# Gender differences in top leadership roles: Does worker backlash matter?*

Priyanka Chakraborty†
Southern Methodist University

Danila Serra‡
Southern Methodist University

This version: February 26, 2019

**Abstract**

Top leadership positions involve the necessity of making decisions, like promotions, demotions and dismissals, which please some employees and upset others. Backlash from unhappy employees may therefore arise. We examine whether the anticipation of such backlash induces women, more than men, to select out of top leadership roles and to perform differently when/if they become leaders. We conduct a novel laboratory experiment that simulates corporate decision-making. We find that women are significantly less likely to self-select into a managerial position when facing the possibility of receiving angry messages from employees. However, once in a leadership role, women perform no differently than men and are unaffected by the possibility of worker backlash. We also find that male and female managers have different leadership styles, i.e. they motivate their employees differently, and that female managers receive significantly more angry messages from employees.

JEL Codes: C92, D91,J16

Key words: Gender differences, leadership, backlash, experiment

*We thank Jeff Carpenter, Lata Gangadharan, Karen Hauge and Tim Salmon for useful comments and discussions. We thank seminar participants at SMU and conference participants at the Southern Economic Association conference, the ESA North-American and World meetings, the 2018 Texas Experimental workshop, and the 2019 ASSA conference.

†Email: pchakraborty@mail.smu.edu.

‡Corresponding author. Email: dserra@smu.edu.

# 1 Introduction

Numerous studies have shown that, holding performance constant, women are evaluated more negatively than men. This is true in politics (Beaman et al., 2009; Branton et al., 2018; Rheault et al., 2019), business (Egan et al., 2017; Elsesser and Lever, 2011), academia (Boring, 2017; Mengel et al., 2017; Hengel, 2017; Sarsons, 2017) and laboratory settings (Ayalew et al., 2018; Grossman et al., 2016; Shurchkov and van Geen, 2017).

In this paper, we examine whether the anticipation of (harsher) negative judgment from subordinates may induce women to select out of top leadership positions, and to perform differently when in a leadership role.

The problem of missing top female leaders exists in all spheres of life. Only 19% of firms worldwide have female top managers and only 6% of CEOs at S&P 500 companies are women. In politics, women hold only 23% of seats in national parliaments worldwide.[1] In the US, only 10% of governors and 20% of the mayors of the 100 largest American cities are women. In academia, averaging across all fields, less than one third of full professors are women.[2] This percentage falls to 13.9% in economics[3] and 10.5% in engineering (Yoder, 2018).[4]

A number of studies have identified behavioral or preference-based constraints to women's self-selection into top leadership roles.[5] There is evidence that risk aversion (see, e.g., Eckel and Grossman, 2008), reticence to initiate negotiations (e.g., Bowles et al., 2007; Babcock and Laschever, 2009), aversion to competitive environments (e.g., Gneezy et al., 2003; Niederle and Vesterlund, 2007; Flory et al., 2014; Preece and Stoddard, 2015),[6] preferences over job attributes (Wiswall and Zafar, 2017), willingness to volunteer for or accept low promotability tasks (see, e.g., Babcock et al., 2017) and self-stereotyping (Coffman, 2014) may hold women back. A recent study by Born et al. (2018) also shows that women are less likely to self-select into a leadership role in male dominated environments.

Here, we ask whether women may be less likely than men to pursue top leadership roles to avoid the social disapproval that they expect to receive from unhappy employees while on the job. This might be because they anticipate receiving more disapproval than men, in line with the literature, or because they are more averse to social judgment. The latter would be the case if women had stronger social image motivations. Such motivations, as modeled in Bénabou and Tirole (2006), capture the role of others' opinions in one's utility, i.e., the desire to be liked and respected by others.[7] A few experimental studies suggest that men and women differ in their image motivations. Jones and Linardi (2014), in the context of a laboratory donation experiment, find that women are more likely to be "wallflowers", i.e., to be averse to any unwanted attention. Alan et al. (2018) find evidence of gender differences in adolescents' – but not in children's – willingness to make risky decisions on behalf of a group, and attribute such gender gap to a significant decline in teenage girls' "social

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

[2] For recent statistics on the gender leadership gap in the US, see Warner and Corley (2017): `https://cdn.americanprogress.org/content/uploads/2017/05/21145352/WomenLeadershipGap2017-factsheet1.pdf`

[3] See the 2017 report from the American Economic Association's Committee of the Status of Women in the Economics Profession (CSWEP): https://www.aeaweb.org/content/file?id=6388

[4] A recent study by Nittrouer et al. (2018) shows that male academics in 6 disciplines are significantly more likely to be invited to be colloquium speakers at prestigious US universities. The observed gender difference is neither due to differences in the gender and rank of the available speakers, nor to women declining invitations more often than men.

[5] 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).

[6] Erkal et al. (2018) 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.

[7] The literature on how social observability and judgment affect behavior is fast growing. See for instance: Andreoni and Bernheim (2009); Andreoni and Petrie (2004); Ariely et al. (2009); Linardi and McConnell (2011); Salmon and Serra (2017); Xiao and Houser (2011). See also the recent overview provided by Bursztyn and Jensen (2017).

confidence", as measured by their willingness to perform a real effort task under public scrutiny. Banerjee et al. (2015) find that while women are less likely to volunteer to act as third part punishers in public goods games, the gender difference disappears when the role of punisher is made anonymous and is therefore shielded from public scrutiny.

We consider a business environment, although our setting could be easily applied or extended to other domains. In a firm setting, top leadership involves decision-making that necessarily makes some workers happy and others unhappy. Think of promotions, demotions and dismissals. This implies that backlash from unhappy employees, in the form of, at the minimum, negative judgment and disapproving messages, is warranted. We investigate whether the possibility of worker backlash deters women more than men from self-selecting into top leadership roles and whether it differentially affects the actions of male and female managers, possibly leading to gender differences in performance and outcomes.

We employ a novel laboratory experiment that simulates managerial decision-making involving rank allocations among employees. 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).[8] In our leadership experiment, the main task of a leader, which we refer to as *manager*, is to promote or demote employees. Managers, who, by design, are the highest earners in their group, have an informational advantage over the productivities of two employees in multiple rounds of a real effort task, and, at the beginning of each round, they have the responsibility of allocating ranks that determine the employees' earnings. In each round, there can only be a high-rank and a low-rank employee; therefore, rank allocation necessarily creates income inequality among workers.

In our baseline treatment, we assign the managerial role based on performance in a preceding real effort task. In our two treatments of interest (*Choice* and *Choice & Messages*), we allow subjects to volunteer for the leadership role.[9] Crucially, in one of these treatments (*Choice & Messages*), we allow managers to send free-form messages to workers following each rank allocation stage, and we allow workers to send messages back; moreover, we allow low-rank workers to send angry emoticons to signal their disapproval of the rank allocation outcome. By comparing our two *Choice* treatments, we are able to clearly examine whether the possibility of worker backlash plays a role in the observed gender leadership gap as opposed to (or in addition to) a mere aversion to creating inequality among employees.[10]

Our design also allows us to test for gender differences in managerial performance, where performance is measured as the propensity to assign the high rank to the best performing worker. By varying our treatment conditions, we are able to assess whether the possibility of worker backlash leads to gender differences in the criteria used when making decisions regarding employees' promotions and demotions. The existing literature assessing the outcomes of female versus male leadership tends to examine the effects of increased gender diversity on corporate boards, typically due to the introduction of gender quotas, with mixed findings. While

[8] 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., 2018), or tasks that require leaders to make decisions on behalf of their group (Alan et al., 2018; Reuben et al., 2012; Born et al., 2018)

[9] The manager is selected among the volunteers based on performance in the preceeding real effort task.

[10] Numerous experimental studies have shown that women have different distributional preferences than men, i.e., they are more egalitarian (e.g., Eckel and Grossman, 1998; Andreoni and Vesterlund, 2001; Dufwenberg and Muren, 2006). This may induce (some) women to avoid jobs that would make them responsible for creating inequalities among employees. If this is the case, we should observe a gender difference in self-selection into leadership already in our *Choice* treatment. Moreover, if aversion to creating inequality is the only driver of gender differences in leadership and worker backlash plays no role, we should see no differences between our *Choice* and *Choice & Messages* treatments.

female leadership seems to reduce firms' short-term profits, due to fewer workforce layoffs (Matsa and Miller, 2013, 2014),[11] there is evidence that it improves employees' working conditions (Devicienti et al., 2016), and it reduces both the gender pay gap among top executives (Matsa and Miller, 2011) and the gender gap in promotions (Kunze and Miller, 2017).[12] The main advantage of our experimental setting is that it allows comparing men's and women's managerial behavior and performance in a controlled environment where male and female leaders are subject to the exact same environment, decision set and incentive systems.

Through the analysis of the messages that managers send to their employees in the *Choice & Messages* treatment we are also able to examine whether men and women have different leadership styles, i.e., whether they communicate with and motivate employees differently. This is an underdeveloped area of research in economics.[13] While there are numerous experimental studies where leaders can communicate with followers, we know of only one study, Timko (2017), which examines gender differences in the language used by leaders. In the context of a minimum effort game[14] where leaders can send free-form messages to group members to induce them to coordinate on the Pareto-efficient equilibrium, Timko (2017) finds that while men and women leaders are equally effective in inducing high effort, men leaders send more assertive messages while women leaders, in treatments where followers can send messages back, express significantly more often that they are part of the group. Studies in psychology also find gender differences in the general use of language, with men using more assertive language - e.g., through imperative statements - and women using more affiliative language - e.g., statements of support, agreement, and acknowledgment.[15]

Finally, by comparing the frequency and number of angry emoticons sent to male and female managers by low-rank employees, we can provide a clean assessment of differences in the attitudes of low-rank employees toward male and female managers. Since the conditions under which low-rank workers can send angry messages to male or female leaders are identical in the experimental setting, we are able to identify differences in workers' attitudes driven purely by the gender of their manager.[16]

Our results can be summarized as follows. First, women are less likely to self-select into a leadership position in a setting where leaders have to make decisions that affect some workers positively and others negatively. This is not due to gender-specific distributional preferences. Rather, women's reluctance to lead is caused by the possibility of worker backlash, i.e., the possibility of receiving angry messages from employees. Second, while women are less likely to volunteer to be managers, once in a leadership position, they do not perform differently than men; in fact, both genders assign ranks based on worker productivity. This holds both in the absence and in the presence of worker backlash. Gender differences in managerial decision-making emerge only when male and female leaders face two workers of equal productivity. In this case, male managers

[11] On the other hand, Wolfers (2006) finds no significant differences to stock returns to firms under female leadership, and both Adams and Ferreira (2009) and Schwartz-Ziv (2017) provide evidence that gender diverse boards are more active in monitoring executives.

[12] Moreover, Flabbi et al. (2016) find evidence that female CEOs are better at evaluating the productivity of female workers, leading to better allocations of female workers across tasks and to wage distributions that more clearly reflect individual productivities. For a recent review of the literature, see Miller (2017)

[13] The operations management literature distinguishes between *transformational* leaders who "transform or change the basic values, beliefs, and attitudes of followers so that they are willing to perform beyond the minimum levels specified by the organization" and *transactional* leaders, who "are founded on an exchange process in which the leader provides rewards in return for the subordinate's effort" (Podsakoff et al., 1990). Based on a meta-analysis of 45 studies, Eagly et al. (2003) conclude that female leaders tend to be more transformational than male leaders, although the difference is small in magnitude.

[14] In a minimum-effort game, members of a group simultaneously choose an effort level, with higher effort being associated with higher individual costs. Payoffs depends positively on the lowest effort level chosen within a group and negatively on the effort chosen by the decision-maker. In the game employed by Timko (2017), payoffs for individual $i$ are equal to: $200 - 5effort_i + 6effort_{\min}$.

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

[16] 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).

tend to keep the ranking status quo (i.e., the worker that was high-rank in the previous round stays high-rank), whereas female leaders are more inclined to switch ranks, therefore promoting the worker previously assigned the low rank. Third, the analysis of the messages sent by managers to workers suggests, in line with the existing literature, that men and women have different leadership styles. While male and female leaders are equally encouraging, men's messages are more likely to induce competition between workers, whereas women's messages tend to emphasize team building, give practical suggestions on how to better solve the real effort task, and are more likely to contain greetings and apologies. Finally, male and female leaders are equally likely to receive worker backlash (the extensive margin), yet female managers receive significantly more angry messages from low-rank employees (the intensive margin). This last finding confirms that the observed gender gap in willingness to assume leadership roles in our setting may be due to women's correct anticipation of more severe backlash from unhappy employees rather than by a greater absolute aversion to negative judgment.

# 2 The Leadership 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 we 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 asked for their age, gender, field of study, and previous participation in an experiment. The answer to the gender question led to a pre-determined list of either male or female names, which we took from Bertrand and Mullainathan (2004)'s correspondence study of race-based discrimination.[17] The male subjects saw a list of male names and the female subjects saw a list of female names. We informed subjects that for the duration of the experiment they would be identified with a fictitious name, and we invited them to pick a name from the gender-specific list they saw on their screen.[18] We did not allow two or more subjects to choose the same name, so each name disappeared from the list in real time when picked by another participant.

Stage 1 followed. In this stage, and in the following five stages, subjects engaged 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 consisted in finding a 4-letter word in a 6x6 letter matrix in 5 minutes, for a maximum of 20 matrices.[19] Subjects played individually. They received an endowment of 40 ECU and

[17] As our focus is on gender differences, we used 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.

[18] 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.

[19] The decision screen was divided in two halves, as shown in Appendix. On the left, subjects saw the matrix and on the right, they saw a list of 40 words. Each puzzle contained two words that appeared on the list. In order to earn money, subjects had to identify one word per puzzle. We used the website http://tools.atozteacherstuff.com/word-search-maker/wordsearch.php to

earned 2 ECU for each puzzle they solved correctly in 5 minutes. At the end of Stage 1, subjects received feedback on their performance and were provided instructions on the following 5 stages (Stages 2 to 6) of the experiment. Crucially, they were randomized into groups of 3 and they were 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.[20] Subjects remained in the same group for the duration of the experiment.

In Stages 2 to 6, two group members played in the role of workers and one in the role of manager. The roles of manager and worker were assigned at the end of Stage 1 according to treatment-specific rules, and retained through Stage 6. The manager got a fixed wage of 100 ECU and his/her main task was to decide, at the beginning of each stage of the experiment, which worker would be Rank A and which worker would be Rank B in the following stage. The Rank A worker got a wage of 80 ECU, while the Rank B worker got a wage of 20 ECU. After the rank allocation, all members of the group engaged in a similar puzzle task as in Stage 1 of the experiment.[21] Each correctly solved puzzle generated 2 ECU in addition to the initial wage. Moreover, each puzzle solved correctly by the Rank A worker generated 2 ECU also to the Manager. Therefore the earnings from each of the 5 active stages (Stages 2 to 6) of the experiment were determined as follows:

- The Manager got 100 ECU + 2 ECU per puzzle + 2 ECU per puzzle solved by Rank A worker;
- Rank A worker got 80 ECU plus 2 ECU per puzzle;
- Rank B worker got 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 was informed about the performances of the current Rank A and Rank B workers and had to decide whether to keep or reassign ranks before the next stage began.

Following Stage 6, subjects filled in a post-experiment questionnaire, where we elicited 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).[22]

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 two employees. This allows us to define and clearly measure managerial efficiency as the likelihood that the manager will assign ranks based purely on workers' relative performances. Moreover, we designed the task so that it would lead as much as possible to constant relative performances within a group across the 5 stages (Stages 2 to 6) of the experiment. This is to limit the occurrence of rank-switching behavior driven purely by changes in relative performances, as our aim is to examine rank-switching that is instead due to distributional concerns and/or anticipation of worker backlash. Of course, in many settings workers' performances cannot be objectively or

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 configurarion 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.

[20] We ended up having 67% of the groups made of one woman and two men, 27% made of two women and one man, and 6% made of men only.

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

[22] 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.

precisely measured, and 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. Finally, in our setting, we can also investigate rank allocation decisions when the two workers performed equally well in the task.[23] This way, we are able to assess whether there are differences in the *subjective* criteria used by male and female managers to allocate ranks when no objective distinction could be made between the two workers.

## 2.2 Treatments

In our *Baseline* (T1) treatment, at the end of Stage 1, in each group of 3 participants the manager is chosen based on performance in Stage 1.[24] 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 will 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.[25] 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* (T2) 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 volunteer for the manager role, we choose the manager based on performance in Stage 1, as in Baseline.

Finally, in our *Choice & Messages* (T3) treatment, we still allow for self-selection into the leadership role, yet we also allow for two-way free-form communication between the manager and each of the two workers at each rank allocation stage. Specifically, at the beginning of each stage (Stages 2 to 6), after assigning ranks A and B to the workers and before the real-effort task begins, the manager needs to send a free-form message to the Rank A worker and a free-form message to the Rank B worker. The manager can write anything he or she wishes to communicate to each worker. After the manager submits the individual messages, each worker sees the message sent to him or her and has to send a message back to the manager. Importantly, the Rank B worker can also send up to 5 angry emoticons to the manager to express disapproval of the ranking decision. The messages sent by the two workers, including the angry emoticons, 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 emoticons from Rank B workers, before they are

[23] This happened about 16% of the time, averaging across all treatments.

[24] 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 – we break ties in favor of women. Subjects are unaware of this.

[25] While there is a large literature documenting gender differences in competitiveness and self-confidence (especially in math-based tasks) we wanted to abstract from both factors in our experiment, in order to be able to isolate the role played by the possibility of worker backlash in the origination of gender differences in leadership.

asked whether they would like to be the manager of their group.

Overall, our design allows us to clearly investigate: 1) gender differences in self-selection into leadership roles in settings where leadership involves decision-making that makes some people happy and others unhappy: 2) gender differences in managerial performance, as measured by efficiency in the rank allocation decisions; 3) gender differences in leadership styles (in the *Choice & Messages* treatment); and 4) differences in workers' attitudes towards male and female leaders, as measured by the frequency and number of angry emoticons sent by rank B workers (in the *Choice & Messages* treatment).

Our payoff structure makes it optimal for each subject to volunteer for the manager role and for each manager to always assign ranks based on past performance (since the nature of the real effort task remains constant across stages and it is only the performance of the Rank A worker that generates additional earnings to the manager). Therefore, if individuals are purely money-maximizers we should see no gender differences in volunteering and rank-allocations, with all subjects volunteering in T2 and T3, and all leaders assigning Rank A to the best performing worker in each stage of the experiment in all treatments. However, if managers also care about their workers' payoffs and have distributional concerns – as shown by a large experimental literature – they may not allocate ranks based purely on task performance. Instead, they may alternate ranks between the two workers in order to equalize their earnings. If women are more inequality averse than men, we may see more arbitrary rank switches, and hence lower efficiency in rank allocation, from female managers in all treatments. Gender differences in distributional preferences may also cause women to be less willing than men to manage their team in both T2 and T3.

The comparison between the *Choice & Messages* and the *Choice* treatments allows us to test whether the possibility of worker backlash plays a role in the decision of men and women to assume a leadership role. In particular, an aversion to the possibility of receiving angry messages from rank B employees should lead to a decline in volunteering for the managerial role in T3 as compared to T2. Moreover, if women are more averse to worker backlash or expect more severe backlash than men, we should see a larger gender gap in volunteering in T3 than in T2. Finally, aversion to workers' backlash may also reduce efficiency in rank allocation in T3 as compared to T2 and T1.

## 2.3 Implementation

We conducted 20 experimental sessions at the Laboratory for Research In Experimental Economics (LREE) at Southern Methodist University. We involved a total of 306 participants, of which 41% are women, as shown in Table 1. Each subject participated in only one session and one treatment. In each session, we had between 3 and 7 groups of three subjects (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. 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.[26]

Only one randomly selected active stage of the experiment was used for actual payments. Experimental earnings were converted from ECUs 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.

[26] The instructions employed in the *Choice & Messages* treatment are provided in Appendix.

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.

# 3 Results

We start by describing our subject pool and conducting balance tests across treatments. A total of 182 men and 124 women participated in the experiment. In Table 10, 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 speakers, whether they reported having held a leadership position, and their average Big 5 Agreeableness Index. The average age is 22.7, with no significant differences across treatments and between men and women. Most of our participants are STEM, Business or Economics majors, although the percentage of men majoring in these fields (85%) is significantly higher than the percentage of women (68%, $p = 0.000$). Most subjects reported having held a leadership role in the past, and most of them are not native speakers (36% of men and 47% of women are native speakers, $p = 0.066$). In line with the existing literature, our female participants score significantly higher than men in the Big 5 Agreeableness Index ($p = 0.007$). Balance tests reveal no statistically significant differences in individual characteristics across treatments in the female sample, whereas in the male sample the only significant difference is in the percentage of native speakers, which is significantly higher ($p = 0.020$) in T3 than in T2.

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. Table 2 reports the average number of puzzles correctly solved by men and women in Stage 1 of the experiment (which determined participants' chances of becoming managers later on). The lack of statistically significant differences in the performances of men and women in all treatments suggests that we succeeded in employing a gender-neutral task. Table 2 also shows that there are no statistically significant differences in performances across treatments for both the male and female samples. This is confirmed by regression analysis for both Stage 1 and all the active stages of the experiment. The corresponding estimates are reported in Table 11 in the Appendix.

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. Descriptive statistics[27] show that 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.

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 willingness to volunteer for the manager position (Section 3.1), and on their performance once in the leadership role (Section 3.2). We then present our findings on male and female managers' leadership styles (Section 3.3). We conclude by reporting on the attitudes of Rank B workers toward male and female leaders, as measured by the number of angry messages sent in our *Choice & Messages* treatment (Section 3.4).

[27] Not presented here but available upon request.

### 3.1 Gender differences in the decision to be manager

Almost all subjects volunteered to be a manager in our *Choice* treatment, with no significant differences between men and women, as shown in Table 3 and Figure 2. However, a large and statistically significant gender gap emerges in our *Choice & Messages* treatment, where we see 78% of women volunteer to be a manager as opposed to 95% of men ($p = 0.007$). This suggests that women are no less willing to assume the leadership role in our setting absent the possibility of worker backlash. In other words, the gender leadership gap that we observe is due purely to gender differences in the reaction to the possibility of receiving angry messages from Rank B workers.

This is confirmed by regression analysis, as shown in Table 4. We report estimates generated by linear probability models where the dependent variable is a dummy equal to 1 if the subject stated the he or she would like to be the manager of his/her group.[28] We first examine the male and female samples separately (columns 1 to 6). We start by testing for treatment effects by including our treatment variables only. We then gradually add demographics and other individual characteristics. We also control for individuals' performance in the real effort task in Stage 1, as this may affect subjects' perceived probability of being chosen as the manager if they volunteered. Finally, following Born et al. (2018), which show that the gender composition of a group may affect women's self-selection into leadership, we include a dummy equal to 1 if there were two women in the group. Recall that by design we aimed to have only groups with 1 woman and 2 men. However, we ended up having 2 women and 1 man in 26% of our groups. Note that we lose a few observations when controlling for past leadership positions and the Big 5 Agreeableness index. This is due to a software glitch that prevented us from conducting the post-experiment survey in one of our sessions.

The estimates in columns 1 to 6 confirm that women, but not men, are less likely to volunteer as managers when facing the possibility of worker backlash. We also see that performance in the task is a strong predictor of volunteering for women. As for the individual characteristics affecting self-selection into the manager role, they seem to differ by gender. 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 volunteer for the manager role. This does not apply to the female sample. Among 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. The gender composition of the group does not seem to affect individuals' decision to self-select into the leadership role in our setting, regardless of their gender.[29]

Our main finding is confirmed when pooling the male and female samples and introducing both a female dummy and its interaction with our *Choice&Messages* treatment. The estimates in column 9, in particular, confirm that while men are not less likely to volunteer to be managers in T3 than in T2, women are both less likely to volunteer in T3 than men (see the estimated coefficient and p-value obtained for *Female*x*T3*) and less likely to volunteer in T3 than T2 ($p - value = 0.039$, Wald test for the sum of coefficients of Female and

[28] We employ linear probability models because it has been shown (Norton et al., 2004) that interpreting interaction terms in non-linear models is not straightforward. Importantly, Norton et al. (2004) show that the marginal effect of an interaction term may not be the same as the estimated coefficient, and further that the standard t-test is inaccurate. Nevertheless, we report estimates from probit regressions in Table 12 in Appendix for comparison. The results are qualitatively the same.

[29] However, we should note that we did not randomize the number of women across treatments (as we aimed to have groups of 2 men and one woman only). Any difference in the number of women per group is purely due to the gender composition of the subjects that showed up for the experiment on a given day. As a result, we have significantly fewer groups with two women and one man in the Choice&Messages treatment (17%) than in the other two treatments (36% in Baseline and 35% in Choice).

its interaction with T3). We summarize our first result as follows:

**Result 1** *Women are less likely to volunteer for the leadership position when facing the possibility of receiving angry messages from employees.*

Given the observed positive correlation between performance in the task in Stage 1 and women's decision to volunteer to be manager, we examine whether gender differences in the decision to volunteer are more extreme for bottom performers than for the middle and top performers. Figure 3 reports the percentages of volunteers among men and women in T2 and T3 by performance terciles.[30] In the *Choice* treatment, we do not see evidence of significant gender differences in self-selection into the manager role for any tercile ($p = 0.311$ for the bottom tercile, $p = 0.674$ for the middle tercile, $p = 0.277$ for the top tercile). In contrast, under *Choice&Messages*, significant gender differences in volunteering exist among both bottom ($p = 0.065$) and middle performers ($p = 0.009$). Indeed, in both groups, we observe a 30 percentage point gender leadership gap. 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 volunteering among performance terciles in the male sample for either treatment, female top performers are significantly more likely to volunteer than female bottom performers, especially in the *Choice & Messages* treatment ($p = 0.058$ in *Choice&Messages*, $p = 0.104$ in *Choice*).[31] This indicates that, although in our setting the performance of the manager is never disclosed to the employees and the manager is never in competition with his or her workers, women feel that, in order to be managers of their group, they need to be extremely good at the task. Men do not seem to have the same concerns.

## 3.2 Gender differences in managers' decision-making

Before testing for gender differences in rank-allocation decisions, we examine whether men and women who become managers through self-selection (in *Choice* and *Choice&Messages*) differ in some salient individual characteristics. We find that the only significant difference between male and female managers is their score in the Big 5 Agreeableness index, with women being more "agreeable" than men, as shown in Table 13 in Appendix. Note that we do not see any significant difference in the percentages of men and women majoring in STEM, business or economics, which indicates that, as expected, our female manager sample is more likely to be majoring in male-dominated fields as compared to the general female population.

Next, we look at the behavior of male and female managers in Stages 2 to 6 of the experiment. Our primary outcome variable is the likelihood that the manager assigns ranks based on worker productivity, i.e., the number of correctly solved puzzles in the previous stage. Specifically, in each stage, we define a rank-allocation by a manager as *efficient* if the manager assigns Rank A to the worker that solved more puzzle in the previous stage. Note that such definition only applies to cases where there is a clear difference between the performances of the two workers. In what follows, we restrict the analysis to such cases, while in Section 3.2.2 we examine rank allocations when the managers faced two equally productive workers.

[30] The bottom tercile is made of subjects who completed 12 or less puzzles correctly in Stage 1 of the experiment. The middle tercile consistes 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.

[31] Female top performers are also more likely to volunteer than female middle performers, albeit the difference is not statistically signficant ($p = 0.142$ in *Choice&Messages*, $p = 0.259$ in *Choice*).

### 3.2.1 Efficiency in rank-allocations

Table 5 displays the percentage of efficient rank-allocations made by male and female managers in our three treatments, i.e., the percentage of times that Rank A was assigned to the worker who performed best in the previous stage (when the difference in the performances of the two workers was non-zero). In all of our treatments, we find no evidence of gender differences in the efficiency of rank allocations. It seems that, in the female sample, self-selected managers are more likely to allocate ranks based on performance as compared to exogenously chosen managers (compare T2 and T3 to T1), while we see no such difference in the male sample. Finally, the possibility of worker backlash does not seem to significantly decrease the efficiency of the decision-making of both male and female managers.[32]

One important factor that may not be constant across treatments and managers' gender is the difference in the performances of the two workers. For instance, it is possible that in the *Choice&Messages* treatment, managers were more likely to face similarly skilled workers, which may have led to more frequent rank switches. In order to control for the difference in workers' performances and other individual characteristics, we estimate a set of linear probability models, where the dependent variable is a dummy equal to 1 if the manager assigned Rank A to the best performing worker and 0 otherwise. Our results are displayed in Table 6.[33] We first look at the male manager and female manager samples separately (columns 1 to 6) and then pool the samples while introducing a female manager dummy and its interaction with our treatment dummies (columns 7 to 9). In all specifications, we cluster the standard errors at the group level. We always start by including only our treatment dummies and a stage dummy. We then control for the difference between the previous stage's performances of the worker currently assigned rank A and the worker currently assigned Rank B $[(R_A - R_B)_{lag}]$. We also control for the manager's own performance in the previous stage ($Perf_{lag}$). In the most comprehensive specification, we include demographics and other individual characteristics (same as in Table 4).

The estimates show that there are no significant differences in the likelihood of making efficient rank allocations across treatments. This is true for both male and female managers. The difference between the performances of the two workers is an important predictor of the likelihood of allocating rank A to the best performing worker. Indeed, the positive and highly significant estimated coefficient of $(R_A - R_B)_{lag}$ indicates that a manager is significantly more likely to allocate Rank A to a worker the greater the difference between such worker's and the other workers' numbers of completed puzzles. Importantly, the estimates in columns 7 to 9 show no evidence of gender differences in the efficiency of rank allocations in all treatments. Our second result follows:

**Result 2** *Female managers are as likely as male managers to allocate ranks efficiently, both in the absence and in the presence of the possibility of worker backlash.*

### 3.2.2 Rank allocation when workers perform equally

An important feature of our real effort task is that it allows for a clear and objective assessment of workers' performance, thus making it possible to examine managers' decision-making, i.e. their likelihood of allocating ranks based on relative worker performances in the different treatments. In this section, we examine cases

[32] Figure 6 in Appendix breaks down the data by stage, i.e., it shows the percentage of efficient rank allocations by male and female managers in each stage (when there was a clear best performer among the employees).

[33] Probit regressions generate similar results, except that, when estimating probit regressions we are unable to control for the difference in the performances of rank A and rank B workers, since any positive difference predicts the dependent variable perfectly.

where managers face equally performing workers and therefore, when allocating ranks, are forced to use criteria other than objective assessments of workers' productivities. This way, we can investigate whether male and female managers use different *subjective* criteria in their rank allocation decisions.

Managers faced equally performing workers in 16% of the cases over the five stages of the experiment, with no signficant differences across treatments.[34] In Table 7, we estimate linear probability models where the dependent variable is equal to 1 if a worker is assigned Rank A in a given stage.[35] Since the only information managers have about their workers is their gender and the ranks they held in the previous stage (besides their productivities), we include both variables in our empirical specification. We first include only the rank status of the worker in the previous stage, i.e., whether he/she was Rank A, the gender of the worker, the gender of the manager and our treatment dummies (column 1). We then include the interaction between previous rank A status and the gender of the manager (column 2).

The estimates show that workers under male and female leadership are allocated ranks differently. In particular, workers under a male manager are more likely to be allocated Rank A if they were Rank A in the previous stage. In other words, male managers, when facing equally performing workers, tend to keep the ranks assigned in the previous stage. Female managers, on the other hand, are less likely to keep the status quo and more likely to switch ranks ($p - value = 0.020$, Wald test for the sum of coefficients of Female Manager and its interaction with the worker's lagged Rank A status). Our next result follows.

**Result 3** *When facing equally productive workers, male managers are more likely to keep the previous stage's ranks whereas female managers are more likely to switch ranks.*

In column 3, we also include interactions with our treatment dummies. This allows us to test whether male and female managers allocate ranks differently to equally performing workers depending on the treatment. What emerges is that while male managers are more likely to keep current ranks (i.e., less likely to switch ranks) in the *Baseline* and the *Choice* treatments, they behave similarly to female managers – i.e., they tend to switch ranks – when facing the possibility of worker feedback in the *Choice&Messages* treatment.

### 3.2.3 Managers' responsiveness to angry messages in T3

In our *Choice & Messages* (T3) treatment, Rank B workers can send up to 5 angry emoticons to their manager after learning their rank, at the beginning of each stage (Stages 2 to 6). We analyze workers' likelihood of sending angry emoticons, and the number of emoticons sent, in Section 3.4. Here, we examine an important aspect of a managers' behavior and performance, i.e., how they respond to worker backlash. In particular, are managers more likely to assign Rank A to a less productive Rank B worker after the receipt of angry messages? And does the receipt of angry messages affect male and female managers' future rank-allocation decisions differently? The answer to both questions is no, according to our data. If we replicate the analysis displayed in Table 6 while adding the number of angry emoticons received by the Rank B worker in the previous stage, we find that worker backlash, i.e. the number of angry emoticons received, has no significant impact on a manager's likelihood of allocating Rank A to the best performing worker. This is true for both male and female managers.[36] Furthermore, the analysis of the likelihood that a Rank B worker is promoted to Rank A conditional on the angry emoticons sent in the previous stage – see Table 14 in Appendix – shows that, no matter the gender of the manager, the likelihood of being promoted is not affected by the number of

[34] The two workers in a group performed equally 14% of the times in Baseline, 13% of the times in Choice and 20% of the time in Choice&Messages, with no signfiicant differences across treatments.

[35] Probit regressions generate the same results and are available from the authors upon request.

[36] The corresponding regression table – not displayed here – is available from the authors upon request.

angry emoticons sent. If anything, sending angry emoticons to a manager may lower the likelihood of being promoted.

**Result 4** *Male and female managers' ranking decisions are equally unaffected by the receipt of angry messages from employees.*

### 3.3 Gender differences in leadership styles

Recall that in our *Choice & Messages* treatment, the manager had to send a free form message to each worker following the rank allocation decision and before the real effort task began. In this section we examine whether the messages sent by male and female managers to Rank A and Rank B workers differ in content and tone. We enlisted two independent coders, who classified the messages sent by managers over the 5 stages of the experiment as: 1) encouraging; 2) inducing competition among workers; 3) mentioning the number of puzzles solved by the worker; 4) mentioning fairness; 5) providing suggestions on how to solve the task; and 6) using cordial words like "thank you" and "sorry."[37]

Both male and female managers sent primarily encouraging messages to both Rank A and Rank B workers, as shown in Figure 4. Among the other categories, we see most messages being either meant to induce competition among workers or suggesting how to better do the task or using cordial words. We sum up the messages in the latter two categories and refer to them as "nice" messages. We focus only on these main three message categories - encouraging, competitive and "nice" messages - in the analysis below.

We find that over the five stages of the experiment, male managers sent significantly more messages inducing competition among workers. This is true both for the messages sent to Rank A workers ($p = 0.077$ two-side ttest, $p = 0.114$ Wilcoxon rank-sum test) and those sent to Rank B workers ($p = 0.005$ two-side ttest, and $p = 0.002$ Wilcoxon rank-sum test). On the other hand, women sent more "nice" messages although the difference is not statistically significant for the messages sent to the rank B workers, and borderline significant at the conventional level for the messages sent to Rank A workers ($p = 0.117$ two-side ttest, $p = 0.09$ Wilcoxon rank-sum test). Our findings are in line with studies of language in psychology ( e.g., Kern et al., 2016), showing that women are more likely to use affiliative language, i.e., statements of support and acknowledgment. We summarize our findings as follows:

**Result 5** *Male and female managers communicate differently to employees, with men being more likely to use language aimed at inducing competition among workers, and women being more likely to use words signaling support and acknowledgment.*

### 3.4 Gender differences in the extent of received worker backlash

In this section, we examine the angry emoticons that Rank B workers sent to male versus female managers in our *Choice&Messages* treatment. This is what we call *worker backlash*. A first look at the data shows that male and female managers are equally likely to receive at least one angry emoticon from rank B workers, as shown in the first row of Table 8. Over the 5 stages of the experiment, both male and female managers received at least one angry emoticon about 45 percent of the times. Figure 7, in 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

[37] For each category, we computed the reliability coefficient of intercoder agreement (Cohen, 1960). Categories with $\kappa \geqslant 0.3$ are considered fair or better (Landis and Koch, 1977). The computed $\kappa$ scores indicate that all categories have been coded with at least a fair agreement between the two coders. We therefore aggregate the responses of the two coders by taking their averages.

rank B worker, the difference is not statistically significant. Figure 5 and the second row of Table 8, however, show that, conditional on receiving at least one emoticon, female managers receive a higher number of angry emoticons on average.

Next, we conduct regression analysis on Rank B workers' decisions to send one or more angry emoticons to their manager over the 5 stages of the experiment. We first look at the decision to send at least one angry emoticon (i.e., worker backlash at the extensive margin)[38] and then turn to the number of emoticons sent, conditional on sending at least one (i.e., worker backlash at the intensive margin). By conducting regression analysis, we are able to examine the impact of the gender of the manager, the gender of the worker and their interaction. In columns 3 and 7 of Table 9 we add the worker's performance and his/her rank status in the previous stage, and in columns 4 and 8 we add workers' demographics and individual characteristics. In all regressions, we cluster the standard errors at the group level.

The estimates in columns 1 to 4 confirm the lack of statistically significant differences in worker backlash toward male versus female managers at the extensive margin. The estimates in columns 5 to 8 show that female managers are subject to more severe worker backlash at the intensive margin, and this is due primarily to the behavior of male Rank B workers. In fact, while female managers do not seem to receive more angry emoticons when averaging across the behaviors of male and female Rank B workers (column 5), gender differences emerge when looking at the decisions of male and female Rank B workers separately. The significant coefficient of the "To Female Manager" dummy in columns 6 to 8 indicate that male workers send significantly more angry messages to female managers. Interestingly, the positive and significant coefficient of the female dummy indicates that female Rank B workers send more angry emoticons to male managers. Since, by design, we have more male workers than female workers in our setting to simulate a male-dominated work environment, female managers end up receiving more negative messages in total. Our findings suggest that a more gender equal workforce would likely annul the gender difference in worker backlash. We summarize our results below.

**Result 6** *a) Female and male managers are equally likely to receive worker backlash;*

*b) Female managers receive more severe backlash, due to male rank B workers' tendency to send more angry messages to female than male managers.*

Recall that in our *Choice&Messages* treatment, workers also sent free-form messages to their manager at the rank allocation stage. Two independent coders categorized the messages sent by the Rank A workers and by Rank B workers as: 1) expressing commitment to hard work; 2) containing jokes; 3) justifying past performance; 4) requesting the high rank; 5) challenging the manager, either by asking information about the relative performances in previous round, or by boasting about own performance, or by expressing anger about the ranking. We report the average numbers of messages of each kind received by male and female managers in Tables 15 and 16 in Appendix. While it seems that female managers are more likely to be challenged by Rank B workers and more likely to receive messages in which workers (of both ranks) justify their performances in the task, the differences are not statistically significant at the conventional level. Overall, we do not find any evidence of differences in the language used by Rank A and Rank B workers when communicating with their male or female managers.

[38] As before, for ease of interpretation of the estimated coefficients, we report findings from linear probability models. Probit regressions generate qualitatively the same results and are available from the authors upon request.

# 4 Conclusion

The literature on gender differences in leadership is still in its nascent stage. We contribute to this literature by conducting an experimental study aimed at enhancing our understanding on whether and why men and women differ in their willingness to assume leadership roles and in their behaviors as leaders. We ask whether women are less likely to self-select into managerial positions that require decision-making generating inequalities among employees, possibly leading to worker backlash. We find strong evidence of gender difference in willingness to assume a managerial role. This is not due to gender differences in aversion to generating inequality among employees. Rather, it is due to men's and women's differential responses to the possibility of worker backlash. The analysis of the angry messages sent by employees to their managers shows that female managers receive more severe backlash, suggesting that women's reluctance to assume leadership positions may be due to the correct anticipation of the harsher negative judgment they would receive from their subordinates, rather than or in addition to a greater absolute aversion to social disapproval.

By employing a controlled experiment, we are also able to clearly assess the performances of male and female managers under identical incentive systems and decision sets. In our setting, managers have the task to assign either a high rank or a low rank to two employees, where ranks determine the employees' earnings, and workers' productivities are only visible to the manager. We do not find any significant gender differences in managerial performance, with both genders assigning ranks based on workers' relative performance, both in the absence and in the presence of the possibility of worker backlash. Moreover, the receipt of angry messages does not seem to affect the subsequent rank-allocation decisions of either (self-selected) male or female managers. The only gender difference we observe in managers' behaviors concerns the allocation of ranks when the two workers are equally productive, hence equally deserving of the high rank. In these cases, we see that male managers are more likely to keep the status quo, i.e. they tend to keep ranks as they were in the past (when possibly the two workers performed differently). In contrast, female managers are more likely to switch ranks, therefore promoting the worker who was previously assigned the low rank. This is the only instance where gender differences in distributional preferences seem to lead to differential rank-allocation decisions.

Finally, we find evidence of gender differences in the language used by male and female managers to communicate with and motivate their workers. In line with the existing literature, we find that men are more likely to use language aimed at inducing competition among workers, whereas women are more likely to use words signaling support and acknowledgment. In contrast, we do not find any significant differences in the language used by employees when communicating with male versus female managers.

Overall, our study offers important insights into an under-studied behavioral constraint that may prevent women from self-selecting into top leadership roles that involve the management of lower rank employees, including the necessity to promote, demote or dismiss members of the workforce. In evaluating the relevance and implications of our findings, it is important to note that our analysis is based on a laboratory setting where decisions are anonymous and worker backlash takes the form of angry emoticons sent by only one other subject via a computer terminal, rather than via personal face to face interaction. We may therefore be underestimating the role that the possibility of worker backlash plays in contributing to the gender leadership gap in field settings, where managers have to face more direct and personal expressions of anger from multiple unhappy subordinates.

Our study could be extended in many interesting ways. First, future work could examine whether and how the gender (im)balance existing within the workforce may affect women's willingness to assume the managerial role. If, as suggested by our data, female managers are more likely to receive worker backlash from male workers and this contributes to their reluctance to self-select into the leadership role, having a more

gender balanced workforce may reduce the gender leadership gap. By design, in our experiment we aimed to have all groups made of two men and one woman to simulate male dominated work environments. We ended up having some (26%) of the groups composed of two women and one man. While we did not see the gender composition of the groups affect subjects' self-selection into the leadership role, we are unable to draw definite conclusions as we did not purposely randomize the gender composition of the groups across our treatments. Another interesting extension of our study would be to examine the relationship between the possibility of worker backlash and leadership decisions in a setting where managers can also receive approval messages from happy employees. In particular, future work could assess how strong or numerous the expected approval messages should be to compensate for the expected negative messages and therefore induce more women to self-select into top leadership roles.

# References


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

Alan, Sule and Ertac, Seda and Kubilay, Elif and Loranth, Gyongyi. (2018). Understanding gender differences in leadership.

Andreoni, James and Bernheim, B Douglas. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica 77*(5), 1607–1636.

Andreoni, James and Petrie, Ragan. (2004). Public goods experiments without confidentiality: a glimpse into fund-raising. *Journal of public Economics 88*(7-8), 1605–1623.

Andreoni, James and Vesterlund, Lise. (2001). Which is the fair sex? gender differences in altruism. *The Quarterly Journal of Economics 116*(1), 293–312.

Ariely, Dan and Bracha, Anat and Meier, Stephan. (2009). Doing good or doing well? image motivation and monetary incentives in behaving prosocially. *American Economic Review 99*(1), 544–55.

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

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

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

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

Banerjee, Debosree and Ibañez, Marcela and Riener, Gerhard and Wollni, Meike. (2015). *Volunteering to take on power: Experimental evidence from matrilineal and patriarchal societies in India*. Number 204. DICE Discussion Paper.

Beaman, Lori and Chattopadhyay, Raghabendra and Duflo, Esther and Pande, Rohini and Topalova, Petia. (2009). Powerful women: female leadership and gender bias. *Quarterly Journal of Economics 124*(4), 1497–1540.

Bénabou, Roland and Tirole, Jean. (2006). Incentives and prosocial behavior. *American economic review 96*(5), 1652–1678.

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

Bertrand, Marianne and Mullainathan, Sendhil. (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, Anne. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics 145*, 27–41.

Born, Andreas and Ranehill, Eva and Sandberg, Anna. (2018). A man's world? the impact of a male dominated environment on female leadership.

Bowles, Hannah Riley and Babcock, Linda and Lai, Lei. (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, Jordi and Cooper, David J. (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, Regina and English, Ashley and Pettey, Samantha and Barnes, Tiffany D. (2018). The impact of gender and quality opposition on the relative assessment of candidate competency. *Electoral Studies 54*, 35–43.

Bursztyn, Leonardo and Jensen, Robert. (2017). Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual Review of Economics 9*, 131–153.

Chaudhuri, Ananish and Li, Yaxiong and Sbai, Erwann. (2018). (un)willingness to lead? men, women and the leadership gap.

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

Cohen, Jacob. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*(1), 37–46.

Devicienti, Francesco and Grinza, Elena and Manello, Alessandro and Vannoni, Davide. (2016). Which are the benefits of having more female leaders? evidence from the use of part-time work in italy.

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

Dufwenberg, Martin and Muren, Astri. (2006). Generosity, anonymity, gender. *Journal of Economic Behavior & Organization 61*(1), 42–49.

Eagly, Alice H and Johannesen-Schmidt, Mary C and Van Engen, Marloes L. (2003, July). Transformational, transactional, and laissez-faire leadership styles: a meta-analysis comparing women and men. *Psychological Bulletin 129*(4), 569–591.

Eckel, Catherine C and Grossman, Philip J. (1998). Are women less selfish than men?: Evidence from dictator experiments. *The economic journal 108*(448), 726–735.

_____ (2008). Men, women and risk aversion: Experimental evidence. *Handbook of experimental economics results 1*, 1061–1073.

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

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

Erkal, Nisvan and Gangadharan, Lata and Xiao, Erte. (2018). Leadership selection: Can changing the default break the glass ceiling?

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

Flabbi, Luca and Macis, Mario and Moro, Andrea and Schivardi, Fabiano. (2016). Do female executives make a difference? the impact of female leadership on gender gaps and firm performance. Technical report, National Bureau of Economic Research.

Flory, Jeffrey A and Leibbrandt, Andreas and List, John A. (2014). Do competitive workplaces deter female workers? a large-scale natural field experiment on job entry decisions. *The Review of Economic Studies 82*(1), 122–155.

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

Grossman, Philip J and Eckel, Catherine and Komai, Mana and Zhan, Wei and others. (2016). It pays to be a man: Rewards for leaders in a coordination game. Technical report, Monash University, Department of Economics.

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

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

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

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

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

Kern, Margaret L and Park, Gregory and Eichstaedt, Johannes C and Schwartz, H Andrew and Sap, Maarten and Smith, Laura K and Ungar, Lyle H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological methods 21*(4), 507.

Kunze, Astrid and Miller, Amalia R. (2017). Women helping women? evidence from private sector data on workplace hierarchies. *Review of Economics and Statistics 99*(5), 769–775.

Landis, J. Richard and Koch, Gary G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics 33*(2), 363–374.

Leaper, Campbell and Smith, Tara E. (2004). A meta-analytic review of gender variations in children's language use: talkativeness, affiliative speech, and assertive speech. *Developmental psychology 40*(6), 993.

Linardi, Sera and McConnell, Margaret A. (2011). No excuses for good behavior: Volunteering and the social environment. *Journal of Public Economics 95*(5-6), 445–454.

Matsa, David A and Miller, Amalia R. (2011). Chipping away at the glass ceiling: Gender spillovers in corporate leadership. *American Economic Review 101*(3), 635–39.

_____ (2013). A female style in corporate leadership? evidence from quotas. *American Economic Journal: Applied Economics 5*(3), 136–69.

_____ (2014). Workforce reductions at women-owned businesses in the united states. *ILR Review 67*(2), 422–452.

Mengel, Friederike and Sauermann, Jan and Zölitz, Ulf. (2017). Gender bias in teaching evaluations. *Journal of the European Economic Association*.

Miller, Amalia R. (2017). Women and leadership.

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

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

Nittrouer, Christine L and Hebl, Michelle R and Ashburn-Nardo, Leslie and Trump-Steele, Rachel CE and Lane, David M and Valian, Virginia. (2018). Gender disparities in colloquium speakers at top universities. *Proceedings of the National Academy of Sciences 115*(1), 104–108.

Norton, Edward C and Wang, Hua and Ai, Chunrong and others. (2004). Computing interaction effects and standard errors in logit and probit models. *Stata Journal 4*, 154–167.

Podsakoff, Philip M and MacKenzie, Scott B and Moorman, Robert H and Fetter, Richard. (1990). Transformational leader behaviors and their effects on followers' trust in leader, satisfaction, and organizational citizenship behaviors. *The leadership quarterly 1*(2), 107–142.

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

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

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

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

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

Salmon, Timothy C and Serra, Danila. (2017). Corruption, social judgment and culture: An experiment. *Journal of Economic Behavior & Organization 142*, 64–78.

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

Schwartz-Ziv, Miriam. (2017). Gender and board activeness: the role of a critical mass. *Journal of Financial and Quantitative Analysis 52*(2), 751–780.

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

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

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

Warner, Judith and Corley, D. (2017). WomenŠs leadership by the numbers.

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

Wolfers, Justin. (2006). Diagnosing discrimination: Stock returns and ceo gender. *Journal of the European Economic Association 4*(2-3), 531–541.

Xiao, Erte and Houser, Daniel. (2011). Punish in public. *Journal of Public Economics 95*(7-8), 1006–1017.

Yoder, BL. (2018). Engineering by the numbers. 2018. In *American Society for Engineering Education*.

# Tables and figures

| | Sessions | Groups | Participants | | |
|---|---|---|---|---|---|
| | | | Men | Women | Total |
| Baseline (T1) | 7 | 34 | 61 | 41 | 102 |
| Choice (T2) | 6 | 33 | 57 | 42 | 99 |
| Choice&Messages (T3) | 7 | 35 | 64 | 41 | 105 |
| Total | 20 | 102 | 182 | 124 | 306 |

Table 1: Sessions and treatments

| | Men | Women | $H_0$: M=W p-value |
|---|---|---|---|
| Baseline (T1) | 13.77 | 13.83 | 0.96 |
| Choice (T2) | 13.44 | 14.57 | 0.34 |
| Choice&Messages (T3) | 14.16 | 14.70 | 0.64 |
| $H_0$: T1=T2 (p-value) | 0.75 | 0.59 | |
| $H_0$: T1=T3 (p-value) | 0.72 | 0.53 | |
| $H_0$: T2=T3 (p-value) | 0.50 | 0.92 | |

P-values are generated by double-sided tests of equality of means.

Table 2: Number of correctly solved puzzles in Stage 1

| | Men<br>% | Women<br>% | $H_0$: M=W<br>p-value |
|---|---|---|---|
| Choice (T2) | 94.74 | 92.86 | 0.698<br>[0.696] |
| Choice&Messages (T3) | 95.31 | 78.05 | 0.007***<br>[0.010***] |
| $H_0$: T2=T3 (Chi-square test: p-value) | 0.884 | 0.055* | |
| [Fisher: p-value] | [1.00] | [0.067*] | |

P-values in column 3 are generated by Chi-square tests. P-values from Fisher exact tests are in brackets. *** p<0.01, ** p<0.05, * p<0.1.

Table 3: Percentages of subjects who volunteered to be managers

| | Dep. Variable: Subject wants to be manager | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Men | | | Women | | | All | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Choice&Messages(T3) | 0.006 | 0.018 | 0.043 | -0.148* | -0.134* | -0.154* | -0.057 | 0.006 | 0.044 |
| | (0.886) | (0.627) | (0.339) | (0.057) | (0.060) | (0.063) | (0.151) | (0.886) | (0.347) |
| Performance | | 0.007 | 0.008 | | 0.025*** | 0.027*** | | | 0.016*** |
| | | (0.202) | (0.154) | | (0.003) | (0.003) | | | (0.002) |
| Age | | 0.003 | 0.005 | | 0.002 | 0.005 | | | 0.004 |
| | | (0.250) | (0.193) | | (0.800) | (0.552) | | | (0.332) |
| STEM-bus-eco | | 0.082 | 0.060 | | 0.158* | 0.210** | | | 0.148** |
| | | (0.333) | (0.609) | | (0.086) | (0.033) | | | (0.047) |
| Native | | -0.034 | -0.058 | | -0.117 | -0.100 | | | -0.074 |
| | | (0.495) | (0.338) | | (0.124) | (0.296) | | | (0.143) |
| Leadership | | | 0.077 | | | 0.087 | | | 0.075 |
| | | | (0.463) | | | (0.477) | | | (0.327) |
| big5-agree | | | -0.008** | | | 0.004 | | | -0.002 |
| | | | (0.046) | | | (0.537) | | | (0.669) |
| Two women in group | | | -0.029 | | | -0.064 | | | -0.035 |
| | | | (0.690) | | | (0.442) | | | (0.521) |
| Female | | | | | | | -0.097** | -0.019 | 0.040 |
| | | | | | | | (0.028) | (0.708) | (0.529) |
| Female x T3 | | | | | | | | -0.154* | -0.182** |
| | | | | | | | | (0.077) | (0.044) |
| Constant | 0.947*** | 0.726*** | 0.814*** | 0.929*** | 0.468** | 0.165 | 0.981*** | 0.947*** | 0.487*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.037) | (0.623) | (0.000) | (0.000) | (0.005) |
| Fem.+Fem.xT3=0 | | | | | | | | 0.015** | 0.039** |
| Observations | 121 | 121 | 110 | 83 | 83 | 76 | 204 | 204 | 186 |

Estimates generated by linear probability models. Robust pvalues in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 4: Regression analysis of the decision to be manager

| | % of Efficient Rank Allocations | | |
|---|---|---|---|
| | Male Manager | Female Manager | $H_0$: M=W p-value |
| Baseline (T1) | 81.33 | 76.06 | 0.436 |
| Choice (T2) | 83.82 | 90.67 | 0.218 |
| Choice&Messages (T3) | 84.29 | 87.14 | 0.629 |
| $H_0$: T1=T2 (p-value) | 0.695 | 0.017** | |
| $H_0$: T1=T3 (p-value) | 0.638 | 0.090* | |
| $H_0$: T2=T3 (p-value) | 0.941 | 0.499 | |

We report the percentage of times the best performing worker was selected as Rank A by a male or a female manager. p-values are generated by Chi-square tests.

Table 5: Efficiency in rank allocations

| | Dep. Var: Manager allocates Rank A to best performer | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Male Managers | | | Female Managers | | | All | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Choice (T2) | 0.029 | 0.066 | 0.043 | 0.149* | 0.066 | 0.064 | 0.028 | 0.067 | 0.058 |
| | (0.727) | (0.273) | (0.347) | (0.062) | (0.274) | (0.281) | (0.731) | (0.254) | (0.292) |
| Choice&Mess. (T3) | 0.028 | 0.035 | 0.024 | 0.113 | 0.061 | 0.060 | 0.028 | 0.036 | 0.035 |
| | (0.746) | (0.524) | (0.619) | (0.151) | (0.304) | (0.332) | (0.742) | (0.518) | (0.496) |
| $(\mathrm{R}_A - \mathrm{R}_B)_{lag}$ | | 0.042*** | 0.043*** | | 0.041*** | 0.043*** | | 0.041*** | 0.043*** |
| | | (0.000) | (0.000) | | (0.000) | (0.000) | | (0.000) | (0.000) |
| $\mathrm{Perf}_{lag}$ | | -0.000 | -0.003 | | 0.012 | 0.021** | | 0.005 | 0.006 |
| | | (0.992) | (0.717) | | (0.105) | (0.023) | | (0.321) | (0.327) |
| Female | | | | | | | -0.053 | -0.011 | -0.012 |
| | | | | | | | (0.549) | (0.836) | (0.836) |
| Female x T2 | | | | | | | 0.122 | -0.005 | -0.008 |
| | | | | | | | (0.288) | (0.953) | (0.920) |
| Female x T3 | | | | | | | 0.085 | 0.025 | 0.019 |
| | | | | | | | (0.463) | (0.762) | (0.807) |
| Stage | -0.046** | -0.023* | -0.022* | -0.033** | -0.020* | -0.020* | -0.039** | -0.021** | -0.020** |
| | (0.024) | (0.066) | (0.083) | (0.025) | (0.080) | (0.087) | (0.002) | (0.014) | (0.021) |
| Constant | 0.997*** | 0.706*** | 1.035*** | 0.891*** | 0.455*** | 0.261 | 0.970*** | 0.602*** | 0.677** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.005) | (0.510) | (0.000) | (0.000) | (0.011) |
| T2=T3 | 0.993 | 0.649 | 0.732 | 0.545 | 0.931 | 0.941 | 0.999 | 0.119 | 0.424 |
| Fem+FemxT2=0 | | | | | | | 0.340 | 0.373 | 0.758 |
| Fem+FemxT3=0 | | | | | | | 0.669 | 0.371 | 0.406 |
| Observations | 213 | 213 | 209 | 216 | 216 | 205 | 429 | 429 | 414 |
| Controls | No | No | Yes | No | No | Yes | No | No | Yes |
| Clusters | 47 | 47 | 46 | 53 | 53 | 50 | 100 | 100 | 96 |
| $R^2$ | 0.032 | 0.523 | 0.601 | 0.047 | 0.413 | 0.447 | 0.039 | 0.468 | 0.486 |

Estimates generated by linear probability models. Standard errors clustered at the group level. $(\mathrm{R}_A\text{-}\mathrm{R}_B)_{lag}$ is the difference in the performances of Rank A and Rank B workers in the previous stage. $\mathrm{Perf}_{lag}$is the performance of the manager in the previous stage. Robust pval in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 6: The allocation of Rank A to the best performing worker

| | Dep. Var: Worker is assigned Rank A in current stage | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Rank $A_{lag}$ | -0.020 | 0.574** | 0.881*** |
| | (0.907) | (0.025) | (0.000) |
| Choice (T2) | 0.009 | 0.006 | -0.058 |
| | (0.745) | (0.804) | (0.742) |
| Choice & Messages (T3) | 0.013 | 0.009 | 0.237 |
| | (0.740) | (0.801) | (0.215) |
| Female worker | 0.047 | 0.033 | 0.054 |
| | (0.734) | (0.798) | (0.643) |
| Female Manager | 0.013 | 0.417*** | 0.418*** |
| | (0.742) | (0.008) | (0.005) |
| Female Manager x Rank $A_{lag}$ | | -0.817** | -1.022*** |
| | | (0.011) | (0.000) |
| Rank $A_{lag}$x T2 | | | 0.037 |
| | | | (0.848) |
| Rank $A_{lag}$x T3 | | | -0.765** |
| | | | (0.030) |
| Female Manager x Rank $A_{lag}$x T2 | | | 0.123 |
| | | | (0.626) |
| Female Manager x Rank $A_{lag}$x T3 | | | 0.425 |
| | | | (0.228) |
| Stage | 0.000 | 0.000 | 0.007 |
| | (0.820) | (0.839) | (0.381) |
| Constant | 0.480*** | 0.192 | 0.081 |
| | (0.000) | (0.101) | (0.564) |
| Fem.Manager + Fem.Manager x Rank$A_{lag}$ =0 | | 0.020** | 0.001*** |
| Fem.Manager + Fem.Manager x Rank$A_{lag}$+Fem.Manager x Rank$A_{lag}$ x T2=0 | | | 0.011** |
| Fem.Manager + Fem.Manager x Rank$A_{lag}$+Fem.Manager x Rank$A_{lag}$ x T3=0 | | | 0.538 |
| Observations | 138 | 138 | 138 |
| Clusters | 41 | 41 | 41 |
| $R^2$ | 0.002 | 0.135 | 0.214 |

Estimates generated by linear probability models. Sample restricted to cases where the two workers performed equally. Standard errors clustered at the group level. Robust pval in parentheses.*** p<0.01, ** p<0.05, * p<0.1.

Table 7: Worker's likelihood of being assigned Rank A under equal performance

| | Male Manager | Female Manager | $H_0$: M=W p-value |
|---|---|---|---|
| Received at least 1 angry emoticon (% of times) | 43.75 | 46.32 | 0.734 |
| Number of angry emoticons received (conditional on receiving at least 1) | 2.86 | 3.41 | 0.156 (one-sided: 0.078) |

P-values are generated by Chi-square test (row 1) or a double-sided test of equality of means (row 2).

Table 8: The receipt of angry emoticons

| | Worker Sent at least 1 angry emoticon | | | | Number of angry emoticons sent | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| To Female Manager | 0.077 | 0.011 | 0.025 | 0.007 | 0.710 | 1.120* | 1.123** | 1.384* |
| | (0.472) | (0.929) | (0.862) | (0.961) | (0.151) | (0.066) | (0.037) | (0.074) |
| Female | 0.281* | 0.149 | 0.162 | 0.112 | 0.753 | 1.421* | 1.675** | 1.802*** |
| | (0.063) | (0.461) | (0.473) | (0.638) | (0.163) | (0.082) | (0.017) | (0.009) |
| Female x Fem.Manager | | 0.360 | 0.394 | 0.403 | | -1.384 | -0.524 | -1.058 |
| | | (0.150) | (0.179) | (0.201) | | (0.168) | (0.562) | (0.267) |
| $\text{Performance}_{lag}$ | | | 0.015 | 0.013 | | | 0.105*** | 0.079 |
| | | | (0.140) | (0.222) | | | (0.009) | (0.138) |
| Rank $A_{lag}$ | | | -0.127 | -0.106 | | | -0.699 | -0.740 |
| | | | (0.223) | (0.294) | | | (0.180) | (0.135) |
| Round | 0.009 | 0.012 | 0.002 | -0.001 | 0.313** | 0.303** | 0.219 | 0.207 |
| | (0.707) | (0.642) | (0.941) | (0.986) | (0.020) | (0.033) | (0.266) | (0.285) |
| Constant | 0.327*** | 0.358*** | 0.217 | 0.500 | 1.602*** | 1.386*** | -0.000 | -1.676 |
| | (0.002) | (0.001) | (0.311) | (0.245) | (0.003) | (0.010) | (1.000) | (0.617) |
| Observations | 175 | 175 | 140 | 140 | 79 | 79 | 64 | 64 |
| Controls | No | No | No | Yes | No | No | No | Yes |
| Clusters | 35 | 35 | 35 | 35 | 29 | 29 | 27 | 27 |
| R-squared | 0.049 | 0.067 | 0.088 | 0.110 | 0.117 | 0.151 | 0.255 | 0.369 |

OLS regressions. Standard errros clustered at the group level. Robust pval in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 9: Regression analysis of worker backlash

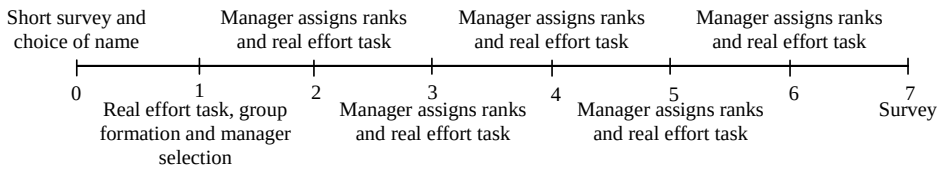



Figure 1: Stages of the experiment

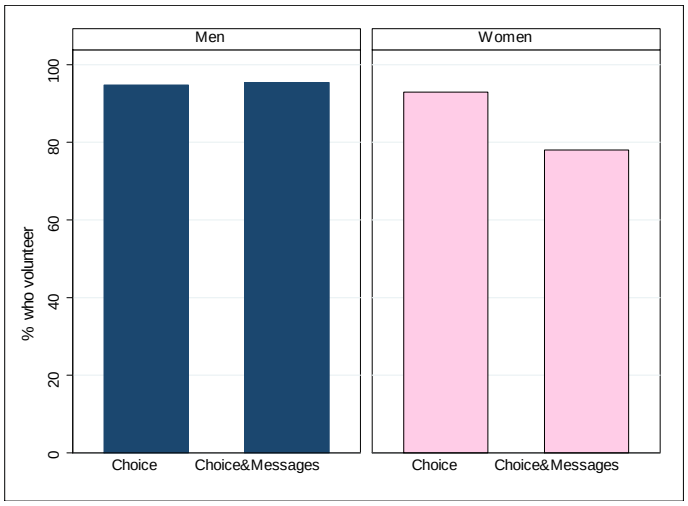



Figure 2: The decision to be manager by treatment and gender

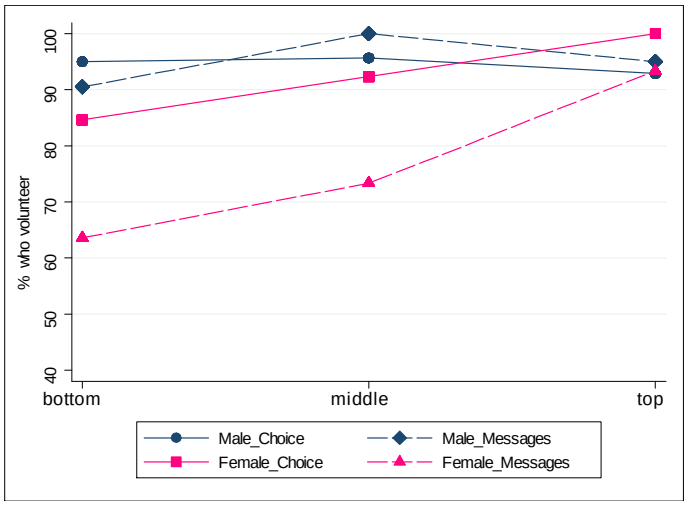



Figure 3: The decision to be a manager by performance tercile

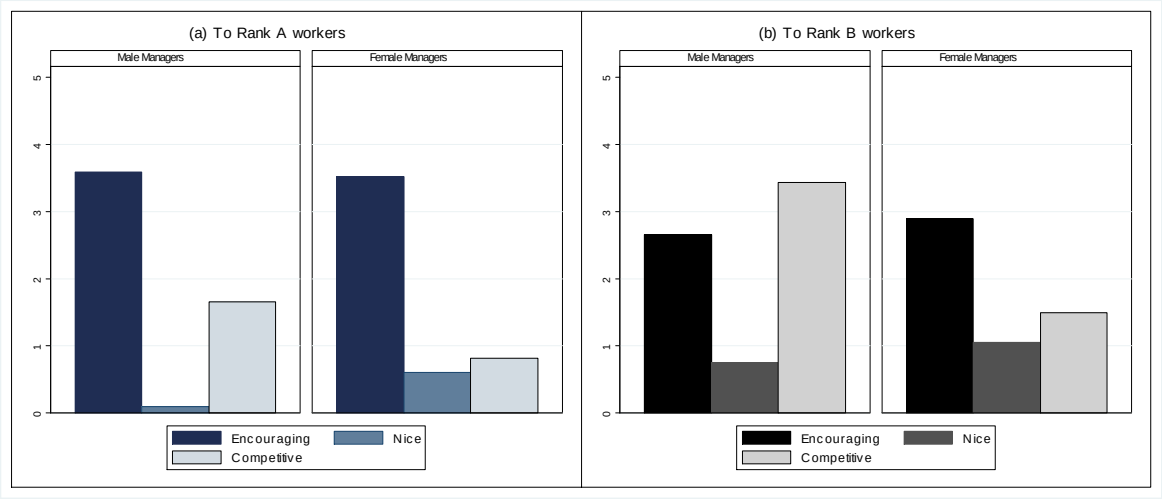



Figure 4: Average number of messages of each kind sent to Rank A and Rank B workers.

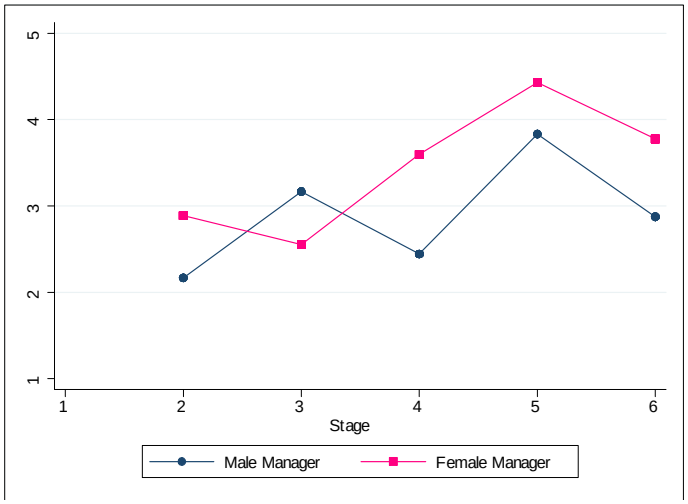



Figure 5: Average number of angry emoticons received by managers

# APPENDIX TABLES AND FIGURES

| | Male | | | | Female | | | | M=F (All) |
|---|---|---|---|---|---|---|---|---|---|
| | T1 | T2 | T3 | All | T1 | T2 | T3 | All | p-value |
| Age | 22.61 | 22.84 | 22.55 | 22.66 | 23.32 | 22.19 | 22.00 | 22.50 | 0.712 |
| | (2.49) | (2.87) | (4.63) | (3.48) | (4.51) | (3.09) | (4.27) | (4.01) | |
| STEM-Bus-Econ (frequency) | 0.85 | 0.91 | 0.80 | 0.85 | 0.71 | 0.67 | 0.66 | 0.68 | 0.000*** |
| | (0.36) | (0.29 | (0.41) | (0.36) | (0.46) | (0.48) | (0.48) | (0.47) | |
| Native speaker (frequency) | 0.34 | 0.26 | 0.47 | 0.36 | 0.41 | 0.43 | 0.56 | 0.47 | 0.066 |
| | (0.48) | (0.44) | (0.50) | (0.48) | (0.50) | (0.50) | (0.50) | (0.50) | |
| Past Leadership (frequency) | 0.80 | 0.85 | 0.83 | 0.82 | 0.90 | 0.83 | 0.83 | 0.85 | 0.498 |
| | (0.40) | (0.36) | (0.38) | (0.38) | (0.30) | (0.38) | (0.38) | (0.35) | |
| Big 5 Agreeableness Index | 27.80 | 26.67 | 27.70 | 27.44 | 29.44 | 28.34 | 29.95 | 29.29 | 0.007*** |
| | (5.05) | (5.53) | (5.19) | (5.24) | (7.26) | (6.78) | (5.21) | (6.44) | |

Table 10: Subjects' characteristics

| | Dep. Variable: Number of correctly solved puzzled | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Female | 0.569 | 0.578 | 0.059 | -0.064 | -0.732 |
| | (0.418) | (0.411) | (0.963) | (0.882) | (0.306) |
| Choice (T2) | | 0.112 | -0.332 | -0.133 | -0.399 |
| | | (0.894) | (0.755) | (0.798) | (0.561) |
| Choice&Messages (T3) | | 0.584 | 0.386 | 0.410 | -0.118 |
| | | (0.488) | (0.717) | (0.419) | (0.865) |
| Female x T2 | | | 1.074 | | 0.664 |
| | | | (0.539) | | (0.524) |
| Female x T3 | | | 0.492 | | 1.332 |
| | | | (0.778) | | (0.180) |
| Manager | | | | 3.891*** | 3.882*** |
| | | | | (0.000) | (0.000) |
| Stage | | | | 0.380*** | 0.380*** |
| | | | | (0.000) | (0.000) |
| Constant | 13.802*** | 13.562*** | 13.770*** | 13.528*** | 13.800*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Observations | 306 | 306 | 306 | 1,836 | 1,836 |
| $R^2$ | 0.002 | 0.004 | 0.005 | 0.158 | 0.161 |

OLS regressions. Robust pvalues in parentheses. In Columns 1 to 3, we restrict the analysis to Stage 1. In Columns 4 and 5 we include all stages and we cluster the standard errors at the individual level.
*** p<0.01, ** p<0.05, * p<0.1.

Table 11: Performance in the real effort task

| | Dep. Variable: Dummy equal to 1 if subject wants to be manager | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Men | | | Women | | | All | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Choice&Messages (T3) | 0.056 | 0.270 | 0.891 | -0.691* | -1.018** | -1.175** | -0.351 | 0.056 | 0.840* |
| | (0.885) | (0.435) | (0.121) | (0.059) | (0.037) | (0.036) | (0.189) | (0.885) | (0.099) |
| Performance | | 0.067* | 0.090** | | 0.184*** | 0.187*** | | | 0.125*** |
| | | (0.076) | (0.026) | | (0.001) | (0.001) | | | (0.000) |
| Age | | 0.043 | 0.046 | | 0.027 | 0.059 | | | 0.045 |
| | | (0.213) | (0.197) | | (0.701) | (0.461) | | | (0.241) |
| STEM-bus-eco | | 0.776 | 0.945 | | 1.024** | 1.320** | | | 1.165*** |
| | | (0.116) | (0.142) | | (0.028) | (0.015) | | | (0.003) |
| Native | | -0.285 | -0.705 | | -1.218* | -1.079 | | | -0.691* |
| | | (0.542) | (0.236) | | (0.092) | (0.158) | | | (0.099) |
| Leadership | | | 0.937 | | | 0.772 | | | 0.750* |
| | | | (0.148) | | | (0.105) | | | (0.061) |
| big5-agree | | | -0.097*** | | | 0.021 | | | -0.025 |
| | | | (0.008) | | | (0.613) | | | (0.386) |
| Two women in group | | | -0.520 | | | -0.276 | | | -0.310 |
| | | | (0.379) | | | (0.530) | | | (0.348) |
| Female | | | | | | | -0.592** | -0.155 | 0.619 |
| | | | | | | | (0.024) | (0.700) | (0.254) |
| Female x T3 | | | | | | | | -0.747 | -1.692** |
| | | | | | | | | (0.160) | (0.013) |
| Constant | 1.620*** | -0.753 | 0.795 | 1.465*** | -1.116 | -3.123 | 1.849*** | 1.620*** | -1.704 |
| | (0.000) | (0.490) | (0.580) | (0.000) | (0.516) | (0.169) | (0.000) | (0.000) | (0.225) |
| Fem.+Fem.xT3=0 | | | | | | | | 0.009*** | 0.021** |
| Obs. | 121 | 121 | 110 | 83 | 83 | 76 | 204 | 204 | 186 |

Probit regressions. Robust pval in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 12: The decision to be manager: Probit regressions

| | Men | Women | $H_0$: M=W p-value |
|---|---|---|---|
| Age | 21.48 | 21.16 | 0.640 |
| Stem-bus-econ major | .74 | .63 | 0.291 |
| Native speaker | .55 | .65 | 0.400 |
| Past leadership role | .87 | .94 | 0.307 |
| Big 5 - Agreeableness | 26.9 | 29.88 | 0.067 |

P-values are generated by double-sided tests of equality of means for Age and Big 5 AI. p-values from single sided tests in parentheses. For the remaining variables, we conducted Chi-square tests.

Table 13: The characteristics of self-selected managers

| | Dep. Var: Rank B worker is promoted to Rank A in T3 | | | |
|---|---|---|---|---|
| | Male Manager | Female Manager | All | All |
| | (1) | (2) | (3) | (4) |
| $(R_B\text{-}R_A)_{lag}$ | 0.035*** | 0.047*** | 0.041*** | 0.041*** |
| | (0.005) | (0.000) | (0.000) | (0.000) |
| female worker | 0.286 | 0.052 | 0.198 | 0.259 |
| | (0.166) | (0.764) | (0.145) | (0.199) |
| $(\text{Angry emoticons})_{lag}$ | -0.037 | -0.060 | -0.052* | -0.044 |
| | (0.245) | (0.145) | (0.055) | (0.187) |
| Female Manager | | | 0.281*** | 0.280*** |
| | | | (0.001) | (0.001) |
| Female Manager x $(\text{Angry emoticons})_{lag}$ | | | | -0.029 |
| | | | | (0.611) |
| Stage | -0.051 | -0.002 | -0.026 | -0.026 |
| | (0.308) | (0.973) | (0.421) | (0.429) |
| Constant | 0.598** | 0.775*** | 0.553*** | 0.543*** |
| | (0.032) | (0.001) | (0.002) | (0.002) |
| Observations | 64 | 76 | 140 | 140 |
| Clusters | 16 | 19 | 35 | 35 |
| $R^2$ | 0.307 | 0.299 | 0.321 | 0.323 |

Estimates generated by linear probability models. Standard errors are clustered at the group level.

Robust pval in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 14: Rank B worker's likelihood of being promoted conditonal on the angry emoticons sent

| | Avg. number of messages of each type from Rank A workers over the 5 stages | | | | | |
|---|---|---|---|---|---|---|
| | Commitment to work hard | Jokes | Justification of performance | Rank request | Thankful | Challenging |
| Male Manager | 2.187 | .906 | .125 | .094 | 2.344 | .656 |
| Female Manager | 2.132 | .974 | .316 | .210 | 2.763 | .868 |
| $H_0$: M=F ttest p-value | 0.883 | 0.949 | 0.172 | 0.345 | 0.423 | 0.514 |
| [Wilcoxon rank-sum p-value] | 0.893 | 0.986 | 0.103 | 0.319 | 0.385 | 0.600 |

Table 15: Messages received from Rank A workers

| | Avg. number of messages of each type from Rank B workers over the 5 stages | | | | | |
|---|---|---|---|---|---|---|
| | Commitment to work hard | Jokes | Justification of performance | Rank request | Thankful | Challenging |
| Male Manager | 1.031 | .812 | .281 | .562 | .312 | 2.094 |
| Female Manager | .868 | .605 | .553 | .500 | .553 | 2.684 |
| $H_0$: M=F ttest p-value | 0.666 | 0.531 | 0.154 | 0.799 | 0.350 | 0.292 |
| [Wilcoxon rank-sum p-value] | 0.797 | 0.231 | 0.131 | 0.384 | 0.583 | 0.161 |

Table 16: Messages received from Rank B workers

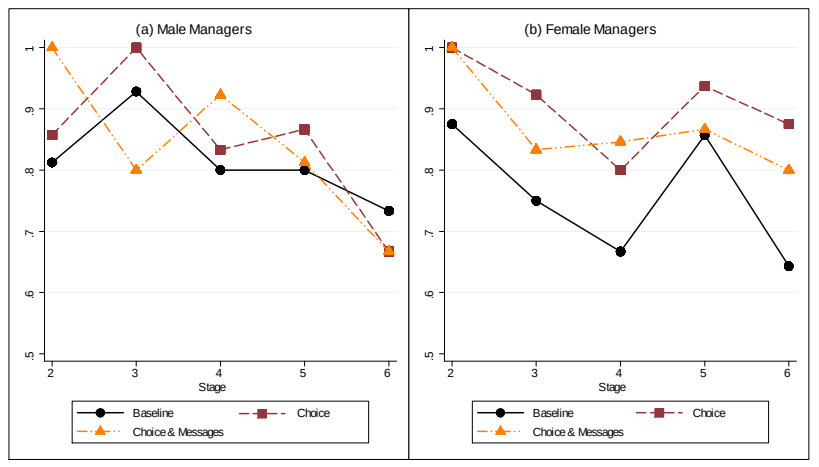



Figure 6: Percentage of efficient rank allocations over time

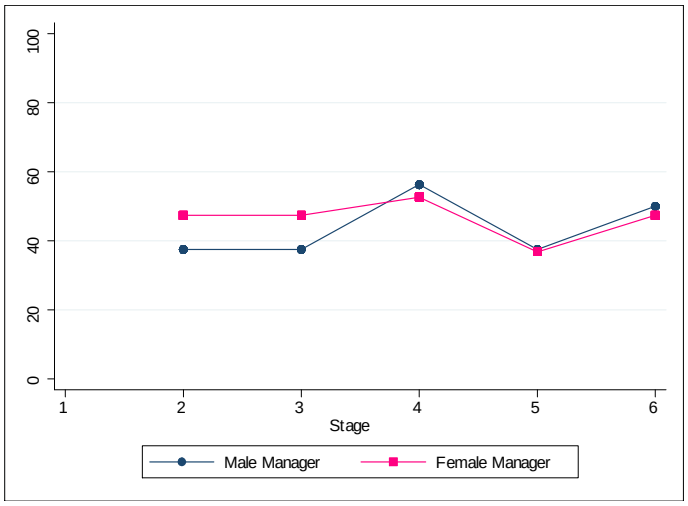



Figure 7: Percentage of managers receiving at least one angry emoticon

# ONLINE APPENDIX

**EXPERIMENTAL INSTRUCTIONS:** *Choice & Messages* treatment

**General instructions**

Thank you all for coming today. You are here to participate in an experiment. In addition to a $10 participation fee, you will be paid any money you accumulate from the experiment. You will be paid privately, by check, at the conclusion of the experiment.

The experiment will consist of six stages and the instructions will be provided separately on your screen at the beginning of each stage. You will have the chance to earn money in each stage of the experiment. Earnings during the experiment will be denominated in Experimental Currency Units, or ECU. At the end of the session one stage of the experiment will be randomly selected for payment and your earnings in that stage will be converted to dollars at the exchange rate of $1 for 6 ECU. After participating in all the stages of the experiment you will be asked to complete a brief questionnaire. You will then be paid the money your earned in the selected stage of experiment.

This study has been reviewed and approved by the SMU Human Subjects Committee. If you have any questions during the experiment, please raise your hand and wait for an experimenter to come to you. Please do not talk, exclaim, or try to communicate with other participants during the experiment. Participants intentionally violating these rules may be asked to leave the experiment and may not be paid.

Please read and sign the Consent Form that you found on your desk. Please raise your hand if you have any question about any of the information on the Consent form. We will proceed with the experiment once we have collected all signed consent forms.

[Collect consent forms. Start program. When everybody is on Screen 3, distribute Puzzle Example]

**PUZZLE EXAMPLE** (*Handout 1*)

During the experiment, you will engage in multiple rounds of a puzzle-solving task. Please refer to the paper you have been given to see an example of the task. Each task consists of finding a 4-letter word in a 6x6 matrix. As you can see on the example you have been given, the screen will be divided in two halves. On the left, you will see the matrix and on the right, you will see a list of 40 words. Each puzzle has two words that appear on the list. In order to earn money, you will have to identify one word per puzzle. Once you identify the word, you will have to enter the number next to that word in the list. You will then have to press "submit" to move to the next puzzle.

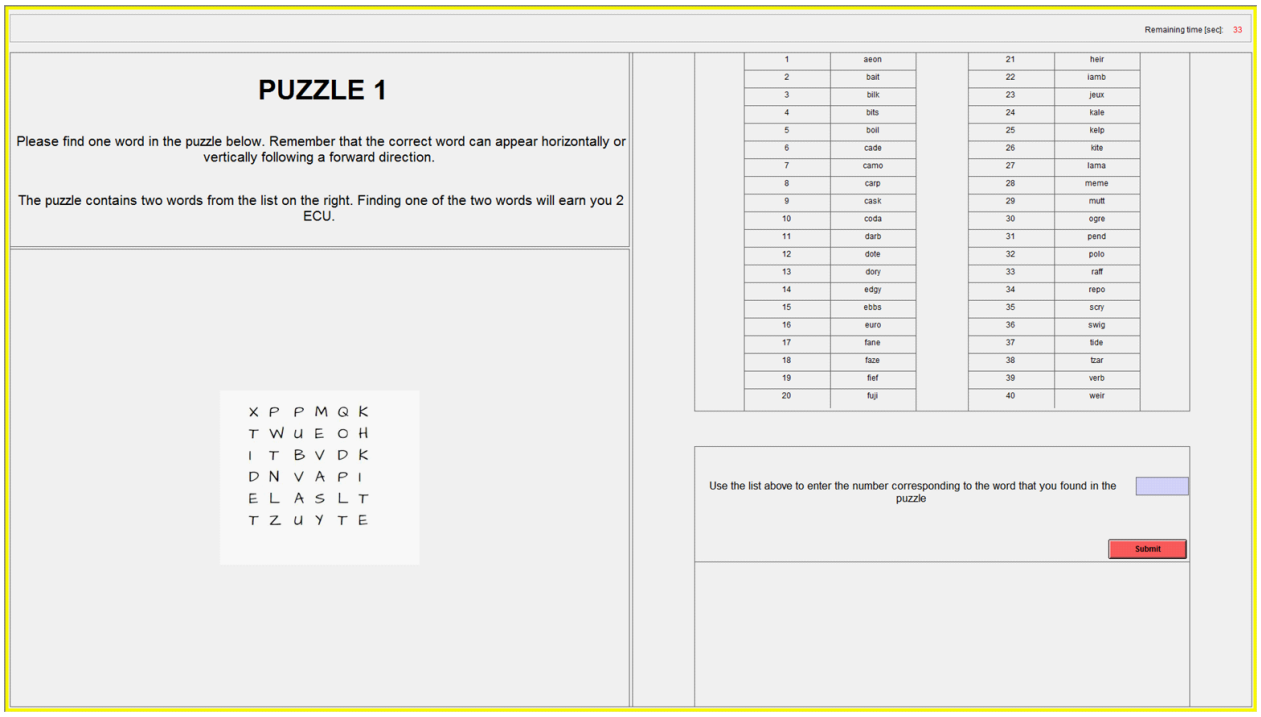
Please note that the word you are looking for can appear horizontally or vertically in the matrix, following a forward direction. You should ignore words that are read backward or diagonally. You should also ignore words that do not appear in the list.

Look at the example you have been given. In order to earn points, you would have to find either the word "tide" or the word "kite" and enter the corresponding number. The word "tide" appears vertically on the first column. The word "kite" appears vertically on the sixth column.

Can you all see the two words in the puzzle? Raise your hand if you cannot see them.

Note that there are other words that you may identify in the matrix. For instance, the word "sale" [appears horizontally on the fifth row, but it reads backward] and the word "bale" [appears diagonally]. These words would not be valid entries, since they either appear backward or are not on the list. Remember that for an entry to be valid, it MUST be on the list to the right of the matrix.

Can you all see the two words in the puzzle? Raise your hand if you cannot.




Do you have any questions or doubts about the puzzle-solving task?

**Instructions for Stages 2 to 6** *(Handout 2)*
Stage 2 of the experiment is about to begin.

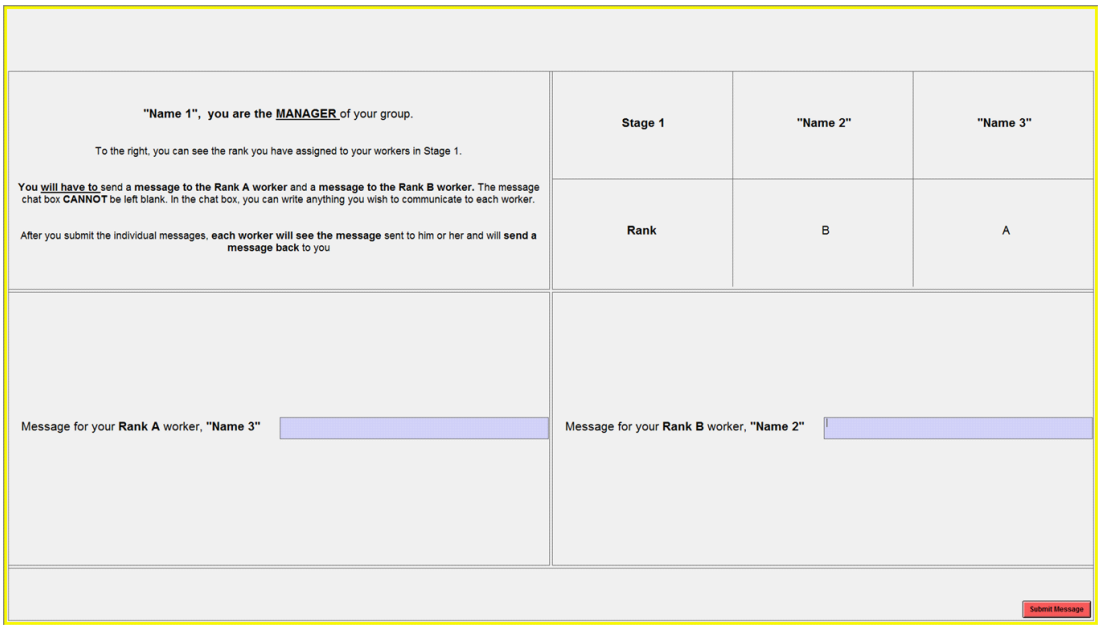
- In this stage and in the following 5 stages of the experiment you will be part of a group, together with two other participants. One group member will assume the role of Manager and the other two group members will assume the role of Worker.
- The manager gets a wage of 100 ECU. The main role of the manager is to decide which worker will be rank A and which worker will be rank B in the group, in this stage of the experiment.
- A Rank A worker gets a wage of 80 ECU. A Rank B worker gets 20 ECU.
- After the rank allocation, all members of the group will engage in a similar puzzle task as in Stage 1 of the experiment. Each correctly solved puzzle generates 2 ECU in addition to the initial wage. Moreover, each puzzle solved correctly by the Rank A worker generated 2 ECU also to the manager.
- Therefore the earnings from this stage of the experiment are determined as follows:
  - The Manager gets 100 ECU + 2 ECU per puzzle + 2 ECU per puzzle solved by Rank A worker
  - Rank A worker gets 80 ECU plus 2 ECU per puzzle
  - Rank B worker gets 20 ECU plus 2 ECU per puzzle
- Stages 3 to 6 will be identical to Stage 2. However, at the end of each stage of the experiment, the Manager will be informed about the performance of Rank A and Rank B workers and will have the chance to reassign ranks before the next stage begins, at his or her discretion.

Please turn this page around to have a look at the screen that the Manager will see when making the rank allocation decision.

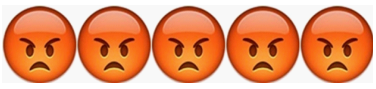
The Manager will have to decide which worker will be Rank A and which worker will be Rank B. After the Manager makes the allocation decision, each worker will be informed about the Rank they have been assigned, either rank A or rank B.

After assigning ranks A and B to the workers, the Manager will have to send a message to the Rank A worker and a message to the Rank B worker. The message chat box CANNOT be left blank. In the chat box, the Manager can write anything he or she wishes to communicate to each worker. After the Manager submits the individual messages, each worker will see the message sent to him or her and will have to send a message back to the Manager.

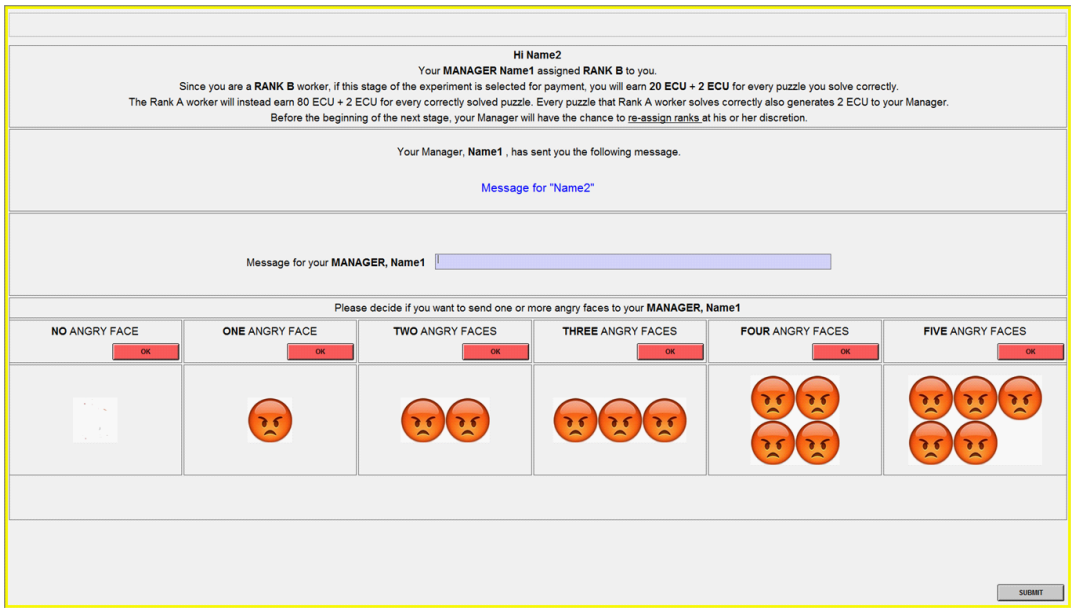
The Manager will see the following screen and will have to send a message to each of the two workers. As before, we are referring to the Manager as "Name 1" and the workers as "Name 2" and "Name 3" but in the actual experiment the names of the three group members will be displayed.



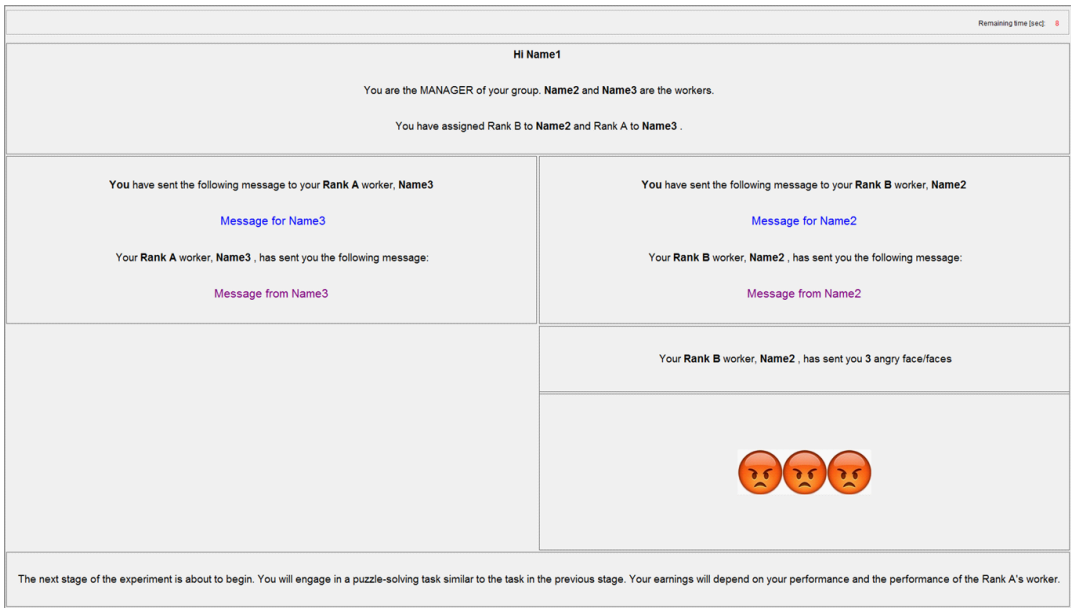

Each worker will see the message sent to him or her and will have to send a message back to the Manager. In addition, the Rank B worker can send one or more angry faces to the Manager to express their disapproval of the ranking decision. In particular, the Rank B's worker can send up to 5 angry faces to the Manager, as shown below.



Please turn this page to see the screen that Worker Bs and the Manager will see. Rank B worker will see the following screen and will have to decide how many angry faces, if any, he or she will want to send to the Manager. Rank B worker will also have to write a message in the chat box, in response to the Manager's message. Rank A worker will see a similar screen, except that he or she will not be able to send angry faces to the Manager.




The messages sent by the workers will be displayed to the Manager as shown in the screen below before the next stage begins. [Please note that in this example there is no actual text displayed in the Message Box.] In the example below, Rank B worker has sent 3 angry faces. Remember that Rank B can send between 0 and 5 angry faces.




At the end of each of the next 5 stages of the experiment, the Manager will have to decide whether to keep or re-allocate ranks A and B to the two workers, at his or her discretion. The Manager will also have to send messages to Rank A and Rank B worker before the beginning of each stage of the experiment, and the workers will have to reply to those messages. At the beginning of each stage, the Rank B workers will also have to decide whether to send angry faces to the Manager.

Is the role of the Manager clear? Please raise your hand if you have any questions about the next 5 Stages of the experiment.

In the next screen, you will be asked **whether you want to be the Manager of your group.**