Gender and leadership in organizations: The threat of backlash*

Priyanka Chakraborty† Danila Serra‡

June 20, 2022

Abstract

Decisions made by leaders please some people and upset others. We examine whether having to communicate such decisions to the affected parties, and knowing that there might be backlash, have a differential impact on men's and women's self-selection into leadership roles, and their decisions as leaders. In a laboratory experiment that simulates corporate decision-making, we find that women are significantly less likely to self-select into a leadership position when they can receive backlash from employees. Once in a leadership role, women get more backlash. There is some evidence of gender differences in leaders' decision-making and communication styles under the threat of backlash, but little difference in final outcomes. Data from an online experiment where third parties are asked to predict behavior in the laboratory experiment, shed light on possible mechanisms.

JEL Codes: C92, D91, J16

Key words: Gender differences, leadership, backlash, experiment.

*We thank Manuela Angelucci, Jeff Carpenter, Catherine Eckel, Lata Gangadharan, Karen Hauge, Alan Love, Elira Kuka, Shanthi Manian and Tim Salmon for useful comments and discussions. We thank seminar participants at the American University, FAIR - The Choice Lab, Georgia State University, Southern Methodist University, Universidad de Los Andes, University of Queensland, University of Texas at Austin, Virginia Tech University, Washington State University, and the Discrimination and Disparities Online Seminar series. We obtained ethical clearance for the laboratory experiment from the SMU Institutional Review Board, and for the online experiment from the Texas A&M Institutional Review Board.

†Frost School of Business, Centenary College. Email: pchakraborty@centenary.edu

‡Department of Economics, Texas A&M University. Email: dserra@tamu.edu.

1 Introduction

Individuals in a position of leadership have to make decisions knowing that some people will be negatively affected by their choices and will therefore be unhappy. In any organization, managerial decisions concerning promotions, salary raises, demotions and dismissals necessarily generate negative judgment and possibly anger among some employees. Numerous studies have shown that, holding performance constant, women in a position of power are judged more negatively than men. This is true in politics, business, academia, as well as laboratory and online experiments (Abel, 2022; Ayalew et al., 2018; Boring, 2017; Egan et al., 2017; Grossman et al., 2019; Shurchkov and van Geen, 2017).[1] There is also evidence that women are more sensitive to public scrutiny (Alan et al., 2020; Jones and Linardi, 2014) and negative feedback (Mayo et al., 2012; Johnson and Helgeson, 2002), and are more severely discouraged by such feedback in male-dominated fields or stereotypically male tasks (Ellison and Swanson, 2018; Kugler et al., 2017).

In this paper, we ask whether and to what extent the possibility of backlash from subordinates induces women, more than men, to select out of leadership positions, and to perform differently when in a leadership role. We employ a laboratory experiment that simulates a business environment where managerial decisions, such as rank, job title or task allocations, affect the earnings of some workers positively and the earnings of others negatively.[2] This implies that backlash from unhappy employees, in the form of, at the minimum, negative judgment and disapproving messages, is likely.

In the experiment, a manager is matched with the same two employees for multiple decision-making rounds. This is to simulate the repeated interactions that occur within a team whose members differ in their job and leadership statuses. Following an initial real effort task, the manager's main responsibility is to decide which employee will be high rank and which employee will be low rank, where ranks determine employees' earnings. There can only be one high-rank and one low-rank employee; therefore, rank allocation necessarily creates income inequality between workers. This way, the experiment simulates an environment where the leader/manager has control over some financially meaningful aspects of his or her subordinates' jobs, e.g., their wages, job titles, or yearly bonuses. The manager needs to make the

[1] See also: Branton et al. (2018); Rheault et al. (2019); Elsesser and Lever (2011); Hengel (2017); Mengel et al. (2017); Sarsons (2017).

[2] This is a departure from the existing experimental studies of leadership, which have typically employed sequential public goods games or coordination games where leaders can induce followers to increase their contributions through leading by example (e.g., Güth et al., 2007; Grossman et al., 2015; Jack and Recalde, 2015) or through the use of messages suggesting contributions (e.g., Brandts and Cooper, 2007; Reuben and Timko, 2017). Other important studies of leadership have employed minimum-effort games or real effort tasks where leaders incentivize (Shurchkov and van Geen, 2017) or suggest the effort to be put in by followers (Chaudhuri et al., 2018; Erkal et al., 2022), or tasks that require leaders to make decisions on behalf of their group (Alan et al., 2020; Born et al., 2022; Reuben et al., 2012).

same decision - for instance whom to assign a bonus to - involving the same team members periodically, e.g. at the end of each fiscal year. Hence, in the experiment, at the beginning of each time period, the manager has to decide whether he of she wants to keep the workers' ranks or switch ranks, i.e., promote the employee who was previously assigned the low rank (and low salary). Importantly, workers engage in a real effort task in every time period, and the manager receives money only from the product of the worker who is assigned the high rank. This generates incentives to assign ranks based on the workers' relative productivity.

In our baseline condition (*No Choice*), we assign the managerial role based on performance in the preceding real effort task. In our treatments, we allow subjects to self-select into the managerial position (*Choice* treatment), and we progressively add the necessity to communicate with employees when assigning ranks (*Choice & Talk* treatment) and the possibility for employees to talk back and send angry emojis to express their disapproval of the rank allocation (*Choice & Backlash* treatment).[3] The use of angry emojis, in particular, allows us to quantify backlash towards managers in the experiment.

Through our treatment manipulations, we are able to isolate the impact of the possibility of backlash on men's and women's self-selection into leadership. Through the analysis of rank allocations and of the messages that managers send to their employees, we are also able to examine whether men and women make different decisions when in a leadership role, conditional on the possibility of backlash. Finally, by examining the angry emojis sent to male and female managers in the *Choice & Backlash* treatment, we can test whether women receive more backlash than men.

We find evidence of a gender gap in self-selection into leadership only when leaders face the possibility of backlash. Neither being responsible for generating income inequality between employees nor having to talk to employees when assigning ranks discourage women, as long as employees cannot send angry messages back. When backlash is possible, men are unaffected, while women are significantly less likely to want to be managers. In line with the existing literature, we find that women leaders get more backlash than men leaders.

When looking at the decisions made by men and women once in a leadership position, we find that the possibility of backlash increases the likelihood of rank-switching for women, but not men. However, this does not lead to profit losses, i.e., we do not see any significant differences in final outcomes across treatments and by gender. Finally, the analysis of the messages sent by managers to workers in the *Choice & Backlash* treatment suggests, in line with the existing literature, that men and women have different communication styles. Female managers praise employees more and use cordial words like thank you and sorry, whereas male managers are more likely to explain the ranking decision and to foster competition among

[3]The manager is selected among those who want to be managers based on performance in the preceding real effort task.

workers. However, these differences only emerge when workers can talk back to managers, which suggests that they reflect men's and women's different strategies aimed at minimizing worker backlash.

In order to investigate possible mechanisms behind the observed differences in men's and women's responsiveness to the possibility of worker backlash, we conducted an experiment on the online platform Prolific. The online experiment allows us to: 1) test whether men and women significantly differ in their aversion to negative feedback; and 2) test whether men and women differ in their beliefs regarding the backlash received by men and women managers in the laboratory experiment. This way, we aim to assess whether women's higher responsiveness to backlash at the self-selection stage and the rank-switching stage in the laboratory experiment are due to women's expectations of more severe backlash, as compared to men, and/or more aversion to any backlash. The data show that women score significantly higher in Brief Fear of Negative Evaluation (BFNE) psychological index (Carleton et al., 2011; Rodebaugh et al., 2004), which indicates that they are significantly more averse to the possibility of negative judgment than men. We also found that, while women (but not men) expect women managers to get more backlash than men managers, there are no statistical significant differences in the likelihood and severity of backlash that men and women expect their own gender to receive. This suggests that the higher responsiveness of women to the possibility of backlash in the laboratory experiment is likely driven by higher sensitivity to (any) backlash.

Our study contributes to the growing literature on the gender leadership gap, which has been documented in all spheres of life. For instance, only 19 percent of firms worldwide have female top managers and only 6 percent of CEOs at S&P 500 companies are women.[4] In politics, women hold only 23 percent of seats in national parliaments worldwide.[5] In academia, averaging across all fields, less than one third of full professors are women. The existing literature has identified behavioral or preference-based constraints to women's self-selection into top leadership roles.[6] These include risk aversion (see, e.g., Eckel and Grossman, 2008), reticence to initiate negotiations (e.g., Bowles et al., 2007; Babcock and Laschever, 2009; Exley et al., 2020), aversion to competitive environments (e.g., Gneezy et al., 2003; Niederle and Vesterlund, 2007; Flory et al., 2014; Preece and Stoddard, 2015),[7] preferences over job attributes (Wiswall and Zafar, 2017), time spent on low promotability tasks (Babcock et al.,

[4]For recent statistics on the gender leadership gap in the US, see Warner and Corley (2017)

[5]World Bank DataBank: https://data.worldbank.org/indicator/SG.GEN.PARL.ZS

[6]External demand-side constraints, such as taste-based or statistical discrimination stemming from traditional gender stereotypes concerning men and women's productivities, skills and family constraints, are of course also important. The existing evidence suggests that women are likely to be discriminated against in higher-status jobs, particularly in male-dominated fields. For a review of the literature, see Riach and Rich (2006), Azmat and Petrongolo (2014) and Bertrand and Duflo (2017).

[7]**?** show that women are more likely to compete for leadership roles if there is a system in place that, by default, enroll individuals in the competition, while allowing them to opt-out.

2017) and self-stereotyping (Coffman, 2014).[8] We add to this literature by showing that the threat of backlash contributes to the leadership gap by inducing women, but not men, to select out of leadership roles.

Our study also relates to the small literature in economics on gender differences in leaders' communication styles (Timko, 2017; Manian and Sheth, 2020), which shows that female leaders send less assertive messages than male leaders (Timko, 2017), even when they know that such messages are more effective in impacting others' behavior (Manian and Sheth, 2020). This is in line with studies in psychology showing that men tend to use more assertive language - e.g., through imperative statements - while women use more affiliative language - e.g., statements of support, agreement, and acknowledgment.[9] We contribute to this literature by showing that gender differences in the language used by male and female leaders to motivate employees do not reflect innate gender differences. Rather, they emerge as different strategies used by male and female managers to minimize backlash.[10] Finally, we add to the growing set of studies providing evidence of significant differences in individual attitudes toward male and female leaders (see, e.g., Abel, 2022; Grossman et al., 2019) by showing that women managers get more backlash than man managers from unhappy employees.[11]

2 The Experiment

2.1 Design

The experiment consists of 6 active stages (Stages 1 to 6), followed by a survey, as shown in Figure 1. An important feature of our design is the method used to reveal subjects' genders to other participants without making gender artificially salient in the game. We achieved this by asking subject to fill in a brief survey at the very beginning of the session, before Stage 1. The survey registers subjects' age, gender, field of study, and previous participation in an experiment. The answer to the gender question leads to a pre-determined list of either male or female names, which we took from Bertrand and Mullainathan (2004)'s correspondence

[8] A recent study by Born et al. (2020) also shows that being in a male-dominated team reduces women's willingness to lead, due to lower confidence and expectations of team support.

[9] See, for instance, Leaper and Smith (2004) and Kern et al. (2016).

[10] We also indirectly contribute to the literature on gender differences in executives' performance (see, e.g., Adams and Ferreira, 2009; Matsa and Miller, 2014). In line with existing studies, we find that female leaders are more likely to make decisions that benefit low-earning employees, but this holds only if managers have to personally communicate their decisions to their subordinates, and if they can receive angry messages back as a result.

[11] There is also a large literature on attitudes toward male and female leaders in psychology, sociology and management. These studies typically either provide written description of leadership situations, varying the sex of the leader, or use trained actors to lead, allowing the experimenters to control the degree of success the leader achieves (see, e.g., Swim et al., 1989).

study of race-based discrimination.[12] The male subjects see a list of male names and the female subjects see a list of female names. We inform subjects that for the duration of the experiment they will be identified with a fictitious name, and we invite them to pick a name from the gender-specific list they see on their screen.[13] We do not allow two or more subjects to choose the same name, so each name disappears from the list in real time when picked by another participant.

Stage 1 follows. In this stage, and in the following five stages, subjects engage in a real effort task. Previous studies of leadership have typically used public goods games or coordination games. In order to resemble firm environments where managers are chosen based on qualifications, in our study we wanted a game/task that would allow us to clearly assess participants' relative performance and select the best performing subject in a group as the leader. Specifically, we wanted a gender-neutral task requiring cognitive thinking and focus, where men and women would be equally confident and would perform equally well. We therefore chose a language task, as it has been shown (e.g., Dreber et al., 2014; Niederle, 2016) that language-based tasks are less likely than math-based tasks to generate gender differences in both self-confidence and performance in competitive environments.

In Stage 1, our real-effort task consists in finding a 4-letter word in a 6x6 letter matrix in 5 minutes, for a maximum of 20 matrices.[14] Subjects play individually. They receive an endowment of 40 Experimental Currency Units (ECU)[15] and earn 2 ECU for each puzzle they solve correctly in 5 minutes. At the end of Stage 1, subjects receive feedback on their absolute performance and are provided instructions on the following 5 stages (Stages 2 to 6) of the experiment. Crucially, they are randomized into groups of 3 and they are shown the fictitious names of their group members. In order to simulate male-dominated environments, the randomization algorithm created groups of 2 men and 1 woman, whenever possible.[16]

[12] As our focus is on gender differences, we use Bertrand and Mullainathan (2004)'s list of distinctively white sounding names only. Distinctive names are those that have the highest ratio of frequency in the corresponding racial group.

[13] We did not ask subjects to use their real names as we did not want to lift anonymity nor did we want the potential confounding bias of race, nationality or ethnicity associated with the actual name of the subject to play a role in the experiment.

[14] The decision screen is divided in two halves, as shown in the Online Appendix. On the left, subjects see the matrix and on the right, they see a list of 40 words. Each puzzle contains two words that appear on the list. In order to earn money, subjects have to identify one word per puzzle. We used the website http://tools.atozteacherstuff.com/word-search-maker/wordsearch.php to create the puzzles and the website http://www.thefreedictionary.com to find words of varying lengths. We ran some pilots of the puzzle task with varying levels of difficulty with different sizes of the matrix, different word lengths, and different ways in which words could be identified in the puzzle (forward, backward, up, down, diagonal etc). We found the configuration of finding 4-letter words that appear horizontally or vertically in a 6X6 matrix with a time of 5 minutes in Stage One to be optimal in creating enough heterogeneity in performance among subjects.

[15] Earnings from the experiment were later converted to dollars at the exchange rate of 6 ECU for 1 USD, as explained in Section 2.3.

[16] We ended up having 63% of the groups made of one woman and two men, 30% made of two women and one man, and 7% made of men only.

Since our interest is in work environments where manager and workers interact multiple times, so that a manager is responsible for the long-run income of his or her employees, subjects remain in the same group for the duration of the experiment. In Stages 2 to 6, two group members play in the role of workers and one in the role of manager. The roles of manager and worker are assigned at the end of Stage 1 according to treatment-specific rules, and retained through Stage 6. The manager gets a fixed wage of 100 ECU and his/her main task is to decide, at the beginning of each stage of the experiment, which worker will be rank A and which worker will be rank B in the following stage. The rank A worker gets a wage of 80 ECU, while the rank B worker gets a wage of 20 ECU. After the rank allocation, all members of the group engage in a similar puzzle task as in Stage 1 of the experiment.[17] Each correctly solved puzzle generates 2 ECU in addition to the initial wage. Moreover, each puzzle solved correctly by the Rank A worker generates 2 ECU also to the manager. This is to incentivize the manager to take the workers' relative productivities into account when making the rank allocation decision. In sum, the earnings from each of the 5 active stages (Stages 2 to 6) of the experiment are determined as follows:

- The manager gets 100 ECU + 2 ECU per puzzle + 2 ECU per puzzle solved by the rank A worker;
- The rank A worker gets 80 ECU plus 2 ECU per puzzle;
- The rank B worker gets 20 ECU plus 2 ECU per puzzle.

Stages 3 to 6 are identical to Stage 2. However, at the end of each stage of the experiment, the manager is informed about the performances of the current rank A and rank B workers and has to decide whether to keep or reassign ranks before the next stage begins.

Following Stage 6, subjects fill a post-experiment questionnaire,[18] where we elicit demographics, previous leadership experiences, and answers to personality questions that allow us to generate the Big 5 Agreeableness Personality Index, which has been shown to be significantly higher in women than men (Schmitt et al., 2008) and to correlate negatively with leadership ambition (e.g., Ertac and Gurdal, 2012).[19]

One feature of our design requires further discussion. In our setting, the nature of the task is such that the manager is always able to accurately assess the relative performances of the

[17] In order to account for learning effects, while in Stage 2 we keep the time limit to solve the 20 matrices equal to 5 minutes, we reduce the time to 4 minutes in Stages 3 and 4, and to 3.5 minutes in Stages 5 and 6.

[18] Due to a computer glitch, we were unable to conduct the post-experiment questionnaire on 12 participants.

[19] The Agreeableness Index measures the tendency to be kind, altruistic, trusting and trustworthy, and cooperative. There is evidence from psychology studies(e.g., Judge and Bono, 2000) that agreebleness predicts transformational leadership, i.e. leadership that operates through inspiration, intellectual stimulation and individual consideration.

two employees. This allows us to define and clearly measure managerial performance as: 1) the likelihood that the manager makes the first rank allocation based on performance in Stage 1, and 2) the likelihood that the manager makes an ex-post mistake in the rank allocation, by assigning rank A to a worker who ends up not performing as well as the other worker. We acknowledge that, in many settings, workers' performances cannot be objectively or precisely measured, and a manager's rank allocation decisions are at least partly discretionary; it is the lack of transparency and the subjectivity of the decision process that may be especially conducive to worker backlash. Note that even though our task generates objective workers' rankings, we still allow for lack of transparency and perceived subjectivity of the manager's decisions by not disclosing relative performances to the workers. In other words, the workers do not know how they compare to each other, and do not know what criteria the manager followed to allocate ranks.

2.2 Treatments

In our *No Choice* (NC) treatment, at the end of Stage 1, in each group of three participants the manager is chosen based on performance in Stage 1.[20] Recall that subjects participate in the Stage 1 real effort task individually without knowing anything about Stages 2 to 6, and therefore ignoring the fact that their performance would determine their role in the subsequent stages of the experiment. This prevents competition-driven anxiety from playing a role in determining subjects' performance and subsequent chances of becoming the manager of the group. While there is a large literature documenting gender differences in competitiveness and self-confidence (especially in math-based tasks), we abstract from both factors by design, in order to be able to isolate the role played by the possibility of worker backlash in the origination of gender differences in leadership.

At the beginning of Stage 2, subjects are informed that the manager was chosen based on performance rather than randomly. This is important, as we aimed to simulate an environment where employees could not doubt the qualifications of their manager. This way, any differences in workers' attitudes toward male versus female managers could not be attributed to differential subjective beliefs about the right of the manager to hold his or her role in the group.

In our *Choice* (C) treatment, we allow subjects to self-select into the leadership position. At the end of Stage 1, after receiving information about the next five stages of the experiment and the fictitious names of their group members, we ask subjects to state whether they would like to be the manager of their group. From the subset of those who want to be manager, we choose the manager based on performance in Stage 1, as in the *No Choice* treatment. Since,

[20] Since, by design, we have more men than women participating in each session of the experiment – due to the objective of having male-dominated groups – in all treatments we break ties in favor of women. Subjects are unaware of this.

by design, the manager is always the highest earner of the group, everybody should want to be manager.[21]

In the *Choice & Talk* (CT) treatment, we inform participants that the manager will have to assign ranks to workers in each stage, as before, and will also have to communicate with each worker when assigning ranks by sending a free-form message to the rank A and the rank B workers before the beginning of the next stage of the real effort task. Subjects are told that the manager can write anything he or she wishes to communicate to each worker. The process is repeated at each rank-allocation stage. As in the *Choice* treatment, after providing information about the next stages of the game, including tasks and payoffs of manager and workers, as well as the fictitious names of their group members, we ask subjects to state whether they want to be the manager of their group. If more than one group member wants to be manager, we select the manager following the same procedure as in the *Choice* treatment. In this treatment, an aversion to having to repeatedly provide feedback to subjects who are affected either positively or negatively by the rank allocation decision, may lead subjects to select out of the manager role[22] and impact men and women differentially.

Finally, in our *Choice & Backlash* (CB) treatment, the manager still has to send a free-form message to the rank A and the rank B workers, yet in this case, each worker can send a free-form message back to the manager. Moreover, the worker assigned rank B can send up to 5 angry emojis to the manager to express disapproval of the rank allocation. The messages sent by the two workers, including the angry emojis, are displayed to the manager before the next real-effort task begins. The process is repeated at each rank-allocation stage. As before, participants receive information about the rules governing Stages 2 to 6 of the experiment, including the presence of two-way communication and the possibility of receiving angry emojis from Rank B workers, before they are asked whether they would like to be the manager of their group.

The *Choice & Backlash* treatment most closely simulates a work environment where managers have to communicate their promotion or task allocation decisions to employees, knowing that employees will be able to talk back and express their discontent. By comparing men's and women's behaviors in this treatment versus the *Choice* and the *Choice & Talk* treatments

[21] Individuals may perceive the manager selection as a competitive process, hence, it is possible that gender differences in competitiveness (Niederle and Vesterlund, 2007) may hold women back. However, contrary to standard competition tasks, in our setting not becoming manager, i.e., losing the competition, does not come with a loss, since the consequence would be that the subject plays the game as a worker, i.e., same outcome as if he or she did not self-select into leadership. In other words, there is no risk in wanting to be manager. Still, there may be a gender difference in willingness to receive negative feedback about the outcome of the competition, as it could reveal information about one's relative ability in the real effort task. Even if that were true, it would apply to all our self-selection treatments; therefore, it would not act as a confound in our study of the behavioral mechanism of interest.

[22] For instance, subjects may feel uncomfortable at the thought of having to talk to workers assigned to be rank B, possibly multiple times during the course of the experiment.

we are be able to identify the role of anticipation of worker backlash on self-selection into a leadership role. Note that the decision to use angry emojis in the *Choice & Backlash* treatment follows the aim to make both workers and manager aware that expression of disapproval of the manager's decision is possible and allowed in the experiment - which is not easy to achieve with the use of free-text messages only, and without framing the instructions in a way that may prime or lead subjects to say specific things to each other. Moreover, emojis have become a widespread form of communication among the younger generations, and the meaning of an angry emoji is universally understood. Therefore, the use of emojis allows us to: 1) hold the boundaries of potential worker discontent constant (between 0 and 5); 2) quantify the attitudes of workers toward managers; and 3) examine how male and female managers respond to angry messages from employees, holding such messages constant.[23]

In addition to self-selection into a manager role, our design allows us to investigate gender differences in managers' decision-making. By design, only the performance of the rank A worker generates additional earnings to the manager. Therefore, the manager should assign rank A to the worker that he or she expects to be the best performer in the upcoming stage. The only information available to the manager is the workers' past performances, so we expect the difference in the workers' past productivities to play a primary role in the rank allocation decisions. However, managers' distributional concerns and/or the desire to avoid the discomfort of having to tell the same worker that he or she has been assigned rank B multiple times, and/or the objective to minimize the receipt of angry emojis, may lead managers to switch ranks between the two workers over the course of the experiment in order to equalize their earnings. Our treatment comparison allows us to examine gender differences in managers' rank switching behavior, and the mechanisms leading to them. We are also able to assess whether differential rank switching propensities lead to different outcomes, i.e., mistakes in rank allocations and profit losses.

Finally, by examining the free-form messages sent by managers to workers in the *Choice & Talk* and in the *Choice & Backlash* treatments, not only can we study gender differences in communication styles, but we can also examine the extent to which the languages used by male and female managers change under the threat of worker backlash. Furthermore, the analysis of both free-form messages and angry emojis sent by workers to managers provides a comprehensive assessment of differences in workers' attitudes toward male and female managers.

[23]We do not allow rank A employees to send happy emojis. Although this could be an interested extension of our study, as we discuss further in the Conclusion, the focus here is the anticipation of negative judgment and hostile messages from unhappy employees, hence the use of angry emojis only.

2.3 Implementation

We conducted 29 experimental sessions at the Laboratory for Research In Experimental Economics (LREE) at Southern Methodist University. We involved a total of 417 participants, of which 41% are women, as shown in Table 1. We employed a between-subject design, with each subject participating in only one session and one treatment. In each session, we had multiple independent groups of one manager and two workers. Groups were fixed for the duration of the experiment, and members of each group made decisions independently from all the other groups participating in a session.

As described in Section 2.2, the experiment consisted of an initial brief survey and name-assignment stage, followed by six active stages plus a post-experiment survey. Subjects were presented with the instructions for each stage on their computer screen immediately before that stage began. Additionally, at the beginning of Stage 1, they received hand-outs and verbal instructions about the puzzle-solving task, and at the end of Stage 1 they received further verbal and written instructions (hand-outs) about the rules applying to Stages 2 to 6 of the experiment.[24]

Only one randomly selected active stage of the experiment was used for actual payments. Experimental earnings were converted from ECU to dollars at the exchange rate of $1 for 6 ECU. The experiment was programmed in z-Tree (Fischbacher, 2007) and subjects were recruited among pre-registered LREE students.

In order to guarantee anonymity, at the beginning of each session subjects were randomly assigned an identification number, which they kept for the duration of the experiment. At no point during the experiment did we ask subjects to reveal their names and, although actual names were used during the payment process for accounting purposes, we informed subjects that we would not register their names and therefore would not be able to link them to the choices made in the experiment. Each session lasted between 60 and 90 minutes, with average earnings of $28 per subject, including a $10 show-up fee.

2.4 Estimation Strategy

We test the effects of the treatments on men's and women's decision to become a manager by estimating equation 1 below, separately for men and for women:

$$Y_i = \beta_0 + \beta_2 T_{3i} + \beta_3 T_{4i} + \delta \mathbf{X}_i + u_i \tag{1}$$

where Y_i is participant i's stated desire to be the manager of the group. T_{3i} is a dummy equal to 1 if i is in the *Choice & Chat* treatment and T_{4i} is a dummy equal to 1 if i is in

[24]The experiment instructions are provided in the Online Appendix.

the *Choice & Backlash* treatment. The *Choice* treatment is the excluded category. In our most comprehensive specification, we include X, a vector of demographic controls, i.e., age, field of study, being a native English speaker, having held a leadership position outside the lab and the Big 5 agreeableness measure. We also control for the performance in Stage 1 of the experiment. We estimate equation (1) using linear probability models, since interpreting interaction terms in non-linear models is not straightforward, as the marginal effect of an interaction term may not be the same as the estimated coefficient, and further the standard t-test is inaccurate (Norton et al., 2004). Even though we employ a between-subject design and groups are independent from each other, the gender composition of the group may affect self-selection into leadership, as recently shown by Born et al. (2020); therefore, we control for the number of women in the group and we cluster the standard errors at the group level.

In order to test whether there are significant gender differences in self-selection into leadership by treatment, we also estimate equation 2 below, where we pool the men and women samples and include interactions between a gender dummy and each treatment:

$$Y_i = \beta_0 + \beta_1 T_{3i} + \beta_2 T_{4i} + \beta_3 F_i * T_{3i} + \beta_4 F_i * T_{4i} + \gamma F_i + +\delta \mathbf{X}_i + u_i \tag{2}$$

where F_i is a dummy variable equal to 1 if subject i is a woman. In this specification, β_1 and β_2 indicate the impacts of the *Choice & Talk* and *Choice & Backlash* treatments on men, while β_3 and β_4 capture the differential impacts of the two treatments on women. We also report p-values generated by Wald tests for linear combinations of the estimated coefficients of interest, e.g. for $\beta_1 + \beta_3 = 0$, to assess whether the treatments had a significant impact on women.

We adopt the same estimation strategy when examining gender difference in managers' decision-making. However, this analysis also includes the *No Choice* treatment and explores the fact that managers made active decisions affecting the workers in their group multiple times. In these regressions, we still keep the *Choice* treatment as our benchmark category and we either split the sample by gender, as shown in equation 3, or we use the full sample and include interactions between gender and treatment indicators, as shown in equation 4:

$$Y_{it} = \beta_0 + \beta_1 T_{1i} + \beta_2 T_{3i} + \beta_3 T_{4i} + \delta \mathbf{X}_i + \eta DP_{it-1} + \lambda_t + u_{it} \tag{3}$$

where Y_{it} is manager i's decision in Stage t. T_{1i} is a dummy equal to 1 if the manager is in the *No Choice* treatment, T_{3i} is a dummy equal to 1 if the manager is in the *Choice & Talk* treatment, T_{4i} is a dummy equal to 1 if the manager is in the *Choice & Backlash* treatment. Here, we control for the difference in the performances of the two workers assigned to manager

i in the previous Stage, DP_{it-1}, as it is likely to affect manager decision-making, and we include stage fixed effects. As before, we cluster the standard errors at the group level, which in this case is the equivalent of clustering the standard errors at the individual level, as each group has the same manager for the duration of the experiment. We employ three measures of managerial decision-making: 1) whether the manager switches worker ranks after the first rank allocation; 2) Whether the manager makes an ex-post mistake in the rank allocation by promoting a worker who ends up performing worse than the other worker, and 3) the amount of profits that are lost due to mistaken rank allocations.

We formally test for differential responses to treatments by gender by estimating equation 4 below, where we include interactions between gender and treatment indicators:

$$\begin{aligned} Y_{it} = \beta_0 + \beta_1 T_{1i} + \beta_2 T_{3i} + \beta_3 T_{4i} + \beta_4 F_i * T_{1i} + \beta_5 F_i * T_{3i} + \beta_6 F_i * T_{4i} \\ + \gamma F_i + \eta DP_{it-1} + \delta \mathbf{X}_i + \lambda_t + u_{it} \end{aligned} \tag{4}$$

Given that we have four outcome variables measuring individuals' decision-making as a result of our treatments, hence multiple hypotheses, we correct the p-values associated to individual hypotheses by employing the step-down multiple testing method developed by Romano and Wolf (2005). We present the results for the split samples in the main text (equations 1 and 3), including the multiple hypothesis correction. We report estimates from the full sample with interaction terms (equations 2 and 4) in the appendix.

When analyzing the number of angry emojis that rank B workers send to male versus female managers in the *Choice & Backlash* treatment, we estimate equation 5 below,

$$A_{jit} = \beta_0 + \beta_1 FM_i + \beta_2 P_{jt-1} + \beta_3 RankB_{jt-1} + \delta \mathbf{X}_j + \lambda_t + u_{jit} \tag{5}$$

where A_{jit} is the number of angry emojis, including 0, that the rank B worker j sent to manager i at the beginning of Stage t, FM_i is a dummy equal to 1 if manager i is a woman, P_{jt-1} is the performance of worker j in the task in Stage $t-1$, and $RankB_{jt-1}$ is a dummy equal to 1 if worker j was rank B also in Stage $t-1$. We include our set of controls, stage fixed effects and we cluster the standard errors at the group level. To account for the fact that rank B workers can send angry emojis *after* reading the free-form message sent by their manager, we categorize the manager messages based on their content and in our most comprehensive specification we also control for the type of message sent by the manager.

Finally, in order to test whether male and female managers are more or less likely to promote a rank B worker after receiving angry emojis, we restrict the analysis to the *Choice & Backlash* treatment and estimate equation 6:

$$Y_{it} = \beta_0 + \beta_1 A_{it-1} + \beta_2 F_i * A_{it-1} +_i + \beta_3 DP_{it-1} + +\gamma F_i + \delta \mathbf{X}_i + \lambda_t + u_{it} \quad (6)$$

where Y_{it} is a dummy equal to 1 if manager i, at the beginning of Stage t promotes the worker who was rank B in Stage $t-1$, F_i is a dummy equal to 1 if manager i is a woman, A_{it-1} is the number of angry emojis that i received from the rank B worker when allocating ranks at at $t-1$, DP_{t-1} is the difference between the performance of the rank A and the rank B workers at $t-1$. Importantly, to test whether male and female managers respond differently to the receipt of angry emojis, we include the interaction between the gender of the manager and the number of angry emojis received from the rank B worker at $t-1$. As before, we include stage fixed effects and cluster the standard errors at the group level.

3 Results

We start by describing our subject pool and conducting balance tests across treatments. A total of 245 men and 172 women participated in the experiment. In Table A1, in the Appendix, we report descriptive statistics for our male and female sample pools, i.e. their average age, whether they were majoring in STEM, Business or Economics or in a different field, whether they were native English speakers, whether they reported having held a leadership position, and their average Big 5 Agreeableness Index. The average age is 22.9, with no significant gender differences across treatments and genders. Most of our participants are STEM, Business or Economics majors, although the percentage of men majoring in these fields (83%) is significantly higher than the percentage of women (64%, $p = 0.000$). The percentage of men studying STEM or Economics is higher in the *Choice* treatment than in the *Choice & Talk* ($p = 0.044$) treatment. There are no significant differences across treatments for female participants. Most subjects reported having held a leadership role in the past, with no significant differences across treatments and genders. About 41% of men and 55% of women are native English speakers, ($p = 0.008$). We do not see any significant differences in the percentages of native speakers across treatments for women, whereas we see fewer native speakers among men in *Choice* than in the *Choice & Talk* ($p = 0.011$) and in the *Choice & Backlash* ($p = 0.048$) treatments. In line with the existing literature, our female participants score significantly higher than men in the Big 5 Agreeableness Index ($p = 0.004$), with no significant differences across treatments for both the male and the female samples.

Before presenting and discussing our main findings, we assess possible gender differences in the performance in the real effort task employed in the study. Recall that we aimed to design a real effort task that would be as much as possible gender-neutral. On average, subjects solved 14 out of 20 puzzles correctly in Stage 1. The only difference we see across genders

and treatments is that women are better at the task in the *Choice & Talk* treatment than in the other treatments ($p = 0.04$). Table A2 in Appendix shows estimates from regression analysis where the dependent variable is the number of correctly solved puzzles per stage. The table shows that there are no statistically significant gender differences within and across treatments, which suggests that we were successful in employing a gender-neutral task.

We also wanted our real effort task to lead to stable performance rankings within a group. In other words, we wanted to employ a task whereby being the best(worst) performer in the task in Stage 1 would be a good predictor of the likelihood of being at the top(bottom) of the group in the subsequent stages. This is what we see in the data, i.e., if a subject is the top(bottom) performer in Stage 1, he or she is the top(bottom) performer in about 80% of the following stages. When restricting the analysis to the two workers, the best performer in Stage 1 is the best performer in 77% of the following stages, i.e., about 4 of the following 5 stages.

In what follows, we present and discuss the core results of the paper, i.e., the effects of our treatments on men's and women's self-selection into the manager position (Section 3.1), and on their decisions once in the manager role (Section 3.2). We then present evidence of the backlash received by men and women managers, and of their reactions to such backlash (Section 3.3). We conclude by conducting robustness checks on our main findings (Section 3.4).

3.1 Self-selection into the leadership role

Almost all subjects wanted to be a manager in both our *Choice* and our *Choice & Talk* treatments, with no significant differences between men and women, as shown in Table 2 and Figure 2. However, a large and statistically significant gender gap emerges in our *Choice & Backlash* treatment, where we see 78% of women self-select into the manager role as opposed to 95% of men ($p = 0.007$). The decline in managerial ambition among women is significant compared to both the *Choice* ($p = 0.055$) and the *Choice & Talk* treatments ($p = 0.011$). On the other hand, we do not see a decline in self-selection into the manager role in the *Choice & Talk* as compared to the *Choice* treatment. This suggests that, absent the possibility of worker backlash, women are not less willing than men to assume the leadership role in our setting.

This is confirmed by the regression analysis displayed in columns 1 (for men) and 5 (for women) of Table 3, which present estimates from equation 1 of Section 2.4, and by the analysis on the full sample, including interactions between gender and treatment indicators, displayed in Table A3. The estimates confirm that women, but not men, are less likely to self-select into the manager role only when facing the possibility of backlash. Importantly, only for women,

the likelihood of self-selection is also significantly lower in the *Choice & Backlash* treatment when compared to the *Choice & Talk* treatment, suggesting that it is the possibility of backlash and not the necessity to give feedback to employees that deters women from wanting to be managers. Table A3 in Appendix also provides evidence of a 16 percentage point gender leadership gap in the *Choice & Backlash* treatment only.

Among the controls, we find that performance in the task is a strong predictor of (only) women's managerial ambition.[25] We therefore examine whether gender differences in the decision to volunteer are more extreme for bottom performers than for the middle and top performers. Figure A1 in Appendix reports the percentages of volunteers among men and women in all treatments by performance terciles.[26] In the *Choice* and *Choice & Talk* treatments, we do not see evidence of significant gender differences in self-selection into the manager role for any tercile. In contrast, under *Choice & Backlash*, significant gender differences in volunteering exist among both bottom ($p = 0.065$) and middle performers ($p = 0.009$). Importantly, the gender gap closes among the very top performers ($p = 0.883$), i.e. the men and women who solved 20 out of 20 puzzles in Stage 1. The figure also shows that while there are no significant differences in self-selction among performance terciles in the male sample for either treatment, in the *Choice & Backlash* treatment, female top performers are significantly more likely to volunteer than female bottom performers, treatment ($p = 0.058$). This suggests that, contrary to men, women feel that, in order to be managers of their group in a setting when employees can express negative feedback, they need to be extremely good at the task.

3.2 Managers' decision-making

We start by examining the likelihood that the first rank assignment, at the beginning of Stage 2, is based on workers' relative productivity, i.e., the number of correctly solved puzzles in the previous stage. Overall, 94% of the managers assigned rank A to the best performing worker in Stage 1,[27] with no statistical significant differences by manager gender and/or by treatment.

Next, we examine managers' tendency to switch worker ranks across stages. As previously noted, we designed the task with the aim of having workers' relative performances be stable

[25] For women, the only robust determinant of volunteering decision is the field of study, with women majoring in STEM, Business and Economics being more likely to want to be managers. This is essentially indicating that women who have already self-selected into fields that are typically male-dominated and conducive to competitive high-paying jobs, are also more likely to self-select into leadership roles in the experiment. In the male sample, the only significant variable is the Big 5 Agreeableness score, which, in line with the existing studies, appears with a negative sign, suggesting that less agreeable, hence more competitive and aggressive men, are more likely to self-select into the manager role.

[26] The bottom tercile is made of subjects who completed 12 or less puzzles correctly in Stage 1 of the experiment. The middle tercile consists of students who completed more than 12 puzzles but less than 20 puzzled in Stage 1. The top tercile is made of students who completed 20 out of 20 puzzles in Stage 1.

[27] If the two workers solved the same number of puzzles, we code our outcome variable as a 1, no matter the worker who was chosen to be rank A.

over time. On average, the best worker in stage 1 outperforms the other worker in nearly 4 of the 5 subsequent stages. Therefore, once ranks are established based on past performance at the beginning of Stage 2, switching ranks and promoting the rank B worker may generate less money than if the ranks had been kept unchanged. However, a desire to equalize workers' payoffs, or to reward the rank B worker if he or she performs equally or better than the rank A worker in a given stage, or to reduce the likelihood of backlash, may induce managers to switch ranks across stages.

Overall, averaging all treatments, male and female managers switch ranks 1.03 and 1.44 times, respectively (*p-value*=0.053). Workers' relative past performance is a strong predictor of a manager's decision to switch ranks. Rank B workers who outperform rank A workers are promoted 56% and 65% of the times by male and female managers respectively (*p-value*=0.215). Panel (a) of Figure 3 displays the average number of rank switches by treatment and suggests that the possibility of backlash increases the number of rank switches by women managers. We estimate equation 3 of Section 2.4 and report the results in column 2 (for men) and column 6 (for women) of Table 3. Among the controls, we include the difference in the performances of the two workers, which is a strong predictor of manager ranking decisions. Estimates for the full sample (equation 4 of Section 2.4) are reported in Table A4 in the appendix. The estimates show that while the treatments do not affect men's tendency to switch workers' ranks, the possibility of backlash induces women managers to switch ranks more often. However, since rank switching is highly responsive to workers' relative performances, the higher likelihood of switching observed among women managers in the *Choice & Backlash* treatment does not lead to a higher likelihood of mistakenly assigning rank A to a worker who ends up under-performing the other worker. This applies to both male and female managers, as shown in columns 3 and 7 of Table 3. The only significant difference emerging from column 7 of Table 3 is among self-selected and not self-selected female managers. The latter are more likely to make mistakes in rank allocations.

As a final assessment of managerial performance, we look at lost profits caused by mistakes in rank allocations. Recall that the manager receives earnings from the number of puzzles that are successfully completed by the rank A worker only. If a manager assigns rank A to a worker who ends up performing worse than the other worker, the manager suffers a loss in profit equal to the difference between the profit that the he or she would have made by allocating rank A to the best performing worker (i.e., the counterfactual) and the profits that he or she actually made. Lost profits are equal to zero if the manager was not mistaken in the rank allocation, and equal to the difference in the number of puzzles solved by the two workers, multiplied by the piece rate (2 ECU), if he or she was.

Lost profits are low on average, i.e., below 2 ECU, as shown in panel (b) of Figure 3. This is both because mistakes only happen 22 percent of the times on average, and because, when

they do happen, the difference in the performances of the two workers tends to be small, i.e., less than 4 puzzles on average. Results from regression analysis, displayed in columns 4 and 8 of Table 3, and Table A4 in the appendix, show no evidence of treatment or gender effects on lost profits: neither having to talk to employees when assigning ranks, or facing the possibility of receiving angry messages from low-ranked workers leads managers to make decisions that ultimately result in financial losses. This is true for both male and female managers. In Table 4, we compare the total profits that male and female managers made over the 5 stages of the experiment from the productivity of the workers assigned rank A. We find no evidence of significant differences by treatment and by manager gender. Table 4 also displays the difference in the total earnings of the two workers. While there are no significant differences in inequality between workers' earnings by manager gender in any of the treatments, the inequality tends to be reduced when the manager is a woman and there is the possibility of backlash.

Finally, we assess gender and treatment differences in manager leadership styles, as measured by the free-form messages that managers sent to the workers at the rank assignment stage in the *Choice & Talk* and the *Choice & Backlash* treatments. We conducted text analysis of the messages sent by managers over the 5 stages of the experiment and identified seven primary types of messages: 1) praising the worker for past performance, e.g., "amazing work!"; 2) using motivating words, e.g., "You got this!"; 3) providing an explanation for the rank allocation, e.g., "the other worker did better"; 4) inducing competition among workers, e.g., "the other worker is catching up so keep it up"; 5) mentioning fairness, e.g., "trying to be fair"; 6) using team building words, e.g., "let's go team!"; and 7) using cordial words like "thank you" and "sorry".[28] Table A6 shows the percentages of messages within each category sent by male and female managers to rank B workers over the five stages of the experiment, in the *Choice & Talk* and the *Choice & Backlash* treatments.[29] In the *Choice & Talk* treatment, where workers could not send messages back, we do not see any gender difference in message types. Gender differences however emerge when worker backlash is possible, as male managers praise less, explain more and become more likely to induce competition between workers. Female managers also praise employees less and explain themselves more, but they also become more likely to use cordial words, such thank you and sorry.[30] While these gender differences are in line with previous studies of differences in the language used by men and women, both

[28]Table A5 in the appendix provides additional examples of the messages of each kind sent by managers to workers.

[29]Statistics for the messages sent to rank A workers are shown in Table A6 in the appendix. It shows evidence of no statistical significant differences in the types of messages sent by male and female managers in the *Choice & Talk*. Similarly to what we observe for the messages to the rank B workers, when worker backlash is possible, we see that men become more likely to explain their decision and induce competition among workers, whereas women become more likely than men to send praising messages.

[30]We do not find any gender difference, or treatment effects, in the propensity to use team building words, or to mention fairness.

in psychology (e.g., Kern et al., 2016) and economics (Timko, 2017), our treatments reveal that such gender differences are generated by differential responses to the possibility of worker backlash, rather than by innate gender differences in communication styles.

3.3 Backlash and manager reaction to backlash

Over the 5 stages of the experiment, male and female managers received backlash from the rank B worker - in the form of at least one angry emoji - around 45 percent of the times.[31] Figure 4 shows that female managers tended to receive more angry emojis. In Table 6, we estimate equation 5 of Section 2.4. In particular, we conduct regression analysis of the decision of Rank B workers to send one or more angry emojis to their manager over the 5 stages of the experiment. We progressively add controls, and, in column 4, we also include dummy variables capturing the type of free-form message (see Table 5) sent by the manager,[32] since the rank B worker sees such message immediately before he or she is given a chance to send angry emojis. The estimates show that female managers receive more angry emojis from rank B workers. Controlling for the type of message sent by the manager strengthens this result. Rank B workers tend to send 1 more angry emoji to a female manager than to a male manager.

We also conduct text analysis of the free-form messages that rank B workers sent to managers in the *Choice & Talk* and *Choice & Backlash* treatments, as shown in Tables A8 in the appendix. The most common messages can be categorized as follows: 1) thanking the manager or joking; 2) expressing approval of the ranking decision; 3) expressing commitment to work hard; 4) apologetic; 5) containing negative feelings, i.e., upset; 6) questioning the ranking decision.[33] We do not see significant differences in the messages sent by rank B workers, with one exception. Female managers are more likely to receive messages in which rank B workers question the ranking decision. This happens 17 percent of the times for female managers, versus 6 percent for male managers (p=0.032).[34]

Overall, the analysis of worker backlash suggests that rank B workers are both more likely to verbally question the ranking decision when it comes from a female manager, and more likely to express disapproval of the choice made by a female manager by sending more angry emojis. This is despite the fact that female managers switch ranks more often and that the

[31] Figure A2, in the appendix, shows the percentages of male and female managers who received at least one angry emoticon in each stage of the experiment. While female managers started off by being slightly more likely to receive at least one angry message from their rank B worker, the difference is not statistically significant.

[32] We include dummy variables for each category of message described in Section 3.2 and displayed in Table 5. Note that the dummies are not mutually exclusive, since the same message can for example be both praising and motivating.

[33] Table A7 in the appendix provides examples of the messages belonging to each category.

[34] There are no significant differences in the messages sent by rank A workers to male versus female managers, which belong to three main categories: 1) thanks or jokes (56 %), 2) approval of rank assignment (7%), 3) commitment to work hard (23%).

propensity to send angry emojis declines when workers' ranks are switched, as suggested by the positive and significant sign of the dummy indicating that the worker was not switched, i.e., he or she was rank B also in the previous round.

Finally, we examine how male and female managers react to the receipt of angry emojis in the *Choice & Backlash* treatment. We report our estimates of equation 6, presented in Section 2.4, in Table 7. The dependent variable is the decision to switch ranks and promote the rank B worker at the beginning of stage t. Since the analysis is restricted to the *Choice & Backlash* treatment, we now include the number of angry emojis that the rank B worker sent to the manager when he or she was notified of the assigned rank in the previous stage. Importantly, we interact the number of angry emojis received with the gender of the manager. While the receipt of angry emojis does not affect male managers' ranking decisions, it actually lowers the likelihood that a female manager promotes the rank B worker by 7 percentage points, everything else being equal ($p = 0.002$, Wald test of the linear combination of angry emojis and its interaction with the female dummy). This indicates that the observed higher propensity of female managers to promote the rank B worker in the *Choice & Backlash* treatment as compared to the other *Choice* treatments is not a response to *received* angry emojis and is not due to a desire to appease employees who have expressed disapproval of the ranking decision.

3.4 Robustness: The gender composition of the group

When recruiting subjects to participate in the experiment, we aimed to have groups of two men and one woman to simulate male-dominated environments. About 63 percent of our groups ended up being of the desired gender composition, with the remaining groups consisting primarily of two women and one man. When examining self-selection into leadership in Section 3.1, we controlled for the number of women in the group, and we clustered the standard errors at the group level. In the regression analysis, the number of women in the group did not seem to impact the decision to become the manager of the group.

In Figure A3 in Appendix, we reproduce Figure 2 for the restricted sample of groups composed of two men and one woman only. The pattern we saw for the full sample remains, with women less likely than men to self-select into leadership in the *Choice & Backlash* treatment only (p=0.009). Among women, we see a lower percentage of women wanting to be manager in the *Choice & Backlash* treatment than in the *Choice & Talk* treatment, although the difference is not statistically significant in the restricted smaller sample (p=0.123). We do not see treatment-driven significant differences among male participants.

The gender composition of the group, combined with the gender of the manager, has immediate consequences for the genders of the workers a manager is matched with. In our setting, averaging across all treatments, 77 percent of male managers are matched with one

male and one female worker, versus 47 percent of female managers (p=0.000). We conduct robustness checks of managers' decision-making, restricting the analysis to groups where both male and female managers are matched with one male and one female worker. The results obtained for rank-switching, ex-post mistakes and lost profits are displayed in Table A9 in the appendix. Even though the sample is smaller by about one third, the signs of the estimates are consistent with those obtained in Table A4, with male and female managers more willing to switch ranks when having to talk with employees, and overall no significant differences in the likelihood of ex-post mistakes and lost profits.[35]

4 Mechanisms: The online experiment

The primary findings generated by the laboratory experiment could be driven by women's expectations of more backlash or by women's higher sensitivity to any backlash, or both. Our lab experimental data do not allow us to test for the mechanisms driving the results, as we did not collect data on beliefs about backlash. This is because we did not want belief elicitation to prime participants to consider the severity of backlash at the self-selection into leadership stage, in the *Choice & Backlash*, more than they would have naturally done. We also worried that eliciting incentivized beliefs on backlash in only one of the treatments - the one where backlash was possible - would introduce methodological differences between the treatments that would compromise the internal validity of the experiment.

In order to generate insights on the mechanisms that could drive our findings - aversion to backlash and expectation of backlash - we conducted an online experiment on Prolific in Fall 2021. We involved a sample of over 500 participants of similar age and education as our laboratory subjects. We restricted recruitment to individuals aged between 18 and 30, either currently enrolled in college or with a college degree. The online survey had three main parts: 1) A brief section on demographics (to check that the recruitment requirements were met); 2) A set of questions used to construct the Brief Fear of Negative Evaluation (BFNE) psychology index (Rodebaugh et al., 2004); 3) The description of the laboratory experiment and the incentivized elicitation of beliefs regarding the likelihood and severity of backlash that men or women managers received.

The survey lasted 15 minute and subjects earned an average of 4.5USD. A total of 572 subjects participated, of which 46% women. Participants were 23.5 years old on average, about half of them were enrolled in college at the time of the survey, and the vast majority (64%) of them held a leadership position in the past.[36]

[35]The higher likelihood of female managers to switch ranks in the *Choice & Backlash* treatment disappears in the restricted sample, but we may be underpowered to detect the treatment effect.

[36]Compared to the participants in our laboratory experiment, the Prolific subjects are about a year older and have experienced leadership less frequently. See Table A1 for the descriptive statistics of our lab sample.

4.1 The BFNE Index: A measure of aversion to backlash

In the online survey we include 8 questions that allow us to construct the "straightforward" version of the Brief Fear of Negative Evaluation (BFNE) index (Rodebaugh et al., 2004; Carleton et al., 2011). This psychological index is a short version of the 30-item index originally developed by Watson and Friend (1969) and often used in personality and social psychology to measure anxiety associated with the possibility of being evaluated or judged negatively by others.[37]

Figure 5 shows the distribution of the BFNE scores for men and women, with a higher value indicating higher apprehension at the prospect of negative evaluation. Women score significantly higher than men on average (29.36 versus 27.24, with $pvalue$=0.0012). The two distributions are also significantly different from each other ($pvalue$=0.005 from Kolmogorov-Smirnov test for equality of distribution functions). This indicates that, at least among individuals in this age group (18 to 30) and education level (college graduates or college enrolled), women are significantly more sensitive to the possibility of being evaluated negatively by others. This suggests that a greater aversion to backlash is a plausible driver of our laboratory findings.

4.2 The experiment: Predictions of backlash

A crucial objective of the online study was to elicit subjects' beliefs about the backlash that men and women managers got in the laboratory experiment. To this end, after explaining the rules of the leadership game under the *Choice & Backlash* treatment, we randomly assigned participants either to a Male Manager treatment or a Female Manager treatment. Subjects in the Male Manager treatment had to first guess the percentage of male managers that received at least one angry emoji from the rank B worker after the rank allocation (using 10-point percentage ranges, e.g. 0-10%, 11-20% etc.), and then the average number of angry emojis received by male managers (conditional on receiving at least one). Participants assigned to the Female Manager treatment answered the same questions but for female managers.[38] Subjects earned 0.50 USD for each correct guess.

As a result of our between-subject design, we have both men and women being randomized to guess the backlash received by either male or female managers. Figure 6 displays our primary findings. First, women believe that: 1) female managers are more likely to receive any backlash

[37] Rodebaugh et al. (2004); Weeks et al. (2005) have shown that the 8-straightforwardly-worded-item index is more reliable than the original score, which also employed reverse-scored items. See the Online Appendix for details information on the 8 questions forming the BFNE index employed here.

[38] To simplify the setting and make belief elicitation easier, we presented the game as a one-shot game, and we incentivized beliefs based on the occurrence of backlash in stage 2 of the lab experiment, i.e., the first stage in which backlash could occur.

compared to male managers (i.e., the extensive margin of backlash), and 2) female managers receive more angry messages (i.e., the intensive margin of backlash) than men. Second, men do not hold such beliefs. Third, both men and women overestimate the occurrence of backlash by about 20 to 30 percentage points on average, and the severity of backlash by about 1 angry emoji.[39] Finally, while women expect women to get more backlash than men, there is no statistical significant difference in the likelihood and severity of backlash than men and women expect their own gender to receive. In fact, men expect between 70 and 80 percent of male managers, and women expect between 70 and 80 percent of female managers, to get backlash. The same applies to the severity of backlash: men and women expect male and female managers, respectively, to receive between 3 and 4 angry emojis on average.

Overall, the online experiment suggests that, while women expect women to receive more backlash then men, the impact of the backlash treatment (only) on women in the laboratory experiment - both at the self-selection into leadership and rank assignment stages - is unlikely to be driven by gender differences in expectations of backlash. Sensitivity to backlash seems particularly important.

5 Conclusion

The literature on gender differences in leadership is still in its nascent stage. In this paper, we asked whether women are less likely to self-select into leadership positions that require discretionary decision-making likely resulting in backlash. By employing a controlled experiment, we were able to assess also the causal impact of the threat of backlash on leader decision-making and final outcomes.

We found strong evidence of gender differences in individuals' reaction to the possibility of backlash. While men are unaffected, women are less likely to want to assume a manager role under the threat of backlash. No gender differences are observed when backlash is not allowed, with over 90 percent of both men and women wanting to assume the leadership role. We also find that the possibility of worker backlash leads women managers to switch worker ranks, by promoting the low-rank worker, more often, but this does not lead to losses in profits or to gender differences in final outcomes. Further analysis shows that the higher likelihood of rank-switching among women under the threat of backlash is not a reaction to received backlash, i.e., a desire to appease a worker who expressed discontent. Rather, rank-switching is likely used by women as a strategy to minimize future backlash.

We also found gender differences in the messages sent by managers to workers. Female managers are more likely to praise employees and use cordial words, while male managers are

[39] See Figure A2 in the appendix for the actual extensive margin of backlash by stage, and panel b of Figure 4 for the actual intensive margin of backlash by stage.

more direct and more likely to foster competition among employees. However, these differences emerge only under the possibility of worker backlash. Finally, our laboratory findings confirm previous studies on differential attitudes toward male and female leaders. In our setting, female managers receive more backlash, and are more often questioned by workers regarding their ranking decisions.

In order to assess whether gender differences in expectations of backlash and/or in aversion to any backlash are the main drivers of our laboratory findings, we conducted an online experiment on Prolific. We recruited a sample of survey respondents of comparable age and education level as our laboratory subjects. We elicited survey questions aimed at constructing the Brief Fear of Negative Evaluation psychological index. We then presented subjects with the rules of the leadership game in the *Choice & Backlash* treatment, and randomly assigned them to predict the backlash received by either male or female managers in the laboratory experiment. We find evidence of significant gender differences in the Brief Fear of Negative Evaluation index, with women being more averse to the possibility of negative judgment. Results from the experimental elicitation of beliefs about backlash show that, while women expect women to receive more backlash than men, there is no gender differences in the backlash that men and women expect their gender to receive. This suggests that expectations of backlash are unlikely to be the driving mechanism behind our primary lab findings.

In evaluating the relevance and implications of our results, it is important to note that the analysis is based on a laboratory setting where decisions are anonymous and worker disapproval takes the form of angry emojis sent by only one other subject via a computer terminal, rather than via personal face to face interactions. We may therefore be underestimating the role that the threat of backlash plays in contributing to the gender leadership gap in field settings, where leaders have to face more direct and personal expressions of anger from multiple unhappy subordinates.

References

Abel, M. (2022). Do workers discriminate against female bosses? *Journal of Human Resources*, 1120–11318R3.

Adams, R. B. and D. Ferreira (2009). Women in the boardroom and their impact on governance and performance. *Journal of financial economics 94*(2), 291–309.

Alan, S., S. Ertac, E. Kubilay, and G. Loranth (2020). Understanding gender differences in leadership. *The Economic Journal 130*(626), 263–289.

Ayalew, S., S. Manian, and K. Sheth (2018). Discrimination from below: Experimental evidence on female leadership in ethiopia. Technical report, mimeo.

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

Babcock, L. and S. Laschever (2009). *Women don't ask: Negotiation and the gender divide*. Princeton University Press.

Babcock, L., M. P. Recalde, L. Vesterlund, and L. Weingart (2017). Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review 107*(3), 714–47.

Bertrand, M. and E. Duflo (2017). Field experiments on discriminationa. In *Handbook of Economic Field Experiments*, Volume 1, pp. 309–393. Elsevier.

Bertrand, M. and S. Mullainathan (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *The American Economic Review 94*(4), 991–1013.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics 145*, 27–41.

Born, A., E. Ranehill, and A. Sandberg (2020). Gender and willingness to lead does the gender composition of teams matter?

Born, A., E. Ranehill, and A. Sandberg (2022). Gender and willingness to lead: Does the gender composition of teams matter? *Review of Economics and Statistics 104*(2), 259–275.

Bowles, H. R., L. Babcock, and L. Lai (2007). Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask. *Organizational Behavior and Human Decision Processes 103*(1), 84–103.

Brandts, J. and D. J. Cooper (2007). It's what you say, not what you pay: an experimental study of manager-employee relationships in overcoming coordination failure. *Journal of*

the European Economic Association 5(6), 1223–1268.

Branton, R., A. English, S. Pettey, and T. D. Barnes (2018). The impact of gender and quality opposition on the relative assessment of candidate competency. *Electoral Studies 54*, 35–43.

Carleton, R. N., K. C. Collimore, R. E. McCabe, and M. M. Antony (2011). Addressing revisions to the brief fear of negative evaluation scale: Measuring fear of negative evaluation across anxiety and mood disorders. *Journal of anxiety disorders 25*(6), 822–828.

Chaudhuri, A., Y. Li, and E. Sbai (2018). (un)willingness to lead? men, women and the leadership gap.

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

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

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

Egan, M. L., G. Matvos, and A. Seru (2017). When harry fired sally: The double standard in punishing misconduct. Technical report, National Bureau of Economic Research.

Ellison, G. and A. Swanson (2018). Dynamics of the gender gap in high math achievement. Technical report, National Bureau of Economic Research.

Elsesser, K. M. and J. Lever (2011). Does gender bias against female leaders persist? quantitative and qualitative data from a large-scale survey. *Human Relations 64*(12), 1555–1578.

Erkal, N., L. Gangadharan, and E. Xiao (2022). Leadership selection: Can changing the default break the glass ceiling? *The Leadership Quarterly 33*(2), 101563.

Ertac, S. and M. Y. Gurdal (2012). Deciding to decide: Gender, leadership and risk-taking in groups. *Journal of Economic Behavior & Organization 83*(1), 24–30.

Exley, C. L., M. Niederle, and L. Vesterlund (2020). Knowing when to ask: The cost of leaning in. *Journal of Political Economy 128*(3), 816–854.

Flory, J. A., A. Leibbrandt, and J. A. List (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.

Gneezy, U., M. Niederle, and A. Rustichini (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics 118*(3), 1049–1074.

Grossman, P. J., C. Eckel, M. Komai, and W. Zhan (2019). It pays to be a man: Rewards for leaders in a coordination game. *Journal of Economic Behavior & Organization 161*, 197–215.

Grossman, P. J., M. Komai, and J. E. Jensen (2015). Leadership and gender in groups: An experiment. *Canadian Journal of Economics/Revue canadienne d'économique 48*(1), 368–388.

Güth, W., M. V. Levati, M. Sutter, and E. Van Der Heijden (2007). Leading by example with and without exclusion power in voluntary contribution experiments. *Journal of Public Economics 91*(5), 1023–1042.

Hengel, E. (2017). Publishing while female. are women held to higher standards? evidence from peer review.

Jack, B. K. and M. P. Recalde (2015). Leadership and the voluntary provision of public goods: Field evidence from bolivia. *Journal of Public Economics 122*, 80–93.

Johnson, M. and V. S. Helgeson (2002). Sex differences in response to evaluative feedback: A field study. *Psychology of Women Quarterly 26*(3), 242–251.

Jones, D. and S. Linardi (2014). Wallflowers: Experimental evidence of an aversion to standing out. *Management Science 60*(7), 1757–1771.

Judge, T. A. and J. E. Bono (2000). Five-factor model of personality and transformational leadership. *Journal of applied psychology 85*(5), 751.

Kern, M. L., G. Park, J. C. Eichstaedt, H. A. Schwartz, M. Sap, L. K. Smith, and L. H. Ungar (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological methods 21*(4), 507.

Kugler, A. D., C. H. Tinsley, and O. Ukhaneva (2017). Choice of majors: Are women really different from men? Technical report, National Bureau of Economic Research.

Leaper, C. and T. E. Smith (2004). A meta-analytic review of gender variations in children's language use: talkativeness, affiliative speech, and assertive speech. *Developmental psychology 40*(6), 993.

Manian, S. and K. Sheth (2020). Follow my lead: Assertive cheap talk and the gender gap.

Matsa, D. A. and A. R. Miller (2014). Workforce reductions at women-owned businesses in the united states. *ILR Review 67*(2), 422–452.

Mayo, M., M. Kakarika, J. C. Pastor, and S. Brutus (2012). Aligning or inflating your leadership self-image? a longitudinal study of responses to peer feedback in mba teams. *Academy of Management Learning & Education 11*(4), 631–652.

Mengel, F., J. Sauermann, and U. Zölitz (2017). Gender bias in teaching evaluations. *Journal of the European Economic Association*.

Niederle, M. (2016). gender effects in economic experiments. *Handbook of experimental economics. Princeton: Princeton university press. Google Scholar*.

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

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

Preece, J. and O. Stoddard (2015). Why women don't run: Experimental evidence on gender differences in political competition aversion. *Journal of Economic Behavior & Organization 117*, 296–308.

Reuben, E., P. Rey-Biel, P. Sapienza, and L. Zingales (2012). The emergence of male leadership in competitive environments. *Journal of Economic Behavior & Organization 83*(1), 111–117.

Reuben, E. and K. Timko (2017). On the effectiveness of elected male and female leaders and team coordination.

Rheault, L., E. Rayment, and A. Musulan (2019). Politicians in the line of fire: Incivility and the treatment of women on social media. *Research & Politics 6*(1), 2053168018816228.

Riach, P. A. and J. Rich (2006). An experimental investigation of sexual discrimination in hiring in the english labor market. *Advances in Economic Analysis & Policy 5*(2).

Rodebaugh, T. L., C. M. Woods, D. M. Thissen, R. G. Heimberg, D. L. Chambless, and R. M. Rapee (2004). More information from fewer questions: the factor structure and item properties of the original and brief fear of negative evaluation scale. *Psychological assessment 16*(2), 169.

Romano, J. P. and M. Wolf (2005). Stepwise multiple testing as formalized data snooping. *Econometrica 73*(4), 1237–1282.

Sarsons, H. (2017). Recognition for group work: Gender differences in academia. *American Economic Review 107*(5), 141–45.

Schmitt, D. P., A. Realo, M. Voracek, and J. Allik (2008). Why can't a man be more like a woman? sex differences in big five personality traits across 55 cultures. *Journal of personality and social psychology 94*(1), 168.

Shurchkov, O. and A. van Geen (2017). Why female decision-makers shy away from promoting competition.

Swim, J., E. Borgida, G. Maruyama, and D. G. Myers (1989). Joan mckay versus john mckay: Do gender stereotypes bias evaluations? *Psychological Bulletin 105*(3), 409.

Timko, K. (2017). Gender, communication styles, and leader effectiveness.

Warner, J. and D. Corley (2017). Womens leadership by the numbers.

Watson, D. and R. Friend (1969). Measurement of social-evaluative anxiety. *Journal of consulting and clinical psychology 33*(4), 448.

Weeks, J. W., R. G. Heimberg, D. M. Fresco, T. A. Hart, C. L. Turk, F. R. Schneier, and M. R. Liebowitz (2005). Empirical validation and psychometric evaluation of the brief fear of negative evaluation scale in patients with social anxiety disorder. *Psychological assessment 17*(2), 179.

Wiswall, M. and B. Zafar (2017). Preference for the workplace, investment in human capital, and gender. *The Quarterly Journal of Economics 133*(1), 457–507.

Figures and Tables

Figure 1: Stages of the experiment

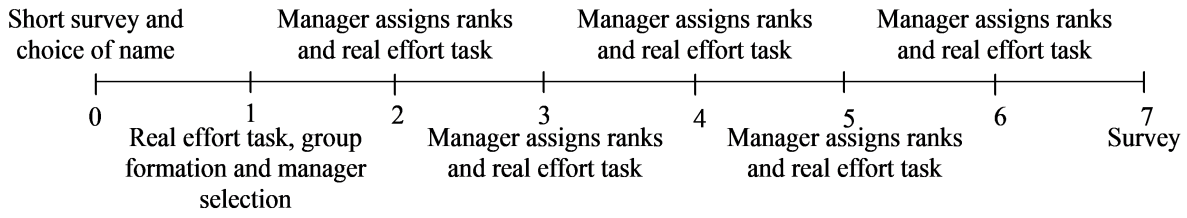
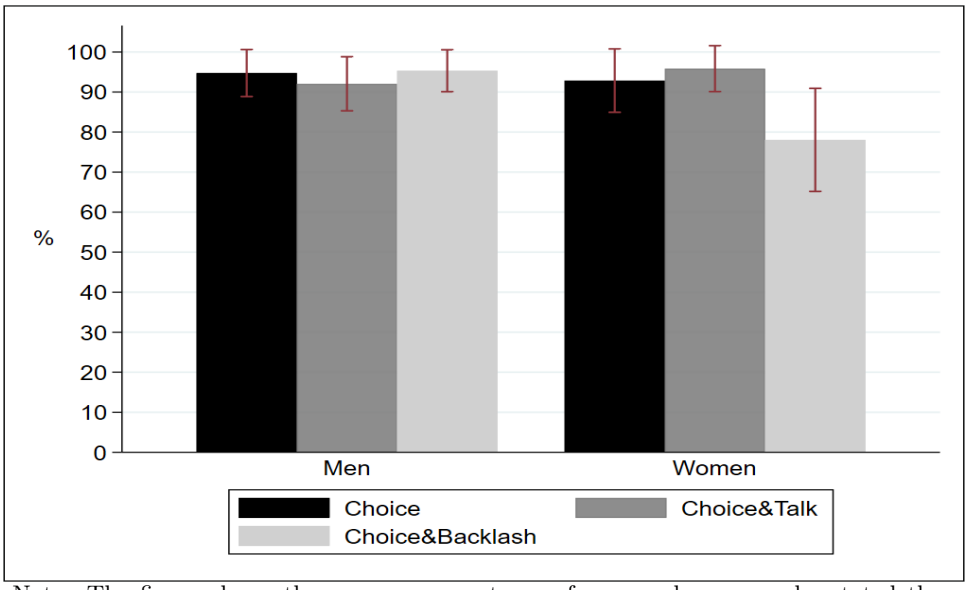
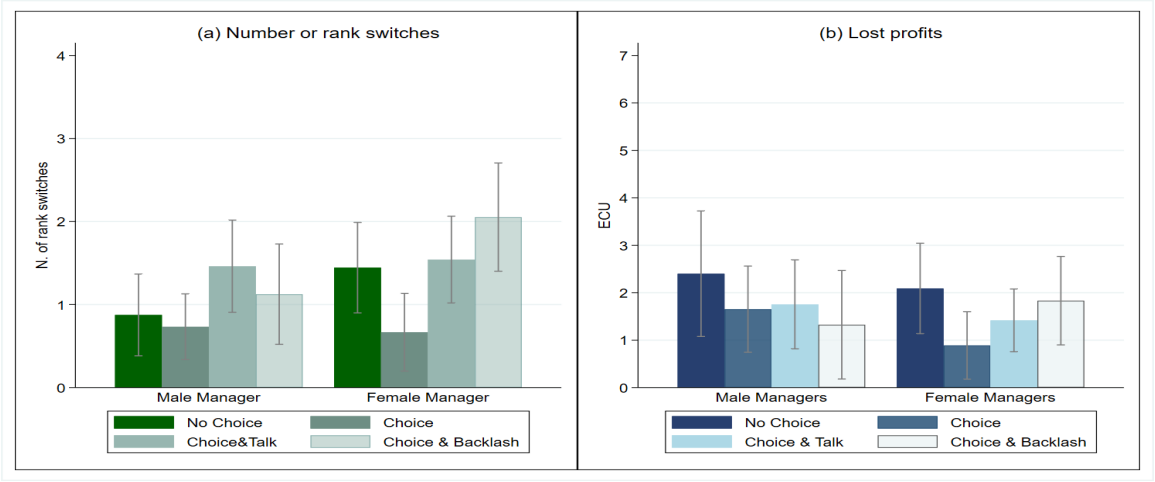


Figure 2: Self-selection into the manager position



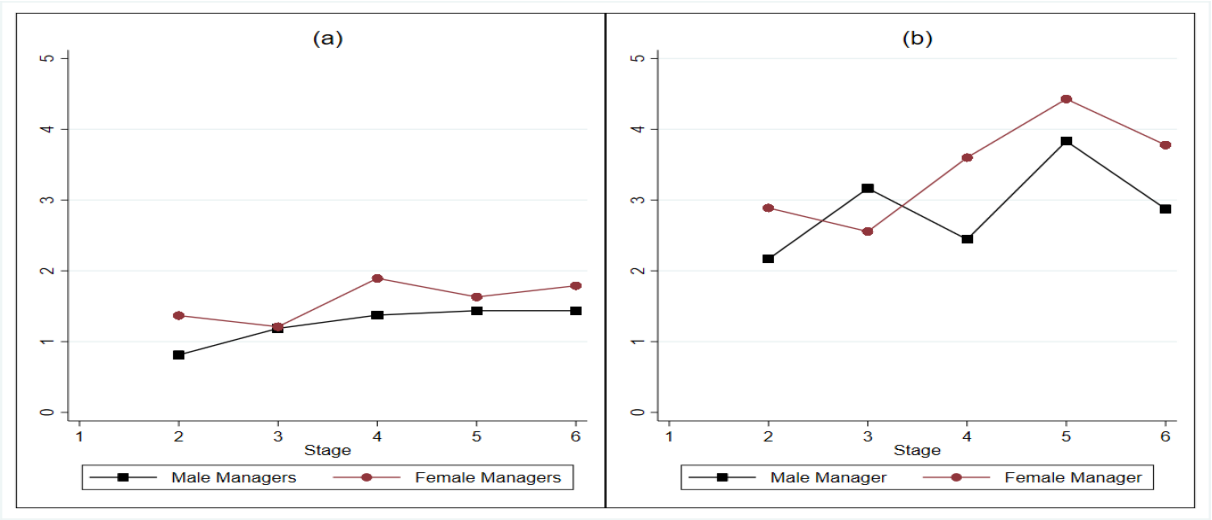
Note: The figure shows the average percentages of men and women who stated they wanted to be the manager of their group, by treatment and by gender.

Figure 3: Manager rank switching and lost profits



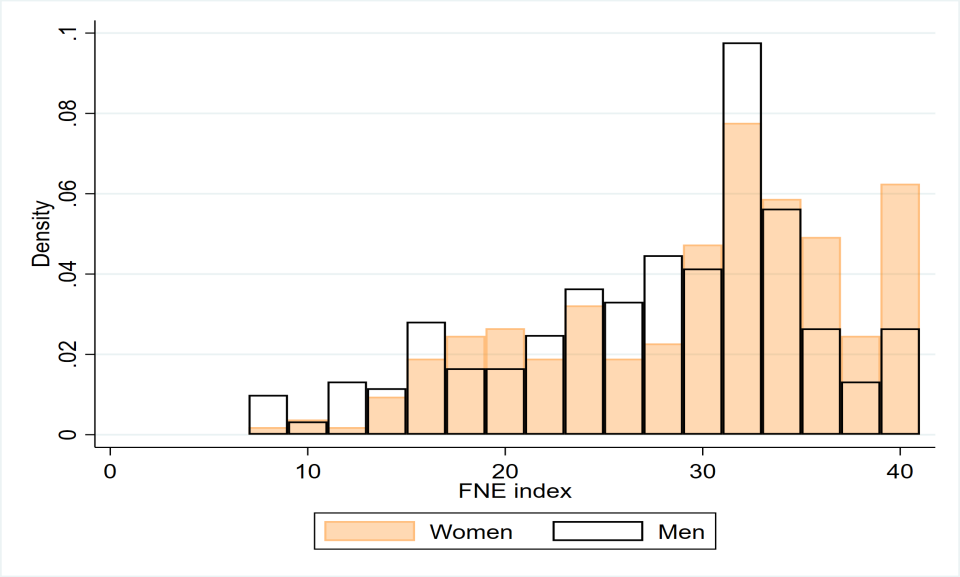
Note: Panel (a) shows the average number of rank switching by treatment for men and women managers. Managers could switch workers' ranks at most 4 times during the experiment. Panel (b) shows average lost profits by treatment and by manager gender. Lost profits refer to the difference between the highest profit that the manager could have made in a Stage (by hiring the worker who ended up performing the best) and the profits that the manager actually made. Lost profits are set to 0 if the manager assigned rank A to the worker who ended up being the best performer in a given stage. They are greater than 0 if the worker who was assigned rank A performed worse than the worker who was assigned rank B. Lost profits are expressed in Experimental Currency Units (ECU).

Figure 4: Angry emojis sent to managers



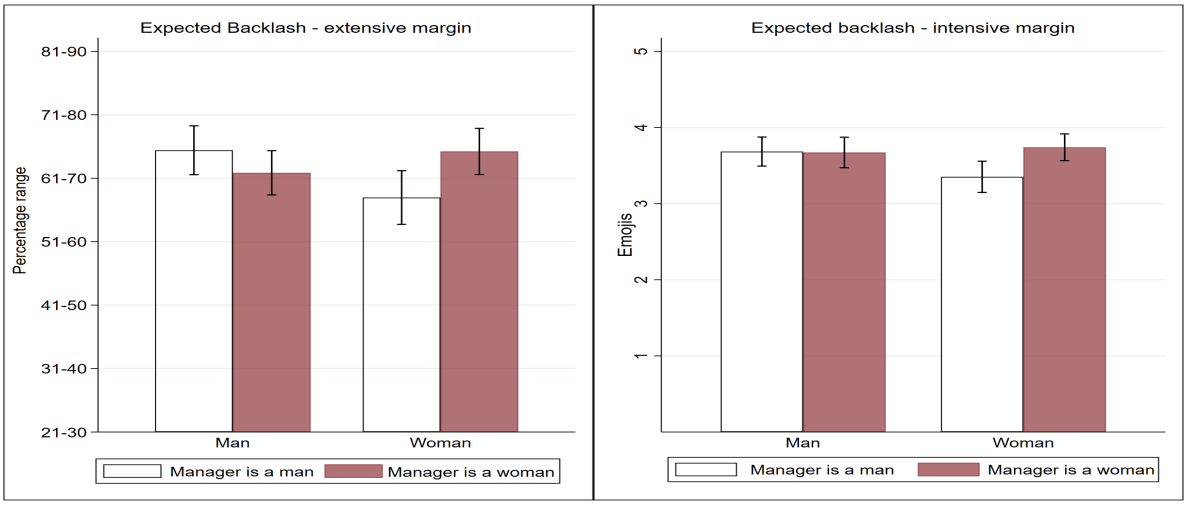
Note: Panel (a) shows the average number of angry emojis sent by rank B workers to male and female managers in the *Choice & Backlash* treatment. Panel (b) shows the emojis sent, conditional on sending at least one.

Figure 5: Online experiment: The BFNE Index



Note: The figure displays how men and women who participated in the online experiment scored in the Brief Fear of Negative Evaluation (BFNE) index. A larger value indicates higher fear of negative evaluation.

Figure 6: Online experiment: Predictions of backlash



Note: Participants in the online experiment were asked to predict the percentage of male or female managers (experimentally manipulated through a between-subject design) who received angry emojis from the rank B worker. The left panel shows the average percentage range of man/woman managers that participants guessed received at least one angry emoji. The right panel shows the average number of angry emojis that participants guessed man/woman managers received from rank B workers, conditional on receiving at least one.

Table 1: Groups and Treatments

	Groups	Men	Women	Total
No Choice (NC)	34	61	41	102
Choice(C)	33	57	42	99
Choice & Talk (CT)	37	63	48	111
Choice & Backlash (CB)	35	64	41	105
Total	139	245	172	417

Note: Subjects participated in one of four treatments following a between-subject design. Participants were placed in a group of three people and remained in that group for the duration of the experiment.

Table 2: Men and Women who want to be managers

	Men	Women	$H_0 : M = W$
Choice(C)	94.74	92.86	0.70
Choice & Talk (CT)	92.06	95.83	0.42
Choice & Backlash (CB)	95.31	78.05	0.01
$H_0 = C = CT$ (p-value)	0.56	0.54	.
$H_0 = C = CB$ (p-value)	0.88	0.06	.
$H_0 = CT = CB$ (p-value)	0.45	0.01	.

Note: We report the percentages of participants who wanted to be the managers of their group. P-values are generated by Chi-square tests.

Table 3: Self-selection into the manager role and manager decision-making by treatment and gender

	Men				Women			
	Self-select	Switch	Mistaken	Lost Profit	Self-select	Switch	Mistaken	Lost Profit
No Choice (NC)		0.085	-0.017	0.931		0.095	0.161	0.992
		(0.066)	(0.081)	(0.826)		(0.079)	(0.061)**	(0.710)
		[0.465]	[0.990]	[0.594]		[0.337]	[0.010]**	[0.287]
Choice&Talk (CT)	-0.010	0.145	-0.004	0.189	0.030	0.134	0.078	0.352
	(0.051)	(0.075)*	(0.088)	(0.699)	(0.048)	(0.076)*	(0.061)	(0.673)
	[0.990]	[0.119]	[0.990]	[0.990]	[0.634]	[0.149]	[0.317]	[0.634]
Choice&Backlash (CB)	0.011	0.113	-0.117	-0.269	-0.149	0.268	0.093	0.621
	(0.050)	(0.080)	(0.074)	(0.643)	(0.073)**	(0.076)***	(0.065)	(0.811)
	[0.990]	[0.366]	[0.257]	[0.990]	[0.089]*	[0.010]**	[0.287]	[0.614]
Observations	184	240	240	240	131	316	316	316
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of Clusters	105	60	60	60	100	79	79	79
$H_0 : NC = CT$		0.435	0.870	0.377		0.639	0.178	0.259
$H_0 : NC = CB$		0.729	0.109	0.124		0.042**	0.311	0.582
$H_0 : CT = CB$	0.630	0.709	0.116	0.500	0.013**	0.120	0.820	0.697

Note: LPM in columns 1 to 3, and 5 to 7; OLS in columns 4 and 8. The dependent variables are: i) a dummy equal to 1 if the subject stated that he or she wanted to be the manager of the group (columns 1 and 5); ii) a dummy equal to 1 if the Manager switched ranks (columns 2 and 6); iii) a dummy equal to 1 if the the worker chosen to be rank A ended up performing worse than the other worker (columns 3 and 7); iv) the difference between the profits that the manager would have made by assigning rank A to the best performing worker, and the profits actually made in the current Stage (columns 4 and 8). The analysis of manager decision-making is restricted to Stages 3 to 6, since Stage 3 is the first stage when rank switching is allowed. Controls are: age and STEM or economics field of study. The analysis of leader decision-making also controls for the difference in the performances of the two workers in the previous stage. Robust standard errors, clustered at the group level, in parentheses. Romano-Wolf corrected p-values presented in brackets. *** p<0.01, ** p<0.05, * p<0.1.

Table 4: Worker-generated Manager earnings and inequality in workers' earnings

	Total earnings generated by Rank A workers			Inequality in workers' earnings		
	Man Manager	Woman Manager	$M = W$ pvalue	Man Manager	Woman Manager	$M = W$ pvalue
No Choice (NC)	166.25	172.44	0.49	261.75	193.89	0.11
Choice(C)	163.33	178.44	0.16	246.40	278.44	0.44
Choice & Talk (CT)	177.69	179.83	0.84	176.77	212.75	0.42
Choice & Backlash (CB)	181.38	168.53	0.20	240.63	180.10	0.15
$\mathrm{H}_0 : C = NC$	0.79	0.48	.	0.72	0.04	.
$\mathrm{H}_0 : NC = CT$	0.28	0.42	.	0.09	0.62	.
$\mathrm{H}_0 : NC = CB$	0.05	0.71	.	0.62	0.74	.
$\mathrm{H}_0 : C = CT$	0.27	0.88	.	0.16	0.09	.
$\mathrm{H}_0 : C = CB$	0.08	0.34	.	0.89	0.02	.
$\mathrm{H}_0 : CT = CB$	0.68	0.28	.	0.20	0.40	.

Note: We report the total earnings generated by rank A workers to the manager, and the difference in the total earnings of the highest and the lowest earning worker within each group. Both total profits and differences in workers' earnings are computed over the 5 stages of the experiment. p-values are generated by tests of equality of means. p<0.01, ** p<0.05, * p<0.1.

Table 5: Manager messages to rank B workers

	Praise	Motivate	Explain	Compete	Fairness	Team	Cordial
Choice & Talk (CT)							
Male Manager	0.52	0.29	0.31	0.09	0.06	0.00	0.25
Female Manager	0.57	0.36	0.13	0.04	0.06	0.03	0.15
$\mathrm{H}_0 : M = F(p - values)$	(0.57)	(0.36)	(0.00)	(0.16)	(0.93)	(0.14)	(0.11)
Choice & Backlash (CB)							
Male Manager	0.28	0.21	0.51	0.19	0.07	0.03	0.20
Female Manager	0.42	0.28	0.31	0.06	0.08	0.04	0.33
$\mathrm{H}_0 : M = F(p - values)$	(0.04)	(0.28)	(0.01)	(0.01)	(0.82)	(0.54)	(0.06)
For M $\mathrm{H}_0 : CT = CB(p - values)$	(0.00)	(0.05)	(0.00)	(0.00)	(0.64)	(0.87)	(0.76)
For F $\mathrm{H}_0 : CT = CB(p - values)$	(0.04)	(0.39)	(0.04)	(0.90)	(0.44)	(0.33)	(0.01)

Note: We report the percentage of messages of each type sent by male and female managers to the the rank B workers over 5 stages of the experiment in the *Choice & Talk* and the *Choice & Backlash* treatments. For examples of messages in each category, see Table A5. Since the categories are not mutually exclusive, the percentages do not sum up to 1. P-values are generated by Chi-square tests.

Table 6: Received Backlash

	Dep. Var: Angry emojis sent to Manager			
To Female Manager	0.823*	0.744*	0.980*	1.053**
	(0.452)	(0.440)	(0.490)	(0.494)
Performance$_{t-1}$	0.097**	0.101**	0.090**	0.072*
	(0.040)	(0.038)	(0.041)	(0.040)
RankB$_{t-1}$	0.918**	0.777*	1.067*	1.075**
	(0.430)	(0.391)	(0.411)	(0.415)
Constant	-0.982	-0.346	-0.097	-0.991
	(0.816)	(1.616)	(1.621)	(2.080)
Observations	140	140	140	140
Controls	No	Yes	Yes^{+}	Yes^{+}
Stage FE	Yes	Yes	Yes	Yes
Manager's message type	No	No	No	Yes
Clusters	35	35	35	35

Note: OLS regressions. Robust standard errors, clustered at the group level, in parentheses. The dependent variable is the number of angry emojis, including zeros, that the rank B worker sent to his or her Manager in the *Choice & backlash* treatment. Performance$_{t-1}$ is the number of correct puzzles solved by the rank B worker in the previous Stage. RankB$_{t-1}$ is a dummy equal to 1 if the rank B worker was rank B also in the previous Stage. Controls are: age, gender and STEM or economics field of study. Additional controls in columns 3 and 4 are: native English speaker, having held a leadership position, and Big5 agreeableness measure. In column 4, we add measures of for the type of message that the manager sent to the B worker when assigning ranks. The seven not-mutually exclusive message types are presented in Table A5, i.e., they are either praising, motivating, explaining the ranking decision, inducing competition, mentioning fairness, team building or cordial. p<0.01, ** p<0.05, * p<0.1.

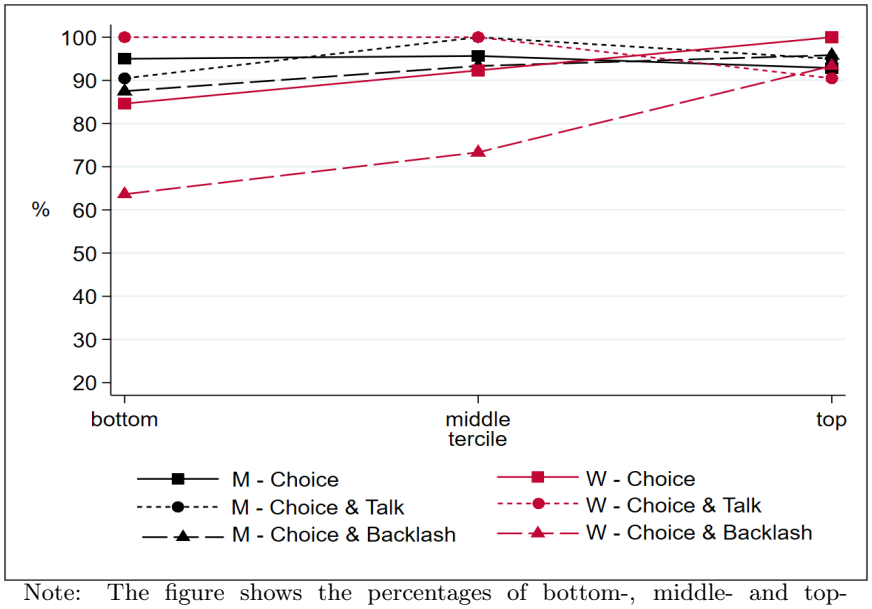
Table 7: Manager reaction to received backlash

	Dep. Var: Switched ranks and promoted the B worker			
Female	0.182*	0.263**	0.234**	0.225**
	(0.093)	(0.105)	(0.099)	(0.098)
Angry Emojis$_{t-1}$	-0.055***	-0.019	-0.037	-0.028
	(0.020)	(0.024)	(0.025)	(0.029)
ΔPerformance$_{t-1}$	-0.047***	-0.047***	-0.045***	-0.045***
	(0.006)	(0.006)	(0.006)	(0.006)
Female $\times$ Angry Emojis$_{t-1}$		-0.061*	-0.034	-0.039
		(0.035)	(0.033)	(0.040)
Constant	0.442***	0.396***	1.284***	1.390**
	(0.097)	(0.097)	(0.366)	(0.539)
H_0 : Emojis$_{t-1}$ + $Fem. \times$ Emojis$_{t-1} = 0$		0.003	0.001	0.003
Observations	140	140	140	140
Controls	Yes	Yes	Yes	Yes$^+$
Stage FE	Yes	Yes	Yes	Yes
Clusters	35	35	35	35

Note: Linear probability models. Robust standard errors, clustered at the group level in parentheses. The dependent variable is a dummy equal to 1 if the Manager switched ranks and promoted the rank B worker at the beginning of the current Stage in the *Choice & Backlash* treatment. Angry Emojis$_{t-1}$ is the number of angry emojis sent by the rank B worker in the previous Stage. Δperformance$_{t-1}$ is the difference in the performances of the rank A and rank B workers in the previous Stage. Controls are: age and STEM or economics field of study. Additional controls in column 4 are: native English speaker, having held a leadership position, and Big5 agreeableness measure. p<0.01, ** p<0.05, * p<0.1.

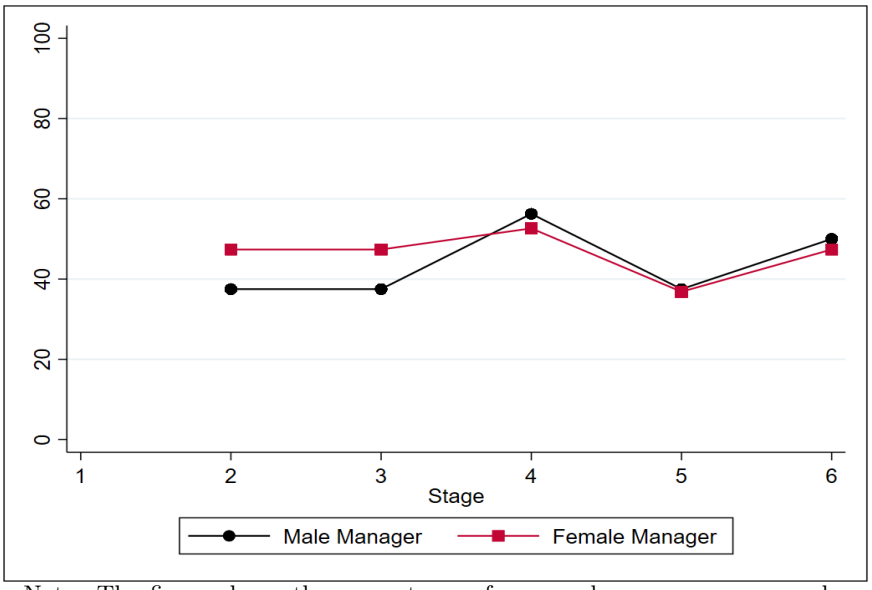
APPENDIX FIGURES AND TABLES

Figure A1: The decision to be Manager by performance terciles



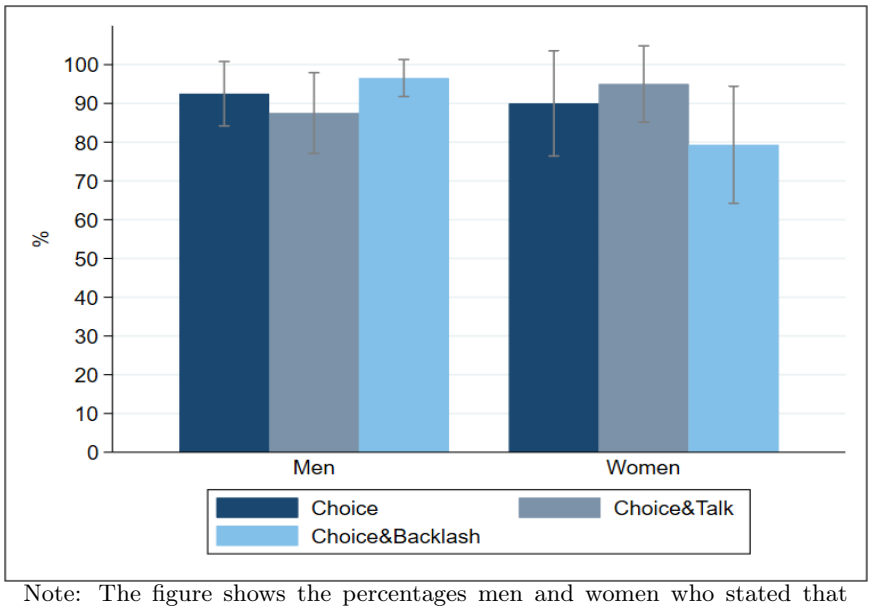
Note: The figure shows the percentages of bottom-, middle- and top-performers who wanted to be manager by gender and treatment.

Figure A2: Percentage of managers receiving at least one angry emoji



Note: The figure shows the percentages of men and women managers who received at least one angry emoji in each stage when backlash was possible.

Figure A3: Self-selection into leadership in groups of 2 men and 1 woman



Note: The figure shows the percentages men and women who stated that they wanted to be the manager of their group, when restricting the sample to groups formed by 2 men and 1 woman.

Table A1: Subjects' characteristics

	Male				Female				M=F (All)
	NC	C	CT	CB	NC	C	CT	CB	p-value
Age	22.61	22.84	24.38	22.55	23.32	22.19	23.27	22.00	0.414
	(2.49)	(2.87)	(6.98)	(4.63)	(4.51)	(3.09)	(6.59)	(4.27)	
STEM-Bus-Econ (frequency)	0.85	0.91	0.78	0.80	0.71	0.67	0.54	0.66	0.000
	(0.36)	(0.29	(0.42)	(0.41)	(0.46)	(0.48)	(0.50)	(0.48)	
Native speaker (frequency)	0.34	0.29	0.52	0.47	0.41	0.51	0.67	0.56	0.008
	(0.48)	(0.46)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	
Past Leadership (frequency)	0.80	0.85	0.89	0.83	0.90	0.83	0.85	0.83	0.724
	(0.40)	(0.36)	(0.32)	(0.38)	(0.30)	(0.38)	(0.36)	(0.38)	
Big 5 Agreeableness Index	27.80	26.67	27.92	27.70	29.44	28.34	29.35	29.95	0.004
	(5.05)	(5.53)	(6.98)	(5.18)	(7.26)	(6.78)	(5.94)	(5.21)	

Table A2: Performance in the real effort task

	Number of correctly solved puzzles		
	(1)	(2)	(3)
No Choice (NC)	0.396	0.269	0.130
	(0.739)	(0.706)	(0.704)
Choice&Talk (CT)	0.735	0.894	0.172
	(0.746)	(0.696)	(0.696)
Choice&Backlash (CB)	0.230	0.019	-0.441
	(0.762)	(0.737)	(0.711)
Female	0.574	0.119	-0.500
	(0.858)	(0.851)	(0.863)
Female x NC	-0.620	-0.198	0.171
	(1.193)	(1.115)	(1.138)
Female x CT	0.707	0.620	1.074
	(1.101)	(1.068)	(1.027)
Female x CB	0.855	1.015	1.322
	(1.153)	(1.129)	(1.073)
Constant	13.172***	19.253***	14.558***
	(0.585)	(1.174)	(1.717)
H_0: NC = CT	0.642	0.359	0.948
H_0: NC = CB	0.824	0.729	0.390
H_0: CT = CB	0.502	0.221	0.330
H_0: NC + Fem. x NC=0	0.811	0.935	0.736
H_0: CT + Fem. x CT=0	0.075*	0.057*	0.098*
H_0: CB + Fem. x CB=0	0.210	0.229	0.272
Controls	No	Yes	Yes$^+$
Stage FE	Yes	Yes	Yes
Observations	2,502	2,502	2,430
Clusters	417	417	405

OLS estimates. Note: each subject solved up to 20 puzzles in each of 6 rounds. Robust standard errors, clustered at the individual level, in parentheses. Controls are: performance in the task, age, and STEM or economics field of study. Additional controls in column 4 are: native English speaker, having held a leadership position, Big5 agreeableness measure. The number of observations is lower in column 4 because a computer glitch preventing us from collecting post-experiment survey data from 12 participants in the *Choice* treatment. *** p<0.01, ** p<0.05, * p<0.1.

Table A3: Self-selection into the manager position - Pooled sample

	Dep. Var: Wants to be manager			
	(1)	(2)	(3)	(4)
Choice&Talk (CT)	-0.002	-0.027	-0.009	-0.005
	(0.037)	(0.055)	(0.051)	(0.052)
Choice&Backlash (CB)	-0.055	0.006	0.010	0.023
	(0.038)	(0.047)	(0.046)	(0.048)
Female	-0.049	-0.019	-0.007	-0.008
	(0.036)	(0.057)	(0.055)	(0.064)
Female x CT		0.056	0.045	0.046
		(0.077)	(0.074)	(0.079)
Female x CB		-0.154*	-0.160*	-0.157*
		(0.092)	(0.090)	(0.095)
Constant	0.960***	0.947***	0.907***	0.810***
	(0.032)	(0.038)	(0.131)	(0.147)
Controls	No	No	Yes	Yes^{+}
Observations	315	315	315	303
Clusters	105	105	105	101
H_0: CT = CB	0.158	0.499	0.639	0.505
H_0: CT + F. x CT=0		0.548	0.455	0.441
H_0: CB + F. x CB=0		0.047	0.037	0.084
H_0: CT + F. x CT=CB + F. x CB		0.011	0.008	0.013

Note: Linear probability models. Robust standard errors, clustered at the group level, in parentheses. Controls are: performance in the task in Stage 1, age, and STEM or economics field of study. Additional controls in column 4 are: native English speaker, having held a leadership position, Big5 agreeableness measure and number of women in the group. The number of observations is lower in column 4 because a computer glitch prevented us from collecting post-experiment survey data from 12 participants in the *Choice* treatment. p<0.01, ** p<0.05, * p<0.1.

Table A4: Manager decision-making - Pooled sample

	Switched ranks			Ex-post mistake			Lost profit		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
No Choice (NC)	0.104	0.101	0.081	-0.021	-0.020	-0.048	0.866	0.852	0.618
	(0.067)	(0.069)	(0.070)	(0.081)	(0.081)	(0.080)	(0.817)	(0.829)	(0.835)
Choice&Talk (CT)	0.172**	0.182**	0.165**	0.034	0.015	-0.024	0.497	0.343	-0.003
	(0.068)	(0.074)	(0.078)	(0.090)	(0.088)	(0.089)	(0.708)	(0.700)	(0.746)
Choice&Backlash (CB)	0.149*	0.129	0.108	-0.121	-0.109	-0.128*	-0.318	-0.346	-0.459
	(0.084)	(0.083)	(0.082)	(0.076)	(0.077)	(0.077)	(0.671)	(0.657)	(0.661)
Female	0.042	0.024	-0.025	-0.089	-0.081	-0.109	-0.179	-0.238	-0.495
	(0.069)	(0.072)	(0.074)	(0.075)	(0.076)	(0.078)	(0.695)	(0.732)	(0.792)
ΔPerformance$_{t-1}$	-0.037***	-0.037***	-0.037***	-0.014***	-0.014***	-0.014***	-0.129***	-0.129***	-0.133***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.031)	(0.031)	(0.032)
Female x NC	0.000	0.008	0.060	0.181*	0.176*	0.207**	0.011	0.020	0.329
	(0.104)	(0.104)	(0.108)	(0.100)	(0.100)	(0.103)	(1.078)	(1.094)	(1.139)
Female x CT	-0.033	-0.047	0.001	0.036	0.050	0.086	-0.173	-0.115	0.294
	(0.101)	(0.103)	(0.109)	(0.109)	(0.106)	(0.110)	(0.968)	(0.972)	(1.019)
Female x CB	0.137	0.155	0.204*	0.209**	0.200**	0.214**	0.864	0.919	1.017
	(0.117)	(0.116)	(0.116)	(0.097)	(0.099)	(0.098)	(1.018)	(1.030)	(1.069)
Constant	0.228***	0.492***	0.484**	0.321***	0.105	-0.062	1.672***	0.784	-0.832
	(0.056)	(0.130)	(0.214)	(0.074)	(0.132)	(0.181)	(0.548)	(1.198)	(1.662)
Controls	No	Yes	Yes$^+$	No	Yes	Yes$^+$	No	Yes	Yes$^+$
Stage FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	556	556	540	556	556	540	556	556	540
Clusters	139	139	135	139	139	135	139	139	135
$H_0 : NC = CT$	0.336	0.278	0.272	0.503	0.656	0.756	0.670	0.550	0.467
$H_0 : NC = CB$	0.597	0.737	0.740	0.138	0.178	0.218	0.155	0.138	0.187
$H_0 : CT = CB$	0.799	0.536	0.518	0.047	0.097	0.170	0.262	0.318	0.534
$H_0 : NC + F. \times NC = 0$	0.198	0.166	0.078	0.008	0.001	0.012	0.188	0.195	0.199
$H_0 : CT + F. \times CT = 0$	0.066	0.080	0.031	0.255	0.278	0.331	0.625	0.727	0.688
$H_0 : CB + F. \times CB = 0$	0.001	0.000	0.000	0.153	0.152	0.172	0.471	0.460	0.501
$H_0 : CT + F. \times CT = CB + F. \times CB$	0.077	0.087	0.096	0.781	0.691	0.715	0.738	0.606	0.685

Note: LPM in columns 1 to 6; OLS in columns 7 to 9. The dependent variables are: i) a dummy equal to 1 if the Manager switched ranks (columns 1-3); ii) a dummy equal to 1 if the the worker chosen to be rank A ended up performing worse than the other worker (columns 4-6); iii) the difference between the profits that the manager would have made by assigning rank A to the best performing worker, and the profits actually made in the current Stage (columns 7-9). The analysis is restricted to Stages 3 to 6, since Stage 3 is the first stage when rank switching is allowed. Δperformance$_{t-1}$ is the difference in the performances of the rank A and the rank B workers in previous Stage. Controls are: age and STEM or economics field of study. Additional controls in column 3, 6 and 9 are: native English speaker, having held a leadership position, and Big5 agreeableness measure. The number of observations is lower when adding these control variables because a computer glitch prevented us from collecting post-experiment survey data from 12 participants in the *Choice* treatment. Robust standard errors, clustered at the manager level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A5: Examples of messages sent by managers to workers

Examples of messages sent by managers
Praising
"you are a powerhouse", "Amazing work, you are doing great"
"You're a star. Keep shining.", "Outstanding work as always!"
"you are KILLING IT!", "Manager material! I like the consistency!"
Motivating
"Keep it up!", "You got this dude"
"Keep up the good work!", "I believe in you Jill! You got this."
"I know you can do better than that.", "we are almost there! finish strong!!"
Explaining the rank allocation
"Greg did better in general which is why he is A do better and you will be"
"You are B because Emily did better in 1. If you beat her, I will switch your rank.",
"This is not a tactic: If you do better than Worker A, you will be rewarded as such."
Competition-inducing
"Your coworker outscored you! Earn your spot back!"
"Good job, Carly. Roger is catching up so keep it up"
"your percentage increase was higher than worker A, so beat her this time and Rank A is yours"
Fairness
"Trying to be fair and switch it up"
"I just want to distribute the ranks equally"
"I'll alternate A and B so you have a better chance to be paid more"
Team building
"lets do this together"
"Let's go, team!", "The team needs your best effort, John."
"Let's finish strong here team!"
Cordial
"Sorry", "sorry had to give him a chance"
"Thank you for your hard work!", "Good luck! Hope you make lots of $ today :)"
"Great job! Thank you so much!"

Table A6: Manager messages to rank A workers

	Praise	Motivate	Explain	Compete	Fairness	Team	Cordial
Choice & Talk (CT)							
Male Manager	0.74	0.46	0.12	0.08	0.06	0.02	0.11
Female Manager	0.72	0.35	0.07	0.04	0.03	0.04	0.10
$H_0 : M = F(p - values)$	(0.75)	(0.14)	(0.19)	(0.31)	(0.37)	(0.34)	(0.87)
Choice & Backlash (CB)							
Male Manager	0.51	0.31	0.29	0.17	0.05	0.03	0.07
Female Manager	0.67	0.29	0.18	0.06	0.04	0.06	0.14
$H_0 : M = F(p - values)$	(0.03)	(0.80)	(0.09)	(0.02)	(0.80)	(0.23)	(0.19)
For M $H_0 : CT = CB(p - values)$	(0.00)	(0.23)	(0.00)	(0.00)	(0.81)	(0.75)	(0.48)
For F $H_0 : CT = CB(p - values)$	(0.38)	(0.12)	(0.02)	(0.76)	(0.96)	(0.23)	(0.40)

Note: We report the percentage of messages of each type that male and female managers sent to rank A workers over the 5 stages of the experiment in the *Choice & Talk* and the *Choice & Backlash* treatments. For examples of messages in each category, see Table A5. Since the categories are not mutually exclusive, the percentages do not sum up to 1. P-values are generated by Chi-square tests.

Table A7: Examples of messages sent by rank B workers

Thanks or jokes
"Thanks. Don't agree, but thanks."
"I understand. Thank you."
"haha what?"
Accepts rank
"no problems"
"Fine."
"Fair enough"
Commits to work hard
"I will be better and become rank A"
"I will try my best."
"I will make you proud."
Apologetic
"I apologize for slack last stage and hope to make it up"
"I am disappointed in myself. Need to concentrate more."
"sorry....got stucked..lol"
Upset
"I am mad."
"eat dirt, im out here grinding and u missin out"
"i hate u"
Questions the ranking decision
"why??? i did 20 correct in first and 19 in second?"
"you really can't do better can you"
"are u kidding me... u changed to mine after the failed then u go back to them"

Table A8: Messages sent by Rank B workers to managers

	Thanks or jokes	Approves rank	Commits to work hard	Apologetic	Upset	Questions rank
To Male Manager	0.16	0.59	0.16	0.06	0.09	0.06
To Female Manager	0.17	0.51	0.17	0.09	0.04	0.17
$\mathrm{H}_0 : M = F(p - values)$	(0.92)	(0.28)	(0.92)	(0.43)	(0.22)	(0.03)

Note: We report the percentage of messages of each type that rank B workers sent to managers in the *Choice & Backlash* treatment over the 5 stages of the experiment. For examples of messages in each category, see Table A7. Since the categories are not mutually exclusive, the percentages do not sum up to 1. P-values are generated by Chi-square tests.

Table A9: Robustness: Decision-making in groups with 1 male worker and 1 female worker

	Switched ranks		Mistaken		Lost Profits	
	(1)	(2)	(3)	(4)	(5)	(6)
No Choice (NC)	0.141*	0.147*	0.047	0.043	1.479	1.459
	(0.080)	(0.083)	(0.099)	(0.100)	(0.993)	(0.997)
Choice&Talk (CT)	0.212**	0.204**	0.010	0.001	0.063	-0.207
	(0.086)	(0.092)	(0.116)	(0.119)	(0.885)	(0.909)
Choice&Backlash (CB)	0.152*	0.119	-0.109	-0.103	-0.266	-0.453
	(0.088)	(0.088)	(0.087)	(0.087)	(0.727)	(0.696)
Female	0.081	0.060	-0.023	-0.021	0.559	0.402
	(0.098)	(0.103)	(0.100)	(0.102)	(1.017)	(1.063)
Female x NC	0.029	0.017	0.070	0.068	-1.229	-1.387
	(0.140)	(0.141)	(0.134)	(0.136)	(1.484)	(1.470)
Female x CT	-0.106	-0.109	0.020	0.019	-0.627	-0.682
	(0.149)	(0.151)	(0.150)	(0.149)	(1.384)	(1.371)
Female x CB	0.021	0.039	0.082	0.084	-0.832	-0.612
	(0.165)	(0.164)	(0.118)	(0.119)	(1.155)	(1.193)
ΔPerformance$_{t-1}$	-0.032***	-0.032***	-0.009***	-0.009***	-0.120**	-0.127***
	(0.004)	(0.004)	(0.003)	(0.003)	(0.046)	(0.046)
Constant	0.226***	0.511***	0.293***	0.172	1.784**	1.645
	(0.071)	(0.150)	(0.093)	(0.166)	(0.684)	(1.482)
Observations	332	332	332	332	332	332
Controls	No	Yes	No	Yes	No	Yes
Stage FE	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	83	83	83	83	83	83
$H_0 : NC = CT$	0.448	0.537	0.732	0.699	0.201	0.127
$H_0 : NC = CB$	0.915	0.751	0.048	0.052	0.073	0.038
$H_0 : CT = CB$	0.548	0.353	0.225	0.289	0.701	0.770
$H_0 : NC + F \times NC = 0$	0.145	0.155	0.207	0.243	0.817	0.946
$H_0 : CT + F \times CT = 0$	0.392	0.467	0.759	0.837	0.605	0.392
$H_0 : CB + F \times CB = 0$	0.222	0.237	0.734	0.812	0.214	0.261
$H_0 : CT + F \times CT = CB + F \times CB$	0.640	0.682	0.498	0.650	0.436	0.793

Note: LPM models in columns 1 to 4; OLS in columns 5 and 6. The dependent variables are: i) a dummy equal to 1 if the Manager switched ranks and promoted the worker who was Rank B in the previous Stage (columns 1-2); ii) a dummy equal to 1 if the the worker chosen to be rank A ended up performing worse than the other worker (columns 3-4); iii) the difference between the profits that the manager would have made by assigning rank A to the best performing worker, and the profits actually made in the current Stage (columns 5-6). The analysis is restricted to Stages 3 to 6, since Stage 3 is the first stage when rank switching is allowed. Δperformance$_{t-1}$ is the difference in the performances of the rank A and the rank B workers in previous Stage. Controls are: age and STEM or economics field of study. Robust standard errors, clustered at the manager level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

ONLINE APPENDIX

The Online Appendix contains instructions of both the laboratory and the online experiments.