School of Economics
Working Paper 2021-09

Gender Biases in Performance Evaluation: The Role of Beliefs versus Outcomes

Nisvan Erkal*
Lata Gangadharan**
Boon Han Koh***

*University of Melbourne
** Monash University
***University of East Anglia

University of East Anglia

SCHOOL OF ECONOMICS

School of Economics
University of East Anglia
Norwich Research Park
Norwich NR4 7TJ
United Kingdom
www.uea.ac.uk/economics

Gender Biases in Performance Evaluation: The Role of Beliefs versus Outcomes[*]

Nisvan Erkal[†]
Lata Gangadharan[‡]
Boon Han Koh[§]

Abstract

We investigate whether gender distorts performance evaluation in environments where outcomes are determined by leaders' unobservable effort choices and luck. Evaluators form beliefs about effort choices and make discretionary payment decisions. We find that while the discretionary payments made to male leaders are determined by both outcomes and evaluators' beliefs, those made to female leaders are determined by outcomes only. Hence, beliefs are a source of gender biases in our decision-making environment not because they are biased, but because they play differential roles in female and male leaders' discretionary payments. We label this new source of gender bias as the *gender belief-outcome gap*. These findings further our understanding of the factors driving gender gaps in leadership and performance pay. They imply that in the labor market, good outcomes are necessary for women to get bonuses, but men can receive bonuses for bad outcomes as long as evaluators hold them in high regard.

Keywords: Gender gaps; Performance evaluation; Biases in belief updating; Outcome bias; Social preferences; Laboratory experiments
JEL Classification: C92, D91, J71

[*] We would like to thank Billur Aksoy, Kai Barron, Pol Campos-Mercade, Christine Exley, Philip Grossman, Elif Incekara-Hafalir, Yoshihisa Kashima, Andreas Leibbrandt, Friederike Mengel, Lionel Page, Amrish Patel, Eva Ranehill, Danila Serra, Olga Stoddard, Joseph Vecci, Nina Xue, Xiaojian Zhao, seminar participants at the Centre for Behavioural and Experimental Social Science (CBESS), Institute of Economic Growth, New Zealand Economics eSeminar Series, University of Hawaii, Vienna University of Economics and Business, and participants at the Workshop on Beliefs, Narratives, and Memory (2021), 11th International Conference of the French Association of Experimental Economics (2021), 2021 Subjective Probability, Utility, and Decision Making (SPUDM) Conference, 2021 Australian Gender Economics Workshop, 2020 Behavioral Research in Economics Workshop (BREW), 2020 Annual Meeting of the American Economic Association, 2019 Sydney Workshop on Experimental Economics and Theory, 2019 North American Economic Science Association Meetings, 2019 Lisbon Meetings in Game Theory and Applications, and 14th Annual Australia New Zealand Workshop in Experimental Economics (2019) for their comments and feedback. We gratefully acknowledge funding from the Australian Research Council (DP210102183), the Department of Economics at the University of Melbourne, and the Department of Economics at Monash University. This study has been approved by the Business and Economics Human Ethics Advisory Group at the University of Melbourne (ID 1544873).

[†] Corresponding author. Department of Economics, University of Melbourne, VIC 3010, Australia. n.erkal@unimelb.edu.au.

[‡] Department of Economics, Monash University, VIC 3800, Australia. lata.gangadharan@unimelb.edu.au.

[§] School of Economics and Centre for Behavioural and Experimental Social Science, University of East Anglia, NR4 7TJ, United Kingdom. b.koh@uea.ac.uk.

1 Introduction

Gender gaps in labor market outcomes, such as wages, lifetime earnings, and leadership roles, are well documented (see, e.g., Bertrand and Duflo, 2017; Blau and Kahn, 2017; Eckel et al., 2021). A prominent explanation for these gaps is gender biases in performance evaluation (see, e.g., Goldin and Rouse, 2000; Jensen, Kovacs, and Sorenson, 2018; Grossman et al., 2019; Mengel, Sauermann, and Zölitz, 2019; Régner et al., 2019; Sarsons, 2019; Sarsons et al., 2021; Egan, Matvos, and Seru, forthcoming).[1] Do the observed differences in performance evaluation come from biased beliefs, or are there other forces at play? Understanding the sources of gender biases is important for designing relevant policies to overcome gender discrimination. In this paper, we propose a new channel through which gender biases can arise in performance evaluation. Focusing on the evaluation of leaders, we investigate gender biases in the emphasis evaluators place on the outcomes of leaders versus their beliefs about the leaders' decisions.

We are especially interested in performance evaluation in environments where outcomes are determined by a combination of unobservable actions and luck. Such environments are pervasive in many organizations, and evaluators face the challenging task of assessing performance based on the merits of the actions taken by the leader without being influenced by the outcome of those actions. Outcomes, which are observable, act as signals and can assist evaluators in updating their beliefs about the leader's actions. In the updating process, evaluators have to assess what role unpredictable or unforeseeable circumstances (versus actions) have played in determining the outcomes.

In this process, biases in judgements can influence the decisions taken by evaluators. One such decision is discretionary payments made to leaders. Discretionary rewards, such as bonuses or pay increments, are common features of remuneration packages offered by many organizations. Biases in discretionary payments can emerge, for example, through the beliefs formed about actions taken (i.e., attribution biases) or through the weights given to the informational content of outcomes (i.e., an outcome bias).[2] Such biases cause distortions in the incentive structures provided and their extent can be influenced by subjective and unrelated measures, such as gender and race. For example, leaders may be rewarded excessively for good

[1] Other explanations revolve around gender differences in preferences (see, e.g., Croson and Gneezy, 2009), and institutional factors (see, e.g., Hernandez-Arenaz and Iriberri, 2019; Erkal, Gangadharan, and Xiao, 2021).

[2] According to the informativeness principle in contract theory, signals should affect incentives as long as they are informative about the unobserved actions taken (Bolton and Dewatripont, 2005). A deviation from this principle occurs when an evaluator overweighs a signal relative to its informational content. Such a deviation is known as an outcome bias (Baron and Hershey, 1988).

outcomes (i.e., for having good luck) or penalized disproportionately for bad outcomes (i.e., for having bad luck), and these distortions may vary by the leaders' gender.

Studies which use observational data to analyze gender biases in performance evaluation play an important role in exposing whether the outcomes of men and women are treated differently. As a next step, it is important to establish the drivers of these differences for developing effective strategies against them. Our paper has two research goals. The first research goal is to study whether there are gender biases in the beliefs formed about the actions taken by leaders. The second research goal is to investigate whether there are gender biases in the emphasis given to beliefs versus outcomes in the determination of discretionary payments. Hence, in addition to studying the role of beliefs in gender discrimination (e.g., Bohren, Imas, and Rosenberg, 2019; Coffman, Exley, and Niederle, 2021), we investigate a new source of gender bias in terms of the role beliefs and outcomes play in the performance evaluation process.

Our research strategy relies on laboratory experiments to draw causal links between the gender of the leader, the beliefs of the evaluators, and the discretionary payments made by the evaluators. Using observational data, it is difficult to discern whether any observed gender biases in performance evaluation are due to differences in beliefs about the decision-makers' actions or other factors (such as an outcome bias). This is because the information evaluators have about leaders is not available to researchers. Moreover, the informativeness of the signals observed by evaluators are not known. Experimental methods provide us with more precise and reliable measures of key variables such as beliefs and how they vary with gender. Further, the exogenous allocation of leader roles enables collection of data from female leaders, which can be difficult to achieve using observational data due to the limited number of women in leadership roles.

We consider a setting for decision making in which male and female leaders undertake actions that would affect their own earnings and that of the evaluators. The decisions that leaders take, while costly for them, lead to higher earnings for the evaluators. Leadership, in our study, is hence defined as making decisions for others which inherently involves assuming responsibility for the outcomes of others (Ertac and Gurdal, 2012; Edelson et al., 2018). This is a meaningful paradigm to study as pro-sociality is considered to be a critical characteristic of leaders (Bénabou and Tirole, 2010). Environments where social preferences of leaders and the unobservability of their actions co-exist are common to many workplaces. Consider for example, political leaders who are expected to engage in prosocial activities that impact the welfare of voters who then evaluate their actions, or CEOs whose actions impact the payoffs or reputation of board members who then decide on their compensation.

In the experiment, individuals are divided into groups of three and make an investment choice on behalf of the group. One of them is assigned to be the leader of the group and their investment decisions are implemented, while the other two are appointed as evaluators. The leader's gender is revealed to the group. The group's outcome depends on both the leader's choice, which is unobservable to the evaluators, and luck. A high investment leads to a higher probability of a good outcome for the group, but it comes at a higher private cost to the leader. The group members who evaluate the leader form initial beliefs about the leader's investment choice and update their beliefs after observing the outcome of the leader's decision. Then, they make discretionary payments (which may be positive or negative) to the leaders. Hence, evaluators' beliefs represent their assessment of the leader's intentions. In the theoretical framework we consider, leaders can be rewarded for both their intentions and outcomes (Falk and Fischbacher, 2006).

We find no evidence of gender differences in leaders' effort choices and evaluators' prior beliefs about their choices. Moreover, evaluators attribute male and female leaders' good and bad outcomes similarly. Hence, we do not observe any gender differences in evaluators' beliefs about the leaders' choices. These findings imply that if evaluators were to base their discretionary payment decisions on their beliefs, we should not observe any gender differences in their decisions. Yet, strikingly, we do see gender differences emerging in the discretionary payment decisions. Given a bonus, female leaders receive lower bonuses on average for good outcomes.

Investigating the determinants of discretionary payments, we find that while male leaders' discretionary payments are determined by both the evaluators' assessments of their effort choices and outcomes, female leaders' discretionary payments are predominantly determined by outcomes. Hence, we detect a gender difference in the determinants of discretionary payments. This difference is driven by a *gender belief-outcome gap*. That is, evaluators' beliefs about leaders' choices (i.e., intentions) seem to play a much smaller part in shaping the payments made to female leaders as compared to the payments made to male leaders. Outcomes, however, have a similar influence in the determination of discretionary payments made to male and female leaders.

Our results thus offer a different explanation of the role played by beliefs in creating gender biases in performance evaluation. Beliefs are important in our set-up not because they are biased, but because they play differential roles in the determination of male and female leaders' discretionary payments. The findings from our research have implications for how policies could be designed to counteract gender discrimination relating to performance evaluation. As we discuss in Section 7, organizations could make performance evaluation processes less ambiguous or consider group evaluations to reduce the likelihood of gender biases appearing.

2 Related Literature

Gender discrimination, both the documentation of its evidence and the exploration of its causes, has been a topic of significant research. An important theme in this literature relates to stereotypes and belief formation.[3] Bohren, Imas, and Rosenberg (2019) and Coffman, Exley, and Niederle (2021) distinguish between belief-based and preference-based gender discrimination, and find that discrimination against women tends to be the former rather than the latter. Campos-Mercade and Mengel (2021) consider how irrational updating of beliefs, in the sense of conservatism, can result in gender discrimination. Barron et al. (2020) distinguish between explicit and implicit belief-based discrimination, and find evidence of both. Fenske, Castagnetti, and Sharma (2020) study a principal-agent environment and do not find evidence consistent with attribution biases by gender. Our study complements this emerging literature on the role of beliefs in gender discrimination. In contrast to this literature, our aim is to understand whether there are gender biases in the emphasis given to beliefs versus outcomes in performance evaluation. We find evidence of a gender belief-outcome gap, which can contribute to the persistence of gender gaps in labor markets.

With its emphasis on belief formation, our study is also closely related to the growing literature which examines how individuals update their beliefs in response to feedback, absent gender considerations. A majority of the research in economics on beliefs examines ego-related beliefs.[4] The corresponding literature in psychology has a longer history and examines attribution biases about others in addition to self (e.g., Miller and Ross, 1975). Relatedly, research in psychology finds gender differences in the attribution of outcomes of others to effort, skill, and luck (e.g., Swim and Sanna, 1996), where men are considered to be smart or able in the face of good outcomes, while women are thought to be just lucky. Our study

[3] Coffman (2014), Bordalo et al. (2019), and Coffman, Collis, and Kulkarni (2020) focus on the role of gender stereotypes in driving individuals' prior and updated beliefs about their *own* ability.

[4] Researchers have, for example, investigated whether individuals update their beliefs about their own ability according to the Bayesian benchmark, and whether there is an asymmetry in the response to positive and negative noisy signals about performance (e.g., Eil and Rao, 2011; Ertac, 2011; Grossman and Owens, 2012; Möbius et al., 2014; Benjamin, 2019; Coutts, 2019).

considers an important yet less explored environment in which beliefs are formed about actions shaped by social preferences. Male and female leaders are evaluated about actions in which they face a trade-off between maximizing their own earnings and those of other individuals in their group.[5]

Finally, our study is related to the literature on outcome bias in the compensation of decision makers. Using observational data, Bertrand and Mullainathan (2001), Wolfers (2007), and Gauriot and Page (2019) show that agents are rewarded and penalized for factors beyond their control, such as luck. Research using experimental methods provides mixed evidence on the outcome bias. While Gurdal, Miller, and Rustichini (2013) and Brownback and Kuhn (2019) find evidence that individuals' judgement is biased by luck even when intentions are fully observable, Charness (2004) and Charness and Levine (2007) show that individuals' reciprocal behavior depends on the decision makers' observable intentions.

Combining the insights from this literature and the model proposed by Falk and Fischbacher (2006), we develop a framework where there is uncertainty about the leaders' actions, and the discretionary payments made to them are potentially determined by both outcomes and evaluators' beliefs about their intentions. Our study's novelty is that it aims to identify the role of beliefs versus outcomes in explaining gender differences in performance evaluation. Consequently, this research improves our understanding of the underlying causes of observed gender differences in compensation in different contexts by investigating the role that different channels play in driving these gender differences.

3 Experimental Design

The main task in the experiment is a leadership task which consists of two stages. Table 1 presents a timeline.[6]

3.1 Leadership Task – Stage 1: Investment decision

In Stage 1, all participants make investment decisions. They are informed that once these decisions are made, participants will be assigned to groups of three where the roles will be determined randomly. One person will randomly be assigned to be the leader and the other two group members will be assigned to be evaluators (labelled as "members" in the experiment). Participants are also informed that their decisions will be implemented for their group if they

[5] Attribution biases in situations where social preferences of decision makers can play a role have been the subject of a recent study by Erkal, Gangadharan, and Koh (2021). However, it does not address gender differences in attribution bias.

[6] The instructions used in the experiment can be found in Appendix A.

are assigned to be the leader, and that the evaluators will only learn the final outcome of the investment and not the leader's decisions.

All participants are told to make investment decisions assuming that they will be the leader. As the leader, they are endowed with 300 Experimental Currency Units (ECU) to cover the cost of investing in one of two options. Their decisions affect both their own payoffs and those of the evaluators. Specifically, the leader pays the investment cost, but the return from the investment is shared equally among the three group members (including the leader). Both investment options can either fail (i.e., a bad outcome) or succeed (i.e., a good outcome). Investment X corresponds to a high effort choice while Investment Y corresponds to a low effort choice. Choosing high effort costs the leader 200 ECU and yields a success probability of 0.75, while choosing low effort costs the leader 50 ECU and yields a success probability of 0.25. Each investment leads to a high return if it succeeds and a low return if it fails. The investment options are presented to participants in a diagram (see Figure 1).

In total, participants complete five investment tasks with different parameterizations. The costs and probabilities of success of both investment options are the same in all five tasks. However, the tasks differ in terms of the returns from the investments, leading to different payoffs for the leader and the evaluators. Table 2 presents the leader's and each evaluator's net payoff for each possible outcome of the two investment options across the five tasks. Varying the returns across the tasks allows us to examine whether the leader's investment decisions and the evaluators' beliefs depend on the distribution of returns within the group. As can be seen in Table 2, the expected return to the leader is always higher under Investment Y (low effort), but the expected return to each evaluator is always higher under Investment X (high effort). Hence, leaders face a trade-off between maximizing their own payoff and maximizing the evaluators' payoffs, and we expect social preferences to play an important role in their decisions.

3.2 Treatment: Gender of the leader

Our primary objective is to examine how leaders' outcomes are evaluated and whether these assessments are influenced by the leader's gender. Hence, our treatments relate to the leader's gender. The leader's gender is revealed to the evaluators following the same approach as in Bordalo et al. (2019).

Specifically, participants are randomly assigned to groups of three with either a female leader or a male leader. All sessions have an equal number of female and male leaders. After participants make their decisions in Stage 1 and before Stage 2 begins, they are informed about their group assignment and their roles within the group. Each group is assigned a number and has one leader and two evaluators. Participants find out on their screens both their group number and their assigned role.

Once this information is provided, the experimenter calls out each group separately by their group number and asks the participants in that group to raise their hands. The experimenter also announces the last three digits of the ID number of the group's leader, and the leader is asked to call out "here."[7] This enables the leader's gender to be discreetly revealed to the evaluators. The greeting is brief and standardized across all leaders. We ask evaluators of each group to raise their hands to avoid any obvious attention to the leader's announcement. Since the participants are seated at individual cubicles with sufficiently high partitions facing the computer screens, they are unable to see one another. Moreover, from this point of the experiment onwards, whenever there is a reference to the leader, evaluators see on their computer screens the pronoun corresponding to the leader's gender. Using this protocol, differences in evaluators' evaluations can be attributed to differences in the leader's gender.[8]

3.3 Leadership Task – Stage 2: Elicitation of evaluators' beliefs and discretionary payments

After the groups and roles are revealed, for each investment task, evaluators report two sets of beliefs about their group's leader on two separate screens. First, they report their belief of the likelihood that the leader has chosen Investment X (i.e., a high effort choice). Specifically, they answer the following question based on the gender of the leader: "What is the chance out of 100 that she (he) has chosen Investment X?" We refer to this as the evaluators' *prior* beliefs.

Next, they are asked the same question conditional on each possible outcome of the investment chosen by the leader. That is, they are asked to report their beliefs of the likelihood that the leader has chosen Investment X conditional on the investment being successful and unsuccessful. These correspond to the evaluators' *posterior* beliefs. When reporting their posterior beliefs, evaluators are provided with the prior beliefs they have previously reported. However, we do not impose any restrictions on their posterior beliefs. That is, they can report any belief they want regardless of what their prior beliefs are.

In Stage 2, evaluators are paid for either their prior belief or their posterior belief corresponding to the realized outcome of the leader's investment choice. Beliefs are

[7] The precise protocol we followed is reported at the end of the instructions in Appendix A (under "Experimenter Notes"). All sessions were run according to this protocol and by the same experimenter.

[8] Evaluators are asked to predict the leader's gender and ethnicity in the post-experimental questionnaire. 89.3% of them predict their leader's gender correctly. The accuracy rate is independent of the leader's or the evaluator's gender. Conversely, 41% of evaluators are able to predict their leader's ethnicity correctly. In our main analysis, we categorize the data according to the evaluators' predictions of the leader's gender.

incentivized using the binarized scoring rule (BSR), which is incentive compatible independent of the evaluators' risk preferences (Hossain and Okui, 2013; Erkal, Gangadharan, and Koh, 2020). Specifically, the evaluator's probability of receiving 200 ECU decreases as the distance between their belief report and the leader's actual decision increases.

For each investment task, after evaluators have stated their beliefs, they are asked whether they would like to provide the leader with a discretionary payment. The discretionary payment can be either negative (i.e., a penalty) or positive (i.e., a bonus). In particular, they can choose to adjust the leader's payoff from Stage 1 by an amount between -100 ECU and 100 ECU (in multiples of 10 ECU). Similar to the elicitation of their posterior beliefs, evaluators make two discretionary payment decisions for each investment task, one conditional on the investment being successful and another one conditional on the investment failing. The decisions of one of the two evaluators are randomly chosen to be implemented. The payment made to the leader is based on this evaluator's decision corresponding to the realized outcome of the leader's investment decision. The evaluators' payoffs are not affected by their discretionary payment decisions.[9]

3.4 Procedures

All sessions were conducted at the Experimental Economics Laboratory at the University of Melbourne (E^{2}MU) and programmed using z-Tree (Fischbacher, 2007). Participants were mainly university students recruited across different disciplines using ORSEE (Greiner, 2015). We recruited an equal number of men and women to each session and excluded participants who had previously participated in similar experiments. Each session lasted between 90 and 120 minutes.

Upon registering for an advertised session, each participant was first invited to complete a pre-experimental questionnaire on Qualtrics at least one day before the session.[10] The pre-experiment questionnaire included basic demographic questions, such as the participant's age, gender, ethnicity, nationality, year level and field of study at the university, and past experience with economics experiments. At the end of the questionnaire, they were assigned a six-digit ID number that provided us with information about the participant's gender and enabled us to achieve gender balance in the allocation of leadership roles in the investment task.[11]

[9] We assume that discretionary payment decisions do not have monetary consequences for the evaluators, but there may be non-monetary costs, such as cognitive costs, associated with these decisions. We allow for this possibility in the theoretical framework we consider in Section 4.

[10] We conducted the questionnaire before the main session to reduce the likelihood that participants would perceive the study to be about gender or any of the demographic variables that were elicited in the questionnaire.

[11] A male participant was assigned a random ID number between 100000 and 499999, while a female participant was assigned a random ID number between 500000 and 899999.

During the experiment, once participants have entered their ID numbers, they were provided with printed instructions to the leadership task. Participants also answered a number of control questions for the leadership task to reinforce their understanding of the instructions. The experimenter then read aloud the summary of the instructions before beginning the task. To control for potential order effects, the investment tasks in the leadership task were presented in a random sequence in each session. All participants within the same session saw the five tasks in the same order.

At the conclusion of the leadership task, subjects participated in a dictator game in groups of two. Each participant is endowed with 300 ECU which they can decide to allocate between themselves and their matched partner. While both participants make an allocation decision, one participant's decision is randomly chosen at the end of the session to determine the final earnings for each pair. The decisions in this game provide us with a measure of each participant's social preferences, which is an important motivation underlying their decisions in the leadership task (see Section 4). Upon conclusion of the dictator game, participants completed a questionnaire that included questions relating to their decisions in the experiment, a cognitive reflection task (CRT), and an incentivized risk-elicitation task.

Participants did not receive any feedback during the experiment and were paid for either one randomly chosen investment task in the leadership task or for the dictator game. If participants were paid for a randomly chosen investment task in the leadership task, then the leaders were paid according to the decision they made in that investment task in Stage 1 and any discretionary payments given to them by one of the two evaluators. Evaluators were paid either for their leader's investment decision in Stage 1 or for their reported beliefs in Stage 2. Participants also received additional payments for their decisions in the risk task and the CRT, and a fixed payment of 7 AUD for completing the pre-experimental questionnaire. Earnings were converted into cash at the conclusion of the session at the rate of 10 ECU = 1 AUD. Participants earned 39.78 AUD on average, and earnings ranged between 9 AUD and 77 AUD.

In total, we collected data from 350 participants.[12] Given our sample size, a Type I error rate of 0.05, and statistical power of 0.80, we are able to detect an approximately 6.5 percentage point difference in the proportion of high effort choice between female and male leaders, a 0.168 standard deviation difference in prior beliefs between female and male leaders, and a

[12] 354 participants took part in the experiment. However, one participant (an evaluator) misreported the ID number they received from the pre-experimental questionnaire. We were therefore unable to match their demographic information (including gender) to the experimental data. As such, their data point was dropped. One participant withdrew from the experiment after the session had begun. We removed the entire group from the experiment.

0.195 standard deviation difference in payoff adjustments between female and male leaders for a given outcome. With respect to evaluators' belief-updating behavior, our power calculations rely on simulations using data from Erkal, Gangadharan, and Koh (2021). Our sample size allows us to detect a gender difference of 0.25-0.3 in the estimated parameters in the attribution of outcomes (i.e., differences in the estimated values of γ_G or γ_B based on the econometric framework presented in Section 6.2).

4 Theoretical Framework

In this section, we present a theoretical framework which helps us formulate our hypotheses. In line with our experimental design, we consider an environment where each leader makes a discrete effort choice $e \in \{e_L, e_H\}$ on behalf of a group of N players. The effort cost $c \in \{c_L, c_H\}$ is deducted from an initial endowment $\omega \geq c_H > c_L$ that the leader receives.

Leaders' effort choices are affected in part by their social preferences, i.e., altruism toward the other group members. We assume that leaders are differentiated based on their altruistic preferences and they have private information about their types. Let $\alpha_i \in [0,1]$ stand for the altruistic preference of leader i. It is a private draw from a distribution $F(\alpha)$ with density $f(\alpha)$. $F(\alpha)$ is common knowledge. $\alpha_i = 0$ stands for a purely self-interested leader.

A leader's effort choice results in an output Q, where $Q \in \{Q_L, Q_H\}$ and $Q_H > Q_L$. The realized output level is equally shared between the members of the group (including the leader), although the cost of effort is solely borne by the leader. Output is probabilistic and the leader's effort choice affects the probability of a high-output realization. Specifically, assume that a choice of e_H leads to Q_H with probability $p \in (0.5,1)$ and a choice of e_L leads to Q_H with probability $(1-p)$. Hence, a high effort choice (which is more costly for the leader) leads to a high output level with a higher probability.

4.1 Bonuses and penalties

In our experimental design, each evaluator decides whether they would like to make a discretionary payment to the leader. Although each leader makes decisions for five investment tasks (with different parameterizations) and each evaluator is asked to make five discretionary payment decisions, leaders do not receive any information about the decisions of the evaluators during the experiment. Hence, since the discretionary payments cannot be motivated by an incentive to change future behavior of leaders, we model them as reciprocal actions.

We consider a model of reciprocity where reciprocal actions are potentially determined by both the reciprocator's judgement of the agent's underlying intentions and the consequences of the agent's action (Falk and Fischbacher, 2006).[13] Making this distinction and considering both factors are important in light of the evidence from the literature which shows that players exhibit reciprocal behavior based on the decision makers' intentions (see, e.g., Charness, 2004; Charness and Levine, 2007), and where players have full information about the intentions of actors, they may still reciprocate differently depending on the outcome (e.g., Gurdal, Miller, and Rustichini, 2013; Brownback and Kuhn, 2019).

We assume that evaluators get utility from their material payoff as well as reciprocity. Let $\rho_j > 0$ represent the reciprocity preference of evaluator j. Δ_j stands for the discretionary payment that evaluator j would like to pay to the leader. Δ_j can be positive or negative depending on whether the evaluator decides to award a bonus or impose a penalty. Denote φ_i as evaluator j's perception of leader i's kindness. Each evaluator chooses Δ_j to maximize his/her utility function given his/her reciprocity preference (ρ_j) and his/her perception of the leader's kindness (φ_i).

The kindness term φ_i is key in determining evaluators' discretionary payments, and it depends on the realized output and beliefs about the leader's intentions (effort choice). We let $\sigma_j(e_H|Q)$ stand for the updated (posterior) belief that evaluators have about the leader exerting high effort (e_H) after observing the output Q. The kindness term takes the form

$$\varphi_i = \theta_j\left(1_{Q=Q_H} - 1_{\ Q=Q_L}\right) + \beta_j\left(2\sigma_j(e_H|Q) - 1\right).$$

$1_{Q=Q_H}$ and $1_{Q=Q_L}$ are indicator functions for whether a high or a low output is observed, respectively, and $\theta_j \in [0,1]$ and $\beta_j \in [0,1]$ are the weights evaluator j places on outcomes and beliefs about intentions, respectively, when determining the leader's kindness. This specification implies that $\left(1_{Q=Q_H} - 1_{\ Q=Q_L}\right) \in \{-1,1\}$, $\left(2\sigma_j(e_H|Q) - 1\right) \in [-1,1]$, and, consequently, $\varphi_i \in [-2,2]$. Evaluator j views leader i as kind if $\varphi_i > 0$ and as unkind if $\varphi_i < 0$.

The kindness term indicates that if two evaluators care about intentions only ($\theta_j = 0$ and $\beta_j > 0$) and if they have the same posterior beliefs, then they will choose the same discretionary payment irrespective of the outcome. On the other hand, if two evaluators care about outcomes only ($\beta_j = 0$ and $\theta_j > 0$) and if they observe the same outcome, then they will choose the same discretionary payment even if their posterior beliefs are different.

[13] Unlike Falk and Fischbacher (2006), we have a model where agents have private information about their types. Hence, the reciprocator's judgement of the agent's underlying intention is determined by the reciprocator's belief about the agent's type.

Each evaluator has private information about his/her three-dimensional type represented by ρ_j, θ_j and β_j. Let $G(\rho, \theta, \beta)$ stand for the joint distribution function from which types are drawn. $G(\rho, \theta, \beta)$ is common knowledge.

This framework allows for outputs to have an impact on discretionary payments independent of the beliefs. Hence, outputs potentially affect discretionary payments through two different channels: in addition to the indirect impact they have through evaluators' posterior beliefs, they can also have a direct impact. The direct channel that we allow for means that discretionary payments may differ even if beliefs are the same.

4.2 Utility functions

For a given outcome realization and discretionary payment, leader i's utility is given by

$$U^D = u_i\left(\frac{Q}{N} + \omega - c + \Delta_j\right) + \alpha_i \sum\nolimits_j v_j\left(\frac{Q}{N}\right), \tag{1}$$

where u_i and v_j are twice differentiable utility functions. We assume that u_i represents the direct utility leader i receives from his/her own monetary payoff and v_j is the utility evaluator j receives from his/her own monetary payoff. Leader i's regard for others' utilities depends on his/her social preferences, represented by α_i. Each leader makes the effort choice with the objective of maximizing his/her expected utility (expressed in terms of the priors the leader has over possible outcomes and different evaluator types).

After the output is realized, evaluator j's utility is given by

$$U^M = v_j\left(\frac{Q}{N}\right) + \rho_j \varphi_i \Delta_j - c(\Delta_j), \tag{2}$$

where $\varphi_i = \theta_j(1_{Q=Q_H} - 1_{Q=Q_L}) + \beta_j(2\sigma_j(e_H|Q) - 1)$ as defined above, and $c(\Delta_j)$ stands for the cognitive cost of making a discretionary payment decision. We assume that $c(\Delta_j)$ is increasing and convex in Δ_j.

Evaluator j chooses Δ_j to maximize this utility function. As in Falk and Fischbacher (2006), evaluators maximize their utility by responding to kind actions ($\varphi_i > 0$) with a positive discretionary payment that improves the payoff of the leader and to unkind actions ($\varphi_i < 0$) with a negative discretionary payment that decreases the payoff of the leader.

4.3 Equilibrium

We use Perfect Bayesian Equilibrium as the equilibrium concept. To characterize the equilibrium, we need to specify:

(i) $e(\cdot)$: $[0,1] \rightarrow \{L, H\}$, i.e., the leader's effort as a function of the leader's type;

(ii) $\Delta(\cdot,\cdot)\colon\{Q_L, Q_H\} \to \mathbb{R}$, i.e., the discretionary payment as a function of the evaluator's type;

(iii) $h(\cdot, Q)$, i.e., the probability density function summarizing the posterior beliefs evaluators have after observing the output Q.

The leader makes the effort choice by comparing the expected utility from choosing e_H with the expected utility from choosing e_L. The expected utility from choosing e_H is given by:

$$E_{\rho,\theta,\beta} = \Bigg[p\left(u_i\left(\frac{Q_H}{N} + \omega - c_H + \Delta_j\big(\rho_j, \theta_j, \beta_j, Q_H\big)\right) + \alpha_i \sum\nolimits_j v_j\left(\frac{Q_H}{N}\right)\right)$$

$$+ (1-p)\left(u_i\left(\frac{Q_L}{N} + \omega - c_H + \Delta_j\big(\rho_j, \theta_j, \beta_j, Q_L\big)\right) + \alpha_i \sum\nolimits_j v_j\left(\frac{Q_L}{N}\right)\right)\Bigg]$$

It depends on the discretionary payment that the leader expects to receive. The expected utility from choosing e_L can be written in a similar way. The optimal effort choice partitions the interval $[0,1]$ into two sets:

$$T_L = \{\alpha \in [0,1]\colon e = e_L\} \text{ and } T_H = \{\alpha \in [0,1]\colon e = e_H\}$$

Let α^* denote the type who is indifferent between choosing e_L and e_H. Type α^* stands at the intersection of T_L and T_H such that $T_L = [0, \alpha^*]$ and $T_H = [\alpha^*, 1]$.

After observing Q, evaluators determine their discretionary payments, which are given by

$$\Delta_j \in \arg\max \rho_j \Delta_j\big[\theta_j\big(1_{Q=Q_H} - 1_{Q=Q_L}\big) + \beta_j\big(2\sigma_j(e_H|Q) - 1\big)\big] - c\big(\Delta_j\big)$$

Finally, since leader i will choose high effort in equilibrium if his or her type is above a threshold value α^*, the prior belief an evaluator has about the leader exerting high effort is $1 - F(\alpha^*)$. Then, the probability density function for the posterior beliefs (after observing the output level) will be given by:

$$h(\alpha|Q_H) = \begin{cases} \dfrac{(1-p)f(\alpha)}{\big(1-F(\alpha^*)\big)p + F(\alpha^*)(1-p)} & \text{for } \alpha \in T_L \\ \dfrac{pf(\alpha)}{\big(1-F(\alpha^*)\big)p + F(\alpha^*)(1-p)} & \text{for } \alpha \in T_H \end{cases}$$

$$h(\alpha|Q_L) = \begin{cases} \dfrac{pf(\alpha)}{\big(1-F(\alpha^*)\big)(1-p) + F(\alpha^*)p} & \text{for } \alpha \in T_L \\ \dfrac{(1-p)f(\alpha)}{\big(1-F(\alpha^*)\big)(1-p) + F(\alpha^*)p} & \text{for } \alpha \in T_H \end{cases}$$

5 Hypotheses

Using the framework above, our goal is to investigate the potential channels through which gender biases may affect performance evaluation.

Our first hypothesis is regarding the gender differences we expect to observe in the effort choices of the leaders and the prior beliefs of the evaluators. As shown in (1), leaders' utility is affected by both their own payoff and that of the evaluators. Hence, leaders' effort choices are influenced by their altruism. In their survey papers, Croson and Gneezy (2009), Niederle (2016), and Bilén, Dreber, and Johannesson (2021) report that women tend to be more prosocial than men, although the findings seem to be context-dependent. In our framework, this implies that on average, we would expect women to have a higher α than men, which may create gender differences in effort choices and in evaluators' prior beliefs, as stated in Hypothesis 1.

Hypothesis 1. *(i) Female leaders are (weakly) more likely to choose high effort than male leaders. (ii) Evaluators' prior beliefs are (weakly) higher for female leaders than for male leaders.*

While Hypothesis 1 is about evaluators' prior beliefs, our next hypothesis, Hypothesis 2, is about their posterior beliefs after observing the leader's outcomes. Theoretically, belief updating would be expected to be consistent with Bayes' rule. However, Erkal, Gangadharan, and Koh (2021) show evidence of biases in belief updating. In a set-up where evaluators do not have the option to make discretionary payments, they find that beliefs on average are lower relative to the Bayesian benchmark following good outcomes, implying that leaders do not receive enough credit for good outcomes. The question we address here is whether there are gender differences in posterior beliefs. Evidence from the psychology literature shows that women's outcomes are more likely to be attributed to luck in contexts where outcomes are influenced by ability or skill (see, e.g., Swim and Sanna, 1996). We base Hypothesis 2 on this evidence and investigate attribution biases in an environment where leaders' choices are determined by their social preferences.

Hypothesis 2. *Outcomes of female leaders are more likely to be attributed to luck than those of male leaders.*

Hypothesis 2 implies that evaluators update their beliefs differently depending on the gender of the leader. We refer to the gender difference we may observe in beliefs as the *gender inference gap*.

Finally, we describe our main hypothesis, Hypothesis 3, which is about the sources of gender differences in evaluators' discretionary payments. As discussed in Section 4, we expect both posterior beliefs and outcomes to influence evaluators' discretionary payments. In addition, individuals may differ in the weights they put on intentions (represented by their posterior beliefs) and outcomes in the determination of discretionary payments. These weights are represented by the parameters β_j and θ_j introduced in the expression for the kindness term φ_i above.

Our framework uncovers that it is important to distinguish between biases in beliefs and biases in weights put on beliefs. That is, it implies that gender gaps may emerge in posterior beliefs (represented by σ_j) as well as in the weights individuals put on their beliefs versus outcomes in the determination of discretionary payments. The latter implies that the values of the parameters β_j and θ_j may vary depending on the gender of the individual being evaluated. We refer to the gender differences we may observe in β_j and θ_j as the *gender belief-outcome gap*. Hypothesis 3 summarizes our conjecture on the type of gender biases which may shape discretionary payments.

Hypothesis 3. *Discretionary payments are shaped by both evaluators' evaluation of the leader's intentions (i.e., their posterior beliefs) and the leader's outcomes themselves. Gender differences may emerge in discretionary payments due to (i) a gender inference gap and/or (ii) a gender belief-outcome gap.*

In our analysis, we also consider whether the gender of the evaluator makes a difference. That is, we ask whether female evaluators treat female or male leaders differently from male evaluators. On the one hand, due to homophily, one may expect female evaluators to treat female leaders more favorably. On the other hand, if gender discrimination is the norm and female evaluators choose to conform to social norms, their behavior may not be different from that of male evaluators. In their desire to conform to social norms, it is also possible that women discriminate against women more than men do (e.g., Derks, Van Laar, and Ellemers, 2016; Arvate, Galilea, and Todescat, 2018). Given these potentially conflicting forces, we do not have a clear prediction on the behavior of the evaluators. Hence, we refrain from formulating a hypothesis and prefer to explore this issue empirically.

6 Results

We start in section 6.1 by analyzing the leaders' effort choices and evaluators' prior beliefs (Hypothesis 1). In section 6.2, we examine the evaluators' updating behavior (Hypothesis 2) and in section 6.3, we investigate evaluators' discretionary payment decisions (Hypothesis 3).

6.1 Leaders' effort choices and evaluators' prior beliefs

Figure 2 presents the proportion of high effort choices made by participants (in the role of leaders) in Stage 1 (panel a) and evaluators' average prior beliefs reported in Stage 2 (panel b). In all our analyses, belief is a variable that takes an integer value in [0, 100], where a higher belief implies that the evaluator thinks the leader is more likely to have chosen high effort. Panel (a) of Figure 2 reveals that there are no statistically significant differences between male and female leaders in their effort choices (Fisher's exact test: p-value = 0.626). Correspondingly, panel (b) reveals that there are no statistically significant differences in evaluators' average prior beliefs towards male and female leaders (rank-sum test: p-value = 0.213).[14]

Table 3 and Table 4 present estimates for the participants' effort choices as leaders (marginal effects from a probit model) and evaluators' prior beliefs (ordinary least squares, OLS), respectively. In both tables, in addition to the leader's gender, we also control for the participants' behavior in the risk task, the difference in the return for a good outcome relative to a bad outcome for a given investment task, investment tasks where the investments provide a return of 0 ECU if they fail, order effects, and participants' characteristics (column 2). In Table 3, we control for participants' decisions in the dictator game as a test of our prediction that participants' social preferences are a key driver of their effort choices as leaders. In Table 4, we control for the participants' own effort choices as leaders to examine whether a consensus effect exists.[15] The regression estimates in both tables support our conclusions from the non-parametric analysis (estimated effects of female leader: p-values = 0.474 and 0.472 in columns 1 and 2 of Table 3; p-values = 0.531 and 0.620 in columns 1 and 2 of Table 4, respectively). We also find evidence of a consensus effect in evaluators' prior beliefs. Evaluators who chose high effort as leaders in Stage 1 are more likely to believe in Stage 2 that their leaders have chosen high effort (estimated effect of choosing high effort as leader in both columns of Table 4: p-values < 0.001).

[14] We also do not find any statistically significant differences in prior beliefs towards male and female leaders separately for both female and male evaluators (rank-sum tests: p-values = 0.412 and 0.324, respectively).

[15] The consensus effect is the tendency for individuals to believe that others would act or think in a manner similar to themselves. See, e.g., Engelmann and Strobel (2000, 2012). Erkal, Gangadharan, and Koh (2021) find evidence of the consensus effect in a similar decision-making environment.

Importantly, consistent with our theoretical framework, we find evidence that the leaders' effort choices are indeed motivated by their social preference. There is a statistically significant positive relationship between participants' giving behavior in the dictator game and their effort choices as leaders in the investment task (effect of % endowment transferred in DG in both columns of Table 3: p-values < 0.001). However, examining participants' behavior in the dictator game, we do not observe evidence that women are more prosocial than men in our sample. On average, women contribute slightly more of their endowment in the dictator game relative to men (31% versus 28%), but this difference is not statistically significant (rank-sum test: p-value = 0.318). Hence, the absence of a gender difference in leaders' effort choice may be attributed to men and women being no different in their prosocial preferences in our sample.

We summarize our results relating to Hypothesis 1 as follows.

Result 1. *There are no statistically significant differences in leaders' effort choices and evaluators' prior beliefs about the leader's effort choice between female and male leaders.*

6.2 Evaluators' updating behavior

Estimation strategy

To examine the evaluators' updating behavior, we consider the following econometric specification:

$$\text{logit}\left(\hat{\sigma}_j(e_H|Q)\right) = \delta\,\text{logit}(\hat{\mu}_j) + \gamma_G\, I(Q = Q_H)\cdot\text{logit}(p) + \gamma_B\, I(Q = Q_L)\cdot\text{logit}(1-p) + \varepsilon_j, \quad (3)$$

where $\text{logit}\, x = \log\left(\frac{x}{1-x}\right)$, $I(\cdot)$ is an indicator function for the observed output or return Q from the investment, $\hat{\sigma}_j(e_H|Q)$ and $\hat{\mu}_j$ represent evaluator j's reported posterior beliefs (given Q) and prior beliefs, respectively, and ε_i captures non-systematic errors. Note that a high effort choice leads to a high output (good outcome) with probability p, while a low effort choice leads to a good outcome with probability $1-p$. In our experiment, p = 0.75.

The specification in (3) allows us to determine the weights evaluators place on their prior beliefs and the signals they receive via the observed outcome. It nests the theoretical Bayesian benchmark as a special case with $\delta = \gamma_G = \gamma_B = 1$. Any deviation in the estimated parameters from 1 is interpreted as non-Bayesian updating behavior. Specifically, $\delta < 1$ implies that evaluator j suffers from base-rate neglect while $\delta > 1$ implies that s/he suffers from confirmatory bias. The parameters γ_G and γ_B represent the weights evaluators place on a signal of good and bad outcome, respectively, when updating their beliefs. $\gamma_G < 1$ ($\gamma_B < 1$) implies that the evaluator attributes a good (bad) outcome more to luck relative to a Bayesian, while

$\gamma_G > 1$ ($\gamma_B > 1$) implies that s/he attributes the outcome more to the leader's decision. Finally, a test of $\gamma_G = \gamma_B$ allows us to examine whether there is an asymmetric attribution of good and bad outcomes.[16]

Gender biases in the attribution of outcomes

We estimate equation (3) using OLS and compare the estimated coefficients between male and female leaders to analyze the gender biases that evaluators suffer from when updating their beliefs about the leader's effort choices. Figure B1 of Appendix B shows the distribution of evaluators who update their beliefs inconsistently (i.e., in the opposite direction to that predicted by Bayes' rule) or not at all (i.e., have posterior beliefs equal to prior beliefs). The inclusion of these observations in the analysis may result in biased or incorrect conclusions, particularly if these evaluators are reporting beliefs that do not genuinely reflect their true posterior beliefs. Hence, for the remainder of our analysis, we exclude an evaluator if 25% or more of their posterior beliefs are in the opposite direction to that predicted by Bayes' rule or if all of their posterior beliefs are equal to their prior beliefs. This corresponds to 19.7% and 5.6% of the sample, respectively, which is largely in line with what has been previously found in the literature (Möbius et al., 2014; Coutts, 2019; Barron, 2021; Erkal, Gangadharan, and Koh, 2021). In Table B1 of Appendix B, we present as a robustness check the analysis with the full sample. Our main conclusion in regard to gender differences in the attribution of both good and bad outcomes remains unchanged.

Table 5 presents the regression results separately by the leader's gender. First, we observe that there are no statistically significant differences in the attribution of good and bad outcomes between female and male leaders (comparisons of γ_G and γ_B between columns 1 and 2: p-values = 0.720 and 0.434, respectively). Second, columns (1) and (2) reveal that, while evaluators consistently suffer from base-rate neglect (test of $\delta = 1$: p-values < 0.001 in both columns),[17] they are no different from a Bayesian in their attribution of both good and bad outcomes (tests of $\gamma_G = 1$ and $\gamma_B = 1$: p-values = 0.906 and 0.484, respectively, for female leaders in column 1; and p-values = 0.726 and 0.145, respectively, for male leaders in column 2).[18]

[16] See also, e.g., Grether (1980), Möbius et al. (2014), Ambuehl and Li (2018), Buser, Gerhards, and van der Weele (2018), Coutts (2019), Barron (2021), and Erkal, Gangadharan, and Koh (2021) for similar estimation approaches, and Benjamin (2019) for a recent review of studies which estimate systematic deviations from the Bayesian benchmark.

[17] Note that base-rate neglect is a stylized finding in the literature (Benjamin, 2019).

[18] The results hold even when we consider the analysis separately by both the leader's and evaluator's gender. The estimates in Table C1 of Appendix C reveal that both female and male evaluators are no different from a

We summarize the results for Hypothesis 2 as follows.

Result 2. *There are no statistically significant differences in evaluators' attribution of outcomes between female and male leaders. Moreover, evaluators are no different from a Bayesian in the attribution of the leader's outcomes regardless of the leader's gender.*

6.3 Evaluators' discretionary payments

We next turn to evaluators' discretionary payments to leaders.[19] We first examine the overall payments made to female and male leaders. Figure 3 presents the average discretionary payments by outcome and the leader's gender. The figure shows that on average, both male and female leaders receive negative payments for a bad outcome and positive payments for a good outcome (signed-rank tests of discretionary payments = 0: p-values < 0.001 in all cases). Moreover, there are no differences in the average payments made to male and female leaders, both for a bad outcome and for a good outcome (rank-sum tests: p-values = 0.907 and 0.954, respectively).

Figure 4 presents evaluators' penalty decisions for a bad outcome and bonus decisions for a good outcome.[20] Panel (a) shows the proportion of male and female leaders receiving a penalty for a bad outcome or a bonus for a good outcome. The figure shows that there are no statistically significant gender differences in the penalties and bonuses awarded to leaders on the extensive margin (p-values for Fisher's exact tests: (i) penalties imposed for bad outcomes = 0.783; (ii) bonuses awarded for good outcomes = 0.331).

Panel (b) of Figure 4 shows the average penalty/bonus amounts leaders receive, conditional on receiving a penalty/bonus. The figure reveals gender differences in penalties and bonuses on the intensive margin. Conditional on receiving a penalty/bonus, female leaders receive lower penalties for a bad outcome and lower bonuses for a good outcome as compared to male leaders on average, although the gender difference in the penalty amounts is marginally

Bayesian in their attributions of the leader's outcomes, and that they also do not attribute the outcomes of female and male leaders differently.

[19] As we are interested in examining the link between discretionary payments and evaluators' posterior beliefs, we also exclude evaluators who are classified as either inconsistent or non-updaters in our analysis here. We show in Table B2 and Table B3 of Appendix B that our main conclusions remain largely unchanged with the full sample.

[20] Penalties refer to negative discretionary payments while bonuses refer to positive discretionary payments. Note that evaluators could, in practice, award their leader a bonus for a bad outcome and impose a penalty for a good outcome. We find that the majority of bonus and penalty decisions depend on the outcomes. Only 23% of evaluators award a bonus for a bad outcome (more likely for male leaders than for female leaders, p-value of Fisher's exact test = 0.007) and 14% impose a penalty for a good outcome (no difference between male and female leaders, p-value of Fisher's exact test = 0.140).

statistically significant (p-values for rank-sum tests: (i) penalty amounts = 0.056 and (ii) bonus amounts: 0.025).[21]

Thus, despite not observing any differences in the attribution of outcomes between female and male leaders (Result 2), we observe gender differences in bonuses on the intensive margin. At the first glance, this may appear puzzling. To investigate this further, we turn to the drivers of evaluators' discretionary payment decisions, as hypothesized by our theoretical framework. Figure 5 presents bubble plots of evaluators' discretionary payments against their posterior beliefs separately for female leaders (panel a) and male leaders (panel b). In each panel, the discretionary payments are graphed separately for good outcomes (gray bubbles) and bad outcomes (white bubbles), where the size of each bubble is proportional to the number of observations. We also plot fitted lines using estimates of OLS regressions of evaluators' discretionary payments, as presented in Table 6, along with 95% confidence intervals.

Figure 5 and Table 6 reveal that the channels driving evaluators' discretionary payments depend on the leader's gender. While the payments to male leaders are increasing in evaluators' posterior beliefs (column 2 of Table 6: p-value = 0.007), evaluators' posterior beliefs do not play a role in shaping payments to female leaders (column 1: p-value = 0.506). The difference between female and male leaders in the estimated impact of evaluators' posterior beliefs on discretionary payments is statistically significant (column 1 vs. column 2: p-value = 0.012).

In addition, we observe that outcomes are an important determinant of payments made to leaders (p-values < 0.001 and = 0.001 in columns 1 and 2, respectively). This is consistent with the literature on outcome bias. We find no statistically significant gender difference in the estimated impact of outcomes on discretionary payments (column 1 vs. column 2: p-value = 0.575). Consequently, outcomes play a larger role than beliefs in driving the payments made to female leaders (column 1: p-value = 0.002), but outcomes and beliefs do not play different roles in driving the payments made to male leaders (column 2: p-value = 0.302).[22]

In summary, we find partial support for Hypothesis 3. We observe an outcome bias where, after controlling for beliefs, outcomes still play a role in determining the discretionary payments made to male and female leaders. However, intentions (posterior beliefs) only play

[21] Table C2 of Appendix C reports hurdle model regression estimates of evaluators' penalty decisions for a bad outcome and bonus decisions for a good outcome. The regression estimates are broadly consistent with our conclusions from the non-parametric tests. However, the gender difference in the average penalty sent to leaders for a bad outcome (column 2) is not statistically significant (p-value = 0.256).

[22] This tests the null hypothesis that the coefficient on good outcome is equal to 100 times the coefficient on posterior beliefs. The interpretation of the test is whether there is a difference in the *marginal change* in evaluators' discretionary payments between two scenarios: (i) with respect to a given change in outcome (from a bad outcome to a good outcome), and (ii) with respect to a given change in belief (from a belief that the leader has chosen low effort with certainty to a belief that the leader has chosen high effort with certainty).

a role in shaping the payments made to male leaders. Hence, the determinants of discretionary payments vary by gender. We observe a *gender belief-outcome gap* in the determination of leaders' discretionary payments. We summarize our results as follows.

Result 3.

(a) There are no statistically significant differences in the average discretionary payments and the propensity to receive bonuses and penalties between female and male leaders. However, conditional on receiving a bonus for a good outcome, female leaders receive lower bonuses on average than male leaders.

(b) There exists a gender belief-outcome gap in the determination of discretionary payments. Discretionary payments made to male leaders are increasing in evaluators' posterior beliefs, but the payments made to female leaders do not depend on the evaluators' beliefs.

To shed light on what might be driving the gender biases in discretionary payments, we examine leaders' expectations about the discretionary payments they will receive from the evaluators. Specifically, in Stage 2, while evaluators are making their decisions, we ask the leaders to indicate the average discretionary payment that they expect to receive for a good outcome and for a bad outcome.[23]

Panel (a) of Figure 6 reveals that, relative to male leaders, female leaders are less likely to expect a bonus for good outcomes (Fisher's exact test: p-value = 0.002). Moreover, given a bonus for attaining a good outcome, panel (b) reveals that female leaders expect a lower bonus on average than male leaders (rank-sum test: p-value < 0.001). Hence, the overall observed gender bias in the bonuses leaders receive is consistent with their expectations. This consistency between the leaders' expectations and the evaluators' decisions suggests that both may be shaped by societal norms about the role gender plays in shaping discretionary rewards.

We further investigate whether Result 3 is driven by male or female evaluators. Given the smaller number of observations and therefore lower statistical power, when considering behavior separately by both the leader's and evaluator's gender, the analysis we present here is exploratory.

Figure 7 considers the average bonus/penalty amounts separately by both the leader's and the evaluator's gender. The figure reveals that the gender differences in bonus/penalty amounts we observe in Figure 4 are driven by *female evaluators*. Female evaluators give lower

[23] These beliefs are unincentivized.

bonuses to female leaders than to male leaders on average for a good outcome (rank-sum test: p-value = 0.002). They also impose lower penalties on female leaders than on male leaders on average for a bad outcome (rank-sum test: p-value < 0.001). For male evaluators, there are no differences between female and male leaders in both the average penalty and bonus amounts (rank-sum test: p-values = 0.681 and 0.157, respectively).[24]

Table 7 reports OLS regression estimates of discretionary payments separately for female evaluators (columns 1 and 2) and for male evaluators (columns 3 and 4). The estimates reveal that female evaluators' discretionary payments to male leaders are increasing in their posterior beliefs (column 2: p-value = 0.032), but their beliefs do not have any statistically significant effect on the payments made to female leaders (column 1: p-value = 0.366). For male evaluators, their posterior beliefs do not have any statistically significant effect on the discretionary payments for both female and male leaders (p-values = 0.985 and 0.249 in columns 3 and 4, respectively). The difference between female and male leaders in the impact of posterior beliefs on discretionary payments is statistically significant for female evaluators (column 1 vs. 2: p-value = 0.025) but not for male evaluators (column 3 vs. 4: p-value = 0.371).[25,26]

Hence, we find that female evaluators are driving both the gender differences in bonus and penalty amounts, as well as the gender belief-outcome gap, suggesting that in our sample, female evaluators may be more likely to conform to gender stereotypes than male evaluators. We summarize our results as follows.

Result 4. *Female evaluators award lower bonuses for good outcomes and lower penalties for bad outcomes to female leaders than to male leaders on average, conditional on making a discretionary payment. Moreover, the gender belief-outcome gap in the determination of discretionary payments is predominantly driven by female evaluators.*

[24] These conclusions are supported by hurdle model regression estimates reported in Table C3 of Appendix C.

[25] Note that evaluators' discretionary payments depend on the leader's outcome regardless of the leader's and evaluator's gender (p-values in columns 1 to 4 = 0.006, 0.003, 0.019, and 0.067, respectively). Consequently, outcomes play a bigger role than beliefs in shaping the discretionary payments made to female leaders (column 1: p-value = 0.005), while they do not play different roles in shaping the payments to male leaders (column 2: p-value = 0.644). The difference between female and male leaders in the impact of outcomes is not statistically significant for both female evaluators and male evaluators (column 1 vs. 2, and column 3 vs. 4: p-values = 0.948 and 0.407, respectively).

[26] For the discretionary payments made to female leaders by male evaluators (column 3 of Table 7), we note that the impact of outcomes is larger in magnitude relative to that of beliefs, although this difference is not statistically significant (p-value = 0.140). While this difference becomes statistically significant with the full sample (column 5 of Table B3: p-value = 0.018), there remains no statistically significant difference between female and male leaders in the impact of beliefs (column 5 vs. column 6 of Table B3: p-value = 0.975).

7 Discussion

The key objective of our paper is to examine if female leaders are evaluated differently as compared to male leaders. A distinguishing feature of our research is that we focus on an environment where social preferences play a role in driving the leaders' decisions. We assume that costly effort choices of leaders are not observed, and outcomes are determined by a combination of the choices made and luck. Uncertainty of this kind is ubiquitous in decision making, making performance evaluation a challenging task. While trying to evaluate performance based on the merits of the actions taken by the leader, evaluators need to correctly assess the role of unexpected events in determining the outcomes.

Evaluators in our environment can make discretionary payments, such as bonuses or penalties, to the leaders. We find female leaders receive lower bonuses on average than male leaders for good outcomes. Interestingly, this result cannot be explained by gender biases in beliefs. More specifically, our data on prior beliefs reveal that the observed gender differences in discretionary payments cannot be explained by gender stereotypes to the extent that these stereotypes would show themselves in the prior beliefs. The observed gender differences in discretionary payments cannot be explained by attribution biases in posterior beliefs either (e.g., a tendency to attribute women's outcome more to luck), since we do not find gender biases in the evaluators' posterior beliefs.

Our theoretical model specifies another potential mechanism for gender differences to emerge in discretionary payments. In addition to a gender inference gap, which is driven by gender biases in belief updating, there may be a gender belief-outcome gap, driven by gender differences in the emphasis evaluators put on their beliefs and the outcomes themselves. In our study, we find supporting evidence for such a gender belief-outcome gap. Specifically, while male leaders' discretionary payments are determined by both the evaluators' assessments of their effort choices and outcomes, female leaders' discretionary payments are predominantly determined by outcomes. This result suggests that for both male and female leaders, incentive structures deviate from rewarding them based on the merits of the actions taken. However, in the case of female leaders, surprisingly the deviation is such that beliefs about actions taken do not play any role.

Our findings contribute to the large and influential literature on gender discrimination and gender differences in labor market outcomes by furthering our understanding of the factors that may be driving the observed gender gaps in performance pay. Performance evaluation in many contexts rely on subjective measures of performance, which creates an opportunity for biases to distort incentive structures.[27] Given that gender differences in performance evaluation have been documented across many domains, it is important for organizations to understand the sources of these gender differences and to ensure that their incentive structures are not compromised by biased performance evaluation procedures. Our results can explain the channels behind the observed gender differences in performance evaluation in different contexts. Using observational data, it is difficult to establish the drivers of these differences. Although gender biases in beliefs are a potential explanation of these differences, we show that another contributing factor is that evaluators' beliefs about the actions taken and their outcome biases play differential roles in determining discretionary payments depending on the gender of the leader. This gender belief-outcome gap that we identify implies that in the labor market, good outcomes are necessary for women to get bonuses, but men can receive bonuses for bad outcomes as long as evaluators hold them in high regard.

More broadly, our findings indicate that luck plays a bigger role in female leaders' performance evaluation. While choosing discretionary payments, evaluators put less emphasis on their beliefs about the choices made by female leaders and put more emphasis on the outcomes. The disproportionate emphasis given to outcomes of female leaders has the potential to distort their choices in risky environments and can perpetuate gender gaps. Our results can help inform organizations on how to improve their incentive structures to reduce gender differences.

Organizations may want to consider group-based evaluation to reduce biases in performance evaluation. When evaluation is performed collectively, it is possible that these biases would be reduced.[28] Another approach is to introduce well-defined and structured performance criteria such that there are not many degrees of freedom for biases to emerge. In future research, it would be useful to examine the effectiveness of both the group-based and structured evaluation procedures.

Another fruitful avenue for future research relates to the underlying motivation for decision making. In our research, leaders' choices are determined mainly by their social preferences. There may be a perception that this design favors female leaders. Although we do

[27] See, for example, Eccles and Crane (1988), Hayes and Schaefer (2000), Levin (2003), and Gibbs et al. (2009) who discuss the use of subjective performance evaluations in investment banking, CEO compensation packages, law firms, and auto dealerships. For a comprehensive review of the use of subjective performance measures, see Prendergast (1999).

[28] Mengel (2021), for example, analyzes when committee deliberation may result in gender bias and how the bias can be mitigated through intervention. When it comes to hiring, Bohnet, van Geen, and Bazerman (2016) find that evaluators tend to base their decisions more on performance in joint evaluations, but they rely more on stereotypes in separate evaluations.

not find evidence of this in the prior beliefs, we still observe gender gaps that disadvantage women. It may be interesting to examine if the biases would be attenuated or aggravated when outcomes are instead driven by the decision makers' ability or skill. For example, Bordalo et al. (2019) show that the use of gender-stereotyped tasks changes the prior beliefs men and women have about themselves and of others, while Coffman, Collis, and Kulkarni (2020) show how these stereotypes affect the way individuals update their beliefs about themselves. When evaluating others, would beliefs be updated differently (given new information) depending on the gender of the person being evaluated and the nature of the task being performed?

In summary, our findings point towards an important takeaway. While biases in beliefs may play an important role in some situations, gender discrimination may not be just due to these biased beliefs. Rather, the weight placed on beliefs about intentions versus outcomes may be a source of discrimination, with outcomes playing a disproportionately larger role than beliefs in case of women. This type of gender bias is distinct from biased beliefs and leads to a new channel through which discrimination can occur.

References

Ambuehl, S., Li, S. (2018). Belief updating and the demand for information. *Games and Economic Behavior,* 109:21-39.

Arvate, P.R., Galilea, G.W., Todescat, I. (2018). The queen bee: A myth? The effect of top-level female leadership on subordinate females. *The Leadership Quarterly,* 29(5):533-548.

Baron, J., Hershey, J.C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology,* 54(4):569-579.

Barron, K. (2021). Belief updating: Does the 'good-news, bad-news' asymmetry extend to purely financial domains? *Experimental Economics,* 24:31-58.

Barron, K., Ditlmann, R., Gehrig, S., Schweighofer-Kodritsch, S. (2020). Explicit and implicit belief-based gender discrimination: A hiring experiment. *Working Paper*.

Bénabou, R., Tirole, J. (2010). Individual and corporate social responsibility. *Economica,* 77(305):1-19.

Benjamin, D.J. (2019). Errors in probabilistic reasoning and judgment biases. In B.D. Bernheim, S. DellaVigna, & D. Laibson (Eds.), *Handbook of Behavioral Economics: Applications and Foundations 2* (2:69-186): Elsevier.

Bertrand, M., Duflo, E. (2017). Field experiments on discrimination. In E. Duflo & A. Banerjee (Eds.), *Handbook of Field Experiments* (1:309-393). North Holland: Elsevier.

Bertrand, M., Mullainathan, S. (2001). Are CEOs rewarded for luck? The ones without principals are. *The Quarterly Journal of Economics,* 116(3):901-932.

Bilén, D., Dreber, A., Johannesson, M. (2021). Are women more generous than men? A meta-analysis. *Journal of the Economic Science Association,* 7:1-18.

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

Bohnet, I., van Geen, A., Bazerman, M. (2016). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science,* 62(5):1225-1234.

Bohren, J.A., Imas, A., Rosenberg, M. (2019). The dynamics of discrimination: Theory and evidence. *American Economic Review,* 109(10):3395-3436.

Bolton, P., Dewatripont, M. (2005). *Contract Theory*. Cambridge, Massachusetts: MIT Press.

Bordalo, P., Coffman, K., Gennaioli, N., Shleifer, A. (2019). Beliefs about gender. *American Economic Review,* 109(3):739-773.

Brownback, A., Kuhn, M.A. (2019). Understanding outcome bias. *Games and Economic Behavior,* 117:342-360.

Buser, T., Gerhards, L., van der Weele, J. (2018). Responsiveness to feedback as a personal trait. *Journal of Risk and Uncertainy,* 56(2):165-192.

Campos-Mercade, P., Mengel, F. (2021). Non-Bayesian statistical discrimination. *Working Paper*.

Charness, G. (2004). Attribution and reciprocity in an experimental labor market. *Journal of Labor Economics,* 22(3):665-688.

Charness, G., Levine, D.I. (2007). Intention and stochastic outcomes: An experimental study. *The Economic Journal,* 117(522):1051-1072.

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

Coffman, K., Collis, M., Kulkarni, L. (2020). Stereotypes and belief updating. *Working Paper*.

Coffman, K., Exley, C.L., Niederle, M. (2021). The role of beliefs in driving gender discrimination. *Management Science*.

Coutts, A. (2019). Good news and bad news are still news: Experimental evidence on belief updating. *Experimental Economics,* 22(2):369-395.

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

Derks, B., Van Laar, C., Ellemers, N. (2016). The queen bee phenomenon: Why women leaders distance themselves from junior women. *The Leadership Quarterly,* 27(3):456-469.

Eccles, R.G., Crane, D.B. (1988). *Doing Deals: Investment Banks at Work*. Boston: Harvard Business Press.

Eckel, C., Gangadharan, L., Grossman, P.J., Xue, N. (2021). The gender leadership gap: Insights from experiments. In A. Chaudhuri (Ed.), *A Research Agenda for Experimental Economics*: Edward Elgar.

Edelson, M.G., Polania, R., Ruff, C.C., Fehr, E., Hare, T.A. (2018). Computational and neurobiological foundations of leadership decisions. *Science,* 361(6401).

Egan, M.L., Matvos, G., Seru, A. (forthcoming). When Harry fired Sally: The double standard in punishing misconduct. *Journal of Political Economy*.

Eil, D., Rao, J.M. (2011). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics,* 3(2):114-138.

Engelmann, D., Strobel, M. (2000). The false consensus effect disappears if representative information and monetary incentives are given. *Experimental Economics,* 3(3):241-260.

Engelmann, D., Strobel, M. (2012). Deconstruction and reconstruction of an anomaly. *Games and Economic Behavior,* 76(2):678-689.

Erkal, N., Gangadharan, L., Koh, B.H. (2020). Replication: Belief elicitation with quadratic and binarized scoring rules. *Journal of Economic Psychology,* 81:102315.

Erkal, N., Gangadharan, L., Koh, B.H. (2021). By chance or by choice? Biased attribution of others' outcomes when social preferences matter. *Experimental Economics*.

Erkal, N., Gangadharan, L., Xiao, E. (2021). Leadership selection: Can changing the default break the glass ceiling? *The Leadership Quarterly*.

Ertac, S. (2011). Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization,* 80(3):532-545.

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

Falk, A., Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior,* 54(2):293-315.

Fenske, J., Castagnetti, A., Sharma, K. (2020). Attribution bias by gender: Evidence from a laboratory experiment. *Working Paper*.

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

Gauriot, R., Page, L. (2019). Fooled by performance randomness: Overrewarding luck. *The Review of Economics and Statistics,* 101(4):658-666.

Gibbs, M.J., Merchant, K.A., Van der Stede, W.A., Vargus, M.E. (2009). Performance measure properties and incentive system design. *Industrial Relations,* 48(2):237-264.

Goldin, C., Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review,* 90(4):715-741.

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

Grether, D.M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly Journal of Economics,* 95(3):537-557.

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

Grossman, Z., Owens, D. (2012). An unlucky feeling: Overconfidence and noisy feedback. *Journal of Economic Behavior and Organization,* 84(2):510-524.

Gurdal, M.Y., Miller, J.B., Rustichini, A. (2013). Why blame? *Journal of Political Economy,* 121(6):1205-1247.

Hayes, R.M., Schaefer, S. (2000). Implicit contracts and the explanatory power of top executive compensation for future performance. *The RAND Journal of Economics,* 31(2):273-293.

Hernandez-Arenaz, I., Iriberri, N. (2019). A review of gender differences in negotiation. *Oxford Research Encyclopedia of Economics and Finance*.

Hossain, T., Okui, R. (2013). The binarized scoring rule. *The Review of Economic Studies,* 80(3):984-1001.

Jensen, K., Kovacs, B., Sorenson, O. (2018). Gender differences in obtaining and maintaining patent rights. *Nature Biotechnology,* 36(4):307-309.

Levin, J. (2003). Relational incentive contracts. *American Economic Review,* 93(3):835-857.

Mengel, F. (2021). Gender biases in opinion aggregation. *International Economic Review,* 62(3).

Mengel, F., Sauermann, J., Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association,* 17(2):535-566.

Miller, D.T., Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin,* 82(2):213-225.

Möbius, M.M., Niederle, M., Niehaus, P., Rosenblat, T.S. (2014). Managing self-confidence. *Working Paper*.

Niederle, M. (2016). Gender. In J.H. Kagel & A.E. Roth (Eds.), *The Handbook of Experimental Economics* (2:481-562). New Jersey: Princeton University Press.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature,* 37(1):7-63.

Régner, I., Thinus-Blanc, C., Netter, A., Schmader, T., Huguet, P. (2019). Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature Human Behaviour,* 3(11):1171-1179.

Sarsons, H. (2019). Interpreting signals in the labor market: Evidence from medical referrals. *Working Paper*.

Sarsons, H., Gërxhani, K., Reuben, E., Schram, A. (2021). Gender differences in recognition for group work. *Journal of Political Economy,* 129(1).

Swim, J.K., Sanna, L.J. (1996). He's skilled, she's lucky: A meta-analysis of observers' attributions for women's and men's successes and failures. *Personality and Social Psychology Bulletin,* 22(5):507-519.

Wolfers, J. (2007). Are voters rational? Evidence from gubernatorial elections. *Working Paper*.

List of Tables

Table 1: Timeline of decisions in the leadership task

Stage	Description
Before Stage 1	- Participants are divided into groups of three.
Stage 1	- All participants make investment decisions for five investment tasks. - Participants are informed that decisions are implemented for their group if they are the leader.
Between Stage 1 and Stage 2	- Participants are assigned the roles of "Leader" and "Member" (Evaluator). - Leader's gender is revealed to the group.
Stage 2	For each investment task: - Evaluators state their prior beliefs about the leader's decision. - Evaluators state their posterior beliefs conditional on each possible outcome of the investment. - Evaluators make discretionary payments to the leader conditional on each possible outcome of the investment.

Table 2: Investment tasks

Task	Net Payoff (ECU) in Stage 1					
	Investment X (High Effort)			Investment Y (Low Effort)		
	Succeeds	Fails	Expected	Succeeds	Fails	Expected
Task 1						
Leader	250	100	212.5	400	250	287.5
Each Evaluator	150	0	112.5	150	0	37.5
Task 2						
Leader	300	100	250	450	250	300
Each Evaluator	200	0	150	200	0	50
Task 3						
Leader	350	100	287.5	500	250	312.5
Each Evaluator	250	0	187.5	250	0	62.5
Task 4						
Leader	350	150	300	500	300	350
Each Evaluator	250	50	200	250	50	100
Task 5						
Leader	400	150	337.5	550	300	362.5
Each Evaluator	300	50	237.5	300	50	112.5

Only the returns of both investments to the leader and each evaluator vary across the tasks. The costs of each investment (200 ECU for Investment X and 50 ECU for Investment Y) are fixed for all five tasks. Similarly, the probabilities of each investment succeeding (0.75 for Investment X and 0.25 for Investment Y) are fixed for all five tasks. In addition to the return from the investment, the net payoff to the leader also includes his/her endowment (300 ECU) and the cost of the chosen investment. To illustrate, if the leader chooses Investment X in Task 1, then the cost of 200 ECU is deducted from the leader's endowment of 300 ECU, and the investment provides a return of 150 ECU if it succeeds (75% chance) and 0 ECU if it fails (25% chance). Hence, the expected net payoff for the leader in Stage 1 if s/he chooses Investment X is given by 300 (endowment) – 200 (cost) + (0.75 × 150 + 0.25 × 0) (expected return from Investment X) = 212.5 ECU, while the expected net payoff for each evaluator is given by (0.75 × 150 + 0.25 × 0) (expected return from Investment X) = 112.5 ECU.

Table 3: Probit regressions of leaders' effort choice

Variables	Dependent variable: = 1 if leader chooses high effort	
	(1)	(2)
Female leader	-0.025	-0.025
	(0.036)	(0.035)
% endowment transferred in DG	0.004	0.003
	(0.001)	(0.001)
# risky choices in RT	-0.009	-0.011
	(0.010)	(0.010)
High Return – Low Return	0.002	0.002
	(0.000)	(0.000)
Zero return if investment fails	0.065	0.066
	(0.022)	(0.022)
Individual controls	N	Y
Control for task order	Y	Y
Observations	1,750	1,750

Marginal effects of probit model reported. Robust standard errors clustered at the participant level in parentheses.

DG: Dictator Game; RT: Risk Task.

In column (2), we also control for participants' characteristics, which include their age, whether the participant is pursuing a major in economics, whether the participant is an undergraduate student, whether the participant is Australian, previous experience with economics experiments, and CRT score.

Table 4: OLS regressions of evaluators' prior belief that the leader has chosen high effort

	Dependent variable: Prior belief	
Variables	(1)	(2)
Female leader	1.413	1.151
	(2.250)	(2.319)
Chose high effort as leader	22.790	22.428
	(2.119)	(2.097)
# risky choices in RT	1.338	1.333
	(0.764)	(0.756)
High Return – Low Return	0.078	0.079
	(0.020)	(0.020)
Zero return if investment fails	0.853	0.878
	(1.607)	(1.610)
Constant	16.248	16.848
	(5.279)	(8.892)
Individual controls	N	Y
Control for task order	Y	Y
Observations	1,165	1,165
R-squared	0.186	0.190

Robust standard errors clustered at the participant level in parentheses.
RT: Risk Task.
In column (2), we also control for participants' characteristics, which include their age, whether the participant is pursuing a major in economics, whether the participant is an undergraduate student, whether the participant is Australian, previous experience with economics experiments, and CRT score.

Table 5: OLS regressions of evaluators' posterior belief that the leader has chosen high effort, by the leader's gender

	Dependent variable: Logit(posterior belief)		
	Female Leader	Male Leader	(1) vs. (2)
Variables	(1)	(2)	p-value
δ : Logit(prior belief)	0.602	0.442	0.119
	(0.050)	(0.089)	
γ_G : Good outcome × logit(p)	0.986	1.061	0.720
	(0.114)	(0.173)	
γ_B : Bad outcome × logit($1-p$)	1.088	1.256	0.434
	(0.125)	(0.174)	
$\gamma_G - \gamma_B$	-0.102	-0.195	
	(0.160)	(0.181)	
Observations	840	900	
R-squared	0.530	0.369	

Robust standard errors clustered at the participant level in parentheses. This analysis excludes participants classified as inconsistent or non-updaters. Since the regression specification estimates parameters of an augmented Bayes' rule, no controls can be included as the presence of any controls would invalidate the interpretation of the parameters. Moreover, since $I(\text{Good Outcome}) + I(\text{Bad Outcome}) = 1$, there is no constant term in the regression.

Table 6: OLS regressions of discretionary payments, by the leader's gender

	Dependent variable: Discretionary payments		
	Female Leader	Male Leader	(1) vs. (2)
Variables	(1)	(2)	p-value
Good outcome	49.87	40.325	0.575
	(12.895)	(11.386)	
Posterior belief	-0.139	0.614	0.012
	(0.207)	(0.222)	
Good outcome × Posterior belief	0.127	-0.157	
	(0.238)	(0.243)	
Constant	-69.473	-78.288	
	(30.750)	(23.925)	
Test of Good outcome = 100 × Belief			
p-value	0.002	0.302	
Individual controls	Y	Y	
Control for task order	Y	Y	
Observations	840	900	
R-squared	0.232	0.220	

Robust standard errors clustered at the participant level in parentheses. This analysis excludes participants classified as inconsistent or non-updaters.

In the regressions, we also control for participants' characteristics, which include their age, whether the participant is pursuing a major in economics, whether the participant is an undergraduate student, whether the participant is Australian, previous experience with economics experiments, and CRT score.

Table 7: OLS regressions of discretionary payments, by the leader's and evaluator's gender

	Dependent variable: Discretionary payments					
	Female Evaluator			Male Evaluator		
Variables	Female Leader (1)	Male Leader (2)	(1) vs. (2) p-value	Female Leader (3)	Male Leader (4)	(3) vs. (4) p-value
Good outcome	44.969	43.627	0.948	54.245	31.705	0.407
	(15.522)	(14.087)		(22.206)	(16.809)	
Posterior belief	-0.238	0.566	0.025	0.007	0.489	0.371
	(0.260)	(0.256)		(0.361)	(0.418)	
Good outcome × Posterior belief	0.178	-0.243		0.072	0.134	
	(0.284)	(0.256)		(0.404)	(0.447)	
Constant	-55.329	-101.383		-86.007	-59.515	
	(35.050)	(33.945)		(56.724)	(41.495)	
Test of Good outcome = 100 × Belief						
p-value	0.005	0.644		0.140	0.613	
Individual controls	Y	Y		Y	Y	
Control for task order	Y	Y		Y	Y	
Observations	450	500		390	400	
R-squared	0.237	0.201		0.265	0.301	

Robust standard errors clustered at the participant level in parentheses. This analysis excludes participants classified as inconsistent or non-updaters.

In the regressions, we also control for participants' characteristics, which include their age, whether the participant is pursuing a major in economics, whether the participant is an undergraduate student, whether the participant is Australian, previous experience with economics experiments, and CRT score.

List of Figures

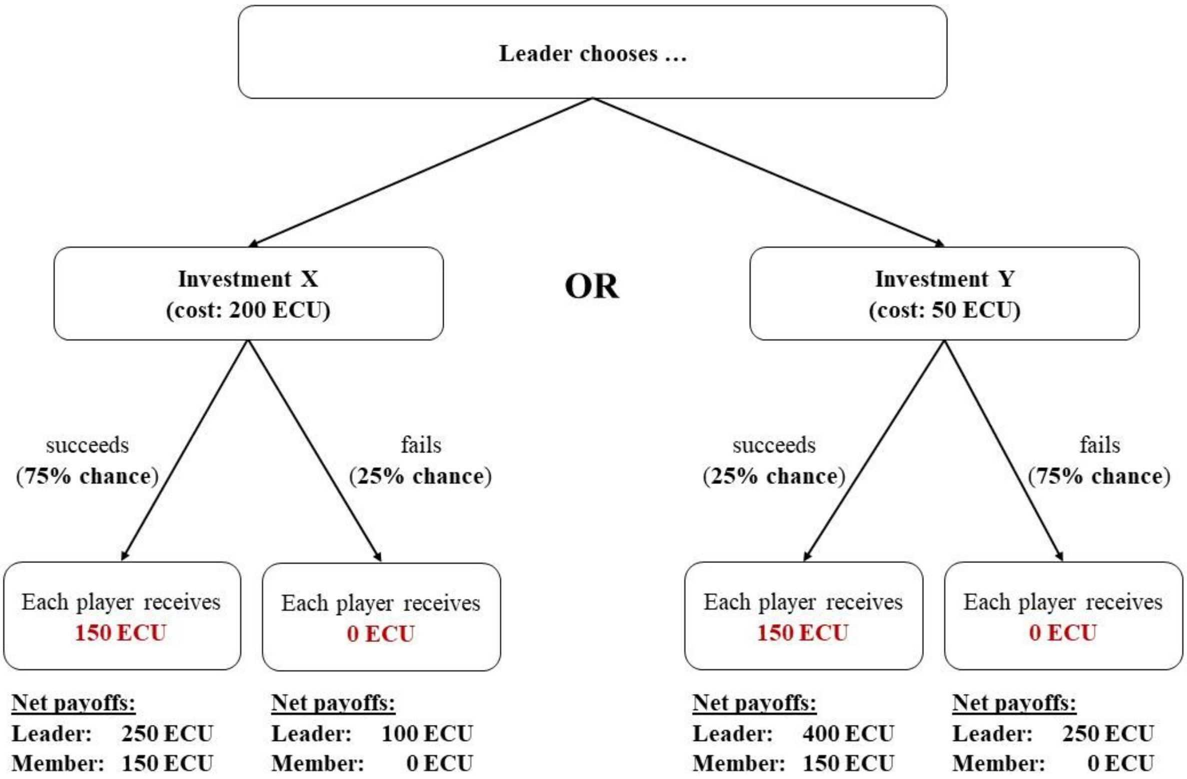
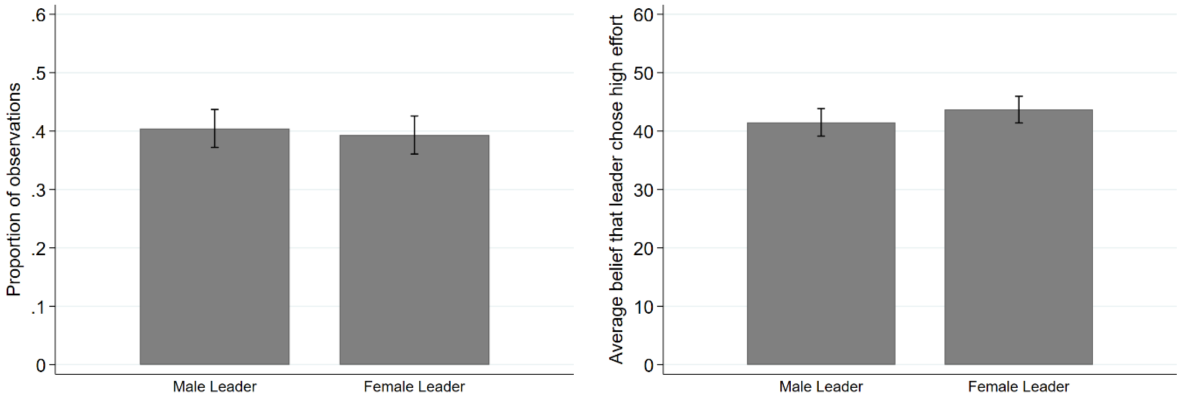


Figure 1: Example of an investment task as presented to participants



(a) Proportion of high effort choices

(b) Evaluators' average prior belief

Figure 2: Leaders' effort choice and evaluators' prior belief that the leader has chosen high effort, by the leader's gender

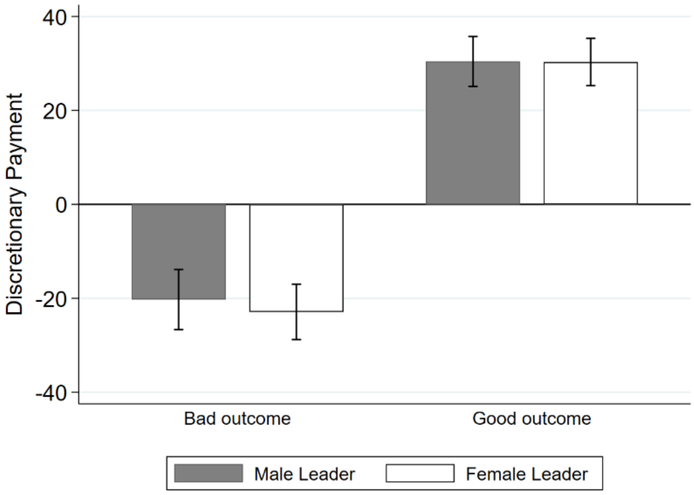
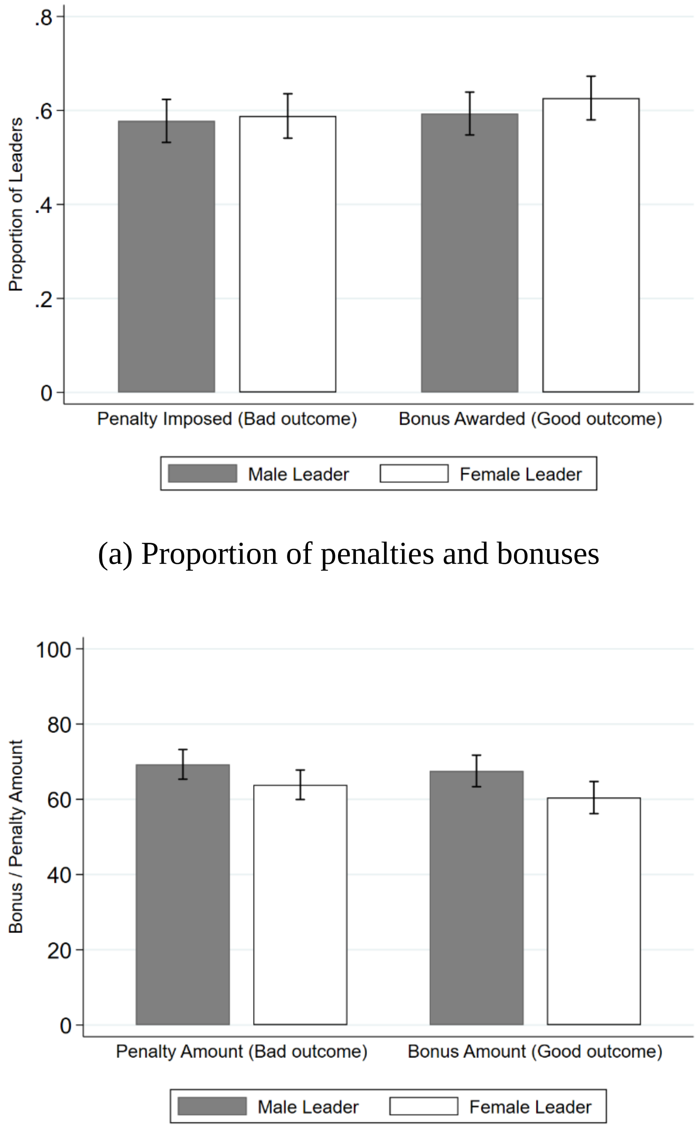


Figure 3: Evaluators' discretionary payments, by outcome and the leader's gender

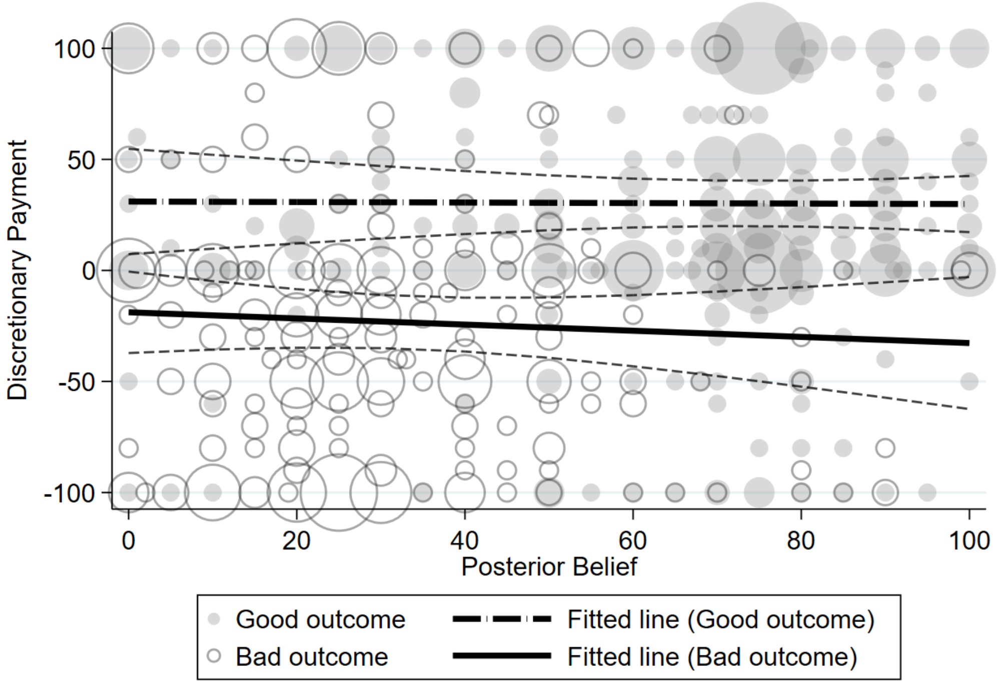


(a) Proportion of penalties and bonuses

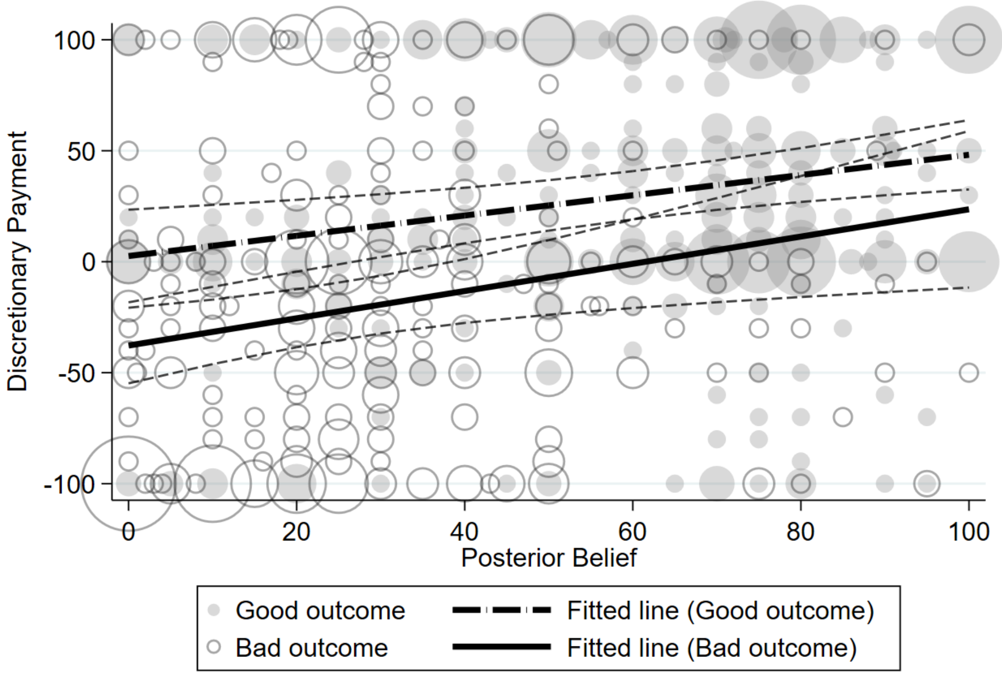
(b) Average penalty and bonus amounts

Figure 4: Evaluators' penalty and bonus decisions, by outcome and the leader's gender

Note: The average penalty and bonus amounts in panel (b) are computed conditional on a penalty being imposed and a bonus being awarded, respectively.



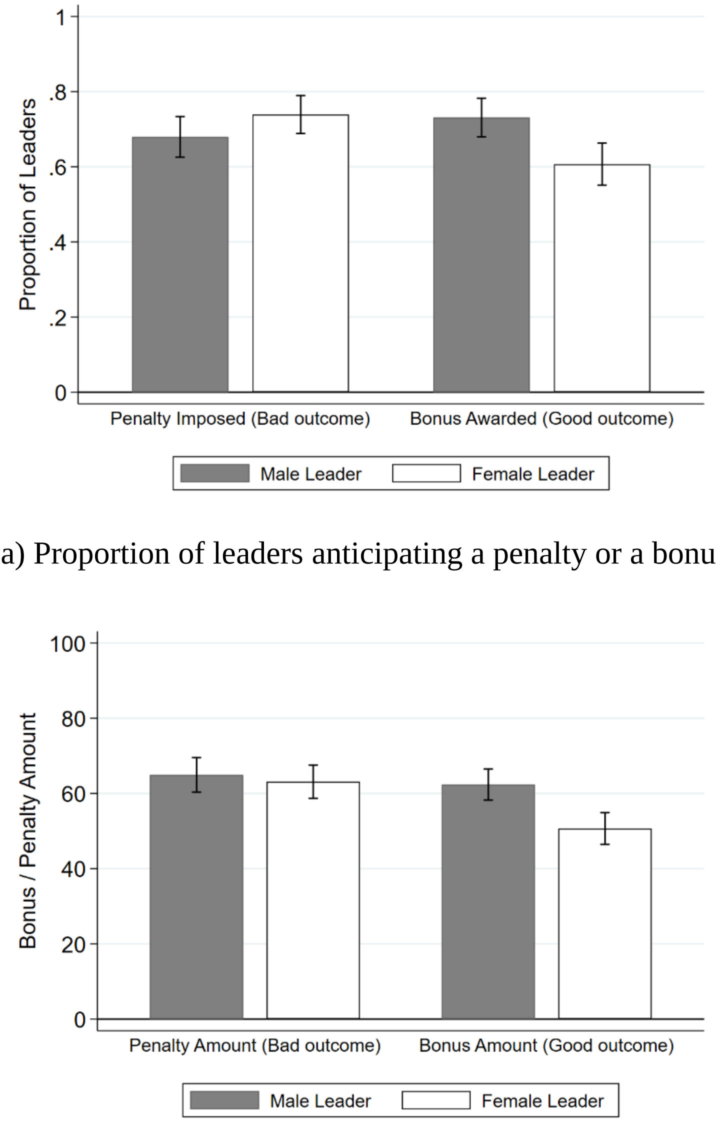
(a) Female Leaders



(b) Male Leaders

Figure 5: Discretionary payments against evaluators' posterior belief that the leader has chosen high effort and leader's outcomes.

Note: Dashed lines above and below each fitted line represent 95% confidence intervals.



(a) Proportion of leaders anticipating a penalty or a bonus

(b) Leaders’ expectations of the average penalty and bonus amounts

Figure 6: Leaders’ beliefs about penalty and bonus decisions, by outcome and the leader’s gender

Note: The average penalty and bonus amounts in panel (b) are computed conditional on a penalty being imposed and a bonus being awarded, respectively.

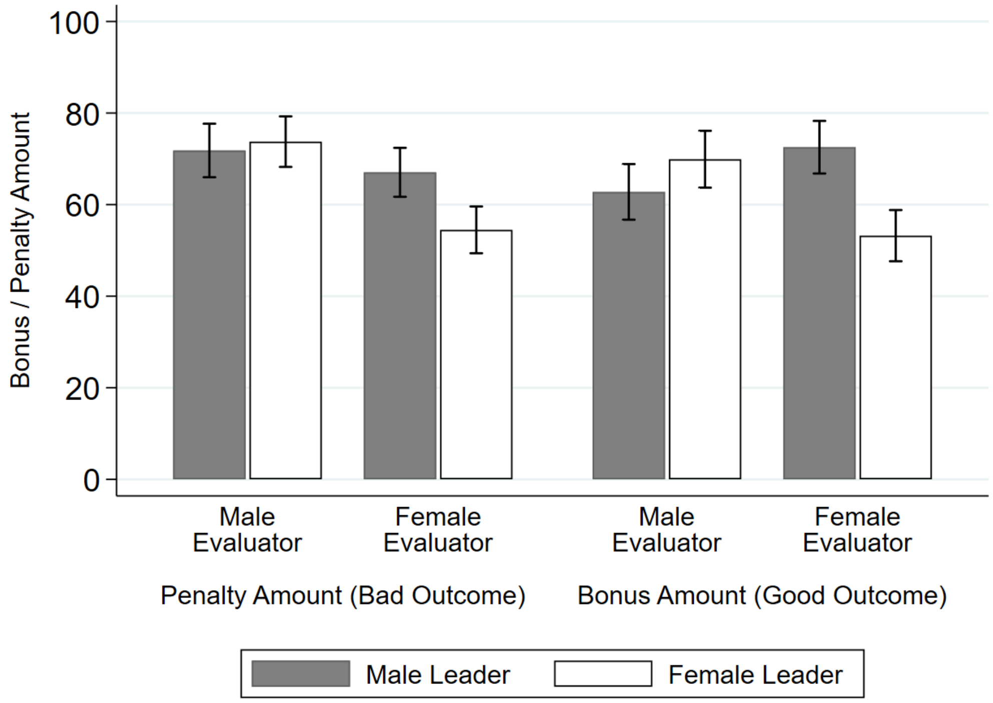


Figure 7: Average penalty and bonus amounts, by both the leader's and evaluator's gender

Note: The average penalty and bonus amounts are computed conditional on a penalty being imposed and a bonus being awarded, respectively

Appendix A Experimental Instructions

This appendix includes the instructions for the experiment reported in the paper. It also includes the practice questions for the leadership task, and the verbal instructions read out by the experimenter between Stage 1 and Stage 2 of the leadership task.

Overview of Experiment

Thank you for agreeing to take part in this study which is funded by the Australian Research Council. Please read the following instructions carefully. A clear understanding of the instructions will increase your earnings from the experiment.

There are two parts in today’s experiment: Part 1 and Part 2. We have provided you with instructions for Part 1, and we will explain them in greater detail shortly. We will hand out instructions for Part 2 at the end of Part 1. At the end of Part 2, you will be asked to complete a post-experimental questionnaire. Please be assured that all your responses and decisions will remain anonymous.

You will be paid for either Part 1 or Part 2 of today’s experiment. Hence, you should carefully consider all the decisions you make in today’s experiment as they may determine your earnings. Whether you will be paid for Part 1 or Part 2 will be randomly determined at the end of today’s session. You will be informed of the outcome of the experiment at the end of the session.

During the experiment, we will be using Experimental Currency Units (ECU). At the end of the session, we will convert the amount you earn into Australian Dollars (AUD) using the following conversion rate: 10 ECU = 1 AUD. You have already earned 50 ECU for completing the pre-experimental questionnaire.

Please do not talk to one another during the experiment, and please refrain from using your mobile phones and/or tablets. We require you to pay attention to the computer screen at all times. If anyone is found using their mobile phones and/or tablets, they may be asked to leave the experiment and may be excluded from future experiments. If you have any questions, please raise your hand and we will come over to answer your questions privately.

Do not turn over to the next page until you have been instructed by the experimenter to do so.

Part 1

You will participate in Part 1 in groups of three. There are two possible roles: Leader and Member. Each group will consist of one Leader and two Members.

Part 1 consists of two stages. In each stage, you will be asked to make decisions relating to five investment tasks. The following two sections explain the decisions that you will make in each stage.

(i) Stage 1: Investment decisions as Leaders

In Stage 1, you will be asked to make a decision for all five investment tasks assuming that you are the Leader of your group.

You will be informed whether you are the Leader at the end of Stage 1. Your decisions will be implemented if you are assigned to be the Leader of your group.

For each investment task, you will be given an endowment of 300 ECU. You will be asked to choose between two investment options. Your choice will affect both your payoff and your Members' payoffs. Each investment can either fail or succeed. The two investment options have different chances of success/failure, as well as different costs to you.

Specifically, the two investments are:

> **Investment X:** This investment will succeed with a 75% chance and fail with a 25% chance, and it costs you 200 ECU.
>
> **Investment Y:** This investment will succeed with a 25% chance and fail with a 75% chance, and it costs you 50 ECU.

Each investment provides you and your Members a high return if it succeeds, and a low return if it fails.

Your payoff and your Members' payoffs are calculated as follows:

> Payoff to you (Leader) = 300 ECU – Cost of investment + Returns on investment
>
> Payoff to each Member = Returns on investment

Note that the returns of the two investments may be <u>different</u> for each task, and this will therefore affect the final payoffs to you and your Members. However, the investments always provide a higher return if they succeed and a lower return if they fail. Please pay attention to these numbers on the screen for each task.

Figure 1 shows an example where Investment X and Investment Y provide you and each Member a return of 275 ECU if they succeed and 50 ECU if they fail. These numbers are shown in red.

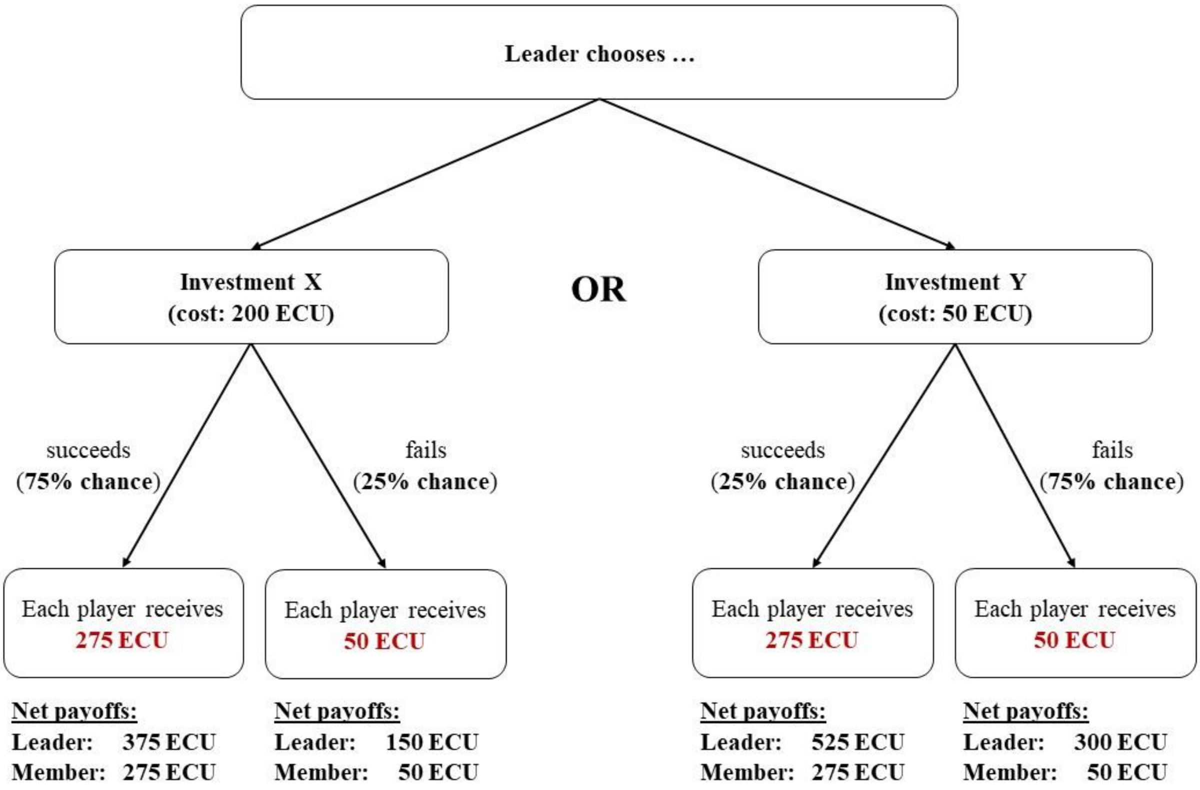


Figure 1: Investment Options in Part 1 (Example of a Task)

Example 1. Suppose in the task depicted in Figure 1, you choose Investment X. Then, the investment costs you 200 ECU, and will succeed with a 75% chance and fail with a 25% chance. If the investment succeeds, then you will receive (300 – 200 + 275) = 375 ECU and each Member will receive 275 ECU.

Example 2. Suppose in the task depicted in Figure 1, you choose Investment Y. Then, the investment costs you 50 ECU, and will succeed with a 25% chance and fail with a 75% chance. If the investment fails, then you will receive (300 – 50 + 50) = 300 ECU and each Member will receive 50 ECU.

The other Members of your group will never learn your investment decisions. At the end of the experiment, they will learn how much they have received from the chosen investment, but they will not learn your investment decision.

(ii) Stage 2: Members' predictions and decisions

At the beginning of Stage 2, you will be provided information about your groups and roles. Hence, you will be informed whether you are the Leader or a Member after you have completed Stage 1, and before Stage 2 begins.

Predictions of Leader's decisions: As a Member, you will be asked to predict your Leader's decisions in Stage 1. Specifically, we would like to know how likely it is in your opinion that the Leader has chosen Investment X in each of the five investment tasks in Stage 1.

For each investment task, the specific questions you will be asked are listed below.

> <u>Question 1</u>
> **How likely do you think it is that your Leader has chosen Investment X? Specifically, what is the chance out of 100 that s/he has chosen Investment X?**

In Question 2, you are given additional information. You are asked to evaluate the same question with this additional information.

> <u>Question 2(a)</u>
> **Suppose you are informed that the investment chosen by your Leader has succeeded. This gives you a high payoff.**
>
> Now consider whether your prediction will be higher than, lower than, or the same as the one you stated in Question 1.
>
> Specifically, given that the investment has succeeded, what is the chance out of 100 that s/he has chosen Investment X?
>
> <u>Question 2(b)</u>
> **Suppose you are informed that the investment chosen by your Leader has failed. This gives you a low payoff.**
>
> Now consider whether your prediction will be higher than, lower than, or the same as the one you stated in Question 1.
>
> Specifically, given that the investment has failed, what is the chance out of 100 that s/he has chosen Investment X?

For both questions, you will need to choose a number between 0 and 100. <u>A higher number means that you think your Leader is more likely to have chosen Investment X.</u>

For your payment, the computer will randomly select one of these two questions and you will be paid for your response to this question. If Question 2 is chosen for payment, then you will be paid for your answer to the scenario that corresponds to the actual outcome, i.e., you will be paid for Question 2(a) if the investment has succeeded or Question 2(b) if it has failed.

To determine your payment, we use a procedure which has been used in many other studies. We explain the procedure in detail, but what is most important is that this payoff structure is designed such that it is in your best interest to report your true belief about the chance that your Leader has chosen Investment X.

Your payment will be determined as follows. You will receive either 0 ECU or 200 ECU. Your chance of receiving 200 ECU depends on your prediction and the actual decision made by your Leader.

Specifically, your chance of receiving 200 ECU is determined by the following formulas:

Chance of receiving 200 ECU if Leader chose **Investment X**

$$= \left[1 - \left(\frac{100-\text{prediction}}{100}\right)^2\right] \times 100$$

Chance of receiving 200 ECU if Leader chose **Investment Y**

$$= \left[1 - \left(\frac{\text{prediction}}{100}\right)^2\right] \times 100$$

Suppose you state a high number as your prediction that your Leader chose Investment X. The formulas above imply that your chance of receiving 200 ECU is high if s/he chose Investment X, and your chance of receiving 200 ECU is low if s/he chose Investment Y. Hence, you should carefully consider how likely it is that your Leader chose Investment X.

To illustrate, suppose your prediction that your Leader chose Investment X is 100. Then, if s/he chose Investment X, your chance of receiving 200 ECU will be $\left[1 - \left(\frac{100-100}{100}\right)^2\right] \times 100 = 100$. If s/he chose Investment Y, your chance of receiving 200 ECU will be $\left[1 - \left(\frac{100}{100}\right)^2\right] \times 0$. Hence, your prediction should depend on whether you think your Leader is more likely to have chosen Investment X or Investment Y.

Here are two more examples explaining how your chance of receiving 200 ECU will be determined based on your prediction and the decision made by your Leader.

Example 1: Suppose you predict 70 as the chance that your Leader chose Investment X. At the end of the experiment, the computer reveals that s/he chose Investment X. Then, your chance of receiving 200 ECU will be $\left[1-\left(\frac{100-70}{100}\right)^2\right] \times 100 = 91$.

Example 2: In the above example, suppose your Leader chose Investment Y. Then, your chance of receiving 200 ECU will be $\left[1-\left(\frac{70}{100}\right)^2\right] \times 100 = 51$.

To determine whether you receive 200 ECU, the computer will randomly draw a number between 0 and 100. Each number between 0 and 100 is equally likely to be picked. If the number drawn by the computer is less than or equal to your chance of receiving 200 ECU as determined by the formulas above, then you will receive 200 ECU. Otherwise, you will receive 0 ECU. Hence, in Example 1 above, if the number randomly drawn by the computer is less than or equal to 91, then you will receive 200 ECU. Otherwise, you will receive 0 ECU.

In summary, your prediction will determine the chance that you receive 200 ECU. The closer your prediction is to the actual decision of your Leader, the higher your chance is of receiving 200 ECU.

Decisions to modify the Leader's payoff: After you state your predictions, you will be asked whether you would like to modify your Leader's payoff from Stage 1, given each possible outcome of the investment chosen by him/her.

Specifically, you may choose to either increase or decrease your Leader's payoff by an amount between 10 ECU and 100 ECU. You may choose any amount in multiples of 10 ECU within this range. You may also choose not to increase or decrease your Leader's payoff, i.e., you may choose to modify your Leader's payoff by 0 ECU.

You will be asked to make this decision both assuming that the investment chosen by your Leader has succeeded, and assuming that the investment chosen by your Leader has failed. Note that your earnings will not be affected by your decisions to increase or decrease your Leader's payoff.

At the end of the experiment, the decisions of one of the two Members within the group will be randomly selected to be implemented. Your Leader's payoff from Stage 1 will then be modified according to the decision that corresponds to the actual outcome of the investment.

Here is a scenario and two examples to illustrate how your decisions as a Member will affect your Leader's payoff from Stage 1.

Scenario: Suppose you choose to increase your Leader's payoff by 60 ECU if the investment succeeds, and decrease it by 80 ECU if the investment fails. The other Member of your group chooses to modify your Leader's payoff by 0 ECU if the investment succeeds, but increase it by 30 ECU if the investment fails.

Example 3: At the end of the experiment, suppose the computer reveals that the investment chosen by your Leader has failed, and that your decisions have been implemented. In this case, your Leader's payoff from Stage 1 will be decreased by 80 ECU.

Example 4: At the end of the experiment, suppose the computer reveals that the investment chosen by your Leader has succeeded, and that the decisions of the other Member have been implemented. In this case, your Leader's payoff from Stage 1 will not be modified.

Payment for Part 1

At the end of the experiment, the computer will randomly determine **one** of the five investment tasks for payment. For that randomly chosen investment task:

1. If you are the **Leader**, you will be paid according to your investment decision and the cost of that decision in Stage 1. Your payoff from Stage 1 may be modified based on your Members' decisions in Stage 2.

2. If you are a **Member**, the computer will randomly determine whether you will be paid for your Leader's investment decision in Stage 1 or your prediction of his/her decision in Stage 2.

Table 1 below summarizes the payoffs of the Leader and Members for each investment task.

Table 1: Payoffs to Leader and Members for each investment task of Part 1

	Paid for:	
	Investment Return	**Prediction**
Leader	Yes	Not applicable
Member	Either one but not both	

Summary

1. In Part 1, you will be divided into groups of three. There are two stages in Part 1. In each stage, you will be asked to make decisions relating to five investment tasks.

2. In Stage 1, you will be asked to make a decision for each investment task assuming that you are the Leader. As a Leader, you will be given an endowment of 300 ECU for each task and asked to choose between two investment options. Your choice will affect both your payoff and the payoffs of the Members you have been matched with. Your decisions in Stage 1 will be implemented only if you are assigned to be the Leader of your group.

3. The returns of Investment X and Investment Y may be different for each task. However, the investments always provide a higher return if they succeed and a lower return if they fail.

4. At the end of Stage 1, you will be provided information about your groups and roles. One participant in the group will be the Leader and the other two participants will be Members. You will be informed whether you are the Leader or a Member of your group after you have completed Stage 1 and before Stage 2 begins.

5. In Stage 2, if you are a Member, you will be asked to predict your Leader's decisions for the five investment tasks in Stage 1. For each investment task, you will be asked two questions.

   In Question 1, you will be asked to predict how likely it is in your opinion that your Leader has chosen Investment X. You will need to choose a number between 0 and 100. A higher number means that you think that s/he is more likely to have chosen Investment X.

   In Question 2, you will be asked the same question under two different scenarios: (i) assuming that the investment has succeeded; and (ii) assuming that the investment has failed. You should consider whether your prediction of your Leader's decision will be higher than, lower than, or the same as the one you stated in Question 1, given that you know the outcome of the investment chosen by him/her.

6. As a Member, the payoff structure used to determine your payment for your pre-dictions is designed such that it is in your best interest to report your true belief about the chance that your Leader has chosen Investment X.

7. After you state your predictions, you will be asked whether you would like to modify your Leader's payoff from Stage 1, given each possible outcome of the investment chosen by him/her.

   You may choose to either increase or decrease your Leader's payoff by an amount between 10 ECU and 100 ECU. You may choose any amount in multiples of 10 ECU within this range. You may also choose not to modify your Leader's payoff. Your earnings as a Member will not be affected by your decisions to modify your Leader's payoff.

8. At the end of the experiment, the computer will randomly select one of the five investment tasks for payment. For the randomly chosen investment task:

    (a) The Leader will be paid for their investment decision in Stage 1, and their payoff may be modified based on the Members' decisions in Stage 2.

    (b) Each Member will be paid either for their Leader's decision in Stage 1 or their prediction of the Leader's decision in Stage 2.

If you have any questions, please raise your hand and an experimenter will come to you to answer your questions privately. Otherwise, please proceed to answer the practice questions on your computer screen. The purpose of these practice questions is to make sure that you understand the experiment.

When you are ready to begin the practice questions, please press the button on your computer screen to launch the practice questions.

Part 2

You will participate in Part 2 in groups of two. The computer will randomly match you with one other person in the room. You will never learn the identity of your partner.

Each of you is given an endowment of 300 ECU, and you are asked to divide this amount between yourself and the person you are matched with.

At the end of today’s session, if Part 2 is picked for payment, then you will be paid either according to your decision or according to the decision made by your randomly matched partner. The computer will randomly determine whose allocation decision will be implemented.

Example. Suppose you choose to divide your endowment by keeping 200 ECU for yourself and giving 100 ECU to your matched partner. Your matched partner decides to keep 130 ECU and give 170 ECU to you. If, at the end of the experiment, the computer randomly determines that it is the allocation of your matched partner that gets implemented, then your payment will be 170 ECU and your matched partner’s payment will be 130 ECU.

Are there any questions? If not, we will proceed with Part 2.

Part 1: Practice Questions

(These are programmed on z-Tree.)

1. Which of the following statements is correct?
   (a) I will be paid for the decisions in both parts of the experiment today.
   (b) I will be paid for the decisions in either Part 1 or Part 2 of the experiment today.

   Answer: (b)

2. We will make decisions relating to 5 investment tasks in Part 1. If we are paid for Part 1, then we will be paid for our decisions for one of the 5 investment tasks.
   (a) True
   (b) False

   Answer: (a)

3. In Stage 1 of Part 1, everyone will make decisions as Leaders.
   (a) True
   (b) False

   Answer: (a)

4. In Stage 2 of Part 1,
   (i) I will learn whether I am the Leader or a Member of my group.
   (ii) everyone will make decisions as Members.

   (a) Both (i) and (ii) are correct.
   (b) (i) is correct but (ii) is incorrect.
   (c) (i) is incorrect but (ii) is correct.
   (d) Both (i) and (ii) are incorrect.

   Answer: (b)

5. Which of the following statements is correct?
   (a) The Members will be informed of the investment chosen by the Leader, but not the outcome of the investment.
   (b) The Members will be informed of the outcome of the investment chosen by the Leader, but not the investment chosen by him/her.
   (c) The Members will be informed of the investment chosen by the Leader, and the outcome of the investment.

   Answer: (b)

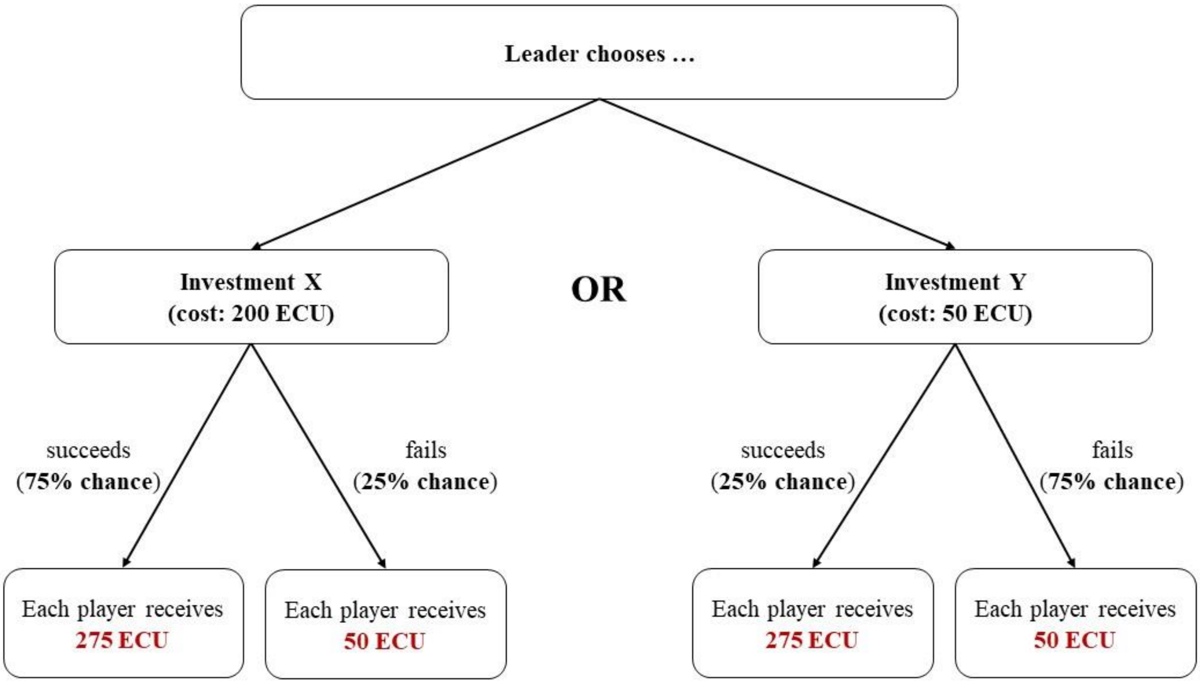
6. If I am a Member, then I will be paid for:
   (a) my Leader's investment decision in Stage 1 only.
   (b) my prediction of my Leader's investment decision in Stage 2 only.
   (c) both my Leader's investment decision in Stage 1 AND my prediction of his/her decision in Stage 2.
   (d) either my Leader's investment decision in Stage 1 OR my prediction of his/her decision in Stage 2, but not both.

   Answer: (d)

7. If I am a Member, I will be asked two questions. If I am paid for my predictions, then I will be paid accordingly to my responses to both questions.
   (a) This statement is true.
   (b) This statement is false. I will be asked only one question as a Member. If I am paid for my predictions, then I will be paid for my response to that question.
   (c) This statement is false. I will be asked two questions as a Member. However, if I am paid for my predictions, then I will be paid for my response to only one of the questions.

   Answer: (c)

Consider the investment options below.



8. Suppose the Leader chooses **Investment X**.

   At the end of the experiment, the computer randomly picks this task for payment and determines that the investment <u>fails</u>. What are the net payoffs of the Leader and each Member in Stage 1?

   Answer:
   Leader: 150
   Each Member: 50

9. Suppose the Leader chooses **Investment Y**.

   At the end of the experiment, the computer randomly picks this task for payment and determines that the investment <u>succeeds</u>. What are the net payoffs of the Leader and each Member in Stage 1?

   Answer:
   Leader: 525
   Each Member: 275

10. Suppose you are a Member. If you strongly believe that your Leader has chosen Investment Y, which of the following statements is true?

    (a) It is in my best interest to choose a high number as my prediction of the chance that my Leader has chosen Investment X.
    (b) It is in my best interest to choose a low number as my prediction of the chance that my Leader has chosen Investment X.
    (c) It is in my best interest to choose 50 as my prediction of the chance that my Leader has chosen Investment X.

    Answer: (b)

11. If I am a Member, then in Stage 2,
    (i) my decisions to increase or decrease my Leader’s payoff from Stage 1 will affect my own earnings.
    (ii) I can choose not to modify my Leader’s payoff from Stage 1.

    (a) Both (i) and (ii) are correct.
    (b) (i) is correct but (ii) is incorrect.
    (c) (i) is incorrect but (ii) is correct.
    (d) Both (i) and (ii) are incorrect.

    Answer: (c)

Experimenter Notes

Before we proceed to Stage 2, we will now announce your groups and roles. Please pay attention to your computer screens.

(LAUNCH NEXT SCREEN)

You can see on your screen the ID number assigned to you prior to this experiment.

Remember that you have been divided into groups of three. One participant in the group is the Leader. The other two participants are Members. In a few moments, you will be informed on your computer screens your group number, and whether you are the Leader or a Member.

To ensure that all participants have been assigned to a group of three, we will announce each group separately. When we call out your group number, please raise your hand (above the partition) so that I can see it.

To ensure that every group has a Leader, we will also announce the Leader in each group by calling out the last three digits of their ID number. If you are the Leader, when I call out the last three digits of your ID number, please loudly and clearly announce “Here”. Please only say “Here” and nothing else.

To maintain your anonymity, please remain seated and face your computer screens.

Does anyone have any questions? If not, we will now begin with Group 1.

(LAUNCH NEXT SCREEN)

If you are in Group X, you will see this information on your screens. Please raise your hand if you are in Group X.

Please put down your hands.

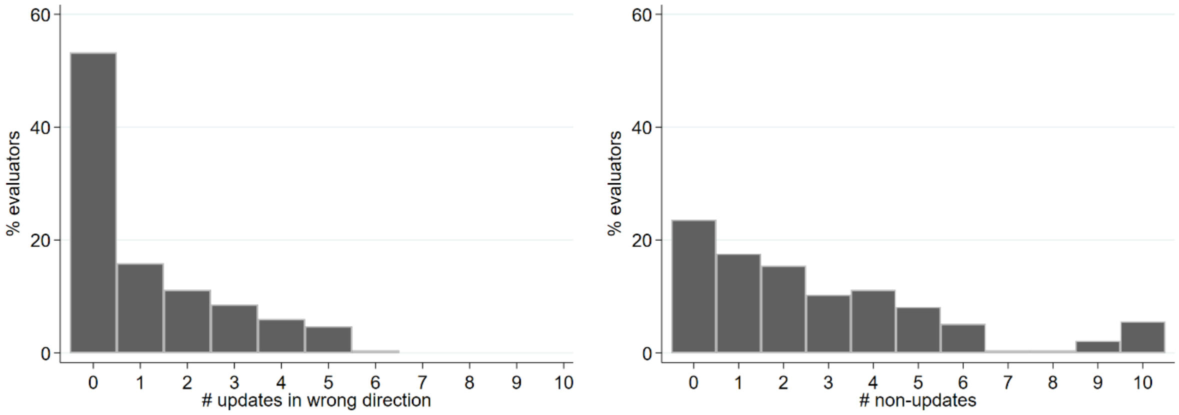
(LAUNCH NEXT SCREEN)

I will now announce the Leader. The Leader in Group X has an ID number ending in: XXX.

(AFTER ALL GROUPS REVEALED)

We will now proceed with Stage 2.

Appendix B Analyses of Updating Behavior and Discretionary Payments using Full Sample



(a) Inconsistent updates

(b) Non-updates

Figure B1: Distribution of inconsistent and non-updates by evaluators

Table B1: OLS regressions of evaluators' posterior belief that leader has chosen high effort, (i) by leader's gender and (ii) by leader's and evaluator's gender (full sample)

	Dependent variable: Logit(posterior belief)								
				Female Evaluator			Male Evaluator		
	Female Leader	Male Leader	p-value	Female Leader	Male Leader	p-value	Female Leader	Male Leader	p-value
Variables	(1)	(2)	(1) vs. (2)	(3)	(4)	(3) vs. (4)	(5)	(6)	(5) vs. (6)
δ : Logit(prior belief)	0.542	0.515	0.788	0.562	0.484	0.561	0.516	0.551	0.817
	(0.062)	(0.079)		(0.081)	(0.106)		(0.095)	(0.118)	
γ_G : Good outcome × logit(p)	0.787	0.804	0.923	0.795	0.902	0.717	0.776	0.704	0.725
	(0.108)	(0.143)		(0.174)	(0.241)		(0.135)	(0.155)	
γ_B : Bad outcome × logit($1-p$)	0.766	1.027	0.185	0.955	1.149	0.487	0.593	0.904	0.266
	(0.118)	(0.157)		(0.169)	(0.224)		(0.167)	(0.224)	
$\gamma_G - \gamma_B$	0.021	-0.224		-0.160	-0.247		0.183	-0.200	
	(0.152)	(0.160)		(0.221)	(0.221)		(0.218)	(0.237)	
Observations	1,160	1,170		560	590		600	580	
R-squared	0.425	0.400		0.457	0.388		0.390	0.418	

Robust standard errors clustered at the participant level in parentheses. This analysis includes the full sample. Since the regression specification estimates parameters of an augmented Bayes' rule, no controls can be included as the presence of any controls would invalidate the interpretation of the parameters. Moreover, since $I(\text{Good Outcome}) + I(\text{Bad Outcome}) = 1$, there is no constant term in the regression.

Table B2: Hurdle model estimations of penalty and bonus decisions conditional on outcomes, (i) at pooled level and (ii) by evaluators' gender (full sample)

	Penalty decisions (Bad outcome)		Bonus decisions (Good outcome)	
Variables	Proportion imposed (1)	Average amount (2)	Proportion awarded (3)	Average amount (4)
(a) Pooled				
Leader is female	0.029	-10.165	0.023	-13.708
	(0.055)	(4.612)	(0.056)	(4.857)
Posterior belief	-0.001	-0.145	0.001	0.127
	(0.001)	(0.080)	(0.001)	(0.074)
Constant		64.725		43.987
		(19.406)		(16.514)
Individual controls	Y	Y	Y	Y
Control for task order	Y	Y	Y	Y
Observations	1,165	694	1,165	676
R-squared		0.123		0.172
(b) Female Evaluators				
Leader is female	0.015	-13.726	0.067	-22.5
	(0.078)	(6.026)	(0.080)	(6.888)
Posterior belief	-0.000	-0.208	0.001	0.112
	(0.001)	(0.115)	(0.001)	(0.097)
Constant		72.678		43.192
		(23.414)		(19.967)
Individual controls	Y	Y	Y	Y
Control for task order	Y	Y	Y	Y
Observations	575	332	575	332
R-squared		0.127		0.181
(c) Male Evaluators				
Leader is female	0.050	-6.643	-0.027	-4.958
	(0.077)	(6.335)	(0.074)	(6.416)
Posterior belief	-0.001	-0.069	0.002	0.131
	(0.001)	(0.105)	(0.001)	(0.105)
Constant		51.642		47.257
		(27.775)		(27.188)
Individual controls	Y	Y	Y	Y
Control for task order	Y	Y	Y	Y
Observations	590	362	590	344
R-squared		0.226		0.241

Marginal effects of a probit model reported in (1) and (3). Robust standard errors clustered at the participant level in parentheses. This analysis includes the full sample.

In the regressions, we also control for participants' characteristics, which include their age, whether the participant is pursuing a major in economics, whether the participant is an undergraduate student, whether the participant is Australian, previous experience with economics experiments, and CRT score.

Table B3: OLS regressions of discretionary payments, (i) by leader's gender and (ii) by leader's and evaluator's gender (full sample)

	Dependent variable: Discretionary payments								
				Female Evaluator			Male Evaluator		
	Female Leader	Male Leader	p-value	Female Leader	Male Leader	p-value	Female Leader	Male Leader	p-value
Variables	(1)	(2)	(1) vs. (2)	(3)	(4)	(3) vs. (4)	(5)	(6)	(5) vs. (6)
Good outcome	43.323	38.649	0.720	25.923	45.155	0.261	63.061	35.226	0.120
	(9.186)	(9.384)		(12.684)	(11.906)		(12.171)	(13.508)	
Posterior belief	-0.048	0.381	0.080	-0.239	0.343	0.057	0.201	0.214	0.975
	(0.145)	(0.201)		(0.200)	(0.237)		(0.191)	(0.343)	
Good outcome × Belief	0.147	-0.009		0.424	-0.105		-0.188	0.132	
	(0.170)	(0.198)		(0.225)	(0.212)		(0.221)	(0.326)	
Constant	-59.661	-50.048		-32.830	-109.061		-88.288	-1.448	
	(23.025)	(21.922)		(31.673)	(30.709)		(33.163)	(31.070)	
Test of Good outcome = 100 × Belief									
p-value	< 0.001	0.979		0.006	0.687		0.018	0.652	
Individual controls	Y	Y		Y	Y		Y	Y	
Control for task order	Y	Y		Y	Y		Y	Y	
Observations	1,160	1,170		560	590		600	580	
R-squared	0.237	0.209		0.220	0.214		0.279	0.282	

Robust standard errors clustered at the participant level in parentheses. This analysis includes the full sample.

In the regressions, we also control for participants' characteristics, which include their age, whether the participant is pursuing a major in economics, whether the participant is an undergraduate student, whether the participant is Australian, previous experience with economics experiments, and CRT score.

Appendix C Additional Analyses

Table C1: OLS regressions of evaluators' posterior belief that the leader has chosen high effort, by both the leader's and evaluator's gender

	Dependent variable: Logit(posterior belief)					
	Female Evaluator			Male Evaluator		
	Female Leader	Male Leader	(1) vs. (2)	Female Leader	Male Leader	(3) vs. (4)
Variables	(1)	(2)	p-value	(3)	(4)	p-value
δ : Logit(prior belief)	0.554	0.362	0.149	0.695	0.572	0.215
	(0.062)	(0.112)		(0.076)	(0.127)	
γ_G : Good outcome × logit(p)	1.084	1.145	0.359	0.878	0.979	0.615
	(0.187)	(0.287)		(0.125)	(0.162)	
γ_B : Bad outcome × logit($1-p$)	1.162	1.308	0.501	0.998	1.167	0.698
	(0.181)	(0.238)		(0.162)	(0.276)	
$\gamma_G - \gamma_B$	-0.078	-0.163		-0.120	-0.188	
	(0.062)	(0.112)		(0.076)	(0.127)	
Observations	450	500		390	400	
R-squared	0.487	0.318		0.618	0.466	

Robust standard errors clustered at the participant level in parentheses. This analysis excludes participants classified as inconsistent or non-updaters. Since the regression specification estimates parameters of an augmented Bayes' rule, no controls can be included as the presence of any controls would invalidate the interpretation of the parameters. Moreover, since $I(\text{Good Outcome}) + I(\text{Bad Outcome}) = 1$, there is no constant term in the regression.

Table C2: Hurdle model estimations of penalty and bonus decisions conditional on outcomes

	Penalty decisions (Bad outcome)		Bonus decisions (Good outcome)	
Variables	Proportion imposed (1)	Average amount (2)	Proportion awarded (3)	Average amount (4)
Leader is female	0.001	-5.804	0.050	-10.534
	(0.066)	(5.085)	(0.064)	(5.604)
Posterior belief	-0.001	-0.172	0.001	0.057
	(0.001)	(0.099)	(0.001)	(0.097)
Constant		74.34		55.03
		(23.450)		(20.086)
Individual controls	Y	Y	Y	Y
Control for task order	Y	Y	Y	Y
Observations	870	507	870	530
R-squared		0.154		0.139

Marginal effects of a probit model reported in (1) and (3). Robust standard errors clustered at the participant level in parentheses. This analysis excludes participants classified as inconsistent or non-updaters.
In the regressions, we also control for participants’ characteristics, which include their age, whether the participant is pursuing a major in economics, whether the participant is an undergraduate student, whether the participant is Australian, previous experience with economics experiments, and CRT score.

Table C3: Hurdle model estimations of penalty and bonus decisions conditional on outcomes, by evaluator's gender

	Penalty decisions (Bad outcome)		Bonus decisions (Good outcome)	
Variables	Proportion imposed (1)	Average amount (2)	Proportion awarded (3)	Average amount (4)
(a) Female Evaluators				
Leader is female	-0.016	-10.082	0.14	-20.365
	(0.087)	(6.843)	(0.084)	(7.688)
Posterior belief	-0.002	-0.271	0.000	0.038
	(0.001)	(0.136)	(0.001)	(0.126)
Constant		76.737		49.046
		(29.274)		(24.068)
Individual controls	Y	Y	Y	Y
Control for task order	Y	Y	Y	Y
Observations	475	266	475	279
R-squared		0.140		0.147
(b) Male Evaluators				
Leader is female	0.023	-5.469	-0.102	0.646
	(0.099)	(7.498)	(0.086)	(8.263)
Posterior belief	-0.001	0.019	0.003	0.060
	(0.002)	(0.150)	(0.001)	(0.146)
Constant		65.541		67.079
		(36.874)		(37.493)
Individual controls	Y	Y	Y	Y
Control for task order	Y	Y	Y	Y
Observations	395	241	395	251
R-squared		0.275		0.239

Marginal effects of a probit model reported in (1) and (3). Robust standard errors clustered at the participant level in parentheses. This analysis excludes participants classified as inconsistent or non-updaters.

In the regressions, we also control for participants' characteristics, which include their age, whether the participant is pursuing a major in economics, whether the participant is an undergraduate student, whether the participant is Australian, previous experience with economics experiments, and CRT score.