0.838*** (0.03)	-0.364*** (0.098)
<i>Arab</i> × <i>Age</i>	0.00 (0.00)	-0.002 (0.00)	0.002 (0.00)
<i>E.Asian</i> × <i>Age</i>	0.003 (0.00)	0.009 (0.00)	-0.041 (0.00)
<i>Black</i> × <i>Age</i>	0.009 (0.00)	-0.008 (0.00)	-0.044 (0.00)
<i>Constant</i>	1.906 (0.00)	-2.413 (0.00)	-1.249 (0.00)
<i>Obs.</i>	633	633	633

Note: ***, **, * denote significance at the 1%, 5% and 10% levels. Standard errors in parentheses. All reported estimates are from Probit regressions. The results remain quantitatively similar if additional control variables are included. The number of observations differ to those reported in Table 1 due to missing entries. The observations from sixteen subjects are dropped, as we were unable to categorise them into one of the three types.

Table 9: Probit Estimates - Identity and Behavioural Types

<i>Table</i>	<i>Observation</i>	<i>Dep. Variable</i>	<i>Marginal Effect</i>	<i>Age Group</i>	<i>p-value</i>	<i>Corrected p-value</i>
<i>Table 3</i>	<i>Observation 1</i>	<i>Egalitarian Type</i>	<i>Age</i>	<i>Children</i>	<i>0.000</i>	<i>0.000</i>
				<i>Teenagers</i>	<i>0.000</i>	<i>0.000</i>
				<i>Students</i>	<i>0.000</i>	<i>0.000</i>
				<i>Adults</i>	<i>0.000</i>	<i>0.000</i>
		<i>Altruistic Type</i>		<i>Children</i>	<i>0.000</i>	<i>0.000</i>
				<i>Teenagers</i>	<i>0.000</i>	<i>0.000</i>
				<i>Students</i>	<i>0.000</i>	<i>0.000</i>
				<i>Adults</i>	<i>0.000</i>	<i>0.000</i>
<i>Table 3</i>	<i>Observation 2</i>	<i>Egalitarian Type</i>	<i>Female</i>	<i>Students</i>	<i>0.0006</i>	<i>0.0288</i>
				<i>Children</i>	<i>0.0005</i>	<i>0.025</i>
<i>Table 5</i>	<i>Observation 3</i>	<i>Altruistic Type</i>	<i>Black</i>	<i>Teenagers</i>	<i>0.0000</i>	<i>0.0004</i>
				<i>Students</i>	<i>0.0000</i>	<i>0.0003</i>

Note: The first column provides information on the Table the p -value is taken from, the second column outlines which Observation the p -value is used to support, the third column the dependent variable and the fourth column gives information on the explanatory variable. The fifth column details the age group the p -value relates to. The sixth column gives the original p -value, and the final column the Holm-Bonferroni corrected p -value. All p -values are 2 sided.

Table 10: Robustness Check – Holm-Bonferroni Corrected p -values

Chapter 2.

Gender Quotas and Task Assignment in Organizations

José J. Domínguez
University of Padova

Natalia Montinari
University of Bologna

Keywords:

Affirmative action, gender quotas, gender gap, task assignment, laboratory experiments.

JEL:

D03, C91, J71

1. Introduction

An increasing number of countries have introduced affirmative action policies during the last decades for combatting gender discrimination in traditionally male-stereotyped occupations. Electoral quotas ensure that a minimum share of females, ranging from one third in Greece and Portugal to one half in Belgium and France, constitutes the body of the public election. In Germany, Italy, Spain and Sweden, political parties introduced quotas in their own lists voluntarily, and Norway also passed a law in 2003 requiring, at least, a 40 percent representation of each gender in the board of public liability firms (Freidenvall & Dahlerup, 2013). Despite the considerable growth of female employment during in Europe this period, gender inequality in the labor market still persists (Bertrand, 2010). Females earn, on average, 14.7% less per hour than males. The gender wage gap is especially higher for managerial positions (23%). Moreover, only the 26.7% of the boards and the 6.5% of the CEOs are female (EU, 2019). Actually, Bertrand et al. (2018) did not find substantial evidence of improvement in females' labor market outcomes in companies subject to the quota in the Norwegian private sector, especially for high-ability females. Mechanisms other than the underrepresentation of females in male-dominated environments need to be revised in order to design efficient policy interventions that help to close the gender gap in labor market inequalities.

In this paper, we aim to investigate how gender quotas affect the task assignment in the organizations. Task allocation has been considered as a fundamental aspect in labor market inequalities. De Pater et al. (2010) suggest that employers' decisions are not gender blind and that discrimination and prior beliefs about workers' skills can relegate females to lower positions in the organization, what consequently increases the gender wage gap. The literature on affirmative action policies has focused on the effect that gender quotas exert in hiring decisions and workers' behavior. Nevertheless, as far of our knowledge, no research has actually focused on the effect of gender quotas on organizational decisions. Then, a better understanding of how organizations allocate workers and how a higher share of females in the organizations affect these decisions is crucial to determine whether gender quotas are desirable for fostering gender equality in the workplace. To answer this question, we propose a laboratory experiment in which employers were asked to create a team of six workers. Then, they had to assign workers to different male-stereotyped tasks in terms of complexity and profitability: four workers to an Easy Task (i.e. solving additions) and two workers to a Hard Task (i.e. solving mathematical problems). We selected two basic tasks that differ in profitability, for both the worker and the employer, to represent an environment in which workers' outcomes depend on the task in which they have been allocated by the employers. To better represent this differential, the tasks will also differ in complexity. While an adding task can be perceived as a basic job that everybody can perform, the math task can be perceived as a job that needs more formation and knowledge in a specific area (e.g. STEM fields).

We contribute to the literature by shedding more light in the debate about 1) the relationship between the task assignment and gender inequalities in the labor market, and 2) the effectiveness of gender quotas by observing the task assignment decisions after the introduction of a gender quota in the hiring stage. We find that female workers have less success than male workers in the hiring but we do not find discrimination in the task assignment. A mandatory quota in the hiring stage, by definition, increases the number of females in the organization. Nevertheless, we do not find an increase in the probability of female workers to be appointed for the more complex task after the policy intervention. Indeed, we find a negative effect of the quota on high-ability females. A quota in the hiring stage produce a decrease in the probabilities of high-ability females to be selected for such task either compared to male workers of similar ability, expanding the gender gap, and compared to female workers in the treatment with no policy intervention, suggesting the ineffectiveness of the quota in improving females' labor market outcomes.

Evidence from laboratory experiments has provided practical insights about the positive effects of gender quotas on both sides of the labor market. On the supply side, the gender differences in taste for competition have been hypothesized as a potential predictor of labor market outcomes (Buser et al.,

2014; Dohmen & Falk, 2011; Heinz et al., 2016). Indeed, Reuben et al. (2015) showed that MBA students who were more competitive during studies were more likely to work in better-paid industries nine years later. Consequently, if females have stronger aversion to competitive workplaces compared to men of similar ability, the low share of females participating in market competitions could explain the gender gap in labor market success (Niederle & Vesterlund, 2007; Flory et al., 2014). In this sense, a number of papers have shown that gender quotas encourage females to self-select into competitive environments (Balafoutas & Sutter, 2012; Calsamiglia, Franke, & Rey-Biel, 2013; Niederle, Segal, & Vesterlund, 2013; Niederle & Vesterlund, 2008; Sutter, Glatzle-Rutzler, Balafoutas, & Czermak, 2016), especially for competing for top positions (Czibor & Domínguez-Martínez, 2018; Maggiani, et al., 2019). In the field, Ibanez and Reiner (2018) showed that job advertisements that specify that females would be favoured attracted more females and helped to close the gender gap in job applications. Moreover, the gains obtained outweighed the losses associated with the decrease in male applications. On the demand side, Beaurain & Masclet (2016) confirmed that voluntary quotas (i.e. organizations must pay a fine if they the quota is not respected) reduce discrimination in the hiring process. Specifically, female candidates were ranked favorably compared with male candidates in settings without policy intervention without affecting the efficiency of the organization. Similar results were observed in Indian village councils, in which the exposure to female leaders was associated with an improvement of the perception of females as leaders and with electoral gains for females (Beaman et al., 2009).

Despite these positive effects of the policy, the gender wage gap and the underrepresentation of females in leadership positions are still questions of concern. Different mechanisms could explain the gender gap in wages and leadership despite the increasing share of females in the organizations. Firstly, it has been argued that a significant fraction of the gender wage gap is explained by the gender differences in salary negotiations. Babcock and Laschever (2003) reported that individuals who did not negotiate first salaries lost more at the end of the work life. Other studies have shown that females are less willing to engage in salary negotiations, especially in environments in which the possibility to negotiate is more ambiguous (Babcock et al, 2006; Small et al, 2007; Leibbrandt & List, 2014). Secondly, females are more willing than males to self-select into tasks that everyone prefers to be completed by someone else and that have little weight for promotion decisions, also called low-promotability tasks (Babcock et al., 2017a).

However, career success not only depend on factors from the supply side. Factors from the demand side such as employers' gender stereotypes and discrimination can also explain a proportion of the wage and leadership gaps. On one hand, the family background seems to exert a different effect between males and females. Mothers are penalized in a set of variables, such as perceived competence

and salary recommendations while fathers do not (Correll et al., 2007). Importantly, under the age of 35, the wage gap between mother and nonmothers is larger than the gap between males and females (Crittenden, 2001). On the other hand, the prior beliefs about workers' skills can also affect females differently. Even if females perform equally than males, employers' gender biases may result in lower roles of females within the organization (De Pater et al., 2010). The experimental evidence suggests that females are discriminated against not only at the vertical hierarchy, having lower probabilities to be selected as a team leader (Reuben et al., 2012; Peterle & Rau, 2017), but also at the horizontal hierarchy, having less probabilities to be assigned to high-promotability tasks within the organization compared to males (Babcock et al., 2017b).

The remainder of this paper is structured as follows: In Section 2, we present the experimental design and procedures. In Section 3, we present the results and Section 4 concludes.

2. Experimental Design

2.1. General Overview

In order to observe how gender quotas affect the task assignment decisions of the organizations we conducted two experiments: one aimed at collecting data on workers (Experiment W), the other aimed at collecting data on employers' decisions (Experiment E). Both Experiment W and Experiment E were divided in 4 parts and only differed for the content of part 3.

At the beginning of the experiments, subjects were told that they were taking part in a labor market experiment and were asked to fill a short CV providing the following information: Gender, Year of Birth and Field of Study. In Part 1, subjects were confronted with the Easy Task, which consisted of summing up as many three three-digit numbers as possible in 6 minutes divided into two sub-parts of 3 minutes each. Subjects received €0.50 per each correct calculation (this task was similar to the one used in Balafoutas & Sutter, 2012). No calculators or electronic devices were allowed. In Part 2, subjects were confronted with the Hard Task, which consisted of solving as many mathematical problems as possible in 10 minutes. Subjects received €1.50 per each correct answer. The set of mathematical problems for the Hard Task was extracted from the set of mathematical questions included in the entry test of the University of Padova's bachelor program in Economics administered in the entry exams of April and August of 2018, to which 1476 students participated¹. We selected a set of problems for which we do not find evidence of gender differences in the probability of answering correctly. Each problem consisted of a multi-choice question with four possible answers, the set of problems used in Part 2 is displayed in the Online Appendix A. In both Part 1 and Part 2 wrong answers

¹ The problems presented to our participants are reproduced in Appendix A.

did not penalize subjects' earnings. In Part 3 we designed a labor market game which was different in the two experiments and it is explained in detail in the next subsection. In Part 4 we elicited risk attitudes with the static version of the Bomb Risk Elicitation Task (Crosetto & Filippin, 2013)². Finally, subjects were provided with feedback about their performances and earnings in each part (no feedback across parts was provided), their final earnings and a post-experimental questionnaire that included a measure of competitiveness (Houston et al. 2002; Harris and Houston 2010) and personality traits (Gosling et al., 2003).

2.2. Experiment E: Part 3- Labor Market Game

In Experiment E, all subjects played the role of employer. Each employer faced a group of, at most, 15 workers that was randomly generated from the information gathered in Experiment W³ and was asked to make two decisions divided in two stages: a hiring stage and a task assignment stage. In the first stage, employers had to hire six workers to form a team. The profiles of the workers in the groups were ordered by the randomly assigned ID number. Employers had to click in the profiles of the six workers they wanted to hire and continue to the second stage^{4,5}.

In the second stage, only the information about the six selected workers was displayed and employers had to assign workers to two different tasks. Specifically, they had to assign four workers to the Easy Task and two workers to the Hard Task, where the Easy Task was both less complex and less profitable compared to the Hard Task.⁶

We provided employers with the information released by the workers when filling the CVs in Experiment W: gender, year of birth and field of study.⁷ Workers were recruited from four different

² In the Bomb Risk Elicitation Task (BRET), subjects were presented with a 10x10 square containing 100 numbered cells in which a bomb were randomly placed behind one of the cells. Subjects were asked to decide how many boxes to collect out of 100. Cells were collected sequentially (i.e. if a subject decided to collect 50 cells, she collected from cell 1 to cell 50). Each cell collected were translated into €0.10 at the end of this part only if the number of cells collected was strictly lower than the number of the cell that contained the bomb. The cell that hidden the bomb was randomly selected for each subject. Subjects received nothing if the number of cells collected was higher than the number of the cell that contained the bomb.

³ In Experiment W, we recruited 120 participants to act in the role of workers and 8 participants to act in the role of employers. In each session, at the beginning of Part 3, 30 participants were randomly assigned to the role of workers and divided into two groups of 15 workers each, while 2 participants were assigned to the role of employers. Participants in the role of employers faced the same choices as in the Experiment E. Workers were asked to decide whether to participate or not in a labor market game (by paying a fixed cost) and obtain different levels of earnings depending on employers' hiring and assignment decisions. Detailed results on the Experiment W are provided in the Online Appendix B.

⁴ Bohnet et al. (2015) showed that employers are less likely to use gender stereotypes in joint than in separate evaluations.

⁵ Screenshots of the stages in Part 3 can be found in the Online Appendix A.

⁶ During the experiment, the tasks were presented as Task 1 (Easy Task) and Task 2 (Hard Task). Information about complexity or profitability was avoided.

⁷ This procedure can be found in other experiments in which employers have to make hiring decisions in order to reduce the salience of gender and possible effects of gender stereotypes. See Bertrand & Mullainathan (2004), Bohnet et al. (2015), Beaurain & Masclet (2016), Heinz et al. (2016), Peterle & Rau (2017) and Paryavi et al. (2019).

fields of studies: Sciences, Social Sciences, Engineering and Foreign Languages. In addition to the information provided in the CV, employers were provided with a signal of performance consisting of the number of correct calculations obtained by the worker in the first half of the Easy Task (Part 1). Each employer evaluated 4 different groups of workers presented in random order, therefore the hiring stage and the task assignment stage were repeated four times (four rounds). The earnings of the employers in each round of Part 3 (Π_E) were computed as estimated by equation (1): employers received €0.50 per each correct answer in Part 2 (CorrectHT) of those workers assigned to the Hard Task (i) plus €0.10 per each correct calculation in Part 1 (CorrectET) of those workers assigned to the Easy Task (j).

$$\Pi_E = 0.50 \sum_{i=1}^2 \text{CorrectHT}_i + 0.10 \sum_{j=1}^4 \text{CorrectET}_j \quad (1)$$

At the end of part 3, we elicited employers' beliefs about their relative performance in the Easy Task and the Hard Task with respect to the other subjects in the same session. Comparing this estimation with the real relative performance allow us the obtain a measure about a subject's degree of self-confidence in both tasks. That is, a subject is underconfident (overconfident) if her expected relative performance is strictly lower (higher) than her real one. Employers were also asked to estimate the average productivity of males and females in both tasks. An English version of the Instructions is reproduced in the Online Appendix A.

2.3. Treatments

In order to observe the effect of the gender quota on workers' and employers' decisions we implemented two treatments. In the Control Treatment, there were no constraints for employers when hiring and assigning workers to the different tasks. Therefore, participants faced the two decisions as described in the previous subsection. In the Quota treatment, employers were told that at least the fifty percent of the workers hired in the first stage (i.e. 3 out of 6 workers) must be women, while there were no constraints when assigning the hired workers either to the Easy Task or the Hard Task in the second stage. In experiment W, subjects were also allocated to one of these treatments. Employers evaluated four groups of workers that were exposed to the same treatment given that the introduction of a quota may affect a worker's decision to participate the labor market game. That is, employers in the Control Treatment (Quota Treatment) evaluated the four groups of workers in the control Treatment (Quota Treatment).

2.4. Procedure and Samples

The experiments were programmed using z-Tree (Fischbacher, 2007) at BLESS, the experimental laboratory of the University of Bologna (Italy). Subjects were recruited by using the information stored in ORSEE (Greiner, 2015)⁸. Experiment E and Experiment W presented similar features: sessions were not gender-balanced, all treatments were run in a between-subjects design and none participated in more than one treatment (i.e. subjects were randomly assigned to treatments). The duration of each session was about 90 minutes. In each session, once arrived at the lab, instructions were read aloud and subjects were informed that Part 4 and one randomly selected part among Part 1, Part 2 and Part 3 would have been relevant for payments. If Part 3 was selected for payments, in Experiment E, one of the four rounds was randomly selected to be relevant for the final payment. The average payment was about €15, including a 5 Euro show-up fee.

From November to December 2018, we run 2 sessions of 32 subjects per treatment in each experiment (8 sessions in total). In Experiment W, 128 subjects from four different fields of studies (Sciences, Social Sciences, Engineering and Foreign Languages) participated: 120 subjects as workers and 8 subjects as employers. We used the information of workers in Experiment W to be shown to employers in Experiment E. We dismissed the decisions made by employers in Experiment W. In Experiment E (our main experiment), 128 subjects from all different schools of the University of Bologna (55% female) participated. In the Control Treatment, 64 employers evaluated 56 workers (3584 observations). In the Quota Treatment, 64 employers evaluated 57 workers (3648 observations).

3. Results

In this section we present our results. First, we check whether female workers are discriminated against in both the hiring and the task assignment stages in the Control Treatment (Section 3.1.). Section 3.2. describes the effectiveness of the gender quota in both the hiring and the task assignment stages and Section 3.3. analyses the effect of the quota on employer's performance/earnings.

3.1. Is there a need for the introduction of a gender quota?

In this subsection, we focus on the Control Treatment to test whether female workers have lower probabilities of success than male workers in both the hiring and the task allocation.

While 52% of the workers are women, the average proportion of women in the teams was 46%, significantly lower than the proportion of men (Two-sided binomial probability test p -value = .000) and

⁸ Descriptive statistics about participants in Experiment E and Experiment W can be found in the Online Appendix C (Table C1).

the 30 per cent of the workers assigned to the Hard Task ($<70\%$; $p\text{-value}=.000$). However, decisions are not only contingent on gender. Employers evaluated four asymmetric groups of workers that differ in several characteristics such as the share of females, the number of workers who decided to participate and workers' characteristics⁹. Since employers' earnings positively depend on workers' performances, we assume that employer made their decisions based on the signal if they believe that it is informative enough in predicting workers' performances¹⁰. Then, we assume that employers will hire, according to the signal, the six best workers and assign to the Hard Task the two best workers if they believe that the signal relative to the Easy Task is also positively and significantly correlated with the performance in the Hard Task.

In our sample, the signal very well predicts the performance in the Easy Task (Pearson's $r=0.935$, $p\text{-value}=.000$) as well as the performance in the Hard Task ($r=0.298$; $p\text{-value}=.000$). Indeed, 86% of the six best workers according to the signal ended in the top-six in the Easy Task. Moreover, the signal and the performance in the Easy Task are equally correlated for male and for female workers (Fisher Z transformation, $Z= 0.57$; $p\text{-value}= 0.568$) evidencing that signal is equally predictive for male and female workers. The signal is also equally correlated with performance in Hard Task for male and female workers (Fisher Z transformation, $Z= -0.77$; $p\text{-value}= 0.441$), even if it has a lower correlation compared with that regarding the Easy Task. 60% of the six best workers according to the signal ended in the top-six in the Hard Task while only 22% of the two best workers according to the signal ended in the top-two of the Hard Task¹¹.

As a consequence, we expect that the lower is the correlation between the signal and the performance in a specific task, the more the employer's evaluation is based on other and more subjective criteria (Grabner & Moers, 2013; Heinz et al., 2016; Paryavi et al., 2019). Moreover, even when the signal is a significant predictor of performance, absolute performances in both tasks are not observable and employers may be guided by gender stereotypes when predicting which workers are the best performers, especially in the Hard Task, where we expect gender stereotyped to play a bigger role given the more advanced level of mathematical knowledge requested¹². For example, employers may believe that is less likely that a woman is among the best performers even when she has a similar signal than that of a male worker. In fact, we find that male workers performed better than female workers in

⁹ Workers from Experiment W who decided not to participate in the hiring game were excluded in the analysis.

¹⁰ Literature on hiring has shown that employers' decisions are conditioned by the signal of performance sent by workers. Azic & Lamé (2018) and Reuben et al. (2014) found that from pairs of workers, employers hired the worker with higher signal 70 and 80 per cent of the cases, respectively. Moreover, as discussed by Bohnet et al. (2015), employers are more likely to focus on individual performance signals in joint evaluations (i.e. when candidates are presented at the same time).

¹¹ We run a number of tobit regressions in Appendix C (Table C2) to more robustly check the correlation between the signal and the number of correct calculations in the Easy Task and the number of problems correctly solved in the Hard Task.

¹² See Spencer et al., (1999), Niederle & Vesterlund (2007), Good et al. (2008).

the Hard Task ($Z=-2,336$; $p\text{-value}=0.019$) but there are no significant gender differences in the signal¹³. It is important to recall that a potential gender discrimination in the task assignment do not necessarily respond to employers' correct priors about performance due to the gender differences in our sample. As stated above, the set of problems was constructed based on the questions that did not show gender differences in an independent, bigger sample of male and female students.

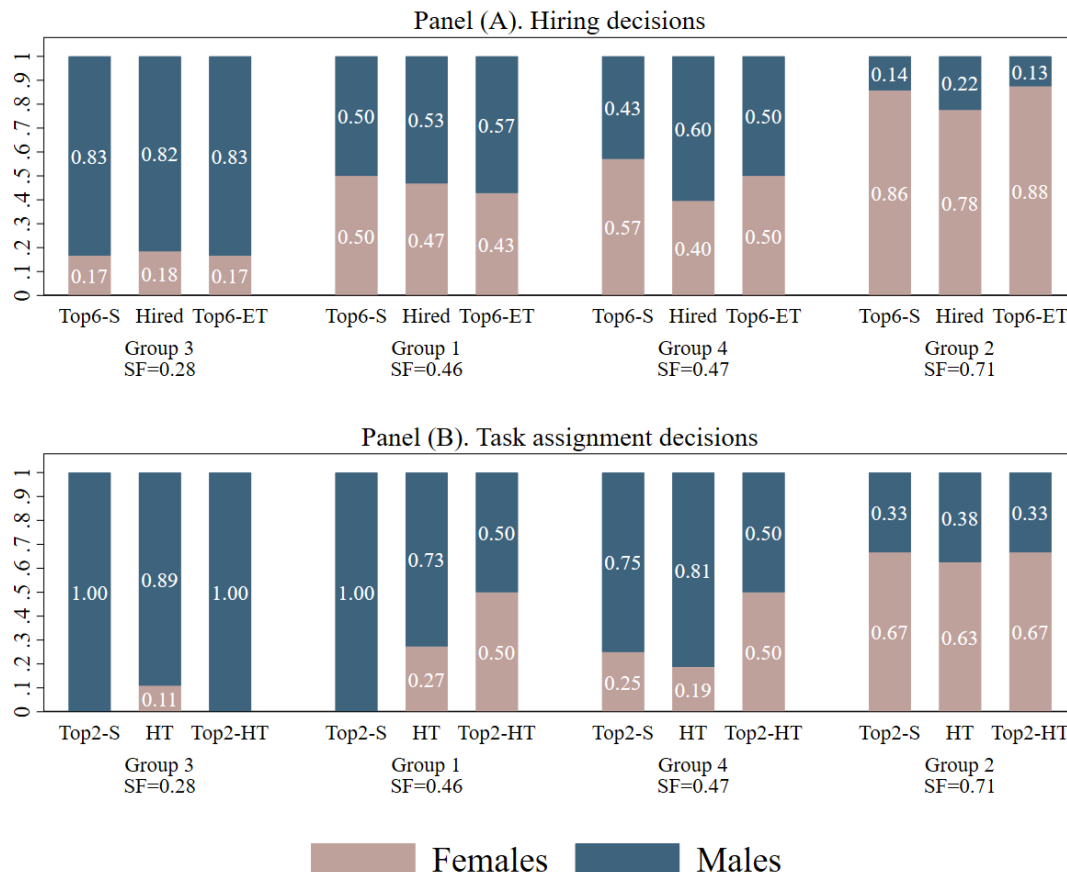
In our two-level analysis, we focus whether high-ability female workers have lower probabilities to being hired and assigned to the Hard Task compared to high-ability male workers. To this end, we first examine employers' decisions based on the observable information they hold. That is, establishing the level of ability based on the signal (ex-ante analysis). This ex-ante analysis studies the potential presence of statistical discrimination which emanates from the limited information that employers have about a specific group of interest (Guryan & Charles, 2013). According to the seminal works of Phelps (1972) and Arrow (1973), employers have to predict the future productivity of a specific candidate by weighting his/her signal of performance and their beliefs about the average performance of workers in the same gender group (Coate & Loury, 1993). And second, we establish the level of ability by means of the whole productivity in both Part 1 and Part 2 (ex-post analysis) in order to test how gender quotas affect the best performers and not the best signal-providers. Because employers' earnings depend on absolute productivities and they have incentives to allocate the best workers to the right position, it is important to understand the degree in which employers are accurate in their decisions. The denomination of "ex-post analysis" could be controversial since the experimental design do not ask workers to perform the tasks in which they have been allocated by the employers. However, we assume that workers would perform similarly in a hypothetical second round of the task. Finally, it is important to recall that establishing the ability according to the whole performance, potential gender differences could not respond to discrimination since employers do not have clues about performance in the Hard Task, but provide important evidence on how employers' beliefs operate.

Figure 1 provides a comparative analysis of the raw data of the Control Treatment. Panel (A) compares the proportion of males and females hired in the first stage with the proportion of males and females belonging to the top-six according to the signal (Top6-S) and according to the performance in the Easy Task (Top6-ET). Similarly, Panel (B) compares the proportion of females assigned to the Hard Task (HT) and the proportion of females that belonged to the top-two according to the signal (Top2-S) and according to the performance in the Hard Task (Top2-HT).¹⁴ It can be noted how, in panel A, the proportion of women hired more or less reflect both the proportion of women in the groups, and the

¹³ The distribution of performances by gender in both tasks can be found in the appendix C (Figure C2).

¹⁴ It is important to note that the top-six (top-two) of the signal can include more than six (two) workers due to the existence of ties. In Figure C1 in Appendix C we present the ranking of workers based on the signal and sorted by gender and group.

proportion of women who belong to the top six according to the performance in the Easy Task (Top6-ET), however, in panel B, the proportion of women assigned to the hard task is always lower than both, the proportion of women in the group, the proportion of women who belong to the top-two according to the performance in the Hard Task (Top2-HT).



Note: Panel (A): Top6-S: Proportion of males and females that are among the six best workers according to the signal. HT: proportion of males and females hired. Top6-ET: proportion of males and females that are among the six best workers according to the performance in the Easy Task.

Panel (B): Top2-S: Proportion of males and females that are among the two best workers according to the signal. HT: proportion of males and females assigned to the Hard Task. Top2-HT: proportion of males and females that are among the two best workers based on the performance in the Hard Task.

In both panels, groups of workers are ordered by the share of females (SF).

Figure 1. Hiring and Task Assignment decisions in the Control Treatment.

According to Figure 1, one could claim that female workers have less chances to be hired in the team and to be assigned to the Hard Task in the second step. However, the existence of gender discrimination in this environment is not so obvious. Therefore, in order to check whether exists discrimination we run several probit regressions controlling for every possible characteristics of the decision making's

environment. For the decisions in the first stage, Table 1 provides four specifications of the following equation:

$$\Pr(\text{Hired}_{ig} = 1) = \alpha + \beta_1 \text{Female}_{ig} + \beta_2 \text{Ability}_{ig} + \beta_4 \text{Ability} \times \text{Female}_{ig} + C'\zeta + u_E + \varepsilon_{ig} \quad (2)$$

The equation regresses the probability of worker i in group g of being hired ($\text{Hired}_{ig}=1$ if the worker is one of the six candidates hired and 0 otherwise) on gender ($\text{Female}_{ig}=1$ if the worker is female and 0 otherwise) and ability. In A different specification of equation (2) we include different measures of ability (Ability_{ig}). In Model 1 and Model 2 we include a dummy variable that equals 1 if the worker belongs to the six best workers according to the signal and 0 otherwise (Top6-S). In Model 3 and Model 4, our measure of ability is a dummy variable that equals 1 if the worker belongs to the best six workers according to the ex-post performance in the Easy Task and 0 otherwise (Top6-ET).

Table 1. Marginal effects, probit regressions on the probability that a worker is being hired in the Control Treatment.

	Dependent variable: Pr (Hired=1)			
	(1)	(2)	(3)	(4)
Female	-0.062 (0.057)	0.183** (0.085)	-0.152*** (0.052)	0.050 (0.080)
Top6-S	2.075*** (0.121)	2.331*** (0.164)		
Top6-S x Female		-0.545*** (0.138)		
Top6-ET			1.584*** (0.091)	1.779*** (0.115)
Top6-ET x Female				-0.392*** (0.108)
Female + Top6 x Female		-0.362***		-0.342***
Full set of controls ^a	Yes	Yes	Yes	Yes
Observations	3,584	3,584	3,584	3,584
N (employers)	64	64	64	64

^a Workers' Controls: Age, Field of study, Position in the pool; Employers' controls: Gender and beliefs; Group controls: Number of workers in the pool and Share of females in the group. In parentheses, robust standard errors clustered at employer level. All regressions contain employers' random effects. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

In models 1 and 3 we present a baseline estimation which isolates the effect of gender and ability. In models 2 and 4 we add an interaction between gender and the corresponding measure of ability ($\text{Ability} \times \text{Female}_{ig}$) in order to check gender differences in hiring among high-ability workers. In all specifications of the equation, we include a vector of covariates ($C'\zeta$) that contains three different sets

of controls: 1) Workers' controls: a dummy variable for fields of study (Sciences, Social Sciences and Engineering, with the omitted variable being Foreign Languages), age and the position in the screen¹⁵ in which the info about the workers are displayed; 2) Employers' controls: gender and beliefs in favor of male workers¹⁶; and 3) Group's controls: order in which the groups were presented to the employer, the number of workers in the group and the share of females in the group. Regressions are estimated including employers' random effects (u_E) and clustering standard errors at employer level (ε_{ig}).

In model (1), the dummy Female has a negative effect on the probability of workers to be hired for the final composition of the team, but the effect is not significant, while being among the six best workers according to the signal significantly increases the probability of being hired.

In Model (2), where the interaction between the signal of ability and the gender is included, splits the effect by ability. In this model, it can be noted that while the coefficient associated to the measure of ability has the same effect as in Model (1), the coefficient associated to the dummy Female turns positive and significant, indicating that low-ability females according to the signal have more probabilities to be hired compared to low-ability males. The interaction between gender and ability presents the opposite effect, meaning that the gender gap (in favor of females) is lower in comparison to the gender gap in low-ability workers. The linear combination of the coefficients Female and the interaction Top6-S x Female in Model (2) calculates the gender gap among high-ability workers. The negative and significant effect suggests that high-ability female workers have lower probabilities of being hired compared to high-ability male workers.¹⁷

In model (3), where we use as measure of ability the whole performance in the Easy Task, being among the six best workers in the Easy Task increases the probability of being hired but being a female worker decreases it significantly.

In model (4), the coefficient for Female is not significant, meaning that there are no gender differences in the probability of being hired among low-ability workers. On the other hand, the negative coefficient of the interaction between gender and ability shows that the gender gap among high-ability workers is lower in comparison to that among low-ability workers. Similar to model (2), the linear combination between Female and the interaction remains negative and significant, meaning that high-ability females have lower chances of begin recruited in comparison to high-ability males. These results justify the introduction of a gender quota to balance the number of males and females in the organizations.

¹⁵ The information of about workers was displayed at the same time and every worker occupied the same position in the screen for all employers.

¹⁶ We asked employers to estimate the average productivity of males and females in a given task at the end of Part 3. We define the beliefs in favor of males if an employer believed that males, on average, outperformed females in a given task.

¹⁷ Significance extracted from Wald tests.

Result 1: In the Control treatment, females have lower probabilities to be hired compared to males. High-ability females have lower probabilities to be hired compared to high-ability males, controlling ability either by the signal provided or the performance in the Easy Task.

In Table 2, we estimate similar specifications of equation (2) but regressing the probability of being assigned to the Hard Task ($HT_{ig}=1$ if the worker is assigned to the Hard Task, 0 otherwise) in order to study employers' decisions in the task assignment stage. In this case, the measures of ability differ from the previous models. We substitute Top6-S by a dummy variable that equals 1 if the worker is among the top-two workers based on the signal and 0 otherwise (Top2-S) and Top6-ET by a dummy variable that equals 1 if the workers is among the two best workers based on the performance in the Hard Task and 0 otherwise (Top2-HT). We also add the share of females hired in the first stage by each employer in the set of controls. Clusters and random effects remain as in the previous estimations.

Table 2. Marginal, effects, probit regressions on the probability of being assigned to the Hard Task in the second stage in the Control Treatment.

	Dependent variable: Pr (HT=1)			
	(1)	(2)	(3)	(4)
Female	0.134*	0.199**	-0.215***	-0.090
	(0.071)	(0.082)	(0.069)	(0.064)
Top2-S	1.074***	1.149***		
	(0.119)	(0.126)		
Top2-S x Female		-0.252		
		(0.172)		
Top2-HT			0.139*	0.350***
			(0.077)	(0.100)
Top2-HT x Female				-0.549***
				(0.166)
Female + Top2 x Female		-0.053		-0.638***
Full set of controls ^a	Yes	Yes	Yes	Yes
Observations	3,584	3,584	3,584	3,584
N (employers)	64	64	64	64

^a Workers' Controls: Age, Field of study, Position in the pool; Employers' controls: Gender, beliefs and share of females hired in the first stage; Group controls: Number of workers in the pool and Share of females in the group. In parentheses, robust standard errors clustered at employer level. All regressions contain employers' random effects. *** p<0.01, ** p<0.05, * p<0.1.

In model (1), being among the best two workers according to the signal significantly increases the probability of being assigned to the Hard Task. The coefficient Female presents a positive and

significant effect, meaning that, overall, females have more chances than males to be assigned to the Hard Task.

In model (2), when splitting the effect of gender by ability, we find similar effects for Female and Top2-S with respect to model (1). In this case, the positive effect of Female suggests that low-ability females, according to the signal, are more likely to be allocated into the more complex task. The interaction between gender and Top2-S shows a negative effect, meaning that the gender gap among high-ability workers (in favor of females) is lower than the gender gap among low-ability workers. However, this effect is not significant. Moreover, the non-significant effect of the linear combination between Female and the interaction suggests that high-ability workers, according to the signal, both males and females have equal chances of being assigned to the Hard Task. These models show that employers do not discriminate females in the task assignment based on the observable information.

When, in models (3) and (4) we include the whole performance as measure of ability, the estimations look very different. In Models (3) being a female worker decreases the probability of being assigned to the Hard Task while the coefficient for the performance has a positive and significant effect.

In model (4), the coefficient of female has a negative but not significant effect, while the interaction shows a negative and significant effect. That is, the differences between the gender gap is lower among high-ability workers. In fact, the linear combination between Female and the interaction suggests that high-ability females, according to the whole productivity in the Hard Task, have less chances to be assigned to more complex tasks compared to high-ability male workers¹⁸.

Result 2:

- In the task assignment, based on the observable information, high-ability females are not discriminated against.
- Females that would be categorized as high-ability, ex-post, have lower probabilities to be assigned to Hard Task compared to high-ability males.

¹⁸ These results are consistent with those observed in similar lab experiments. For gender discrimination in hiring see examples in Bohnet et al. (2015), Reuben et al. (2014), Beaurain & Masclet (2016), Leibbrandt et al. (2017), Azic & Lamé (2018), Coffman et al. (2018). Conversely, Charness et al. (2018) found that females are not discriminated against. In fact, the authors found a slight preference for female workers. For gender discrimination in promotion and assignment to more competitive tasks see Peterle & Rau (2017). On the other hand, we do not find a significant effect of employers' gender neither on hiring decisions nor task assignment decisions. These results are also consistent with other studies assessing the role of evaluators' gender (Bertrand et al. 2014; Reuben et al., 2014; Booth and Leigh 2010; Bohnet et al., 2015; Kunze and Miller 2017, Sandberg 2017). Moreover, we did not find a significant effect of employers' bias.

3.2. The effect of gender quotas on task assignment decisions

In the Quota Treatment, females represented, on average, the 62 % of the participants and the 54% of the composition of the firms (>46%; p-value=.540)¹⁹. However, as in the Control Treatment, the decisions of the employers will take into account several characteristics other than gender. In this subsection, we consider the data from both Treatments to analyze the effect of the gender quota on the task assignment decisions. The set of regressions that confirm that the gender quota, by definition, increases the number of female workers in the firms can be found in the Online Appendix C (Table C3).

$$\Pr(HT_{ig} = 1) = \alpha + \beta_1 Female_{ig} + \beta_2 Treatment_{ig} + \beta_3 Female \times Treatment_{ig} + C'\zeta + u_E + \varepsilon_{ig} \quad (3)$$

Table 3. Treatment effect: Marginal effects, probit regressions on the Probability of being assigned to Hard Task in the second stage on high-ability workers.

	Dependent variable: Pr (HT=1)			
	Top6-S		Top6-HT	
	(1)	(2)	(3)	(4)
Female	-0.222*** (0.065)	-0.349*** (0.114)	-0.371*** (0.070)	-0.199* (0.114)
Treatment	-0.035 (0.042)	-0.131* (0.078)	0.129*** (0.039)	0.224*** (0.057)
Female x Treatment		0.201 (0.138)		-0.276** (0.136)
Female + Female x Treatment		-0.148**		-0.475***
Full set of controls ^a	Yes	Yes	Yes	Yes
Observations	3,648	3,648	3,392	3,392
N (employers)	128	128	128	128

^a Workers' Controls: Age, Field of study, Position in the pool; Employers' controls: Gender, beliefs and share of females hired in the first stage; Experimental controls: order of the presentation of the groups; Group controls: Number of workers in the pool and Share of females in the group. In parentheses, robust standard errors clustered at employer level. All regressions contain employers' random effects. *** p<0.01, ** p<0.05, * p<0.1.

In Table 3, we estimate different specifications of the equation (3) on the sample of high-ability workers²⁰. We employ two measures as a threshold of ability. First, for the ex-ante analysis, we consider the six best workers according to the signal as high-ability workers. Second, for the ex-post analysis, we

¹⁹ The comparative analysis of the raw data of the Quota Treatment can be found in the Appendix C (Figure C3).

²⁰ In Appendix C, similar regressions relative to the low-ability workers can be found in Table C4 and Table C5.

consider as high-ability those workers who ended among the six best workers according to the performance in the Hard Task. We use the six-best workers instead of the two best workers, as shown in the regressions in the previous subsection, to observe how the treatment would affect to those workers who are supposed to be part of the organization and not only those who are supposed to be assigned to the Hard Task.

First, the equation regresses the probability of being assigned to the Hard Task (HT_{ig}) in the second stage on gender ($Female_{ig}$) and a dummy variable that equals 1 if the treatment is the Quota Treatment and 0 if the treatment is the Control Treatment ($Treatment_{ig}$). Additionally, we include an interaction between Female and Treatment ($Female \times Treatment_{ig}$). Similar to previous regressions, we include the full set of controls, employers' random effects and cluster standard errors at employer level.

In model (1), Female has a significant and negative effect while Treatment has no effect on the probability of high-ability workers, according to the signal, of being assigned to the Hard Task.

In model (2), where the effect of gender is splitted by treatment, the coefficient of Female, that captures the gender gap in the control condition, is negative and significant. That is, high-ability females in the absence of quotas have lower chances than high-ability male workers to be assigned to the Hard Task. The coefficient of Treatment, that shows the effect of the quota in high-ability males in comparison to high-ability females is positive and significant, although only at 10 percent level of significance. The interaction in model (2) shows a positive but non-significant effect, meaning that the treatment does not change probabilities of being assigned to the Hard Task of high-ability females. The negative and significant effect of linear combination of the effects of Female and the interaction suggests that, similar to the control, high-ability females have lower chances of assignment to the more complex task in comparison to high-ability male workers in the presence of quotas.

In model (3), when the threshold of ability is set according to the whole performance in the Hard Task, Female has a negative, significant effect while Treatment shows a positive and significant effect. In model (4), where the effect of the gender is splitted by treatment, Female presents a negative and significant effect, meaning that (ex-post) high-ability females have lower probabilities to be assigned to the Hard Task compared to high-ability males in the control treatment. In this model, conversely to model (2), Treatment is positive and significant. That is, in comparison to high-ability females, the treatment exerts a positive effect on high-ability males. The negative and significant effect of the interaction between Female and Treatment, that captures the gender gap in the quota treatment in comparison to that in the control treatment, shows the females have lower probabilities of being assigned to the Hard Task in comparison to high-ability females in the control treatment. The linear combination between Female and the interaction confirm that females are less successful than males

in the task assignment. In sum, when controlling by the whole performance as a measure of ability, the treatment increases the chances of high-ability males, expanding the gender gap.

On the other hand, it is also worthy to evaluate the treatment effect on high-ability males. As noticed in models (2) and (4), the sign of the variable treatment changes as we modify the level of ability. The number of male workers among the six best workers according to signal that ended in the top six of workers in the Hard Task may explain this event. While 74% percent of male workers who were among the best workers according to the signal were among the best workers in the Hard Task, only 46% of female workers did so. This issue seems to increase the probabilities of these male workers to be assigned to the Hard Task since employers have to make less inferences about their whole performance compared to female workers.

Result 3:

- The quota has not a significant effect on high-ability females, when ability is measured by the signal.
- Ex-post, high-ability females under gender quotas have lower probabilities to be assigned to the Hard Task compared to high-ability females in the absence of quotas.

3.3. Employers' Performance.

In this subsection, we look at the earnings of the employers and observe how the gender quota affects the outcome of the decisions. Employers in the Control Treatment (Mean=15.89; SD=2.19) obtained significantly higher earnings than employers in the Quota Treatment (M=14.71; SD=2.21) (Mann-Whitney $Z= 5.470$; $p\text{-value}=0.000$). This approach not only shows what employers obtained according to their decisions, it also provides an understanding of the degree of mistakes they did in selecting and assigning workers to the different tasks suggesting that there were more mistakes in the Quota Treatment compared to the Control Treatment. Nevertheless, the absolute value of the earnings is not an accurate measure of employers' performance. In this sense, we develop a different measure to examine the accuracy of employers' decisions under every Treatment. To better investigate this aspect, we estimate each employers' percentage deviation from the optimal earnings where we define optimal earning those earning that employers could obtain in a given group if they would had assigned the right worker to the right position (identified by looking at the ex-post performances in the Easy Task and the Hard Task). Since employers evaluated four groups of workers, we calculated the average percentage deviation in each group at the individual level. Specifically, the optimal earnings in each group are estimated building two rankings: a ranking according to their performance in the Hard Task

(Ranking HT) and another ranking according to their performance in the Easy Task (Ranking ET). In Ranking HT, ties were broken based on the following premise: the worker with lower productivity in the Easy Task (lower ranking in Ranking ET) had higher ranking in Ranking HT. This premise was built in order to maximize the productivity of workers that will be assigned to the Easy Task. Ties in Ranking ET are broken randomly. The algorithm assigns the two workers with higher Ranking HT to the Hard Task and the four best workers according to Ranking ET to the Easy Task, excluding those workers already assigned to the Hard Task²¹. In the Quota Treatment, if less than 3 females have been hired, the algorithm replaces in the assignment to the Easy Task the male(s) with lower rank by the female(s) that are not among the four best workers in the Easy Task with higher ranking. Once the algorithm has identified who are the right workers for each position, we calculate the optimal earning following equation (1)²². The percentage deviation of each employer (i) in each group (g) is calculated with the following formula:

$$Deviation_{ig} = abs \left(\frac{earnings_{ig} - optimal\ earning_g}{optimal\ earning_g} \right) * 100 \quad (4)$$

Note that we present the percentage deviation in absolute values for making the reading easier. Table 4 provides the average percentage deviation from the optimal earnings for both the Easy and the Hard Task and by treatment²³. The highest average deviation is found in the assignment to the Hard Task in both Treatments, for which it lies above 25 percent with respect to the optimal earnings. That is, employers make more mistakes in their decisions when predicting performance in the Hard Task regardless the treatment they are allocated. This fact can be explained by the lower correlation that the signal provides relative to the performance in the Hard Task compared to the correlation with the performance in the Easy Task, that induce more subjectivity in the decisions. Overall, the Quota Treatment pushes employers to significantly make more mistakes compared to the Control Treatment especially in the assignment to the Easy Task. This result translates into the final average earnings. Even if the quota has positive effect for the female workers, the policy is not optimal in the sense that

²¹ As an illustration, we consider the example in Group 1 of Control Treatment. Worker 1 obtained 13 correct answers and Worker 2, Worker 3 and Worker 4 solved correctly 10 answers in the Hard Task. Worker 1 is directly assigned to the Hard Task. Worker 2 and Worker 3 solved 19 correct calculations in the Easy Task. Worker 4 only solved 9 correct calculations. Then, the triple tie is broken in favour of Worker 4, that is also assigned to the Hard Task. Excluding then Worker 1 and Worker 4 for the assignment of the Easy Task, Worker 2, Worker 3 and Worker 5 (19 correct calculations) and Worker 6 (14 correct calculations) were assigned to the Easy Task.

²² Optimal productivities and optimal earnings for each task calculated with the algorithm can be found in Appendix C (Table C6).

²³ The average percentage deviation from the optimal earning in each group and treatment can be found in Appendix C (Figure C4).

it does not make employers to make more optimal decisions. In fact, employers make more mistakes and earn less in the presence of quotas.

Table 4. Average percentage deviation from the optimal earnings.

Treatment	Easy Task	Hard Task	Total
Control Treatment	9.5 (0.061)	25.9 (0.070)	20.2 (0.052)
Quota Treatment	12.9 (0.055)	27.5 (0.085)	22.6 (0.059)
Difference (Mann-Whitney Z)	-3.650	-1.187	-2.311
p-value	0,000	0,235	0,020

Note: Average deviations presented in absolute values. Each value presents the average deviation from the optimal earnings of the employers. Each employer evaluated four groups, then the percentage deviation of each employer is the average percentage deviation in all groups evaluated calculated at the individual level. Standard deviation in parenthesis.

Result 4: Employers make more mistakes and earn less in presence of gender quotas.

4. Conclusions

In this paper, we aim to study the effect of gender quotas on task assignment decisions in organizations. Our findings contribute to the literature by including the task assignment in the debate about the mechanisms affecting the gender wage gap and the underrepresentation of females in leadership positions. We propose a laboratory experiment to address this question. We asked employers to hire workers to conform a team and then assign them to different tasks: An Easy Task, that is less complex and less profitable, and a Hard Task, more complex and profitable than the Easy Task for both the worker and the employer. We observed signs of gender discrimination in the hiring stage. This result shows that the use of affirmative action policies such as gender quotas are needed for increasing the chances of success of female workers. We also find that quotas have positive effects. The quota increases the number of females in the organization by definition, but it serves to increase the probabilities of high-ability females to be hired. In contrast, we find that quotas have a negative effect on task assignment for female workers. Under gender quotas, high-ability females have lower probabilities to be appointed for the more complex task compared to those female workers in the setting in which there is no policy intervention. The results suggest two important findings. First, males and females are not treated equally in the time they are allocated into different tasks within the organization, what could explain a proportion of the gender gap in wages and leadership. And second, gender quotas are not sufficient in this dimension. In fact, this policy exerts a negative effect on high-ability females

to be assigned to task that are more profitable and with more responsibility. It could be that the reverse discrimination produced by the quota against high-ability male workers could produce a perception of injustice of the policy and the subsequent backlash against females (Leibbrandt et al., 2017). On the other hand, we observe that quotas have additional negative effects. Employers perform worse under gender quotas, contrary to the results in Beaurain & Masclet (2016) in which firm performance is not affected by the policy. In other terms, under gender quotas, employers make more mistakes in selecting the best workers and consequently gain less than with no policy intervention. In sum, the empirical evidence has shown that gender quotas have a number of positive effects, especially on the side of the supply. However, they are ineffective in helping to close the gender gap in career advancement. Then, new policy designs are needed to complement the actual affirmative action policies.

References

- Arrow, K. (1973). The theory of discrimination. *Discrimination in labor markets*, 3(10), 3-33.
- Azic, M., & Lamé, D. (2018). Subjective information in hiring decisions.
- Babcock, L. and S. Laschever (2003). *Women don't ask: Negotiation and the gender divide*. Princeton, NJ: Princeton University Press.
- Babcock, L., Gelfand, M., Small, D., & Stayn, H. (2006). "Gender differences in the propensity to initiate negotiations." In D. D. Crèmer, M. Zeelenberg, & J. K. Murnighan (Eds.), *Social psychology and economics* (pp. 239–259). Mahwah, NJ: Lawrence Erlbaum.
- Babcock, L., Recalde, M. P., & Vesterlund, L. (2017b). Gender Differences in the Allocation of Low-Promotability Tasks: The Role of Backlash. *American Economic Review*, 107(5), 131-35.
- Babcock, L., Recalde, M. P., Vesterlund, L., & Weingart, L. (2017a). Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review*, 107(3), 714-47.
- Bagues, M., Sylos-Labini, M., & Zinovyeva, N. (2017). Does the gender composition of scientific committees matter? *American Economic Review*, 107(4), 1207-38.
- Balafoutas, L., & Sutter, M. (2012). Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science*, 335 (6068), 579-582.
- Beaman, L., Chattopadhyay, R., Duflo, E., Pande, R., & Topalova, P. (2009). Powerful women: does exposure reduce bias? *The Quarterly journal of economics*, 124(4), 1497-1540.
- Beaurain, G., & Masclet, D. (2016). Does affirmative action reduce gender discrimination and enhance efficiency? New experimental evidence. *European Economic Review*, 90, 350-362.

- Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, 94(4), 991-1013.
- Bertrand, M., Black, S. E., Jensen, S., & Lleras-Muney, A. (2018). Breaking the glass ceiling? The effect of board quotas on female labour market outcomes in Norway. *The Review of Economic Studies*, 86(1), 191-239.
- Bertrand, M., Black, S. E., Jensen, S., & Lleras-Muney, A. (2018). Breaking the glass ceiling? The effect of board quotas on female labour market outcomes in Norway. *The Review of Economic Studies*, 86(1), 191-239.
- Bertrand, M., Goldin, C. and Katz, L. F. (2010), “Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors”, *American Economic Journal: Applied Economics*, 2, 228–255.
- Bohnet, I., Van Geen, A., & Bazerman, M. (2015). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, 62(5), 1225-1234.
- Booth, A., & Leigh, A. (2010). Do employers discriminate by gender? A field experiment in female-dominated occupations. *Economics Letters*, 107(2), 236-238.
- Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, 129 (3), 1409-1447.
- Calsamiglia, C., Franke, J., & Rey-Biel, P. (2013). The incentive effects of affirmative action in a real-effort tournament. *Journal of Public Economics*, 98, 15-31.
- Charness, G., Cobo-Reyes, R., & Sanchez, Á. (2018). Anticipated discrimination, choices, and performance: experimental evidence.
- Coate, S., & Loury, G. C. (1993). Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, 1220-1240.
- Coffman, K. B., Exley, C. L., & Niederle, M. (2018). *When gender discrimination is not about gender*. Harvard Business School.
- Correll, S. J., Benard, S., & Paik, I. (2007). Getting a job: Is there a motherhood penalty? *American journal of sociology*, 112 (5), 1297-1338.
- Crittenden, Ann (2001). *The Price of Motherhood: Why the Most Important Job in the World Is Still the Least Valued*. New York: Metropolitan Books.
- Crosetto, P., & Filippin, A. (2013). The “bomb” risk elicitation task. *Journal of Risk and Uncertainty*, 47(1), 31-65.
- Czibor, E. and Dominguez-Martinez, S., (2018), Gender Quotas in a Multistage Tournament. Available at SSRN: <https://ssrn.com/abstract=3037421> or <http://dx.doi.org/10.2139/ssrn.3037421>.
- De Pater, I. E., Van Vianen, A. E., & Bechtoldt, M. N. (2010). Gender differences in job challenge: A matter of task allocation. *Gender, Work & Organization*, 17(4), 433-453.

- Dohmen, T., & Falk, A. (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review*, 101 (2), 556-90.
- European Union (2019). "Report on equality between women and men in the EU". ISBN: 978-92-76-00027-3 doi: 10.2838/395144.
- Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, 10(2), 171-178.
- Flory, J. A., Leibbrandt, A., & List, J. A. (2014). Do competitive workplaces deter female workers? A large-scale natural field experiment on job entry decisions. *The Review of Economic Studies*, 82(1), 122-155.
- Freidenvall, L., & Dahlerup, D. (2013). Electoral gender quota systems and their implementation in Europe: Update 2013. European Union.
- Good, C., Aronson, J., & Harder, J. A. (2008). Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses. *Journal of applied developmental psychology*, 29(1), 17-28.
- Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in personality*, 37(6), 504-528.
- Grabner, I., & Moers, F. (2013). Managers' choices of performance measures in promotion decisions: An analysis of alternative job assignments. *Journal of Accounting Research*, 51(5), 1187-1220.
- Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1), 114-125.
- Guryan, J., & Charles, K. K. (2013). Taste-based or statistical discrimination: the economics of discrimination returns to its roots. *The Economic Journal*, 123(572), F417-F432.
- Harris, P. B., and Houston, J. M. (2010). A reliability analysis of the revised competitiveness index. *Psychological reports*, 106(3), 870-874.
- Heinz, M., Normann, H. T., & Rau, H. A. (2016). How competitiveness may cause a gender wage gap: Experimental evidence. *European Economic Review*, 90, 336-349.
- Houston, J., Harris, P., McIntire, S., and Francis, D. (2002). Revising the competitiveness index using factor analysis. *Psychological Reports*, 90(1), 31-34.
- Ibanez, M., & Riener, G. (2018). Sorting through affirmative action: Three field experiments in Colombia. *Journal of Labor Economics*, 36 (2), 437-478.
- Kunze, A., & Miller, A. R. (2017). Women helping women? Evidence from private sector data on workplace hierarchies. *Review of Economics and Statistics*, 99(5), 769-775.
- Leibbrandt, A., & List, J. A. (2014). Do women avoid salary negotiations? Evidence from a large-scale natural field experiment. *Management Science*, 61 (9), 2016-2024.
- Leibbrandt, A., Wang, L. C., & Foo, C. (2017). Gender quotas, competitions, and peer review: Experimental evidence on the backlash against women. *Management Science*, 64(8), 3501-3516.

- Maggian, V., Montinari, N., & Nicolò, A. (2019). Do quotas help women to climb the career ladder? a laboratory experiment.
- Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, 122 (3), 1067-1101.
- Niederle, M., & Vesterlund, L. (2008). Gender differences in competition. *Negotiation Journal*, 24 (4), 447-463.
- Niederle, M., Segal, C., & Vesterlund, L. (2013). How costly is diversity? Affirmative action in light of gender differences in competitiveness. *Management Science*, 59 (1), 1-16.
- Paryavi, M., Bohnet, I., & van Geen, A. (2019). Descriptive norms and gender diversity: Reactance from men.
- Peterle, E., & Rau, H. A. (2017). Gender differences in competitive positions: Experimental evidence on job promotion.
- Phelps, E. S. (1972). The statistical theory of racism and sexism. *The American Economic Review*, 62(4), 659-661.
- Reuben, E., Rey-Biel, P., Sapienza, P., & Zingales, L. (2012). The emergence of male leadership in competitive environments. *Journal of Economic Behavior & Organization*, 83(1), 111-117.
- Reuben, E., Sapienza, P., & Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 111(12), 4403-4408.
- Reuben, E., Sapienza, P., & Zingales, L. (2015). Taste for competition and the gender gap among young business professionals (Tech. Rep.). National Bureau of Economic Research.
- Sandberg, A. (2017). Competing identities: a field study of in-group bias among professional evaluators. *The Economic Journal*, 128(613), 2131-2159.
- Small, D. A., Gelfand, M., Babcock, L., Gettman, H., (2007). "Who goes to the bargaining table? The influence of gender and framing on the initiation of negotiation." *Journal of Personality and Social Psychology* 93 (4), 600-613.
- Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of experimental social psychology*, 35(1), 4-28.
- Sutter, M., Glatzle-Rutzler, D., Balafoutas, L., & Czermak, S. (2016). Cancelling out early age gender differences in competition: an analysis of policy interventions. *Experimental Economics*, 19 (2), 412-432.

Appendix A. Hard Task's problems, Screenshots and Instructions.

Hard Task's problems

Quale retta passa per l'origine e per (2, - 4)?

- A. $y = -1/2x$
- B. $y = 1/2x$
- C. $y = -12x + 24$
- D. $y = -2x$

Dati i due polinomi $(a^2 + b - 3)$ e $(4 - b)$, il loro prodotto è uguale a:

- A. $4a^2 + ab + b - b^2 - 12$
- B. $4a^2 - a^2b + 7b - b^2 - 12$
- C. $4a^2 + a^2b + 7b - b^2 + 12$
- D. $-4a^2 - a^2b + b + b^2 - 12$

Risolvere la seguente equazione $x(x - 1) = 1 - x$:

- A. Non ha soluzioni reali
- B. Ha infinite soluzioni reali
- C. Ha come soluzione esclusivamente $x = 1$
- D. Ha come soluzioni $x = 1$ oppure $x = -1$

$2y + 3 = 3x - 5$ passa per il punto:

- A. $(1, 5/2)$
- B. $(1, -5/2)$
- C. $(0, 0)$
- D. $(-1, -5/2)$

La decima parte di 10^{14} equivale a:

- A. 10^7
- B. 10^{13}
- C. $10^{1.4}$
- D. 10^4

$2y + 3 = 3x - 5$ passa per il punto:

- A. $(1, 5/2)$
- B. $(1, -5/2)$
- C. $(0, 0)$
- D. $(-1, -5/2)$

La disequazione $5x > -2$ è equivalente (ha lo stesso insieme di soluzioni reali) alla disequazione:

- A. $5x > 2$
- B. $5x < 2$
- C. $-5x < 2$
- D. $-5x > 2$

Trovare le soluzioni della disequazione $(x + 2)^2 - 2x < x^2 - 4x - 3$

- A. $x > 0$
- B. $x < -3$
- C. $x > -7/6$
- D. $x < -7/6$

Indicato con P l'insieme dei numeri primi, indica quali delle seguenti relazioni è corretta.

- A. $10 \in P$
- B. $25 \in P$
- C. $5 \notin P$
- D. $23 \in P$

La disequazione $x > -(7x - 4)$ ha per soluzione:

- A. $x < -1/2$
- B. $x > -1/2$
- C. $x < 1/2$
- D. $x > 1/2$

Il polinomio $x^2 - 7x + 6$ può essere scomposto nei seguenti fattori:

- A. $(x - 1)(x + 6)$
- B. $(x + 1)(x - 6)$
- C. $(x - 1)(x - 6)$
- D. $(x + 1)(x + 6)$

La scomposizione in fattori primi del polinomio $3a^2 - 6ab + 3ac + 3a - 6b + 3c$ è:

- A. $(a + 1)(3a - 6b + 3c)$
- B. $-3(a + 1)(a + 2b + c)$
- C. $3(a - 1)(a - 2b + c)$
- D. $3(a + 1)(a - 2b + c)$

Calcolare: $\sqrt{(3/2)^{-2} + (1/2)^2}$

A. $\frac{7}{6}$

B. $\frac{\sqrt{5}}{3}$

C. $\frac{1}{5}$

D. $\frac{5}{6}$

La rappresentazione grafica della funzione $y = (-2x + 10)^2$:

- A. E' una parabola con la concavità rivolta verso il basso e che è passante per il punto (0,10)
- B. E' una parabola con la concavità rivolta verso l'alto e che è tangente all'asse delle x
- C. E' una parabola passante per l'origine
- D. E' una parabola che ha per asse di simmetria l'asse delle ordinate

Il m.c.m. tra i polinomi $6(x - 1)^2$ e $2(x^2 - 1)$ è:

- A. $2(x - 1)^2 (x^2 - 1)$
- B. $6(x - 1)^2 (x + 1)$
- C. $(x + 1)^2 (x - 1)$
- D. $3(x - 1)(x + 1)$

Eeguire la razionalizzazione di $\frac{1}{\sqrt{3} \cdot (\sqrt{5} - \sqrt{3})}$:

A. $\frac{\sqrt{15}+3}{6}$

B. $\frac{\sqrt{3}}{\sqrt{3}-\sqrt{5}}$

C. $\frac{5\sqrt{3}}{3}$

D. $\sqrt{3} + \sqrt{5}$

La disequazione $2^{x^2-5x+6} > 1$ è soddisfatta per ogni numero reale x tale che:

- A. $2 < x < 3$
- B. $x < 2$ oppure $x > 3$
- C. $x < 3$
- D. $x > 2$

Consideriamo gli insiemi A e B e $C = A \cup B$. Quali delle seguenti affermazioni è falsa?

- A. Se a appartiene ad A, a appartiene a C.
- B. Se b appartiene a B, b appartiene a C.
- C. Se c appartiene a C, c appartiene ad A o a B.
- D. Se c appartiene a C, c appartiene a B e non ad A.

Il M.C.D. tra i polinomi $(x - 1)^3$ e $(x^2 + 1)^2$ è:

- A. $(x - 1)^2(x^2 - 1)$
- B. $(x - 1)^2(x + 1)$
- C. $(x + 1)^2$
- D. 1

$\log 4 + \log 10$ è uguale a:

- A. $\log 7 + 2$
- B. $\log 14$
- C. $3\log 2 + \log 5$
- D. $(2 + \log 3)\log 2$

$\frac{1-\sqrt{3}}{1+\sqrt{3}}$ vale:
















- A. $-2 + 2\sqrt{3}$
- B. $3 + \sqrt{2}$
- C. $\sqrt{2} - 3$
- D. $\sqrt{3} - 2$

Razionalizzare: $\frac{5}{(\sqrt{5}-\sqrt{3})}$

- A. $(3/2)(\sqrt{5} + \sqrt{3})$
- B. $(\sqrt{5} - \sqrt{3})$
- C. $(5/2)(\sqrt{5} + \sqrt{3})$
- D. $(5/2)(\sqrt{8})$

Screenshots







Remaining Time [sec]: 116

 ID 1 Anno di Nascita 1997 Area di studio SCIENZE Segnale 5 <input type="checkbox"/> ASSUMI	 ID 2 Anno di Nascita 1993 Area di studio SCIENZE SOCIALI Segnale 6 <input type="checkbox"/> ASSUMI	 ID 3 Anno di Nascita 1993 Area di studio SCIENZE Segnale 2 <input type="checkbox"/> ASSUMI	 ID 4 Anno di Nascita 1995 Area di studio SCIENZE Segnale 3 <input type="checkbox"/> ASSUMI	 ID 5 Anno di Nascita 1997 Area di studio SCIENZE Segnale 3 <input type="checkbox"/> ASSUMI
 ID 6 Anno di Nascita 1997 Area di studio SCIENZE Segnale 2 <input type="checkbox"/> ASSUMI	 ID 7 Anno di Nascita 1999 Area di studio SCIENZE Segnale 8 <input type="checkbox"/> ASSUMI	 ID 8 Anno di Nascita 1996 Area di studio INGEGNERIA Segnale 5 <input type="checkbox"/> ASSUMI	 ID 9 Anno di Nascita 1997 Area di studio SCIENZE SOCIALI Segnale 7 <input type="checkbox"/> ASSUMI	 ID 10 Anno di Nascita 1992 Area di studio SCIENZE Segnale 8 <input type="checkbox"/> ASSUMI
 ID 11 Anno di Nascita 1995 Area di studio SCIENZE SOCIALI Segnale 8 <input type="checkbox"/> ASSUMI	 ID 12 Anno di Nascita 1995 Area di studio SCIENZE SOCIALI Segnale 6 <input type="checkbox"/> ASSUMI	 ID 13 Anno di Nascita 1995 Area di studio SCIENZE SOCIALI Segnale 4 <input type="checkbox"/> ASSUMI	 ID 14 Anno di Nascita 1995 Area di studio SCIENZE SOCIALI Segnale 4 <input type="checkbox"/> ASSUMI	 ID 15 Anno di Nascita 1993 Area di studio INGEGNERIA Segnale 9 <input type="checkbox"/> ASSUMI

Devi assumere 6 lavoratori per la tua impresa

Figure A1. Hiring stage

Remaining Time [sec]: 64

 ID 1 Anno di Nascita 1997 Area di studio SCIENZE Segnale 5 <input type="checkbox"/> TASK 2	 ID 2 Anno di Nascita 1993 Area di studio SCIENZE SOCIALI Segnale 6 <input type="checkbox"/> TASK 2	 ID 3 Anno di Nascita 1993 Area di studio SCIENZE Segnale 2 <input type="checkbox"/> TASK 2	 ID 4 Anno di Nascita 1995 Area di studio SCIENZE Segnale 3 <input type="checkbox"/> TASK 2	 ID 5 Anno di Nascita 1997 Area di studio SCIENZE Segnale 3 <input type="checkbox"/> TASK 2
 ID 6 Anno di Nascita 1997 Area di studio SCIENZE Segnale 2 <input type="checkbox"/> TASK 2				

Seleziona 2 lavoratori per farli lavorare nel Task 2 (svolto nella Parte 2). I 4 non selezionati lavoreranno nel Task 1 (svolto nella Parte 1).

Figure A2. Task assignment stage

Instructions

Introduction Experiment E

You are taking part in a decision-making study financed by University of Bologna and the University of Lund. During this study you can earn an amount of money according to the rules that will be described in the following pages. The payment will be paid by cash and confidentially.

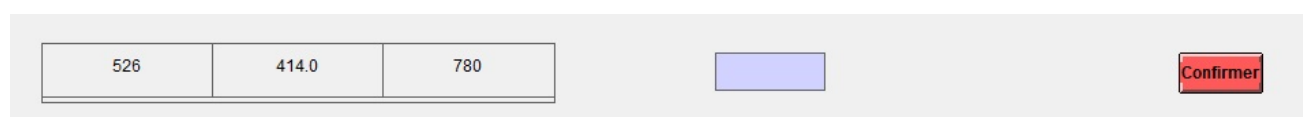
The duration of the present study will be around 1 hour and 30 minutes and is composed by 4 parts. You will be paid for one the first three parts randomly selected by the computer and for Part 4. So, your final earnings in this study will be composed by the earnings of the part selected plus the earnings in Part 4 plus 5 Euro show-up fee.

The rules we will follow to determine your earnings is different in each part. You will receive the instructions of each part sequentially. The instructions of each part will describe in detail how your earnings will be determined in that part.

Please, communicating with other participants during the experiment is forbidden. The use of electronic devices will mean the exclusion for this study. If you have questions during the study, please raise your hand. An assistant will arrive to your station to answer privately.

Instructions Part 1

In Part 1, your mission is to correctly solve the higher number as possible of additions. You will have 6 minutes to solve additions of three three-digit numbers as illustrated below. The numbers to sum will be selected randomly. You will see a scheme like the one represented below.



The screenshot shows a digital interface for a math task. On the left, there are three rectangular boxes containing the numbers 526, 414.0, and 780. To the right of these boxes is a larger, empty blue rectangular box intended for the user's answer. Further to the right is a red rectangular button with the word "Confirmer" written in white text.

Calculators or electronic devices are forbidden. It is possible to use the sheets of paper and the pencil that you will find in your station. When you are ready, you can insert your answer in the available place and click the red button. Immediately, the computer will say if the answer is correct or not. Your answers are anonymous.

Earnings in Part 1

You will earn 50 cents per each correct calculation in the 6 minutes of available time. Your earnings will not decrease with incorrect answers. If Part 1 is selected for payment, you will obtain the money earned solving additions.

What is happening now?

If you have questions about Part 1, please raise your hand. An assistant will arrive to your station to answer privately. Before the start of the study, we will ask you to respond some questions to verify if

you understood the rules correctly.

After the questions, you will have one minute for familiarize with the task. During this period, you can solve additions that will not be considered in the final computation of your earnings. When this minute is over, you will be notified before the start of the first three minutes that composed the first part. You will have 30 second of break before the start of the last 3 minutes.

Instructions Part 2

Your mission in Part 2 will be to solve correctly as many mathematical problems as possible. You will have 10 minutes. You will see and scheme like the one presented below, and you will have to select one the possible answers.

Dati i due polinomi $(a^2 + b - 3)$ e $(4 - b)$, il loro prodotto è uguale a:

- A. $4a^2 + ab + b - b^2 - 12$
- B. $4a^2 - a^2b + 7b - b^2 - 12$
- C. $4a^2 + a^2b + 7b - b^2 + 12$
- D. $-4a^2 - a^2b + b + b^2 - 12$

Navigation icons: back, forward, search, refresh. Submit button.

Calculators or electronic devices are forbidden. It is possible to use the sheets of paper and the pencil that you will find in your station. When you are ready, you can insert your answer choosing one of the answers and click the red button. Immediately, the computer will say if the answer is correct or not. Your answers are anonymous.

Earnings in Part 2

You will earn 1.50 euro per each correct answer in the 10 minutes of available time. Your earnings will not decrease with incorrect answers. If Part 2 is selected for payment, you will obtain the money earned solving problems.

What is happening now?

If you have questions about Part 2, please raise your hand. An assistant will arrive to your station to answer privately.

After the questions, you will have 3 minutes for familiarize with the task. During this period, you can solve problems that will not be considered in the final computation of your earnings.

When the three minutes are over, you will be notified before the start of the 10 minutes that composed the first part.

Instructions Part 3

In Part 3, we will assign you the role of worker and your mission will be to select 6 workers and assign them to two different tasks that correspond to the tasks in Part 1 and Part 2 that you recently participated. The six workers hired composed your team. (Quota Treatment: The half of the team must be female).

Your earnings as employer in this part will depend on the performance that the six workers hired have obtained in one of the tasks: for 4 workers will be relevant their performance in Part 1 (the adding exercises) and for 2 workers will be relevant their performance in Part 2 (the mathematical exercises).

You must take two decisions sequentially:

1. First, you must select the 6 workers to conform your team from a group of candidates (composed by a maximum of 15 workers) that participated in this study previously. Workers will be identified with a number between 1 and 15. (Quota Treatment: at least, 3 females must be present in the team).
2. Then, you must select 2 workers from the 6 already selected (Quota Treatment: with no restrictions by gender). The 2 workers have the mission to solve mathematical exercises for you. The other 4 workers will solve the additions for you.

Your earning will be determined as follows:

$$\text{Earnings} = 0.50 \text{ Euro} \times [E_{W1} + E_{W2}] + 0.1 \text{ Euro} \times [A_{W3} + A_{W4} + A_{W5} + A_{W6}]$$

You will receive:

- 0.50 Euro per each correct mathematical exercise (**E**) that each selected worker employer solved correctly in Part 2, plus
- 0.10 Euro per each correct calculation (**A**) that each selected worker solved correctly in Part 1.

Example

Suppose that you have selected workers W3, W5, W6, W10, W11 and W15 and you have decided that the performance in Part 2 will be relevant for workers W5 and W10, while Part 1 will be relevant for the remaining workers. To determine your earnings, the computer will select the result obtained for each worker in the assigned part. For instance,

	W3	W5	W6	W10	W11	W15
# Correct calculations Part 1 - (A)	10	-	7	-	5	8
# Math exercises solved correctly in Part 2 - (E)	-	6	-	8	-	-

Your earnings will be determined with the following formula:

Employer's earnings= 0,50 Euro x [6 + 8] + 0,10 Euro x[10 + 7 + 5 + 8]= 7 + 3=**10 Eur**

Workers' information

Before selecting workers, you will have the chance to look their CVs by using the information relative to each worker that have already participated in the study. Apart from the information shows in the CV, you will receive information about the productivity of the workers in the first 3 minutes of Part 1 (number of correct calculations). We will call this information SIGNAL. You won't be provided with information relative to the performance in Part 2.

In the previous study, workers were randomly assigned in groups of 15. Each worker could decide whether they want to pay an amount for participating or not in the labor market (may be hired or not by the employer). For this reason, some employers could be provided with groups composed by the less than 15 workers

Repetitions and final earning.

Part 3 will be repeated 4 times and each time you will have to evaluate a different group of workers (i.e. you will select 4 teams, one in each repetition). If Part 3 is selected for payments, the computer will randomly select on the 4 repetitions. The choice of the repetition will finally determine the earnings in Part 3.

What's happening now?

If you have any question about Part 3, we ask you to raise your hand, an assistant will arrive to your station to solve your questions in private. You will be informed when Part 3 starts.

Instructions Part 4

Now, it is starting the fourth part of the study. In this part, we will ask you to make only one decision. The decision you take will influence your earnings and there are no consequences for other participants. In the computer you will see a square composed by 100 cells.

Your mission is to decide how may cells you want to collect.

Earnings Part 4

You will receive 10 cents per each cell collected. These earnings are considered as potential in the sense that after you make your decisions, the computer will randomly place for each participant a bomb behind one of the cells in the square. If you have selected a grey square, you will earn 10 cents per each cell collected. If you have selected a red cell, you will earn 0 cents.

In the decision of how many cells to collect, you have to insert the number in the space available on the right.

Example. If you decide to collect 98 cells and the number 99 is selected by the computer, you will earn 9.8 euro but if the number 10 result selected, you will earn 0 euro. All numbers have the same probability of being selected by the computer.

Final earnings

At the end of Part 4, the computer will randomly select one part between Part 1, Part 2 and Part 3 and you will receive the earnings of the selected part plus the earnings in Part 4 plus the show-up fee.

What is happening now?

If there are no questions, we start with the fourth part. When all participants have finished, we will be ready to proceed with the last part of the study, that consists of a questionnaire and the payment stage.

Appendix B. Experiment W.

Experimental Design: Experiment W- Part 3.

In Part 3, 32 subjects participated in each session. In each session, participants were randomly assigned the role of Employer (N=2) or Worker (N=30). Each Employer was randomly matched with a group of 15 Workers and was asked to make two decisions. First, employers had to select six workers to conform a team. And second, they had to assign the six selected Workers to two different tasks. Specifically, they had to assign four workers to an Easy Task and two workers to a Hard Task. The Easy Task (i.e. a less complex and profitable task) and the Hard Task (a more complex and profitable task) corresponded to the tasks in Part 1 and Part 2, respectively. We provided Employers with Workers' ID, Gender, Year of Birth, Field of Study and a Signal of performance (i.e. number of correct calculations obtained in the first half of the Easy Task). Part 3 comprised two rounds. In each round, employers made the same decisions over the same group but differed in the role of workers. In the first round, workers played a passive role. They received €10 if resulted hired and assigned to the Hard Task, €6 if resulted hired and assigned to the Easy Task and €2 if not hired. In the second round, workers made one decision. They had to decide whether to participate in the hiring process or not. All workers were endowed with €4. Those who decided to opt-out kept the whole endowment and their profile was removed from the pool of workers. The decision to participate in the selection process cost €2 and had three possible outcomes: 1) being not hired, 2) being hired and assigned to the Hard Task, or 3) being hired and assigned to the Easy Task. Workers who participated in the selection and were discarded lost the participation fee and earned the remaining €2. Workers who participated, resulted hired and assigned either to Hard Task or the Easy Task recovered the participation fee and received €8 and €4 respectively. Then, at the end of Part 3, Workers assigned to the Hard Task received €12 and Workers assigned to the Easy Task received €8. The decisions of employers and workers were not simultaneous, workers made their decision first. Therefore, depending on the decision of the workers, the employers may have been confronted with a different number of candidates in the first stage/decision.

Once decisions were made, we elicited workers' beliefs by asking them the number of males and females they expected to be hired and assigned to the Hard Task. Moreover, they were asked whether they expected being hired or not in the first round.

In Part 4 we elicited risk attitudes with the static version of the Bomb Risk Elicitation Task (BRET). Finally, subjects were provided with feedback about their performances and earnings in each part (no feedback across parts was provided), their final earnings and a post-experimental questionnaire that included measures of competitiveness and personality traits.

In this experiment, participants were allocated to two different Treatments: Control Treatment and Quota Treatment. In the Control Treatment, subjects faced the experiment as explained above. In the Quota Treatment, employers were made aware that at least, the fifty per cent of the final composition of the team must be composed by female workers.

The experiment was conducted using z-Tree (Fischbacher, 2007) at BLESS, the experimental laboratory of the University of Bologna (Italy). Subjects were undergraduate students Sciences, Social Sciences, Engineering and Foreign Languages at the University of Bologna, recruited via ORSEE. From November to December 2018, 128 subjects (57% female) participated divided in 2 sessions of 32 subjects per treatment (4 sessions in total). Sessions were not gender-balanced. All treatments were run in a between-subjects design and none participated in more than one treatment. Subjects were randomly assigned to treatments. The duration of each session was about 90 minutes. In each session, once arrived at the lab, instructions were read aloud and subjects were informed that Part 4 and one randomly selected part among Part 1, Part 2 and Part 3 will be relevant for payments, in order to avoid wealth effects. If Part 3 was selected for payments, of the two rounds was relevant for payments in a second random draw. The average payment was about €15.

The group of workers in this experiment were evaluated by the employers of the subsequent experiment (explained in the main text). Groups were not modified, and groups were matched with employers under the same Treatment.

Results and Discussion

We remove the data of employers and focus on workers' decisions in the second round of Part 3 (N=120). Ninety-four per cent of workers in this experiment decided to participate in the hiring process. This percentage corresponds to seven subjects. The high rate of workers participating in the selection process could be explained by the low cost of participation determined in the experimental design. This rate is higher under Quota Treatment. However, the difference is not significant. Figure A1 presents the summary of the entry decisions divided by gender and treatment. We define the entry decision with a dummy variable ($Application_i$) that equals 1 if the decision is positive and 0 if negative. In the Control Treatment, male workers decided to apply significantly more than female workers (Mann-Whitney test $Z = -1.986$; $p\text{-value} = 0.047$). In the Quota Treatment female workers apply more than male workers although the difference is not significant. Across treatments, the difference within genders is not significant. Then, as a result, we can observe even if the quota does not significantly increase the number of females applying for environments that include complex tasks, it seems to be enough to close the gender gap in the willingness to compete.

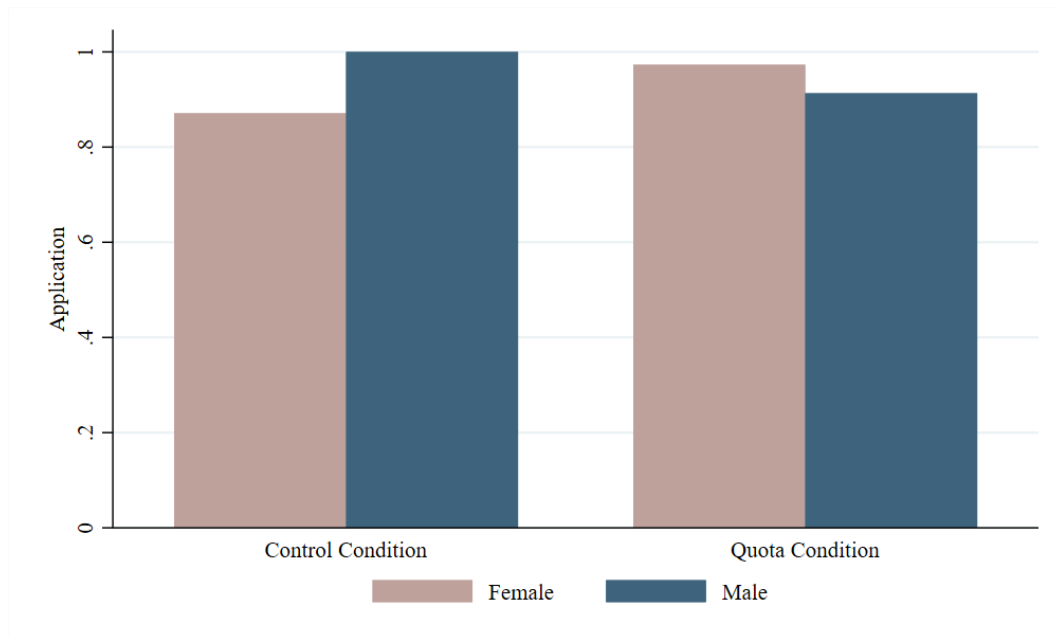


Figure B1. Gender differences in the proportion of applications by treatment.

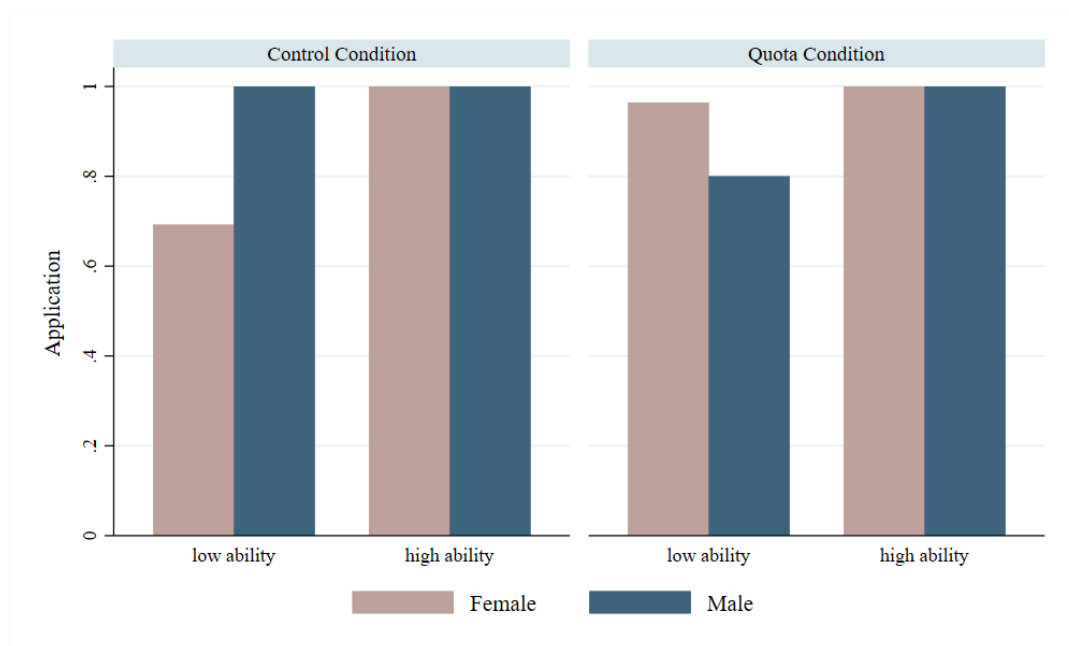


Figure B2. Gender differences in the proportion of applications by treatment and level of ability

On the other hand, Figure A2 presents the same result divided by level of ability. We establish the median in performance in the Easy Task as the threshold of ability (Niederle & Vesterlund, 2007). We consider the performance in the Easy Task because we assume that employers will make their decisions based on the signal of performance in such task. The results do not vary if we determine the threshold of ability in the median value of the performance in the Hard Task. Figure A2 shows that high-ability workers always apply to the hiring process. The sample of workers who decided not to participate is

found among low-ability workers. We find opposite effects of the quota across genders. The quota significantly increases the proportion of female workers who apply ($Z=-2.446$, $p\text{-value}=0.0144$) but decreases the proportion of male workers opting-in ($Z=1.650$, $p\text{-value}=0.098$). In sum, gender quotas have positive effects encouraging females for applying to competitive environments what will result in an increasing number of females in the organizations and leadership positions (Balafoutas & Sutter, 2012; Maggian et al., 2019). Moreover, other positive effect is the discouragement of low-ability males to compete too much, what would diminish their potential earnings. These results are consistent with those observe in the literature (Niederle & Vesterlund, 2007, 2008; Niederle et al., 2013).

Appendix C. Additional results.

Table C1. Descriptive Statistics

	Workers		Employers	
	Female	Male	Female	Male
Age	23,74	23,35	22,35	23,07
Sciences	0,38	0,40	0,06	0,05
Social Sciences	0,49	0,31	0,34	0,40
Engineering	0,06	0,23	0,04	0,18
Letters	0,07	0,06	0,48	0,18
Health	0,00	0,00	0,06	0,16

Table C2. Informativeness of the signal, Tobit regressions.

Dependent variable ^a :	CorrectET		CorrectHT	
	(1)	(2)	(3)	(4)
Signal	1.649*** (0.064)	1.698*** (0.061)	0.326*** (0.108)	0.312*** (0.112)
Sciences	0.000 (0.404)	0.457 (0.489)	3.393*** (0.941)	2.482** (0.991)
Social Sciences	-0.194 (0.347)	0.243 (0.401)	1.027 (0.883)	0.188 (0.789)
Engineering	0.057 (0.478)	0.337 (0.532)	5.269*** (0.982)	4.270*** (1.072)
Female	-0.301 (0.304)	-0.234 (0.323)	-0.402 (0.597)	-0.212 (0.714)
Age	-0.017 (0.044)	0.005 (0.034)	0.146 (0.099)	0.126 (0.097)
Constant	37.161 (87.395)	-5.680 (68.972)	-288.285 (197.145)	-245.871 (193.599)
Unobservable controls ^b	No	Yes	No	Yes
Observations	120	120	120	120

^a CorrectET: Number of correct calculations in the Easy Task; CorrectHT: Number of problems correctly solved in the Hard Task. ^b Risk aversion, underconfidence in Task 2, attitude towards competition and personality traits: Extraverted, Critical, Dependable, Anxious, Open, Quiet, Warm, Careless, Calm, Conventional. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table C3. Treatment effect: Marginal effects, probit regressions on the probability of being hired in the first stage on high-ability workers.

	Dependent variable: Pr (Hired=1)			
	High-ability			
	Top6-S		Top6-ET	
	(1)	(2)	(3)	(4)
Female	-0.102*** (0.039)	-0.430*** (0.073)	-0.274*** (0.040)	-0.388*** (0.058)
Treatment	-0.314*** (0.073)	-0.614*** (0.099)	-0.189*** (0.052)	-0.303*** (0.065)
Female x Treatment		0.514*** (0.089)		0.199** (0.081)
Female + Female x Treatment	-0.102***	0.0832*	-0.274***	-0.188***
Full set of controls ^a	Yes	Yes	Yes	Yes
Observations	3,648	3,648	3,584	3,584
N (employers)	128	128	128	128

^a Workers' Controls: Age, Field of study, Position in the pool; Employers' controls: Gender and beliefs; Experimental controls: order of the presentation of the groups; Group controls: Number of workers in the pool and Share of females in the group. In parentheses, robust standard errors clustered at employer level. All regressions contain employers' random effects. *** p<0.01, ** p<0.05, * p<0.1. Top6-S: high-ability workers according to the signal (i.e. those workers that are among the six best workers according to the signal); Top6-ET: high-ability workers according to the ex-post performance in the Easy Task (i.e. those workers that are among the six best workers according to the performance in the Easy Task).

Table C4. Treatment effect: Marginal effects, probit regressions on the Probability of being hired in the first stage.

Dependent variable: Pr (Hired=1)				
Low ability				
	<Top6-S		<Top6-ET	
	(7)	(8)	(9)	(10)
Female	-0.000 (0.070)	0.119 (0.090)	0.191*** (0.057)	0.155** (0.077)
Treatment	-0.056 (0.109)	0.108 (0.148)	0.289*** (0.082)	0.239** (0.119)
Female x Treatment		-0.277** (0.128)		0.077 (0.119)
Female (net)		-0.158*		0.232***
Full set of controls ^a	Yes	Yes	Yes	Yes
Observations	3,584	3,584	3,648	3,648
N (employers)	128	128	128	128

^a Workers' Controls: Age, Field of study, Position in the pool; Employers' controls: Gender and beliefs; Experimental controls: order of the presentation of the groups; Group controls: Number of workers in the pool and Share of females in the group. Top6-S in models (1) and (2). In parentheses, robust standard errors clustered at employer level. All regressions include employers' random effects. *** p<0.01, ** p<0.05, * p<0.1. <Top6-S: low-ability workers according to the signal (i.e. those workers that are not among the six best workers according to the signal); <Top6-ET: low-ability workers according to the ex-post performance in the Easy Task (i.e. those workers that are not among the six best workers according to the performance in the Easy Task).

Table C5. Treatment effect: Marginal effects, probit regressions on Probability of being assigned to Hard Task in the second stage.

Dependent variable: Pr (Hired=1).				
	Low ability			
	<Top6-S		<Top6-HT	
	(1)	(2)	(3)	(4)
Female	0.039 (0.118)	0.111 (0.128)	-0.097 (0.076)	-0.062 (0.082)
Treatment	-0.304* (0.165)	-0.204 (0.226)	-0.153** (0.067)	-0.104 (0.104)
Female x Treatment		-0.190 (0.232)		-0.073 (0.144)
Female (net)		-0.0786		-0.134
Full set of controls ^a	Yes	Yes	Yes	Yes
Observations	3.584	3.584	3.84	3.84
N (employers)	128	128	128	128

^a Workers' Controls: Age, Field of study, Position in the pool; Employers' controls: Gender, beliefs and share of females hired in the first stage; Experimental controls: order of the presentation of the groups; Group controls: Number of workers in the pool and Share of females in the group. Top2-S in models (1) and (2). In parentheses, robust standard errors clustered at employer level. *** p<0.01, ** p<0.05, * p<0.1. <Top6-S: low-ability workers according to the signal (i.e. those workers that are not among the six best workers according to the signal); <Top6-HT: low-ability workers according to the ex-post performance in the Hard Task (i.e. those workers that are not among the six best workers according to the performance in the Hard Task).

Table C6. Optimal productivities and optimal earnings calculated with the algorithm

Control Treatment	Group 1	Group 2	Group 3	Group 4
Optimal Productivity in the Easy Task	23	28	23	32
Optimal Earning in the Easy Task	11,5	14	11,5	16
Optimal Productivity in the Hard Task	71	70	67	63
Optimal Earning in the Hard Task	7,1	7	6,7	6,3
Optimal Earning (Total)	18,6	21	18.2	22.3

Quota Treatment	Group 1	Group 2	Group 3	Group 4
Optimal Productivity in the Easy Task	26	26	27	24
Optimal Earning in the Easy Task	13	13	13,5	12
Optimal Productivity in the Hard Task	60	70	56	59
Optimal Earning in the Hard Task	6	7	5,6	5,9
Optimal Earning (Total)	19	20	19.1	17.9

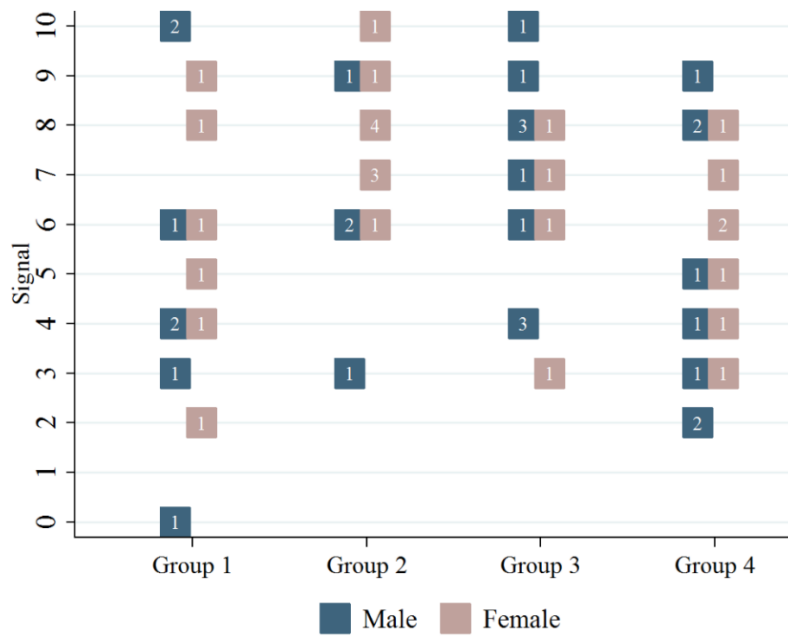


Figure C1. Ranking of workers according to the signal

Note: The numbers inside the items indicate the number of males or females sharing the same position.

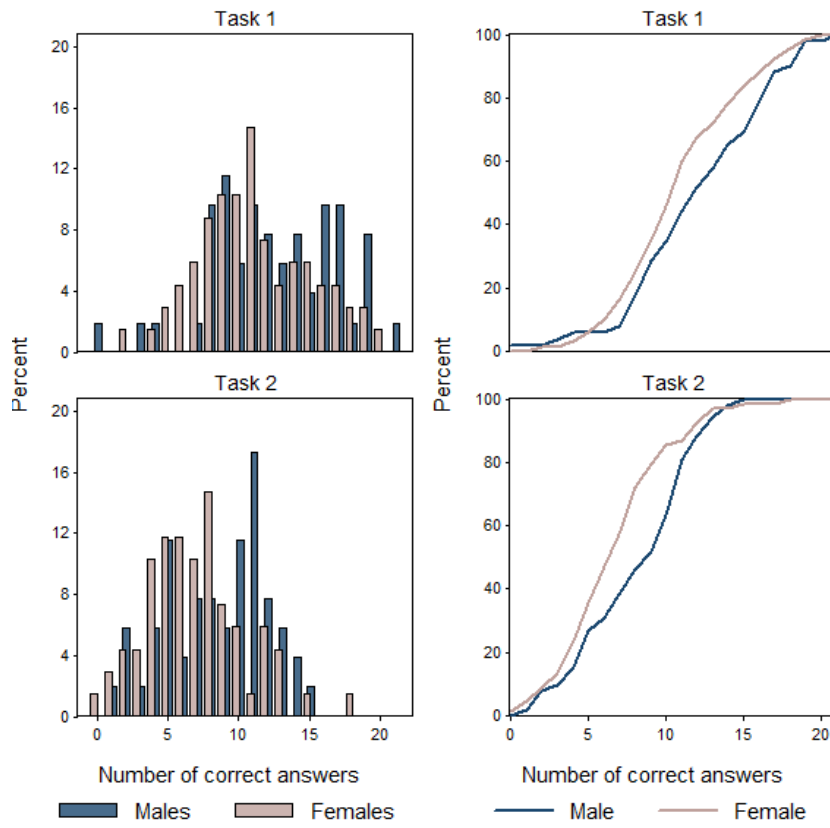


Figure C2. Distribution of performances

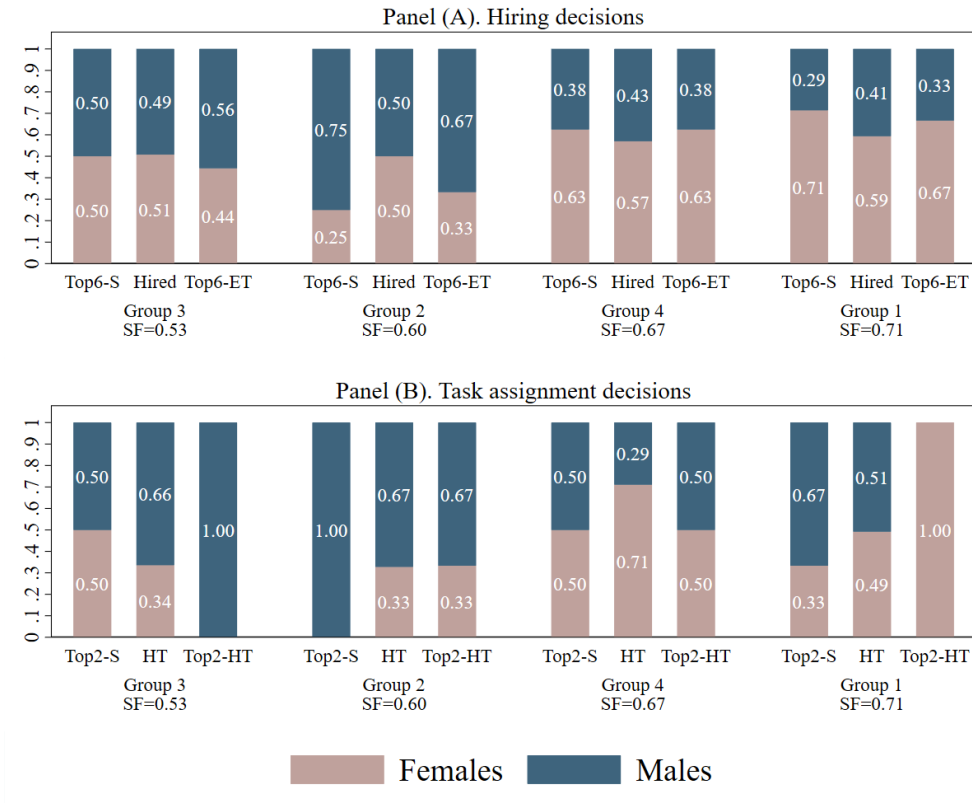


Figure C3. Raw Data: Quota Treatment

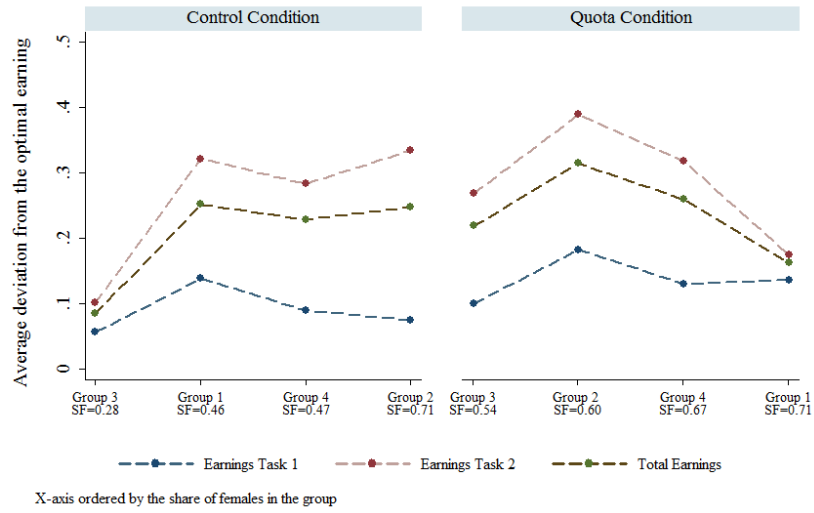


Figure C4. Average deviation from the optimal earning by group

Chapter 3.

Committee Quotas and Gender Gap in Recruitment: A Laboratory Experiment

José J. Domínguez
University of Padova

Keywords:

Committee Quotas; Gender gap; Group Dynamics; Laboratory experiment.

JEL Codes:

D03, C92, J71

1. Introduction

An increasing number of countries have introduced gender quotas for combatting female underrepresentation in traditionally male-stereotyped occupations (Freidenvall & Dahlerup, 2013). Women underrepresentation responds to a number of reasons. Preferences for competition (Niederle and Vesterlund, 2007; Buser et al., 2014), family background (Correll et al., 2007) and employers' stereotypes (Reuben et al., 2014; Williams & Ceci, 2015) are some proposed examples. The underrepresentation of women in the labor market has been also attributed to the lack of women in recruitment committees. In Europe, women held 29% of non-executive positions in the top two decision-making bodies of large companies, and just 17% of executive positions (EU, 2019). Therefore, committee quotas are becoming more widespread. Some governments (e.g. Austria, Finland, Iceland, Norway, Spain and France) and Research Councils (e.g. Norway and Sweden) imposed gender-balanced recruitment committees during the last decades (Frutos et al., 2012; Wallon et al., 2015).

Women benefiting other women is an intuitive and practical assumption, however the evidence on the role of the decision-makers' gender on candidates' outcomes have provided mixed results. The experimental literature has shown evidence on candidates benefiting from the decisions made by evaluators of their same gender (Casadevall & Handelsman, 2014; Kunze & Miller, 2017; Quintana-García & Elvira, 2017; Bossler et al., 2019; Flabbi et al., 2019), evaluations from opposite gender evaluators (Broder 1993; Ellemers et al., 2004; La Mattina et al., 2018) or situations in which the

gender of the evaluator does not have a significant effect (Reuben et al., 2014; Bohnet et al., 2015; Milkman et al., 2015; Williams and Ceci, 2015; Beaurain & Masclet, 2016). In joint decisions, it has been argued that the gender of the decision maker matters. Daskalova (2018) found in-group favoritism in decisions made by groups, while there were no signs of such favoritism in decisions made by individuals. If we then consider that men's preferences for their own gender are the causes of discrimination, a higher share of women in recruitment committees should benefit female candidates. Bagues and Esteve-Volart (2010) made the first steps in observing how a different number of females in committees impact on candidates' chances of success in the exam for the Spanish Judiciary. They found that the higher the number of same-sex evaluators in the committees, the less likely a candidate is to be successful. More recently, research has focused on academic promotions. Paola and Scoppa (2015) found in professorship exams for Economics and Biology in Italy, female candidates benefited from having at least one woman in the committee. Bagues et al. (2017) used a well-adapted administrative dataset from Italy and Spain to show that, overall, there were no significant differences in female candidates' chances of promotion between mixed-gender and all-male committees. The slight preferential treatment of women toward female candidates in mixed-gender evaluations was compensated by the more harshly evaluations of men. In a similar context, Deschamps (2018) also showed that, in the context of the reform made by the French government, more women in committees do not improve the recruitment on female candidates. In summary, the impact of more women in committees is still unclear.

Despite the unclear effect of the gender composition of the committees in the field, there are a number of reasons to believe that more women in committees would be beneficial for female candidates: the existence of gender segregation in specific fields, gendered-networks and gender stereotypes (Bagues et al., 2017). Nevertheless, there are other reasons to consider that committee quotas could be ineffective. The evidence on group dynamics suggests that men and women do not behave similarly in group deliberations, in terms of voice and influence. For instance, women speak less than men and are considered less powerful during group deliberations (Karpowitz et al., 2012). One potential explanation is that powerful women tend to speak less during the deliberations, due to the correctly anticipated backlash triggered by female's volubility (Brescoll, 2011)¹. Furthermore, women's individual decisions are less likely to be incorporated into the group's final decision (Born et al., 2018)². If female evaluators are more inclined for female candidates but they do not contribute in the

¹ In the context of Costa Rica's legislative assembly, Funk & Taylor-Robinson (2014) found that women participate as much as men in committee, even when women are minority with respect to men.

² The gender differences in group dynamics are in line with the literature studying females' underconfidence in male environments and the stereotype threat. Females are less willing to compete (Niederle and Vesterlund, 2007), to contribute

deliberations as much as male evaluators do, the chances of female candidates would depend on the decisions made by men regardless the share of women in the committee.

This paper proposes a laboratory experiment with the aim to test the implicit assumption that more women in committees is beneficial for female candidates, by analyzing the effect of the gender composition of the committees on recruitment decisions. Moreover, it studies how men and women behave in groups deliberations, in terms of voice and influence, as a potential driving-mechanism of the results. The laboratory evidence on gender discrimination usually focuses on situations in which single employers make the hiring, promotion and task assignment decisions. The common result is that female candidates have less chances of success in comparison to male candidates (Reuben et al., 2014; Reuben et al., 2015; Williams & Ceci, 2015; Heinz et al., 2016; Peterle & Rau, 2017; Babcock et al., 2017b; Coffman et al., 2018)³. However, the literature addressing groups decisions in the lab is scarce. As far of my knowledge, this is the first experiment that addresses how groups make decisions in the lab in a hiring setting. This paper contributes to the literature by shedding more light on the inconclusive same-gender preference assumption in a controlled environment by a) studying the trade-off between the individual preferences and group outcomes depending on the share of females in the groups and b) proposing a mechanism for which committee quotas do not exert the expected impact. In the experiment, subjects were allocated into groups of three evaluators and were asked to jointly select two candidates out of six to perform a mathematical task. Since the study of gender discrimination usually focus in male-stereotyped environments, I selected a mathematical task to represent those areas in which females have been traditionally underrepresented and where performance do not show significant gender differences. For the decisions, subjects were provided with different candidates' information: age, gender, field of study and a signal of performance. The groups of subjects (i.e. the committee) in charge of the hiring decision had four different gender compositions: all-male, male-majority, female-majority and all-female. All members of the committee were rewarded according to the productivity of those candidates selected in the mathematical task. In a nutshell, I did not find own-gender preferences in evaluators' individual decisions. Both women and men benefited male candidates over female candidates. According to group decisions, while groups in which men are majority are the most beneficial for female candidates, groups in which females are majority are the most detrimental. I did not find a monotonic improvement of female candidates' chances of selection as the number of women in the committee increases, thus

with their ideas (Coffman, 2014), to lead a team (Born et al., 2018) and more willing to self-select into less demanding tasks within the organizations (Babcock et al., 2017a).

³Interestingly, Charness et al., (2018) exceptionally found signs of positive discrimination in the hiring, being female candidates more likely to be selected.

contradicting the assumption that more females in committees do benefit female candidates. I do not find that women have less voice and influence than men in male-majority groups, which are the most beneficial for female candidates. In contrast, female-majority groups exert a different effect on evaluators' voice and influence. First, I find that females speak more than men. However, men have more influence, measured by the degree in which evaluators are willing to change their initial preferences, and disproportionately opted for male candidates. These results provide more evidence of the inconclusive assumption of the own-gender preferences, suggesting that more women in the committees do not necessarily improve female candidates' outcomes. Anyway, we need to call for more research to understand why and under which conditions women should benefit other women before moving to policy recommendations.

The remainder of this paper is structured as follows: Section 2 presents the experimental design and procedures. Section 3 presents the results and Section 4 concludes.

2. Experimental Design

In the experiment, all subjects played the role of evaluators. Subjects were allocated into groups of 3 subjects. The groups had four different gender compositions: 1) All-male – three males, 2) male-majority – two males and one female, 3) female-majority – one male and two females, 4) All-female – three females. The mission of evaluators in each group was to jointly select two candidates out of six to perform a mathematical task. Candidates were recruited from an independent study in which they were asked to perform two correlated tasks: a mathematical task and an adding task (Pearson's $r = 0.306$; $p < 0.000$). The mathematical task consisted of solving as many mathematical problems as possible in ten minutes⁴. The adding task consisted of summing up as many three three-digit numbers in six minutes. Evaluators were rewarded according to the performance of the selected candidates in the mathematical task plus a fixed compensation. They were also asked to submit an individual decision. In both the individual and the group decision, subjects were provided with the following candidates' information: age, gender, field of study and a signal of performance⁵. The signal of performance consisted of the number of correct calculations in the adding task. In order to familiarize with the task for which subjects had to make decisions, they were asked to participate in a non-

⁴ The mathematical problems set was extracted from the entry test of the University of Padova's bachelor program in Economics. We selected those questions that did not show gender differences in the probability of responding correctly in the editions of April and August of 2018, to which 1476 students participated. Each problem consisted of a multi-choice question with four possible answers.

⁵ This procedure can be found in other experiments in which employers have to make hiring decisions in order to reduce the salience of the gender. See Bertrand & Mullainathan (2004), Bohnet et al. (2015), Beaurain & Masclet (2016), Heinz et al. (2016), Peterle & Rau (2017) and Paryavi et al. (2019).

rewarded trial version of both the adding task and the mathematical task before moving to the selection decisions.

At the end of the experiment, subjects were asked to complete the Implicit Association Test (IAT) to measure gender stereotypes. The IAT is a classification task that provides an indirect measure of association between groups (i.e. “male” and “female”) and career attributes (“math and science” and “humanities”). The IAT score (D-score) takes values between -2 and 2 (Greenwald et al., 2003). A positive score indicates an association between “male” and “math and science” and between “female” and “humanities”. A negative score indicates an association between “female” and “math and science” and between “male” and “humanities”⁶.

2.1. Decision 1: Individual Decision and Beliefs.

Subjects were first asked to individually select two candidates out of six to perform the mathematical task. Each subject faced the same pool of candidates. Subjects received 30 ECU (Experimental Currency Unit) per each correct answer of those candidates selected plus a fixed amount of 350 ECU. Subjects’ beliefs were elicited after the selection decision. Subjects were endowed with 200 ECU. They were asked to invest, for each candidate selected, a number of ECU between 0 and 100 to obtain the double of the amount invested plus the amount not invested if the candidate for which they are investing was among the two best performers in the pool of six the mathematical task. If the candidate was not among the two best performers, they obtained zero plus the amount not invested.

2.2. Decision 2: Group Decision.

As a group, subjects revisited the same pool of candidates and were asked to jointly select two candidates to perform the mathematical task. Each group had 3 minutes to exchange free-format messages using a chat box installed in their own computers. During the chat, subjects were made aware that they cannot reveal personal information in the discussion. The only information that subjects had about other evaluators in their group was gender. Gender was revealed only when exchanging messages without making it salient. In the chat box, each message was labelled with the name of the evaluator who posted it. Subjects were randomly called “Manager 1”, “Manager 2” and “Manager 3”. To subtly provide gender, at the top of the chat box, managers were provided with an additional box containing the names of the managers in the groups together with a standard picture that only differed

⁶ The IAT has been used in other experiments that address gender stereotypes (Reuben et al., 2014; Lowes et al., 2015; Burns et al., 2016, Glover et al., 2017; Carlana, 2017). The experimental procedure of the IAT can be found in Appendix B.

by gender. The picture consisted of a silhouette either of a man or a woman, depending on the gender of the manager. Figure A1 shows the pictures provided for each gender.

After the deliberation stage, subjects had to submit their votes for two candidates (i.e. each subject had 2 votes). The group decision was determined according to a majority voting rule. That is, a candidate needed, at least, two votes to be selected by the group. If the group agreed over two candidates (i.e. two candidates received at least 2 votes), subjects were informed about the outcome and the group decision was made. If the group did not reach full agreement (i.e. only one candidate or no candidate received at least two votes), subjects were informed about the candidates who reached a majority (if any), then the chat opened for an extra minute and subjects voted again. The second voting was not restricted to any worker. At the end of the voting stage, groups could reach three outcomes: 1) two candidates selected and consequently zero vacant positions, 2) one candidate selected and one vacant position, and 3) no candidate selected and two vacant positions. Subjects received 30 ECU per each correct calculation of each candidate selected and 0 ECU per each vacant position, plus fixed 350 ECU.

2.3. Rounds

In both the individual and the group decision making stages, subjects evaluated 3 different pools of candidates (3 rounds). In each round, they were reshuffled into a different group of evaluators keeping constant the gender composition of the group. The order of the pools evaluated was randomly implemented across sessions. At the end of the experiment, every group (individual) evaluated the same pools of candidates, regardless the gender composition of their groups. Subjects were informed about the group decision (i.e. candidates selected) at the end of each round. Information about earnings was displayed at the end of the experiment.

2.4. Candidates

The sample of candidates was obtained from a different experiment. One hundred and twenty subjects were asked to fill a short CV with baseline information: year of birth, gender and field of study.

Participants who participated as workers, first performed an adding task for which they earned 0.5 Euro per each correct calculation. Then, they were asked to perform the mathematical task for which they earned 1.5 Euro per each correct calculation. A random task was selected for payments at the end of the experiment in order to avoid wealth effects.

In order to make subjects to focus in all pieces of information and encourage the debate in the group deliberations, the pools of candidates in this experiment were obtained by selecting the 18 best

performers in the adding task in order to make minimal the variability among performances⁷. The sample of the best 18 performers were splitted in three different pools of candidates according to the ranking based on the performance in the adding task. Ties were broken taking into account the gender composition of the pools so that subjects could evaluated three pools of candidates with a different share of female candidates. In this way, managers evaluated a quite similar pool of candidates in every round. The pools of candidates evaluated in this experiment are reproduced in the Appendix A (Table A1).

2.6. Procedure

The main experiment was conducted using z-Tree (Fischbacher, 2007) at BLESS, the experimental laboratory of the University of Bologna (Italy). The IAT was implemented using Qualtrics (Carpenter et al., 2018). Subjects were undergraduate students from all different schools of the University of Bologna, recruited via ORSEE (Greiner, 2015). None of the participants took part in as a worker in the previous experiment. On June 2019, 120 subjects (50% female) participated divided in 4 sessions of 30 subjects. In each session, subjects conformed 30 different groups (120 groups in total). Sessions were perfectly gender balanced. The duration of each session was about 60 minutes. In each session, once arrived at the lab, instructions were read aloud and subjects were informed that one randomly selected decision (individual or group) in a given round will be relevant for payments, in order to avoid wealth effects. An English version of the instructions are reproduced in Appendix B. The experimental currency was ECU (100 ECU corresponded to 0.50 Euro). The average payment was about 11 Euro including a 5 Euro show-up fee.

3. Results

This section presents the results of the experiment. As general results, Section 3.1. and Section 3.2. focus on the effect of the gender of the evaluators, in both the individual and group decision, on the gender composition of the selection, respectively. Section 3.3. describes the gender differences in group dynamics of the decision-making process. Finally, Section 3.4. analyzes the groups' performance.

⁷ Literature on hiring has shown that employers' decisions are highly conditioned by the signal of performance sent by workers. Azic & Lamé (2018) and Reuben et al. (2014) found that from pairs of workers, employers hired the worker with higher task performances in 70 and 80 % of the cases, respectively. Moreover, as discussed by Bohnet et al. (2015), employers are more likely to focus on individual performance signals in joint evaluations (i.e. when candidates are presented at the same time).

3.1. Individual Decisions: The role of evaluators' gender.

This section addresses whether women, when deciding as individuals, are more inclined toward female candidates⁸. Different measures can be employed to study evaluators' preferences. In one hand, the analysis of the IAT shows that women hold significantly stronger stereotypes (i.e. associate more “math and sciences” with “men” than with “women”) (IAT-score= 0.320) than men (IAT-score= 0.161) (Wilcoxon ranksum test $Z=-1.930$; $p<0.1$). The cumulative distribution of the IAT-score by gender can be found in Appendix A (Figure A2). In the other hand, evaluators' preferences can be observed by attending the degree in which they believe the candidates selected in the individual decision are among the two best performers in the mathematical task. Men invested more in male candidates than in female candidates ($Z= 1.870$; $p<0.01$). Women did not invest more in female candidates than in male candidates but invested more than men ($Z= -2.047$; $p<0.05$). The average investments by gender can be found in Appendix A (Figure A3). The proposition that women prefer female candidates, using these measures, is inconclusive due to the contradiction of the results. It would be then convenient to analyze subjects' individual decisions in order to observe whether female candidates are more benefited in the evaluations made by women.

Table 1. Marginal effects of the probit regression on the probability of candidates of being selected in individual evaluations.

	Dependent variable: Pr (Selected- $I_{ip}=1$)		
	(1)	(2)	(3)
Female candidate	-0.067*** (0.018)	-0.067*** (0.018)	-0.053** (0.026)
Female evaluator		0.007*** (0.002)	0.021 (0.019)
Female candidate x Female evaluator			-0.027 (0.036)
Female candidate + Female candidate x Female evaluator			-0.079***
Additional Controls	Yes	Yes	Yes
Observations	2,124	2,124	2,124
Pseudo R2	0.258	0.258	0.258

^aAdditional Controls: candidates' characteristics (age, field of study, performance adding task and position), share of female candidates in the pool, period and IAT score. Robust standard errors, clustered at evaluator level, in parentheses. ^bThe significance of the linear combination of the coefficients is estimated using Wald tests. ^cTwo evaluators provided invalid results in the Implicit Association Test (IAT). *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

⁸For a better understanding of the results, evaluators in the committee will be referred as “men” and “women” while candidates will be referred as “male candidates” and “female candidates”.

Table 1 estimates the marginal effects of the probability of candidate i in pool p of being selected in the individual decision stage (Selected- I_{ip}) on evaluators' and candidates' gender. All regressions include a set of covariates that contain candidates' characteristics: age, field of study, signal of performance, position in the screen (i.e. candidates were displayed in the screen following an order generated by their randomly assigned ID number), the share of female candidates in each pool and evaluators' implicit bias (IAT score). Standard errors are clustered at evaluator level.

The estimations in column (1) show that female candidates have lower probabilities to be selected for the mathematical task compared to male candidates. The gender gap is estimated 6.7 percentage points in favor of male candidates. In column (2), when controlling for the gender of the evaluators, the coefficient of *Female candidate* remains similar. The dummy for *Female evaluator* presents a positive and significant effect, what could imply that men and women have different individual preferences over candidates. Column (3) includes an interaction between *Female evaluator* and *Female candidate*. The interaction presents a negative but not significant effect, meaning that the gender gap when candidates are evaluated by women, is not significantly different to that when candidates are evaluated by men (captured by the effect of *Female candidate*). The IAT-score does not show a significant effect in any regression. Similar to the gender gap in men's evaluations, estimated in 5.3 percentage points in favor of male candidates, the negative and significant effect of linear combination between *Female candidate* and the interaction *Female evaluator x Female candidate* shows the gender gap in the probability of candidates of being selected in evaluations made by women, estimated in 7.9 percentage points.

Result 1: In individual decisions, female candidates have lower probabilities of being selected than male candidates regardless the gender of the evaluator.

3.2. Group Decisions: The effect of the gender composition of the committees.

Females represent the 50% of the candidates, however, they only represented the 42% of the candidates selected to perform the mathematical task (Two-sided binomial probability test $p\text{-value} < 0.05$). On average, female candidates have a higher percentage of success in male-majority groups (47%) while in female-majority groups, female candidates have the lower percentage of recruitment (38%). Group decisions, however, were not only contingent on gender. Evaluators were also provided with different pieces of information that could affect the selection. Therefore, in order to more accurately observe how the gender gap in the selection varies according to the gender composition of the groups, Figure 1 presents the probabilities of female candidates to be selected in comparison to male candidates in

each type of group accounting for every characteristic of the candidates (i.e. age, field of study, signal of performance, position in the screen and the share of female candidates in the pool). As a referent point, the graph also provides similar estimations regarding the evaluators' individual decisions aggregated by the type of group in which they were allocated. The estimated probabilities for both the group and the individual decisions are extracted from the probit regressions in Table A2 in Appendix A.

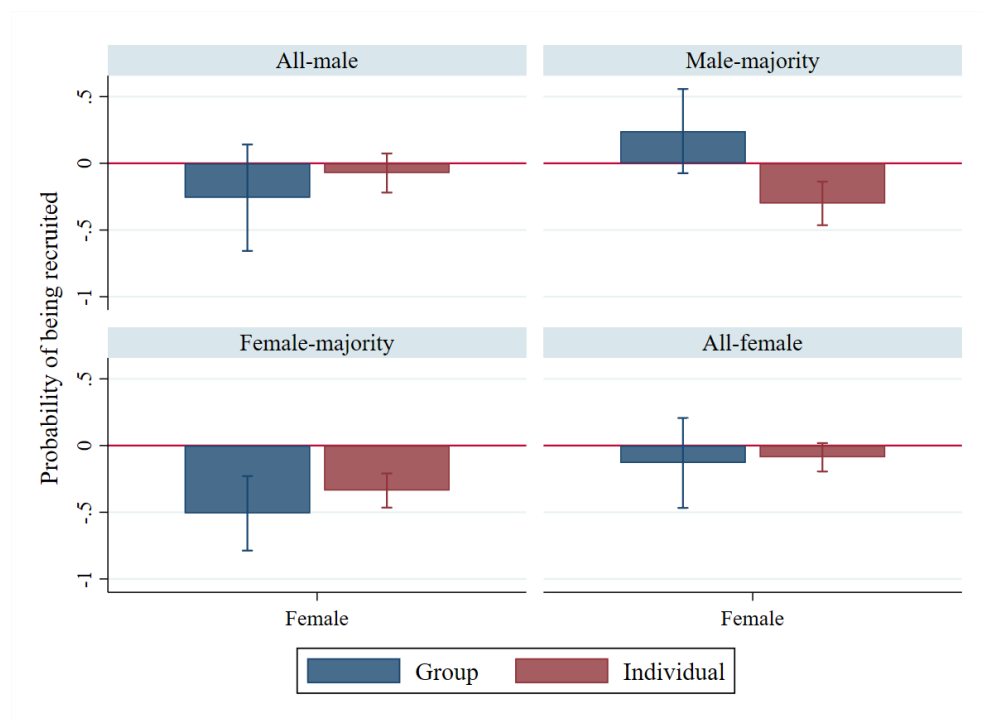


Figure 1. Gender gap in the probability of candidates of being selected in group evaluations.

According to Figure 1, female candidates have less probabilities than male candidates of being selected when they are evaluated individually regardless of the group in which evaluator were allocated. In other words, the aggregated initial preferences of evaluators in any type of group predict lower chances of success of female candidates. However, the gender gap is only significant in mixed-gender groups, similar to that when evaluators make decisions jointly. Mixed-gender groups are those who seem to have an effect when making joint decisions. In Male-majority groups, the probabilities of female candidates of being selected are higher than those of male candidates but this effect is not significant. Anyway, it presents a positive effect in comparison to the aggregated decisions of its members in the individual stage. That is, even if evaluators allocated in male-majority groups, individually, prefer male to female candidates, the deliberation makes that male and female candidates have equal probabilities

of selection. On the other hand, Female-majority groups seem to expand the gender gap in detriment of female candidates with respect to the individual decisions of the evaluators allocated in these groups.

Table 2. Marginal effects of the probit regression on the probability of candidates of being selected by the group stage.

Dependent variable: Pr (Selected- $G_{ip}=1$)			
	(1)	(2)	(3)
Female candidate	-0.064 (0.046)	-0.064 (0.046)	-0.095*** (0.015)
Male-majority		-0.007*** (0.003)	-0.079*** (0.006)
Female-majority		0.008* (0.005)	0.024*** (0.003)
All-female		0.007** (0.003)	-0.013*** (0.002)
Male-majority x Female candidate			0.131*** (0.007)
Female-majority x Female candidate			-0.034*** (0.002)
All-female x Female candidate			0.038*** (0.002)
Female candidate + Male-majority x Female candidate			0.036*
Female candidate + Female-majority x Female candidate			-0.129***
Female candidate + All-female x Female candidate			-0.057***
Additional Controls	Yes	Yes	Yes
Observations	2,160	2,160	2,160
Pseudo R2	0.339	0.340	0.342

Additional Controls: candidates' characteristics (age, field of study, performance adding task and position), votes in the individual stage, share of females in the group and period. Robust standard errors, clustered at composition level, in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

For a better understanding of the probabilities of female candidates according to the gender composition of the groups that make the evaluation, Table 2 presents different estimations of the probability of candidate i in pool p of being selected in the group-decision stage (Selected- G_{ip}). All regressions include candidates' controls and standard errors are clustered at gender composition level. Column (1) regresses the probability of being selected on candidates' gender. The effect of *Female candidate* on the probability of being selected is negative and significant, showing that female candidates, overall, have lower probabilities of being selected for the mathematical task compared to male candidates in group decisions. The gender gap is estimated in 6.4 percentage points, similar to the gender gap observed in the individual decisions. In Column (2), when including dummy variables

to control for the gender composition of the groups, the coefficient *Female candidate* remains constant. Interestingly, the dummies representing the gender composition of the groups present different effects. This suggests that groups post their votes in a different way. Since the pools of candidates are extremely similar, this implies that the gender composition of the groups affects the decisions made. Column (3) include three interactions between *Female candidate* and each type of group to observe the gender gap in each type of group in comparison to the gender gap in the baseline group (All-male). While male-majority and all-female groups exert a positive effect in the chances of female candidates, female-majority groups exert a negative effect. Specifically, the highest chances of success of female candidates appear when they are evaluated by male-majority groups, given the size of the interaction. The linear combinations between each interaction and the variable *Female candidate* (that captures the gender gap in all-male groups) presents the gender gap in each type of group. The estimations show that in all-male, female-majority and all-female, female candidates have less chances of being recruited for the mathematical task in comparison to male candidates. The estimated gender gaps seem to be higher in female-majority groups, with female candidates having almost 13 percent lower probabilities of recruitment. In contrast, female candidates have more chances of recruitment than male candidates when they are evaluated by male-majority groups. The gender gap in male-majority groups is estimated in 3.6 percentage points in favor of female candidates.

Result 2: Female candidates have more probabilities of being selected in male-majority. In contrast, female-majority groups are the most detrimental for female candidates.

3.3. Mechanisms

This section studies how men and women behave in mixed-gender group deliberations in terms of voice and influence and how the individual behavior impact on candidates' probabilities of selection.

3.3.1. Voice in deliberations

This section looks at the relative proportion of words written in the chat by subjects with respect to the total number of words written in the group as a measure of voice. The average number of words written in the deliberations, irrespectively from the gender composition of the groups, is 84,65. This average differs considerably depending on the gender composition of the group. In male-majority groups the deliberations are significantly longer than groups (Mean=106,2; SD=33,23). In All-female groups the deliberations are significantly shorter (Mean= 64,13; SD=33,34) while there are no significant

differences between All-male groups (Mean=82,03; SD=31,46) and Female-majority groups (Mean=86,00; SD=29,16)⁹.

At individual level, men talk significantly more than women in male-majority groups ($Z=4.239$; $p\text{-value}<0.000$) while women talk more than men in female-majority groups ($Z=-2.235$; $p\text{-value}<0.05$). In order to observe the effect of the gender composition of the groups on the gender gap in voice, Columns (1) to (3) in Table 3 estimate different OLS regression models with the proportion talk as dependent variable. All models account for different characteristics other than gender such as being the first evaluator that talks in the group, the period of the evaluation and the big-five personality traits (Guido et al., 2015).

Table 3. Gender differences in voice and influence in mixed-gender groups.

Dependent variable Method	Proportion talk			Pr (Change $V_{vip}=1$)		
	(1)	OLS (2)	(3)	(4)	Probit (5)	(6)
Female evaluator	-0.028 (0.045)	-0.030 (0.051)	-0.095 (0.016)	0.047 (0.032)	0.039 (0.029)	0.024 (0.017)
Female-majority		0.004 (0.020)	-0.063* (0.005)		0.025*** (0.006)	0.011 (0.011)
Female-majority x Female evaluator			0.139* (0.013)			0.028*** (0.002)
Female candidate				0.136*** (0.040)	0.138*** (0.037)	0.137*** (0.037)
Female evaluator + Female-majority x Female evaluator			0.043**			0.052***
Additional controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	180	180	180	360	360	360
R-squared	0.089	0.089	0.151	0.197	0.198	0.199

Additional Controls for models (1)-(3): first evaluator in talk, personality traits and period. For models (4)-(6): candidates' characteristics associated to each vote (age, field of study, performance adding task and position), share of females in the group and period. For models (1)-(6), robust standard errors (in parentheses) are clustered at composition level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Column (1) regresses the proportion that evaluators talk during the deliberation on gender. The effect of *Female evaluator* is not significant, meaning that overall, there were no gender differences in voice. In column (2), when accounting for the type of group, the coefficient for *Female evaluator* remains stable. Column (3) includes an interaction between *Female evaluator* and *Female-majority* to measure the gender gap in voice with respect to the gender gap in male-majority groups (captured by the effect

⁹ Mann-Whitney tests for differences in the number of words written by groups can be found in Table A3 in Appendix A.

of *Female evaluator*). The effect of *Female evaluator* remains insignificant, what implies that the proportion that men and women speak in male-majority groups is similar. The positive and significant effect of the interaction suggests that the gender gap in female-majority groups is higher than that in male-majority groups. That is, women speak more (i.e. post more words) in female-majority groups than in male-majority groups. The positive and significant the linear combination between *Female candidate* and the interaction tells that, moreover, women speak more than men in female-majority groups.

3.3.2. Influence in deliberations.

This section analyzes the influence that evaluators exert in the deliberation process. To this end, I observe evaluators' willingness to change the initial preferences with respect to the final outcome of the group. In this sense, individual decisions matching with the decision made by the group suggest a higher level of influence in the deliberation process. This analysis is focused on vote level. Columns (4) to (6) in Table 3 regress a dummy variable that equals 1 if a vote v submitted by evaluator i in period p in the individual stage is different to the group decision and 0 if the vote were similar to that in the group decision (ChangeV_{vip}). The higher the willingness to change the vote with respect to the final outcome, the lower is the influence over the group. In each model, estimations are controlled by a set of covariates that are associated to the vote posted by each individual in each period: candidates' characteristics, period and share of females in the pools. Table 3 presents the marginal effects of each model.

Column (4) shows that, overall, the votes submitted by female evaluators are equally likely to be removed in the group decision as those made by male evaluators.

In column (5), the effect of *Female evaluators* remains insignificant while the dummy variable that represents female-majority groups shows a positive and significant effect. This suggest that one can find different results depending on the gender composition of the groups.

In column (6), when introducing an interaction between *Female-majority* and *Female evaluator*, the effect of *Female evaluator* still shows a non-significant effect. In this case, capturing the gender gap in male-majority groups. Then, we observe that the votes posted by female evaluators in male-majority groups are equally likely to be changed than the votes posted by male evaluators, suggesting that neither males nor females have more influence than others in male-majority groups. On the other hand, the effect of the interaction is positive and significant. That is, the gender gap in the probability of change the vote in female-majority groups is significantly higher in comparison to that in male-majority groups. This implies that female evaluators are more willing to change their votes in female-majority groups in comparison to male-majority groups. The linear combination between *Female evaluator* and

the interaction also presents a positive and significant effect, what confirms that not only female evaluators change more their vote in female-majority groups, but also that they are more willing to do so more than male evaluators, suggesting that male evaluators have more influence than female evaluators in female-majority groups. Finally, all regressions include the variable *Female candidate*, the measures if the initial vote were posted in favor of a female candidate. The positive and significant effect in all regressions suggests that the votes posted in favor of female candidates are more willing to be changed in comparison to those votes posted in favor of male candidates.

Result 3:

- In male-majority groups, men and women have the same level of voice. In female-majority groups, women have more voice than men.
- Men are more influential than women in female-majority groups. In male-majority groups, men and women have the same level of influence.

For a better evaluation of female candidates' outcomes, it is useful to observe the type of proposals made by evaluators in each type of group by analysing the content of the messages sent during the deliberations. To this end, a message is classified as a proposal if it suggests either one or two candidates to be included in the final decision of the group. Counterproposals are also classified as proposals. Messages in which the evaluator agrees or disagrees with one proposal made by other is not considered as a proposal. Other types of messages (e.g. "well done guys") are not considered as proposals. Only 8 evaluators (7%) did not make any proposal in all three periods. Figure 3 presents the distribution of the proposals according to the gender of the evaluators in each type of group. The graph considers three types of proposals: 1) in favor of female candidates – proposals that include two candidates and both candidates are females and proposals that include only one single female candidate, 2) mixed – proposals that include two candidates and one candidate is male and one candidate is female, and 3) in favor of male candidates – proposals that include two candidates and both candidates are male and proposals that include only one male candidate.

Figure 2 shows that in all-male groups the proposals are slightly inclined for male candidates in comparison to other type of proposals. In all-female groups, almost the 50% of proposals are in favor of female candidates. In mixed-gender groups, the results differ. In male-majority groups, mixed proposals are the most common option, either made by men or women. In contrast, men disproportionately opt for proposals including male candidates in female-majority groups, while only the 15% of the proposals were in favor of female candidates. Women, in female-majority groups, opted mostly for proposals that include only female candidates.

The distribution of the proposals in Figure 2 is corroborated in Table A4 from a multinomial logit model with the type of proposal as dependent variable and with mixed proposals considered the reference type of proposal.

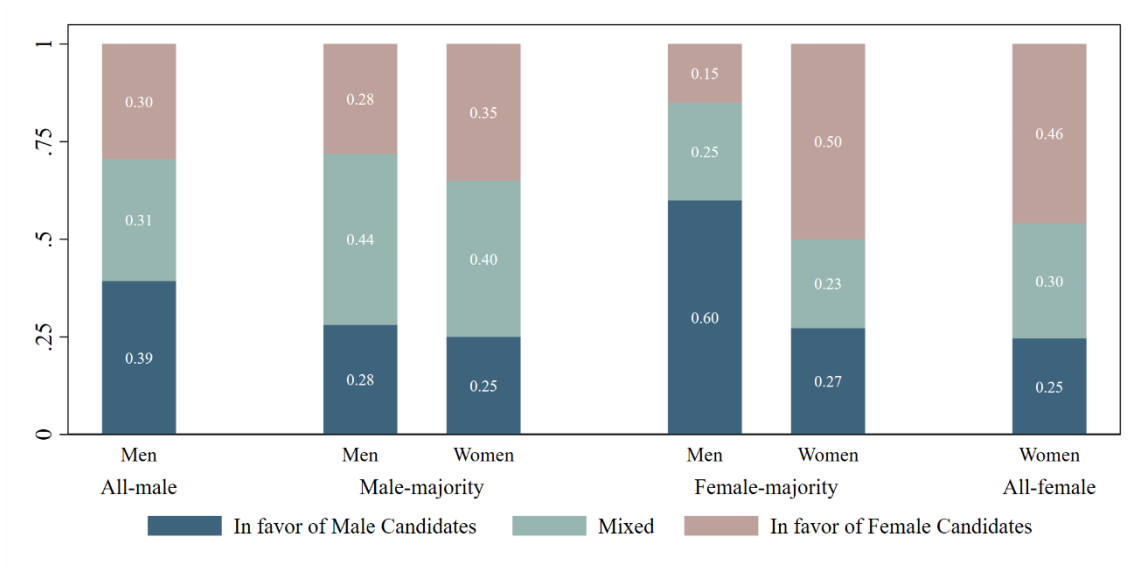


Figure 2. Distribution of the proposals by the gender of the evaluator

Result 4: Men are more likely to propose only male candidates in female-majority groups. In male-majority groups, mixed proposals are the most common either made by men and women.

3.4. Ex-post analysis

This section analyzes the probabilities of the best candidates to be selected according to the gender composition of the groups. A candidate, in each pool, is defined as best candidate if she is in the Top-2 according to the productivity in the mathematical task. If we assume that evaluators will take the performance in the adding task as the best predictor of performance in the mathematical task, it is not trivial to focus on how the signal of performances predicts the performance in the mathematical task. In the sample of 120 candidates in the parallel study, the performances in the adding task are positively and significantly correlated with the performance in the mathematical task for both males (Pearson's $r = 0.381$; $p < 0.01$) and females (Pearson's $r = 0.231$; $p < 0.1$) with no significant differences among the correlation for men and for women (Fisher Z transformation, $Z = 0.880$; $p\text{-value} = 0.378$).

Table 4. Marginal effects of the probit regression on the probability of candidates of being recruited in the group decisions.

Dependent variable: Pr (Selected- $G_{ip}=1$)			
	(1)	(2)	(3)
Best candidate	-0.066*** (0.017)	-0.066*** (0.017)	-0.090*** (0.008)
Male-majority		-0.003** (0.001)	-0.024*** (0.003)
Female-majority		0.006 (0.004)	0.001 (0.006)
All-female		0.005** (0.002)	-0.001 (0.002)
Male-majority x Best candidate			0.066*** (0.001)
Female-majority x Best candidate			0.014** (0.007)
All-female x Best candidate			0.016*** (0.009)
Best candidate + Male-majority x Best candidate			-0.023***
Best candidate + Female-majority x Best candidate			-0.075***
Best candidate + All-female x Best candidate			-0.073***
Additional Controls	Yes	Yes	Yes
Observations	720	720	720
Pseudo R2	0.341	0.341	0.342

Additional Controls: workers' characteristics (gender, age, field of study, performance adding task and position), share of females in the group and period. Robust standard errors, clustered at composition level, in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 4 presents the marginal effects of a number of probit regressions on the probability of the candidate i in pool p of being selected on a dummy variable that equals 1 if the candidate is among the best two candidates according to the productivity in the mathematical task and 0 otherwise. The models in Table 4 are similar than those in Table 2 but considering ability instead of gender. The estimations are controlled by other candidates' characteristics (gender, age, performance in the adding task, position), share of females in the pool and period.

In Column (1), the coefficient *Best candidate* is negative and significant, suggesting that, overall, the best candidates have less chances to be selected in comparison to candidates who are not among the best by 6.6 percentage points. This effect remains stable in column (2) when controlling for the gender composition of the groups. Column (3) adds three interactions between each type of group and *Best*

candidate to measure the gender gap in each type of groups in comparison to all-male groups. All the interactions are positive and significant, meaning that the best candidates have higher probabilities of selection in all groups in comparison to all-male groups. Moreover, the effect of the interaction between *Male-majority* and *Best candidate* is significantly higher than the effect of the interactions between *Best candidate* and *Female-majority* and *All-female*, suggesting that the best candidates have the highest probabilities of being selected in male-majority groups in comparison to the rest of the groups. Anyway, the linear combination of the effect of Best candidate and each of the interactions presents a negative effect in all the cases, suggesting that regardless the gender composition of the groups, the best candidates have lower probabilities to be selected in comparison to candidates who are not among the best in the mathematical task. In comparison to Table 2, the type of group that is most beneficial for females is also the most beneficial for the best candidates.

Result 5: The best candidates have higher probabilities to be selected when they are evaluated by male-majority groups in comparison to other groups.

4. Conclusions

The aim of this paper is twofold. First, it analyzes how the gender composition of the committees affects female candidates' probabilities of selection. To address this question, I designed a laboratory experiment in which subjects were allocated into groups of 3 evaluators and have to select two candidates to perform a mathematical task. The groups had four different compositions according to the number of females in each firm: All-male, Male-majority, Female-majority and All-females. I did not find own-gender preferences of evaluators when they are making decisions individually. Both men and women benefited male over female candidates. As shown by related studies addressing hiring decisions in the lab, this result is driven by the task employed in the design. Prior beliefs associating men with math and science fields, as shown by the analysis of the Implicit Association Test (IAT), generate the gender gap in this setting. On the other hand, group decisions seem to be more driven by the gender composition of the committee. I find that groups in which men are majority are the most beneficial for female candidates. Female-majority groups, in contrast, are the most detrimental. These results suggest that more women in the committees do not necessarily improve female candidates' outcomes, remaining the generalized assumption that women should benefit women still unclear.

An important exercise to understand this assumption is to ask why one is right when considering that more women in committees will automatically improve female candidates' outcomes. It can be that gender roles, peer effects or a pure in-group bias (Eriksson et al., 2015) model this thinking. However,

the evidence on this is very limited. Therefore, before moving to policy recommendations we need more research to understand why and under which conditions females should benefit other females, and b) to better understand how the gender of the evaluators interferes with the chances of candidates. This paper also proposes a mechanism that could explain the counter-intuitive results observed in the group decisions. It analyzes how men and women behave in group deliberations, specifically in terms of voice and influence. The results suggest that in groups that have been more beneficial for females (i.e. male-majority groups), men and women behave similarly: they have the same level of voice and no particular gender is more influential than the other. Interestingly, the most beneficial groups for females coincides with the groups that are more beneficial for the best workers. In other words, male-majority groups are not the most beneficial for female candidates but the more accurate in terms of efficiency. A limited conclusion would be that gender parity in the deliberation exerts the more efficient results. In contrast, in female-majority groups, women speak more than men; however, men are more influential in the sense that their initial beliefs in the individual stage are more likely to be incorporated in the final decision of these groups. From the mechanism proposed, we could conclude that the improvement of female candidates' outcomes do not only come from having more women in committees, but also from group dynamics that present gender equality among its members.

References

- Ambrus, A., Greiner, B., & Pathak, P. (2009). Group versus individual decision-making: Is there a shift. *Institute for Advanced Study, School of Social Science Economics Working Paper, 91*.
- Azic, M., & Lamé, D. (2018). Subjective information in hiring decisions.
- Babcock, L., Recalde, M. P., & Vesterlund, L. (2017b). Gender Differences in the Allocation of Low-Promotability Tasks: The Role of Backlash. *American Economic Review, 107*(5), 131-35.
- Babcock, L., Recalde, M. P., Vesterlund, L., & Weingart, L. (2017a). Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review, 107*(3), 714-47.
- Bagues, M., Sylos-Labini, M., & Zinovyeva, N. (2017). Does the gender composition of scientific committees' matter? *American Economic Review, 107*(4), 1207-38.
- Bagues, M. F., & Esteve-Volart, B. (2010). Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment. *The Review of Economic Studies, 77*(4), 1301-1328.
- Beaurain, G., & Masclet, D. (2016). Does affirmative action reduce gender discrimination and enhance efficiency? New experimental evidence. *European Economic Review, 90*, 350-362.
- Becker, G. S., Block, W., Fraser Institute (Vancouver, C.-B.), Sowell, T., & Vonnegut, K. (1982). *Discrimination, affirmative action, and equal opportunity*. Fraser Institute.

- Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, 94(4), 991-1013.
- Bohnet, I., Van Geen, A., & Bazerman, M. (2015). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, 62(5), 1225-1234.
- Bolton, G. E., Ockenfels, A., & Stauf, J. (2015). Social responsibility promotes conservative risk behavior. *European Economic Review*, 74, 109-127.
- Born, A., Ranehill, E., & Sandberg, A. (2019). A Man's World? The Impact of a Male Dominated Environment on Female Leadership. *The Impact of a Male Dominated Environment on Female Leadership (March 2019)*.
- Bosler, M., Mosthaf, A., & Schank, T. (2019). Are female managers more likely to hire more female managers? Evidence from Germany. *ILR Review*, 0019793919862509.
- Bouas, K. S., and Komorita, S. (1996). "Group Discussion and Cooperation in Social Dilemmas." *Personality and Social Psychology Bulletin* 22: 1144–50.
- Brescoll, V. L. (2011). "Who Takes the Floor and Why: Gender, Power, and Volubility in Organizations." *Administrative Science Quarterly* 56 (4): 622–41.
- Broder, I. E. (1993). "Review of NSF Economics Proposals: Gender and Institutional Patterns." *American Economic Review* 83 (4): 964–70.
- Brunette, M., Cabantous, L., & Couture, S. (2015). Are individuals more risk and ambiguity averse in a group environment or alone? Results from an experimental study. *Theory and Decision*, 78(3), 357-376.
- Burns, J., L. Corno, and E. La Ferrara (2016). Interaction, stereotypes and performance. Evidence from south africa. Working Paper.
- Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, 129 (3), 1409-1447.
- Carbone, E., & Infante, G. (2015). Are groups better planners than individuals? An experimental analysis. *Journal of Behavioral and Experimental Economics*, 57, 112-119.
- Carlana, M. (2017). Stereotypes and self-stereotypes: Evidence from teachers' gender bias. *Opublicerat manuskript*.
- Carpenter, T. P., Pogacar, R., Pullig, C., Kouril, M., Aguilar, S., LaBouff, J., ... & Chakroff, A. (2018). Survey-software implicit association tests: A methodological and empirical analysis. *Behavior Research Methods*, 1-15.
- Casadevall, A. and Handelsman, J. (2014). "The Presence of Female Conveners Correlates with a Higher Proportion of Female Speakers at Scientific Symposia." *mBio* 5 (1): e00846–13.

- Charness, G., Cobo-Reyes, R., & Sanchez, Á. (2018). Anticipated discrimination, choices, and performance: experimental evidence.
- 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., Exley, C. L., & Niederle, M. (2018). *When gender discrimination is not about gender*. Harvard Business School.
- Correll, S. J., Benard, S., & Paik, I. (2007). Getting a job: Is there a motherhood penalty? *American journal of sociology*, *112* (5), 1297-1338.
- Daskalova, V. (2018). Discrimination, social identity, and coordination: An experiment. *Games and Economic Behavior*, *107*, 238-252.
- Deschamps, P. (2018). *Gender Quotas in Hiring Committees: a Boon or a Bane for Women?* (No. 82). Sciences Po.
- Ellemers, Naomi, Henriette Van den Heuvel, Dick de Gilder, Anne Maass, and Alessandra Bonvini. (2004). "The Underrepresentation of Women in Science: Differential Commitment or the Queen Bee Syndrome?" *British Journal of Social Psychology* *43* (3): 315–38.
- European Union (2019). "Report on equality between women and men in the EU". ISBN: 978-92-76-00027-3 doi: 10.2838/395144.
- Experiment Exploring How Pay and Representation Differentially Shape Bias on the Pathway into Organizations." *Journal of Applied Psychology* *100* (6): 1678–1712.
- Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, *10*(2), 171-178.
- Flabbi, L., Macis, M., Moro, A., & Schivardi, F. (2019). Do female executives make a difference? The impact of female leadership on gender gaps and firm performance. *The Economic Journal*, *129*(622), 2390-2423.
- Glover, D., Pallais, A., & Pariente, W. (2017). Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores. *The Quarterly Journal of Economics*, *132*(3), 1219-1260.
- Greenwald, A. G., B. A. Nosek, and M. R. Banaji (2003). Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of personality and social psychology* *85*(2), 197.
- Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114-125.
- Guido, G., Peluso, A. M., Capestro, M., & Miglietta, M. (2015). An Italian version of the 10-item Big Five Inventory: An application to hedonic and utilitarian shopping values. *Personality and Individual Differences*, *76*, 135-140.
- He, H., Martinsson, P., & Sutter, M. (2012). Group decision making under risk: An experiment with student couples. *Economics Letters*, *117*(3), 691-693.

- Heinz, M., Normann, H. T., & Rau, H. A. (2016). How competitiveness may cause a gender wage gap: Experimental evidence. *European Economic Review*, 90, 336-349.
- Karpowitz, C. F., Mendelberg, T. and Shaker, L. (2012). "Gender Inequality in Deliberative Participation." *American Political Science Review* 106 (3): 533-47.
- Kunze, A., & Miller, A. R. (2017). Women helping women? Evidence from private sector data on workplace hierarchies. *Review of Economics and Statistics*, 99(5), 769-775.
- La Mattina, G., Picone, G., Ahoure, A., & Kimou, J. C. (2018). Female leaders and gender gaps within the firm: Evidence from three Sub-Saharan African countries. *Review of Development Economics*, 22(4), 1432-1460.
- Leibbrandt, A., Wang, L. C., & Foo, C. (2017). Gender quotas, competitions, and peer review: Experimental evidence on the backlash against women. *Management Science*, 64(8), 3501-3516.
- Lowes, S., N. Nunn, J. A. Robinson, and J. Weigel (2015). Understanding ethnic identity in africa: Evidence from the implicit association test (iat). *American Economic Review* 105(5), 340-45.
- Mathis, J. (2011). "Deliberation with Evidence." *American Political Science Review* 105: 516-29.
- Masclot, D., Colombier, N., Denant-Boemont, L., & Loheac, Y. (2009). Group and individual risk preferences: A lottery-choice experiment with self-employed and salaried workers. *Journal of Economic Behavior & Organization*, 70(3), 470-484.
- Milkman, K. L., Akinola, M. and Chugh, D. (2015). "What Happens Before? A Field Experiment Exploring How Pay and Representation Differentially Shape Bias on the Pathway into Organizations." *Journal of Applied Psychology* 100 (6): 1678-1712.
- Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, 122 (3), 1067-1101.
- Paola, M., and Scoppa, V. (2015): "Gender Discrimination and Evaluators' Gender: Evidence from Italian Academia," *Economica*, 82(325), 162-188.
- Paryavi, M., Bohnet, I., & van Geen, A. (2019). Descriptive norms and gender diversity: Reactance from men.
- Peterle, E., & Rau, H. A. (2017). Gender differences in competitive positions: Experimental evidence on job promotion.
- Quintana-Garcia, C., & Elvira, M. M. (2017). The effect of the external labor market on the gender pay gap among executives. *Ilr Review*, 70(1), 132-159.
- Reuben, E., Sapienza, P., & Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 111(12), 4403-4408.
- Reuben, E., Sapienza, P., & Zingales, L. (2015). Taste for competition and the gender gap among young business professionals (Tech. Rep.). National Bureau of Economic Research.

Shupp, R., and A. W. Williams. "Risk Preference Differentials of Small Groups and Individuals." *The Economic Journal*, 118(525), 2008, 258–83.

Sutter, M. (2007). Are teams prone to myopic loss aversion? An experimental study on individual versus team investment behavior. *Economics Letters*, 97(2), 128-132

Viscusi, W. K., Phillips, O. R., & Kroll, S. (2011). Risky investment decisions: How are individuals influenced by their groups? *Journal of Risk and Uncertainty*, 43(2), 81.

Williams, W. M., & Ceci, S. J. (2015). National hiring experiments reveal 2: 1 faculty preference for women on STEM tenure track. *Proceedings of the National Academy of Sciences*, 112(17), 5360-5365.

Zhang, J., & Casari, M. (2012). How groups reach agreement in risky choices: an experiment. *Economic Inquiry*, 50(2), 502-515.

Appendix A. Additional Results

Table A1. Pool of workers

Pool	ID Number	Gender	Year of Birth	Field of Study*	Performance Adding Task
1	#1	Female	1992	Sci	17
	#2	Female	1998	SSci	16
	#3	Male	1995	SSci	18
	#4	Female	1995	SSci	18
	#5	Female	1993	Eng	16
	#6	Male	1996	Eng	17
2	#1	Female	1992	Sci	15
	#2	Male	1980	Sci	16
	#3	Male	1997	SSci	16
	#4	Male	1992	Sci	15
	#5	Female	1989	Eng	16
	#6	Male	1994	Eng	15
3	#1	Male	1994	SSci	19
	#2	Male	1998	Sci	19
	#3	Female	1990	Eng	18
	#4	Female	1997	SSci	19
	#5	Male	1993	Eng	19
	#6	Female	1995	SSci	20

* Sci: Sciences; SSci: Social Sciences; Eng: Engineering.

Table A2. Probit regressions on the probability of candidates of being recruited.

Type of decision: Dependent variable:	Group decisions				Individual decisions			
	Pr (Selected-G _{ip} =1)				Pr (Selected-I _{ip} =1)			
	All-male	Male- majority	Female- majority	All-female	All-male	Male- majority	Female- majority	All-female
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female candidate	-0.0793 (0.0954)	0.0661 (0.0709)	-0.168** (0.0735)	-0.0430 (0.0871)	-0.0247 (0.0388)	-0.0998** (0.0424)	-0.118*** (0.0351)	-0.0308 (0.0290)
Performance Adding Task	-0.0214 (0.0185)	0.0279 (0.0214)	0.0377* (0.0211)	0.0289** (0.0113)	0.0243*** (0.00897)	0.0191** (0.00903)	0.0338*** (0.00749)	0.0345*** (0.00895)
Age	-0.0393*** (0.0134)	-0.0216* (0.0111)	0.0146 (0.0149)	-0.0149 (0.0128)	-0.00811 (0.00954)	-0.0156* (0.00875)	-0.00957 (0.00835)	-0.00354 (0.00840)
Sciences	0.381*** (0.119)	0.494*** (0.0864)	0.257* (0.141)	0.342*** (0.119)	0.329*** (0.0865)	0.372*** (0.0882)	0.256*** (0.0773)	0.282*** (0.0851)
Engineering	0.732*** (0.0943)	0.783*** (0.109)	0.477*** (0.103)	0.542*** (0.121)	0.654*** (0.0922)	0.816*** (0.0701)	0.542*** (0.0795)	0.519*** (0.0824)
Share of females	-0.0270 (0.163)	-0.165* (0.0965)	0.377*** (0.136)	0.0118 (0.188)	0.0861 (0.0879)	0.0866 (0.0833)	0.0433 (0.0827)	0.0824 (0.0860)
Position	-0.0314 (0.0374)	-0.00513 (0.0531)	0.0322 (0.0338)	0.0280 (0.0437)	0.0460*** (0.0178)	0.0205 (0.0253)	0.0226 (0.0176)	0.0405* (0.0207)
Period	-0.00582 (0.00842)	0.00531 (0.00503)	0.0233*** (0.00577)	-0.00349 (0.00905)	0.00244 (0.00426)	0.00151 (0.00261)	0.00498*** (0.00186)	-0.00128 (0.00667)
Individual Decision	0.363*** (0.0521)	0.294*** (0.0339)	0.339*** (0.0490)	0.438*** (0.0421)				
Additional Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	540	540	540	540	540	540	540	540

Additional Controls: workers' characteristics (age, field of study, performance adding task and position), votes in the individual stage, share of females in the group and period. Additionally, in models (1) to (4): individual decisions. Robust standard errors, clustered at composition level in (1) to (4) and at evaluator level in (5)-(8), in parentheses. The models present the marginal effects. *** p<0.01, ** p<0.05, * p<0.1.

Table A3. Mann-Whitney tests for group differences in the number of words written in deliberations*

	All-male	Male-majority	Female-majority	All-female
All-male		-5.024 (0.000)	-0.515 (1.000)	3.194 (0.004)
Male-majority			4.980 (0.000)	6.878 (0.000)
Female-majority				4.173 (0.000)
All-female				

*P-values, adjusted for multiple comparisons (Bonferroni), in parentheses.

Table A4. Multinomial logit regression on the type of proposals.

Dependent variable:	Men		Women	
	in favor of males (1)	in favor of females (2)	in favor of males (3)	in favor of females (4)
Male-majority	-0.784*** (0.109)	-0.274** (0.137)	-0.267 (0.879)	-0.212 (0.280)
Female-majority	0.673*** (0.220)	-0.779*** (0.158)	0.861*** (0.162)	0.242*** (0.089)
Constant	2.861 (4.326)	-0.0177 (2.721)	-3.029 (1.143)	-0.275 (0.908)
Controls ^a	Yes	Yes	Yes	Yes
Observations	113	113	125	125

^aAdditional Controls: first in talk, that the proposal is the first proposal made in the group, proportion talk, personality traits and period. Robust standard errors, clustered at composition level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

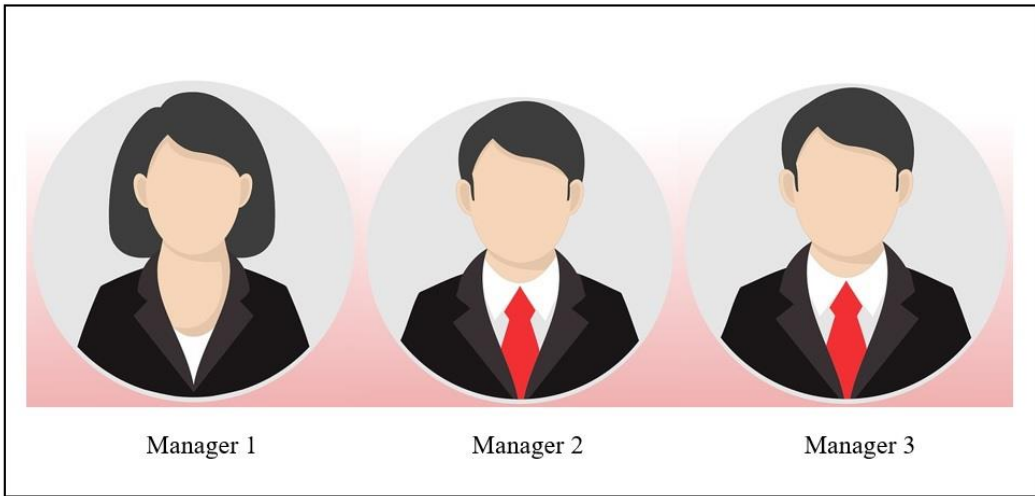


Figure A1. Example of gender displayed in male-majority groups.

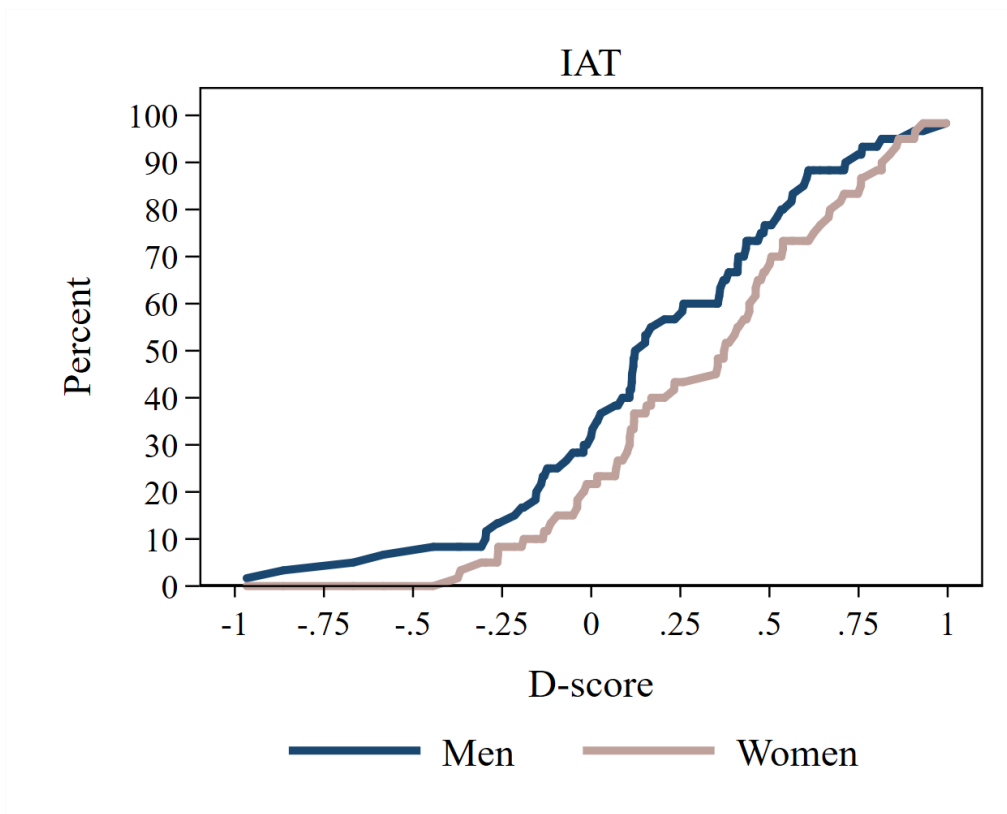


Figure A2. Cumulative distribution of the D-score by gender

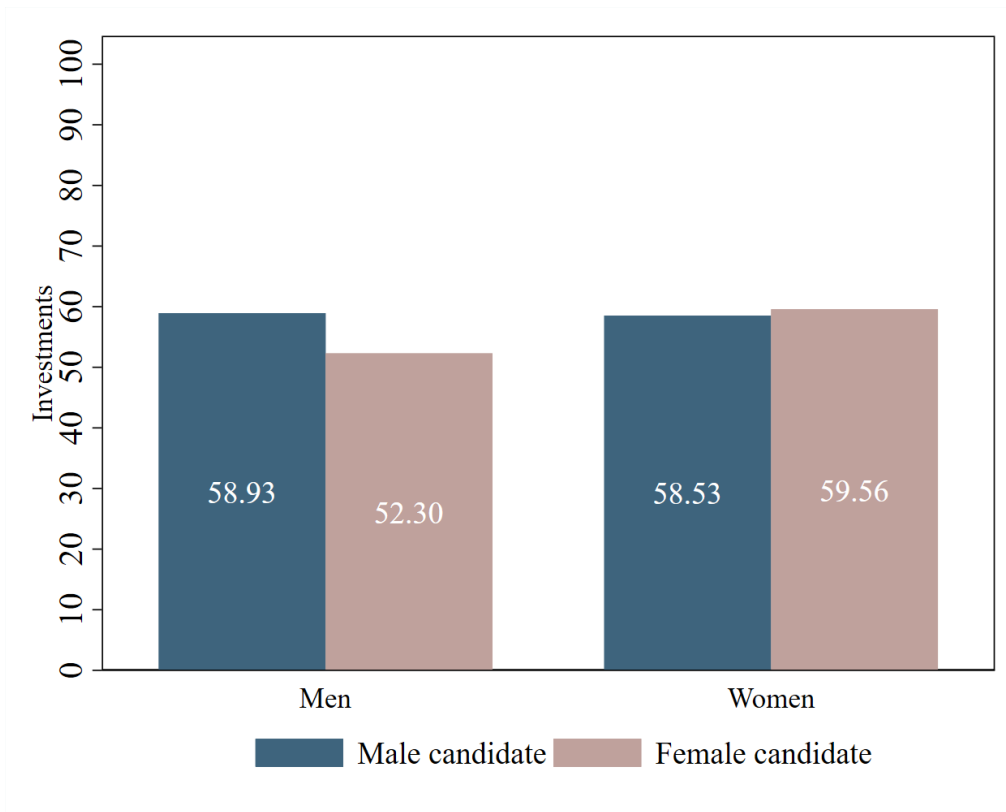


Figure A3. Average investments by gender

Appendix B. Instructions and The Implicit Association Test.

INSTRUCTIONS

Introduction

You are taking part in a decision-making study financed by University of Bologna and the University of Padova. During this study you can earn an amount of money according to the rules that will be described in the following pages. The payment will be paid by cash and confidentially. The duration of the present study will be around 1 hour and 15 minutes.

Today the study and is composed by 3 rounds. In each round, you have to make two decisions. At the end of the study, the computer will randomly select one round and you will be paid one of the two decisions in such round, also randomly determined. So, your final earnings in this study will be composed by the earnings of one decision plus 5 Euro show-up fee. During the study, the currency unit will be ECUs that will be translated into Euro at the end of the experiment at the rate of (100 ECU=0.50 Euro). The rules we will follow to determine your earnings are similar in each round. You will receive the instructions soon.

Communicating with other participants during the study is forbidden. The use of electronic devices will determine the exclusion from the study and from the payment. If you have questions during the study, please raise your hand. An assistant will arrive to your station to answer privately.

General Instructions

In this study, we will assign you the role of evaluator of a firm. The firm is composed by 3 evaluators (including yourself) and 6 workers.

Candidates in the firm face two tasks. The first one is called “Adding Task” and it consists in summing up as many three three-digit numbers as possible in 6 minutes. The second task is called “Problem Task” and it consists in solving as many mathematical problems as possible in 10 minutes. In this session, all participants play the role of evaluators. Candidates were recruited in previous sessions, performed both tasks (the “Adding Task” and the “Problem Task”) and were paid according to their performance.

Your job is to select 2 candidates out of 6 to work in the “Problem task”. Your earnings as evaluator will be composed by the performance that the two candidates selected have obtained in the “Problem task” plus a fixed amount (which you receive for the productivity of the remaining candidates in the “Adding task”).

Before the study starts

Before the study starts, you will face three different screens (Screen 1, Screen 2 and Screen 3). In Screen 1, you will have to fill a CV with your baseline information. Then, in order to familiarize yourself with the tasks for which you have to make a decision you will participate in a non-rewarded stage of the “Adding Task” (Screen 2) and the “Problem Task” (Screen 3).

In Screen 2, your task is to correctly solve the higher number as possible of additions, as the candidates did. You will have 1 minute to solve additions of three three-digit numbers as illustrated below. The numbers to sum will be selected randomly. You will see a scheme like the one represented below.

526	414.0	780
-----	-------	-----

Your task in Screen 3 will be to correctly solve as many mathematical problems as possible. You will have 2 minutes. You will see a scheme like the one presented below and you will have to select one of the possible answers.

Dati i due polinomi $(a^2 + b - 3)$ e $(4 - b)$, il loro prodotto è uguale a:

- A. $4a^2 + ab + b - b^2 - 12$
- B. $4a^2 - a^2b + 7b - b^2 - 12$
- C. $4a^2 + a^2b + 7b - b^2 + 12$
- D. $-4a^2 - a^2b + b + b^2 - 12$

Calculators or electronic devices are forbidden. It is possible to use the sheets of paper and the pencil that you will find in your desk. When you are ready, you can insert your answer choosing one of the answers and click the red button.

Immediately, the computer will say if the answer is correct or not. Your answers are anonymous.

What is happening now?

If you have questions, please raise your hand. An assistant will arrive to your station to answer privately.

Before the start of the study, we will ask you to respond some questions to verify if you understood the rules correctly.

Instructions

In this study, we will assign you the role of evaluator and you will be allocated into a group composed by 6 candidates and 3 evaluators (including yourself). As evaluator, you must take two decisions sequentially:

Decision 1.

You must select 2 candidates out of 6 of your firm to work on the “Problem task”.

- You will earn 30 ECU per each problem correctly solved by the candidates selected in the “Problem task” plus a fixed amount of 350 ECU.
- Example: If you select Candidate1 and Candidate2 and each candidate correctly solved 7 (C_{W1}) and 8 problems in the “Problem Task” (C_{W2}), respectively, your earnings will be computed as follows:

$$\text{Earnings} = 30 \text{ ECU} \times (C_{W1} + C_{W2}) + 350 \text{ ECU.}$$

$$\text{Earnings} = 30 \text{ ECU} \times (7 + 8) + 350 \text{ ECU} = 800 \text{ ECU.}$$

Decision 2.

You must select, jointly with the other two evaluators in your group, 2 candidates of your firm to work on the “Problem task”. This decision is composed by 3 steps.

First step: Communication. You will have 3 minutes to communicate with the other evaluators on your computers using a chat box. You are invited to enter free-format messages in order to reach a group decision. Evaluators will be identified as “Evaluator 1”, “Evaluator 2” and “Evaluator 3”. We invite “Evaluator 1” to start the conversation. We recommend to start the conversation stating your preferences about the candidates that you would like to select. Revealing personal information is forbidden and implies the immediate exclusion from the study and the payment. We also recommend to be respectful with the other evaluators in the group. After the 3 minutes, the chat will end automatically.

Second step: Voting. Once the chat stage is concluded, you will proceed to a voting stage. You have to vote for 2 candidates to work on the “Problem task”. The group outcome will be determined through a **majority voting rule**. That means that a candidate will be selected only if she receives, at least, two votes. We consider the following situations:

- **If two candidates are selected by the group** (i.e. two candidates receive two or three votes), your group will reach an agreement and you will have two positions filled by the candidates with two or more votes. If this is the case, the decision is made.
- **If only one candidate is selected by the group** (i.e. only one candidate has two or three votes), your group will reach an agreement only with respect to one worker. Therefore, your group will have one position filled by the candidate with two or more votes and one vacant position. In this case, the chat box will open again for an extra minute and you will have a second voting stage to fill the vacant position. If the agreement is not reached, you will finally have one filled position by the candidate with two or more votes and one vacant position. If you reach an agreement for two workers, you will finally have two filled positions by the candidates with two or more votes and no vacant positions.
- **If no candidate is selected by the group** (i.e. no candidate receives two or three votes), your group will reach full disagreement. You will have two vacant positions. In this case, the chat box will open again for an extra minute and you will have a second voting stage to fill both positions. If the disagreement still

persists, you will finally have two vacant positions. If you reach an agreement for only one worker, you will finally have one filled position by the candidate with two or three votes and one vacant position. If you reach an agreement for two workers, you will finally have two filled positions by the candidates with two or three votes and no vacant position.

Note: since each evaluator has two votes and candidates need at least two votes to be selected, it is possible that the group selects three candidates with 2 votes each. In this case, the computer will consider this situation as disagreement and the chat will open again and the group will proceed to the second voting.

Third step: Earnings. You will earn 30 ECU per each problem correctly solved by the candidates with two or three votes by the group and 0 ECU per each vacant position, plus a fixed amount of 350 ECU.

- Example: **Suppose the group selects, by majority, Candidate1 and Candidate2**, and each candidate solved 7 and 8 problems correctly in the “Problem task”, respectively.

Earnings for each member in the group = $30 \text{ ECU} \times (7 + 8) + 350 \text{ ECU} = 800 \text{ ECU}$.

- Example: **Suppose the group selects, by majority, only Candidate1** that solved 7 problems correctly in the “Problem task” and consequently one vacant position.

Earnings for each member in the group = $30 \text{ ECU} \times (7 + 0) + 350 \text{ ECU} = 560 \text{ ECU}$.

- Example: **Suppose that no candidate is selected.**

Earnings for each member in the group = $30 \text{ ECU} \times (0+0) + 350 \text{ ECU} = 350 \text{ ECU}$.

Candidates' information

Before selecting workers, you will have the chance to look their CVs by using the information relative to each candidate that have already participated in the study. Candidates will be identified with a number between 1 and 6. You will receive the following information about workers: **ID number, Age, Gender, Field of Study and a Signal of Performance**. The Signal is the number of sums that candidates solved correctly in the “Adding task”. You won't be provided with information relative to the performance in the “Problem Task”.

Take into account that you are selecting candidates to work on the “Problem Task” and that the signal of performance corresponds to the “Adding Task”. That means, as an example, that the best candidate in the “Adding task” may not be the best candidate in the “Problem task”, or that the worst candidate in the “Adding task” may not be the worst candidate in the “Problem task”, because the two tasks are different.

Repetitions and final earnings.

The decision context we just described will be repeated **3 times (which we call rounds)**. In each round you will be allocated into a new group of 6 candidates and 3 evaluators. You will always play the role of evaluator. Candidates and evaluators will be different across rounds. In the whole experiment, you

will select 12 workers, four in each repetition (two individually and two by groups). At the end of the experiment, the computer will randomly select one round for payment. Then, the computer will again select one decision (individual or group) in that round to determine your final payments.

Timing

1. Individual decision.
2. Group chat.
3. First Voting stage.
4. Feedback about the outcome of the first voting.
5. Extra-minute chat (if no agreement in two workers).
6. Second voting stage (if no agreement in two workers).
7. Feedback about the outcome of the second voting (if no agreement in two workers).
8. Repetition 1-7 (3 rounds).
9. Feedback about earnings in each round.
10. Feedback about the final earnings.

What's happening now?

If you have any question about the study, we ask you to raise your hand, an assistant will arrive to your station to solve your questions in private. You will be informed when the study starts.

THE IMPLICIT ASSOCIATION TEST (IAT)

In this task, participants faced a screen where different words appeared randomly and were asked to quickly respond by pressing the left-handed key or the right-handed key depending on the side in which the category or the attribute that the word belonged was placed. Each word was classified into either a category (“male” or “female”) or an attribute (“math and science” or “humanities”). The words used for “math and science” were “biology”, “physics”, “chemistry”, “mathematics”, “geology”, “astronomy” and “engineering”. For “humanities”, “philosophy”, “humanities”, “art”, “literature”, “Italian”, “music”, “history”. The words used for the category “male” were “man”, “father”, “male”, “grandfather”, “husband”, “uncle”. For “female”, “aunt”, “woman”, “wife”, “mother”; “grandmother”; “female”. The task consisted of seven rounds. In each round the position of the attributes, categories or the combination between attributes and categories varied (see Table B). The IAT score (D-score) was constructed by comparing response times in the classification task, taking values between –2 and 2. The D-score of each was calculated according to the subject according to the algorithm developed by Greenwald et al. (2003). This measure is an indirect measure of association between gender groups and career characteristics (i.e. implicit bias or stereotypes) that assumes that the more rapid answer, the stronger the association between the category and the attribute of a given side. A positive score indicates an association of “male” with “math and science” and “female” with “humanities”. A negative score indicates an association of “female” with “math and science” and “male” with “humanities”.

Table B1. IAT rounds

	Left side	Right side
Round 1	Humanities	Math and Science
Round 2	Female	Male
Round 3	Humanities	Math and Science
	Female	Male
Round 4	Humanities	Math and Science
	Female	Male
Round 5	Male	Female
Round 6	Humanities	Math and Science
	Male	Female
Round 7	Humanities	Math and Science
	Male	Female

José Javier Domínguez Ramírez
PhD candidate
Department of Economics
University of Padova
Via del Santo, 33. 35123 Padova (Italy)
josejavier.dominguezramirez@phd.unipd.it