Department of Economics Discussion Papers
ISSN 1473-3307

Discrimination in Evaluation Criteria: The Role of Beliefs versus Outcomes

Nisvan Erkal, Lata Gangadharan, and Boon Han Koh

Paper number 23/16

Discrimination in Evaluation Criteria: The Role of Beliefs versus Outcomes[*]

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

December 2023

Abstract

Using incentivized experiments, we investigate whether different criteria are used in evaluating male and female leaders when outcomes are determined by unobservable choices and luck. Evaluators form beliefs about leaders' choices and make discretionary payments. We find that while payments to male leaders are determined by both outcomes and evaluators' beliefs, those to female leaders are determined by outcomes only. We label this new source of gender bias as the *gender criteria gap*. Our findings imply that high outcomes are necessary for women to get bonuses, but men can receive bonuses for low outcomes as long as evaluators hold them in high regard.

Keywords: Gender gaps; Discrimination; Evaluation criteria; Biases in beliefs; 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, Vera te Velde, Nina Xue, Xiaojian Zhao, seminar participants at the Centre for Behavioural and Experimental Social Science (CBESS) at the University of East Anglia, Institute of Economic Growth, New Zealand Economics eSeminar Series, University of Hawaii, Vienna University of Economics and Business, University of Pittsburgh, Utrecht University, Durham University, Bocconi University, and participants at the Workshop on Identity at TWI (2023), 2022 Maastricht Behavioral Economic Policy Symposium (MBEPS), 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).

[†] 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.

[§] Corresponding author. Department of Economics, University of Exeter, EX4 4PU, United Kingdom. b.koh@exeter.ac.uk.

1 Introduction

Research on gender discrimination has mainly focused on environments where ability shapes individuals' outcomes and has shown that beliefs influenced by gender stereotypes about one's ability can result in discriminatory outcomes. Implicit in the research which investigates belief-based discrimination (see, e.g., Bohren, Imas, and Rosenberg, 2019; Coffman, Exley, and Niederle, 2021) is the assumption that beliefs play a role in the determination of wages, promotion decisions, and other workplace-related outcomes. In this paper, we focus on discrimination in the evaluation of leaders and explore to what extent discretionary payments to leaders are based on the beliefs formed about the leaders' actions. We show that instead of beliefs, gender biases may exhibit themselves in the criteria used for the evaluation of male and female leaders.

Leadership inherently involves assuming responsibility for the outcomes of others (Ertac and Gurdal, 2012; Edelson et al., 2018). Accordingly, we consider a setting where a decision maker in a position of power has to take an action which determines the payoffs of a group of individuals, and where outcomes are determined by a combination of unobservable actions and luck. Such environments are pervasive in many organizations, where evaluators of leaders face the challenging task of assessing them based on the merits of the actions taken without being influenced by the outcomes of those actions. Outcomes, which are observable, act as signals and can assist evaluators in updating their beliefs about the leaders' actions. In the updating process, evaluators have to assess what role unpredictable or unforeseeable circumstances (versus actions) have played in determining the outcomes.

Evaluators often make renumeration decisions based on their evaluation and determine the discretionary payments received by the leaders. Such discretionary rewards can take the form of bonuses or pay increments and are common features of remuneration packages offered by many organizations. In the conceptual framework we develop, we assume that leaders can be rewarded (or punished) for both their intentions and outcomes (which may be high or low). Evaluators' beliefs represent their perceptions of the leaders' intentions.

We collect data on the evaluators' prior and posterior beliefs. This allows us to analyze to what extent evaluators base their payment decisions on their beliefs about the actions taken by their leaders and to what extent is their evaluation also influenced by the outcome itself.[1]

[1] 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).

We also examine if, for the same outcome or for the same beliefs about the leader's intentions, women are evaluated differently than men. In other words, we study the role beliefs versus outcomes play in the evaluation process, and whether these roles differ by gender. If beliefs or outcomes play differential roles in the evaluation of (i.e., payments received by) male and female leaders, this suggests that male and female leaders are subjected to different evaluation criteria by the evaluators. This may cause discretionary payments to differ by gender *even if evaluators hold the same beliefs* about the choices made by male and female leaders.

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 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 is not known, making it difficult to ascertain the process of belief formation. Experimental methods provide us with more precise and reliable measures of key variables, such as prior and posterior beliefs, how they compare to the Bayesian benchmark, and importantly, how they vary with gender. Such an exploration is important for establishing the drivers of gender differences and for developing effective strategies against them.

In the experiment, individuals are divided into groups of three. They all make investment choices on behalf of the group before one of them is assigned to be the leader of the group while the other two are appointed as evaluators. The leader's investment decision is implemented for the group and the leader's gender is revealed to the group. The outcome for the group depends on both the leader's choice, which is unobservable to the evaluators, and luck. Leaders face a trade-off while making their investment decisions. Specifically, a high investment choice leads to a higher probability of a high outcome for the group, but it comes at a higher private cost to the leader. 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. They then make discretionary payments (which may be positive or negative) to the leaders.

Our experimental design mirrors many leadership situations where decision makers face a trade-off between maximizing their own payoff and those of other individuals. Consider, for instance, political leaders who are expected to engage in prosocial activities that impact the welfare of voters who subsequently evaluate the leaders' actions, or CEOs whose actions impact the payoffs or reputation of board members who then decide on their compensation.[2] Effective governance in such situations is as much about competency as about the choices made.[3] Since the choices made are influenced by the decision maker's prosocial preferences, both prosocial preferences and ability are important in producing outcomes which enhance social welfare. The importance of prosocial preferences for leadership outcomes is what we emphasize in our experimental design.

We conjecture that gender discrimination in the discretionary payments awarded to leaders may arise due to gender differences in beliefs and/or gender differences in evaluation criteria (i.e., how much emphasis is placed on the leader's outcomes and the evaluators' beliefs). According to traditional gender stereotypes, prosocial actions are congruent to the expectations from women (see, e.g., Solnick, 2001; Heilman and Chen, 2005; Aguiar et al., 2008; Brañas-Garza, Capraro, and Rascón-Ramírez, 2018).[4] Hence, if evaluators hold the belief that female leaders are more likely to make an altruistic choice, they may want to reward them with higher discretionary payments. Alternatively, evaluators may place different emphasis on the outcomes of leaders depending on their gender. For instance, if evaluators are conditioned to think that favorable outcomes from men are worthy of higher financial renumeration than those from women, then discretionary payments may be higher for male leaders.

Our results reveal gender differences in discretionary payments. Conditional on receiving a bonus or a penalty, female leaders receive lower bonuses for high outcomes (successes) and lower penalties for low outcomes (failures). Strikingly, we do not detect any gender differences in evaluators' beliefs about leaders' investment choices. We conjecture that this is because evaluators do not expect male and female leaders to act differently when they know that their actions may affect the discretionary payments they may receive. However, investigating the determinants of discretionary payments, we detect a *gender criteria gap*. While male leaders' discretionary payments are determined by both the evaluators' beliefs of their investment choices and outcomes, female leaders' discretionary payments are predominantly determined by outcomes. Hence, different criteria are used in the evaluation of male and female leaders.

[2] In addition to leadership, prosocial motivation plays an important role in many other jobs in the economy (Bowles and Polanía-Reyes, 2012; Besley and Ghatak, 2018). For example, teachers and doctors inherently assume responsibility for the outcomes of others.

[3] This is why several researchers in political science and economics emphasize the relationship between accountability and good governance (Persson, Roland, and Tabellini, 1997; Alt and Lassen, 2003; Lederman, Loayza, and Soares, 2005).

[4] In their survey papers, Croson and Gneezy (2009), Niederle (2016), and Bilén, Dreber, and Johannesson (2021) report that in practice, the relative prosociality of women is a context-dependent phenomenon.

Our results thus offer a different explanation for the role played by beliefs in creating gender biases in the evaluation of leaders. By providing evidence for the presence of a gender criteria gap, the findings from our research imply that policies designed to counteract gender discrimination should not only target biases in beliefs. Beliefs are important in our setup not because they are biased, but because they play differential roles in the determination of male and female leaders' discretionary payments.

2 Related Literature

Gender discrimination, both the documentation of its evidence and the exploration of its causes, has been a topic of significant research. Several papers in this literature provide evidence for the existence of 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).[5] We contribute to this literature by exploring gender differences in evaluation criteria. The observed differences in evaluation may come from biased beliefs, or as we posit, it can be that different criteria are being used for different genders. We investigate this issue in a context where we would not expect to observe unfavorable belief formation against women as the leaders' decisions are shaped by their social preferences.

Our study contributes to the emerging literature on the role of beliefs in gender discrimination. Emphasizing the role of stereotypes and belief formation in gender discrimination, 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.[6] Barron et al. (2022) distinguish between explicit and implicit belief-based discrimination, and find evidence of both. Albrecht et al. (2013) and Campos-Mercade and Mengel (2023) consider how irrational updating of beliefs, in the sense of conservatism, can result in gender discrimination. Erkal, Gangadharan, and Koh (2023) examine gender differences in belief updating in a leadership setup where social preferences play a role, but in contrast to the current paper, they consider a setting without discretionary payments. They find gender differences in the attribution of low outcomes but not high outcomes, with the low outcomes of male (female) leaders being

[5] Other explanations for gender gaps include 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, 2022).

[6] 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.

attributed more to their selfish decisions (bad luck).[7] In contrast, we find no gender differences in the evaluators' beliefs when the evaluators make discretionary payment decisions. As our conceptual framework illustrates, discretionary payments can influence the motivations of both leaders and evaluators. The payments may cause leaders to make more prosocial decisions since they act as a channel of reciprocity for the evaluators. Our results reveal that the presence of discretionary payments is sufficient to eradicate gender biases in beliefs. This suggests that evaluators expect male and female leaders to behave similarly in the presence of monetary incentives. However, we show that there are gender biases in the emphasis given to beliefs versus outcomes in evaluation. Hence, we contribute to the literature on beliefs and discrimination by showing that there are gender differences in the weight given to beliefs in the evaluation of leaders.

Our study is also 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 reacts more strongly to intentions than outcomes. As different from the literature, intentions are not directly observed in our context and evaluators form beliefs about the leaders' intentions. Our study contributes to this literature by showing there are gender differences in the role beliefs versus outcomes play in the determination of discretionary payments.

3 Experimental Design

The main task in the experiment is a leadership task which consists of two stages.[8]

3.1 Leadership Task – Stage 1: Investment decision

In Stage 1, all participants make investment decisions. They are informed that they have been assigned to a group of three, that they will remain in the same group for the entire task, and that once these investment decisions are made, their roles within the group will be determined randomly. One person will be assigned to be the leader and the other two group members will

[7] See the related literature in psychology which investigates gender differences in the attribution of outcomes to effort, skill, and luck (e.g., Swim and Sanna, 1996), which shows that positive outcomes of men are more likely to be attributed to their ability while those of women are more likely to be attributed to luck.

[8] Instructions can be found in Appendix A.

be assigned to be evaluators (labelled as “members” in the experiment). All participants are told to make investment decisions assuming that they will be the leader. Their decisions are implemented for their group if they are assigned to be the leader. Evaluators do not learn the leader’s decisions, but they are told the outcome of the investment.

When making decisions as the leader, participants are endowed with 300 Experimental Currency Units (ECU) to cover the cost of investing in one of two options: Investment X or Investment Y.[9] The leader pays the investment cost, but each group member (including the leader) receives the same return from the investment. Both investment options can either fail (leading to a low return) or succeed (leading to a high return). Investment X (Investment Y) costs the leader 200 ECU (50 ECU) and yields a success probability of 0.75 (0.25).

Participants complete five investment tasks with different parameterization as shown in Table 1. The costs and probabilities of success 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. The expected return to the leader is always higher under Investment Y, but the expected return to each evaluator is always higher under Investment X. 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

Participants are randomly assigned to groups of three with either a female leader or a male leader. After participants make their decisions in Stage 1 and before Stage 2 begins, they are informed on their computer screens of their group assignment and their roles within the group.

The leader’s gender is revealed to the evaluators following the approach described in Bordalo et al. (2019). 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”. This enables the leader’s gender to be discreetly revealed to the evaluators. The greeting is brief and standardized across all leaders. We also ask evaluators of each group to raise their hands to avoid any obvious attention to the leader’s announcement. Since the participants are seated in 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

[9] 10 ECU = 1 AUD.

corresponding to their leader's gender. Using this protocol, differences in evaluators' evaluations can be attributed to differences in the leader's gender.[10]

3.3 Leadership Task – Stage 2: Elicitation of 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 *prior* belief of the likelihood that the leader has chosen Investment X. Next, they are asked to report their *posterior* beliefs of the likelihood that the leader has chosen Investment X conditional on the investment being successful and unsuccessful. 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 incentivized using the binarized scoring rule (Hossain and Okui, 2013; Erkal, Gangadharan, and Koh, 2020).

In each investment task, after the evaluators state their beliefs, they can choose to adjust the leader's payoff from Stage 1 by choosing a discretionary payment between -100 ECU and 100 ECU in multiples of 10 ECU. Evaluators make two discretionary payment decisions for each investment task, one conditional on the investment being successful and another conditional on the investment failing. To make each evaluator's decision potentially consequential and to minimize free riding by evaluators on each other's decisions, the choice of one of the two evaluators is randomly chosen to be implemented. The payment made to the leader is based on this evaluator's decision corresponding to whether the leader's investment failed or succeeded. The evaluators' payoffs are not affected by their discretionary payment decisions.

3.4 Procedures

All sessions were conducted at the Experimental Economics Laboratory at the University of Melbourne (E^2MU) and programmed using z-Tree (Fischbacher, 2007). Participants were university students recruited across different disciplines using ORSEE (Greiner, 2015).

Each participant was invited to complete a pre-experimental questionnaire on Qualtrics before attending the session. The pre-experiment questionnaire included basic demographic questions. 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 and evaluator roles in the investment task. In majority of the cases (80%), each leader was matched with one male evaluator and one female evaluator.

[10] 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. In our analysis, we categorize the data according to the evaluators' predictions of the leader's gender.

While the gender of the leader is common knowledge in the group, both the leader and evaluators are not aware of the gender composition of the evaluators in the group.

Participants were provided with printed instructions to the leadership task and answered a number of comprehension questions. A summary of the instructions was then read aloud by the experimenter. In the leadership task, the five investment tasks were randomized across sessions. After completing the leadership task, subjects participated in a dictator game in groups of two. Each participant was endowed with 300 ECU which they were asked to allocate between themselves and their matched partner. Within each pair, one participant's decision was randomly chosen at the end of the session to determine the earnings from the dictator game for both participants. Participants also completed a questionnaire that included questions relating to their decisions in the experiment, as well as incentivized cognitive reflection (CRT) and risk-elicitation tasks, which we use to control for their cognitive ability and risk preferences.

Participants were paid for either the leadership task or for the dictator game, randomly chosen, as well as the incentivized tasks in the questionnaire. If participants were paid for the leadership task, then the leaders were paid according to the decision they made in a randomly chosen investment task in Stage 1 and any discretionary payments given to them by one of the two evaluators in Stage 2. Evaluators were paid either for their leader's investment decision in Stage 1 or for their reported beliefs in Stage 2. Participants earned 39.78 AUD on average.

We collected data from 351 participants. All 351 participants made investment decisions as leaders in the experiment before the roles were revealed. After the roles were revealed, 234 evaluators reported their beliefs and made discretionary payment decisions for 117 leaders.[11] Our power calculations suggest that we are able to detect a 0.168 and 0.195 standard deviation difference in prior beliefs and payoff adjustments, respectively, between female and male leaders. Simulations using data from Erkal, Gangadharan, and Koh (2022) also reveal that we are able 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_H or γ_L based on the econometric framework presented in Section 5.2.1).

4 Conceptual Framework

In this section, we illustrate how we conceptualize the process of performance evaluation in the context of our experiment. This will guide our analysis in the next section.

[11] One participant (an evaluator) misreported their ID number, so their data is dropped from the analysis.

4.1 Leaders

In line with our experimental design, we consider an environment where each leader makes a discrete investment choice of X or Y on behalf of a group of N players. Since leaders' investment 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 which is private information. 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 investment 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 investment cost is solely borne by the leader. Output is probabilistic and the leader's investment choice affects the probability of a high-output realization (Q_H). Specifically, assume that an investment choice of X leads to Q_H with probability $p \in (0.5,1)$ and a choice of Y leads to Q_H with probability $(1-p)$. Hence, investment choice X (which is more costly for the leader) leads to a high output level with a higher probability. The investment cost $c \in \{c_X, c_Y\}$ is deducted from an initial endowment $\omega \geq c_X > c_Y$ that the leader receives.

For a given outcome realization and discretionary payment, we write leader i's utility as:

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

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. Δ_j stands for the discretionary payment that evaluator j pays to leader i. Each leader makes the investment 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).

4.2 Evaluators

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. This implies discretionary payments cannot be motivated by an incentive to change the future behavior of leaders, so 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 leader's underlying intentions and the consequences of the leader's action.[12] Making this distinction and considering both factors are important given the evidence from the literature which shows that reciprocal behavior can be shaped by both intentions (see, e.g., Charness, 2004; Charness and Levine, 2007) and outcomes (e.g., Gurdal, Miller, and Rustichini, 2013; Brownback and Kuhn, 2019).

We assume that evaluators will be motivated to give the leader a bonus ($\Delta_j > 0$) or a penalty ($\Delta_j < 0$) depending on their perception of the leader's kindness. Let φ_i denote evaluator j's perception of leader i's kindness, and $\rho_j > 0$ represent the reciprocity preference of evaluator j. 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 the discretionary payments, and it depends on the evaluator's belief about the leader's intentions (investment choice) and the realized output. We assume the evaluator's belief about the leader's intention is equivalent to the evaluator's belief about the leader's type. The evaluator updates his/her prior belief after observing the output. Let $\sigma_j(X|Q)$ stand for the updated (posterior) belief that evaluator j has about the leader choosing Investment X after observing the output Q.

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

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.

The kindness term indicates that if two evaluators care about intentions only (i.e., $\theta_j = 0$ and $\beta_j > 0$) and if they have the same posterior beliefs, then they will choose the same

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

discretionary payment irrespective of the outcome. On the other hand, if two evaluators care about outcomes only (i.e., $\beta_j = 0$ and $\theta_j > 0$) and if they observe different outcomes, then they will choose different discretionary payments even if their posterior beliefs are the same. Hence, in this framework, 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 independent of the beliefs.

After the output is realized, we assume evaluator j chooses Δ_j to maximize the following utility function:

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

where $\varphi_i = \theta_j(1_{Q=Q_H} - 1_{Q=Q_L}) + \beta_j(2\sigma_j(X|Q) - 1)$ as defined above, and $c(\Delta_j)$ stands for the cost of making a discretionary payment decision. The cost of making a discretionary payment decision may be monetary or non-monetary (e.g., psychological). We assume that $c(\Delta_j)$ is increasing and convex in Δ_j.

4.3 Equilibrium and Hypotheses

We use Perfect Bayesian Equilibrium as the equilibrium concept. In equilibrium, sufficiently altruistic leaders choose Investment X. Evaluators form beliefs about the leaders' type knowing that sufficiently altruistic leaders will choose Investment X.[13] Their payments take the form of bonuses if $\varphi_i > 0$ and penalties if $\varphi_i < 0$. Hence, the bonus may be motivated by a higher belief and/or a high output realization.

Our goal is to investigate whether evaluators use different criteria in the determination of the payments for male and female leaders. In our design, this corresponds to investigating whether θ_j and β_j take on different values depending on the leader's gender. We refer to the gender difference we may observe in the criteria used as the *gender criteria gap*.

Gender differences may also emerge in discretionary payments due to a difference in beliefs. That is, even if the evaluators use the same criteria in the evaluation of male and female leaders (i.e., same whether θ_j and β_j), we may still observe gender differences in the payments if the evaluators' beliefs differ by the leader's gender. Erkal, Gangadharan, and Koh (2023) show in a similar environment but without the threat of discretionary payments that there are gender differences in the attribution of low outcomes but not high outcomes. Given the presence of discretionary payments in our paper, our conjecture is that participants are more likely to choose Investment X and that behavior is not likely to differ by gender since both

[13] More details about the equilibrium can be found in Appendix B.

male and female leaders are likely to react to the financial incentives in a similar way. Anticipating this, we do not expect evaluators to have gender biases in their beliefs, and our design allows us to investigate empirically whether a gender difference exists in beliefs.

To summarize, the key research questions we analyze using this conceptual framework are the following:

(1) Are there gender differences in discretionary payments?

(2) If so, can the differences be explained by gender differences in beliefs?

(3) What are the determinants of discretionary payments? Are they shaped by both beliefs about the leaders' intentions and outcomes?

(4) If so, are there gender differences in the weights evaluators put on beliefs versus outcomes? That is, are evaluators more likely to suffer from an outcome bias or more likely to rely on their beliefs, depending on the gender of the leader?

5 Results

5.1 Evaluators' discretionary payments

We first examine the overall payments made to female and male leaders. Figure 1 presents the average discretionary payments by outcome and the leader's gender.[14] The figure shows that on average, both male and female leaders receive negative payments for a low outcome and positive payments for a high outcome (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 low outcome and for a high outcome (p-values = 0.699 and 0.926, respectively).

Figure 2 presents evaluators' penalty decisions for a low outcome (when the investment fails) and bonus decisions for a high outcome (when the investment succeeds). Penalties refer to negative discretionary payments and bonuses refer to positive discretionary payments. Panel (a) shows that there are no statistically significant gender differences in the proportions of penalties (57.4% and 61.7% for male and female leaders, respectively; p-value = 0.446) and bonuses (57.1% and 59.0% for male and female leaders, respectively; p-value = 0.740) awarded to the leader. Panel (b) reveals that there are gender differences in penalty and bonus amounts. Conditional on receiving a bonus, female leaders receive a bonus that is 13.5% lower relative to male leaders on average (p-value = 0.039), while conditional on receiving a penalty, female

[14] Error bars in all figures represent 95% confidence intervals accounting for standard errors clustered at the participant level. Unless otherwise stated, for all tests reported in text, we report p-values of t-tests with standard errors clustered at the participant level.

leaders receive 15.1% less penalty relative to male leaders on average (p-value = 0.031). Hurdle model regression estimates reported in Table C1 of Appendix C provide similar conclusions.

We summarize as follows.

Result 1. *Male and female leaders are equally likely to receive bonuses and penalties. However, conditional on receiving a bonus or a penalty, female leaders receive lower bonuses and lower penalties on average than male leaders.*

The elicited expectations of the leaders are partly in line with Result 1. Figure C1 in Appendix C presents leaders' beliefs about the penalty and bonus decisions made by evaluators. Relative to male leaders, female leaders are less likely to expect a bonus for high outcomes (p-value = 0.097); and given a bonus for attaining a high outcome, expect a lower bonus on average (p-value = 0.046).[15] This consistency between the leaders' expectations about being discriminated and the evaluators' decisions when it comes to positive payments suggests that both may be shaped by societal norms about the role gender plays in shaping discretionary rewards. For example, if putting others' welfare ahead of one's own is more likely to be expected of women than men, then when they see a high outcome, evaluators may want to pay a higher renumeration to male leaders than to female leaders as a way of rewarding them for being prosocial.

5.2 Determinants of discretionary payments

In this section, we explore potential mechanisms contributing to the observed gender differences in the bonuses and penalties received by leaders.

5.2.1 Evaluators' beliefs

We first examine evaluators' prior and posterior beliefs about the leader. Overall, we do not find any evidence that there are gender differences in evaluators' beliefs.

Figure 3 presents evaluators' prior beliefs that the leader has chosen Investment X. 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 Investment X. Figure

[15] On the other hand, there are no statistically significant gender differences in leaders' expectations of having a penalty imposed for low outcomes (p-value = 0.413) or the average penalty amounts given a penalty for attaining a low outcome (p-value = 0.762).

3 reveals that there are no statistically significant differences in evaluators' average prior beliefs about the investment choices made by male and female leaders (p-value = 0.418).

Ordinary least squares (OLS) estimates reported in Table 2 yield similar conclusions. In column (1) of the table, we control for the leader's investment decisions, evaluators' decisions in the risk task, the difference in investment returns between a successful and failed investment, and whether the investment provides a return of zero should it fail. 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.[16]

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

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

where $\text{logit}\, z = \log\left(\frac{z}{1-z}\right)$,[17] $I(\cdot)$ is an indicator function for the observed return Q from the investment, $\hat{\sigma}_j(X|Q)$ and $\hat{\mu}_j$ represent evaluator j's reported posterior beliefs (given Q) and prior beliefs, respectively, and ε_i captures non-systematic errors. p and $1-p$ denote the probability of a high return under Investment X and Investment Y, respectively. 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 $\delta = \gamma_H = \gamma_L = 1$. Any deviation in the estimated parameters from 1 is interpreted as non-Bayesian updating behavior. The main parameters of interest are γ_H and γ_L, which represent the weights evaluators place on a signal of high and low outcome, respectively, when updating their beliefs.[18] We estimate equation (3) using OLS and compare

[16] Figure C2 and Table C2 in Appendix C reveal that there are no statistically significant differences between male and female leaders in their investment choices. Further examining prior beliefs separately by the evaluator's gender, we do not find any statistically significant differences in the prior beliefs held by female and male evaluators (p-values = 0.633 and 0.497, respectively).

[17] Note that the logit function is only defined for beliefs in (0,100). Instead of excluding observations of evaluators who state 0 or 100 as their prior or posterior belief about the leader, we take the logit of 0.01 or 99.99 as an approximation.

[18] $\gamma_H < 1$ ($\gamma_L < 1$) implies that the evaluator attributes a high (low) outcome more to luck relative to a Bayesian, while $\gamma_H > 1$ ($\gamma_L > 1$) implies that s/he attributes the outcome more to the leader's decision. On the other hand, δ represents the weight evaluators place on their prior beliefs. $\delta < 1$ implies that evaluator j suffers from base-rate neglect while $\delta > 1$ implies that s/he suffers from confirmatory bias. See, e.g., Grether (1980), Barron (2021), Coutts (2019), and Möbius et al. (2022) for similar estimation approaches. Benjamin (2019) provides a review.

the coefficients between male and female leaders to analyze whether evaluators suffer from gender biases while updating their beliefs about the leader's investment choices.

Table 3 presents the regression results separately by the leader's gender. We observe that there are no statistically significant differences in the attribution of high and low outcomes between female and male leaders (comparisons of γ_H and γ_L between columns 1 and 2: p-values = 0.923 and 0.185, respectively).[19,20]

Taken together, we find no gender differences in both prior beliefs and updating behavior. Consequently, we observe no statistically significant gender differences in evaluators' posterior beliefs, both given a high outcome and a low outcome (t-tests: p-values = 0.371 and 0.151, respectively). We summarize as follows.

Result 2. *There are no gender differences in evaluators' prior or posterior beliefs about the leaders' investment decisions.*

5.2.2 Criteria used to determine discretionary payments

Thus far, despite not observing any gender differences in the beliefs and the process of belief formation (Result 2), we observe gender differences in the bonuses and penalties received by leaders (Result 1). To investigate this issue further, we next turn to the drivers of evaluators' discretionary payment decisions.

Figure 4 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 high outcomes (gray bubbles) and low 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 columns (1) and (2) of Table 4, along with 95% confidence intervals. In the regressions in Table 4, we control for participants' characteristics,

[19] The results hold even when the analysis is conducted separately by the leaders' and evaluators' gender (Table C3 in Appendix C). In contrast to this finding, Erkal, Gangadharan, and Koh (2023), who assume that evaluators cannot make discretionary payment decisions, find gender differences in the attribution of low outcomes but not in the attribution of high outcomes. Our results show that, in the presence of discretionary payments, the evaluators do not hold different beliefs because they expect leaders to take into account the potential payment consequences of their actions.

[20] Columns (1) and (2) of Table 3 also reveal that evaluators are more likely to attribute both high and low outcomes of female leaders to luck than a Bayesian, although these effects are marginally statistically significant (tests of $\gamma_H = 1$ and $\gamma_L = 1$: p-values = 0.050 for both). However, they are no different from a Bayesian in their attribution of both high and low outcomes for male leaders (tests of $\gamma_H = 1$ and $\gamma_L = 1$: p-values = 0.174 and 0.862, respectively, for male leaders in column 2).

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.

Figure 4 and Table 4 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 4: p-value = 0.006), evaluators' posterior beliefs do not play a role in shaping payments to female leaders (column 1: p-value = 0.668). 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 = 0.034).

In addition, we observe that outcomes are an important determinant of payments made to leaders (p-values < 0.001 in both columns). This is consistent with the literature on outcome bias. However, the direct impact of outcomes on discretionary payments is not statistically significant between female leaders and male leaders (column 1 vs. column 2: p-value = 0.188). Moreover, we observe that outcomes play a larger role than beliefs in driving the payments made to female leaders (column 1: p-value < 0.001), but outcomes and beliefs do not play different roles in driving the payments made to male leaders (column 2: p-value = 0.970).

In summary, we observe a *gender criteria gap* in the determination of leaders' discretionary payments. That is, the determinants of discretionary payments vary by gender. We summarize our results as follows.

Result 3. *There exists a gender criteria 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.*

The gender criteria gap can thus potentially explain why there are differences in the bonuses and penalties received by female and male leaders despite there being no gender differences in evaluators' beliefs. Specifically, as shown in Figure 4, as posterior beliefs increase, the discretionary payments received by male leaders increase. As a result, although female leaders face higher bonuses and lower penalties than male leaders when beliefs are low, male leaders end up with higher bonuses and lower penalties than female leaders when beliefs are high. Given that beliefs are higher for leaders for high outcomes than for low outcomes, the patterns observed in Figure 4 can explain why male leaders receive higher bonuses than female leaders for high outcomes, but they also receive higher penalties for low outcomes. The figure also reveals that it is even possible for male leaders to receive bonuses for low outcomes if the evaluators have sufficiently high beliefs about the male leaders' decisions. For instance, we observe that 21% of evaluators award a bonus for a low outcome, and this is more likely for male leaders than for female leaders (p-value = 0.088).[21]

We further investigate Result 3 by separating the analysis based on the evaluator's gender (columns 3 to 6 of Table 4). 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 (e.g., Derks, Van Laar, and Ellemers, 2016; Arvate, Galilea, and Todescat, 2018).

The analysis we present here is exploratory given the smaller number of observations and therefore lower statistical power that we have. The results show that, directionally, the gender criteria gap is exhibited by both female and male evaluators. For both female and male evaluators, the discretionary payments they make to female leaders do not depend on their posterior beliefs (columns 3 and 5: p-values = 0.839 and 0.447, respectively), and outcomes play a statistically significant larger role than beliefs in driving discretionary payments for female leaders (p-values = 0.016 and 0.012, respectively). Columns (4) and (6) reveal that the payments made to male leaders are increasing in their posterior beliefs, although the estimated coefficients are not statistically significant (p-values = 0.152 and 0.158, respectively). Nonetheless, outcomes and beliefs do not play any statistically significant role in driving discretionary payments for male leaders (p-values = 0.618 and 0.668, respectively).

Finally, as shown in Figure D1 in Appendix D, some evaluators in our sample 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. To investigate this further, we classify an evaluator as an inconsistent updater if they have 25% or more of posterior beliefs in the opposite direction to that predicted by Bayes' rule, and as a non-updater if they have all their posterior beliefs are equal to their prior beliefs. We find that evaluators classified as either inconsistent updaters or non-updaters answer more comprehension questions incorrectly on the first attempt on average than the rest of the sample (Wilcoxon rank-sum test: p-value = 0.083), suggesting that these evaluators may have a lower

[21] On the other hand, while 15% of evaluators impose a penalty for a high outcome, this does not differ between female and male leaders (p-value = 0.304).

understanding of the instructions during the experiment. As robustness, we restrict our analysis by excluding these evaluators in Appendix D. Our main conclusions remain broadly unchanged in the restricted sample, with the exception of the gender difference in penalties for low outcomes which is no longer statistically significant.

6 Discussion

This paper examines whether different criteria are used in the evaluation of male and female 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 outcomes are determined by a combination of the choices made and luck, and costly investment choices of leaders are not observed. Uncertainty of this kind is ubiquitous in decision making environments, making evaluation a challenging task. While trying to evaluate leaders based on the merits of their (unobserved) actions, evaluators need to correctly assess the role of unexpected events in determining the outcomes.

We find evaluators choose steeper discretionary payments for male leaders than for female leaders in the sense that male leaders receive higher bonuses for high outcomes and higher penalties for low outcomes than female leaders, though this latter finding is less robust. Moreover, leaders' expectations of discretionary payments are consistent with the gender difference observed in bonuses.

We show that the difference in discretionary payments for male and female leaders cannot be explained by gender biases in beliefs. Instead, we find that gender differences emerge in discretionary payments due to a gender criteria gap. Specifically, while male leaders' discretionary payments are determined by both the evaluators' assessments of their investment 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 a role at all.

The finding that beliefs matter more for men could suggest that they are seen to be more fit for leadership roles, as their perceived actions are rewarded and punished more appropriately, whereas beliefs about women's actions in such roles are mostly ignored and not given much weight. Our results are in line with previous research in psychology which has shown that leadership potential is preferred as a criteria when evaluating men while potential is overlooked when evaluating women (and performance is emphasized instead) (see, e.g., Player et al., 2019). Similarly, recent work by Benson, Li, and Shue (2022) shows that women receive lower ratings on their potential despite receiving higher performance ratings. In our context, a possible interpretation of these findings is that we would expect beliefs to play a stronger role in the evaluation of men since beliefs in our experiment reflect evaluators' assessment of each gender's "potential" to act in an altruistic manner.

This gender criteria gap that we identify is a marker of discrimination as it implies that in the labor market, successes (high outcomes) are necessary for women to get bonuses, but men can receive bonuses for failures (low outcomes) as long as evaluators hold them in high regard. More broadly, our findings indicate that luck plays a bigger role in the evaluation of female leaders, both in the domain of bonuses and penalties. While outcomes are determined by both luck and the actions taken, beliefs about the actions that female leaders have taken do not matter. The disproportionate emphasis given to luck (rather than beliefs) has the potential to distort choices in risky environments and can perpetuate gender gaps.

Performance evaluation in many contexts rely on subjective measures, which creates an opportunity for biases to distort incentive structures.[22] Given that gender differences in 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 evaluation procedures. Our findings further our understanding of the factors that may be driving the observed gender gaps in performance pay. In our setting, leaders' decisions can be shaped by their social preference considerations and thus evaluated on this basis. Such leadership environments are commonly observed and relevant to study as leaders often face trade-offs where they can increase the welfare of others at a personal cost. In future research, it would be interesting to examine if the gender criteria gap would emerge and lead to gender discrimination in payments in cases where leaders are evaluated based on their ability only.

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 weights placed on beliefs about intentions versus outcomes may be a source of discrimination, with outcomes playing a disproportionately larger role than beliefs in case

[22] 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).

of women. This type of gender bias is distinct from biased beliefs and leads to a new channel through which discrimination can occur, which we label as the gender criteria gap.

References

Aguiar, F., Brañas-Garza, P., Cobo-Reyes, R., Jimenez, N., Miller, L.M. (2008). Are women expected to be more generous? *Experimental Economics,* 12(1):93-98.

Albrecht, K., von Essen, E., Parys, J., Szech, N. (2013). Updating, self-confidence, and discrimination. *European Economic Review,* 60:144-169.

Alt, J.E., Lassen, D.D. (2003). The political economy of institutions and corruption in American states. *Journal of Theoretical Politics,* 15(3):341-365.

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. (2022). Explicit and implicit belief-based gender discrimination: A hiring experiment. *Working Paper*.

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.

Benson, A., Li, D., Shue, K. (2022). "Potential" and the gender promotion gap. *Working Paper*.

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

Besley, T., Ghatak, M. (2018). Prosocial motivation and incentives. *Annual Review of Economics,* 10:411-438.

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.

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.

Bowles, S., Polanía-Reyes, S. (2012). Economic incentives and social preferences: Substitutes or complements? *Journal of Economic Literature,* 50(2):368-425.

Brañas-Garza, P., Capraro, V., Rascón-Ramírez, E. (2018). Gender differences in altruism on Mechanical Turk: Expectations and actual behaviour. *Economics Letters,* 170:19-23.

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

Campos-Mercade, P., Mengel, F. (2023). Non-Bayesian statistical discrimination. *Management Science*.

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.

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*.

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. (2022). By chance or by choice? Biased attribution of others' outcomes when social preferences matter. *Experimental Economics,* 25(2):413-443.

Erkal, N., Gangadharan, L., Koh, B.H. (2023). Do women receive less blame than men? Attribution of outcomes in a prosocial setting. *Journal of Economic Behavior & Organization,* 210:441-452.

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

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.

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.

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.

Heilman, M.E., Chen, J.J. (2005). Same behavior, different consequences: reactions to men's and women's altruistic citizenship behavior. *Journal of Applied Psychology,* 90(3):431-441.

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.

Lederman, D., Loayza, N.V., Soares, R.R. (2005). Accountability and corruption: Political institutions matter. *Economics and Politics,* 17(1):1-35.

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

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

Möbius, M.M., Niederle, M., Niehaus, P., Rosenblat, T.S. (2022). Managing self-confidence. *Management Science*.

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.

Persson, T., Roland, G., Tabellini, G. (1997). Separation of powers and political accountability. *Quarterly Journal of Economics,* 112(4):1163-1202.

Player, A., Randsley de Moura, G., Leite, A.C., Abrams, D., Tresh, F. (2019). Overlooked leadership potential: The preference for leadership potential in job candidates who are men vs. women. *Frontiers in Psychology,* 10.

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).

Solnick, S.J. (2001). Gender differences in the ultimatum game. *Economic Inquiry,* 39(2):189-200.

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: Investment tasks

	Investment X			Investment Y		
	High Outcome (Succeeds)	Low Outcome (Fails)	Expected	High Outcome (Succeeds)	Low Outcome (Fails)	Expected
Task 1						
Individual investment return	150	0		150	0	
Each evaluator earns	150	0	112.5	150	0	37.5
Leader earns	250	100	212.5	400	250	287.5
Task 2						
Individual investment return	200	0		200	0	
Each evaluator earns	200	0	150	200	0	50
Leader earns	300	100	250	450	250	300
Task 3						
Individual investment return	250	0		250	0	
Each evaluator earns	250	0	187.5	250	0	62.5
Leader earns	350	100	287.5	500	250	312.5
Task 4						
Individual investment return	250	50		250	50	
Each evaluator earns	250	50	200	250	50	100
Leader earns	350	150	300	500	300	350
Task 5						
Individual investment return	300	50		300	50	
Each evaluator earns	300	50	237.5	300	50	112.5
Leader earns	400	150	337.5	550	300	362.5

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. Only the individual investment returns for each group member (including the leader) vary across the five tasks. In each case, the evaluator's earnings are equivalent to the individual investment return. To calculate the net earnings to the leader in Stage 1, we take into account 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 earnings to 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 earnings to each evaluator is given by (0.75 × 150 + 0.25 × 0) (expected return from Investment X) = 112.5 ECU.

Table 2: OLS regressions of evaluators' prior belief that the leader has chosen Investment X

	Dependent variable: Prior belief	
Variables	(1)	(2)
Female leader	1.413	1.151
	(2.250)	(2.319)
Chose Investment X 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
# participants (clusters)	233	233
R-squared	0.186	0.190

Investment X refers to the costlier investment option for the leader but yields a higher success probability. RT stands for the Risk Task. Robust standard errors clustered at the participant level in parentheses. 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 3: OLS regressions of evaluators' posterior belief that the leader has chosen Investment X, 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)	0.542	0.515	0.788
	(0.062)	(0.079)	
γ_H: High outcome × logit (p)	0.787	0.804	0.923
	(0.108)	(0.143)	
γ_L: Low outcome × logit $(1-p)$	0.766	1.027	0.185
	(0.118)	(0.157)	
$\gamma_H - \gamma_L$	0.021	-0.224	
	(0.152)	(0.160)	
Observations	1,160	1,170	
# participants (clusters)	116	117	
R-squared	0.425	0.400	

Investment X refers to the costlier investment option for the leader but yields a higher success probability. Robust standard errors clustered at the participant level in parentheses. 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 $I(\text{High Outcome}) + I(\text{Low Outcome}) = 1$, there is no constant term in the regression.

Table 4: OLS regressions of discretionary payments, by the leader's and evaluator's gender

	Dependent variable: Discretionary payments								
	Pooled			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)
High outcome	49.863	38.285	0.188	43.771	40.854	0.802	54.360	40.717	0.301
	(5.356)	(7.057)		(6.795)	(9.632)		(8.296)	(10.513)	
Posterior belief	0.038	0.376	0.034	0.025	0.278	0.260	0.097	0.290	0.414
	(0.088)	(0.134)		(0.123)	(0.192)		(0.127)	(0.202)	
Constant	-62.221	-49.945		-38.414	-107.679		-83.911	-2.708	
	(22.272)	(21.772)		(31.155)	(30.444)		(32.500)	(30.681)	
Test of High outcome = 100 × Belief [a]									
p-value	< 0.001	0.970		0.016	0.618		0.012	0.668	
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	
# participants (clusters)	116	117		56	59		60	58	
R-squared	0.236	0.209		0.210	0.213		0.278	0.282	

Robust standard errors clustered at the participant level in parentheses.

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.

[a] This tests the null hypothesis that the coefficient on high 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 low outcome to a high outcome), and (ii) with respect to a given change in belief (from a belief that the leader has chosen Investment Y with certainty to a belief that the leader has chosen Investment X with certainty).

List of Figures

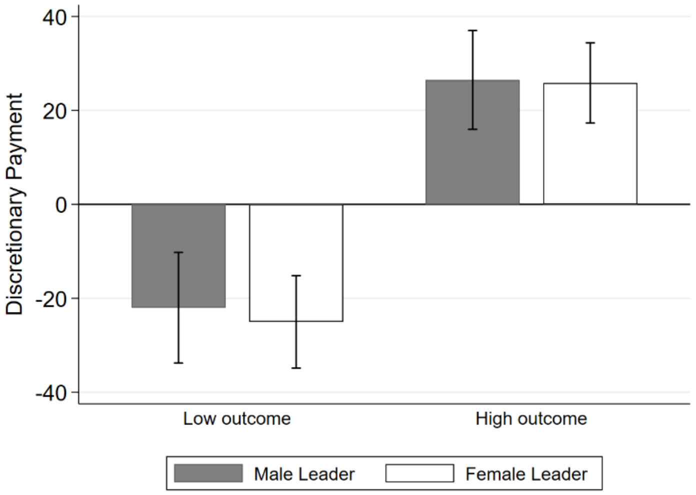
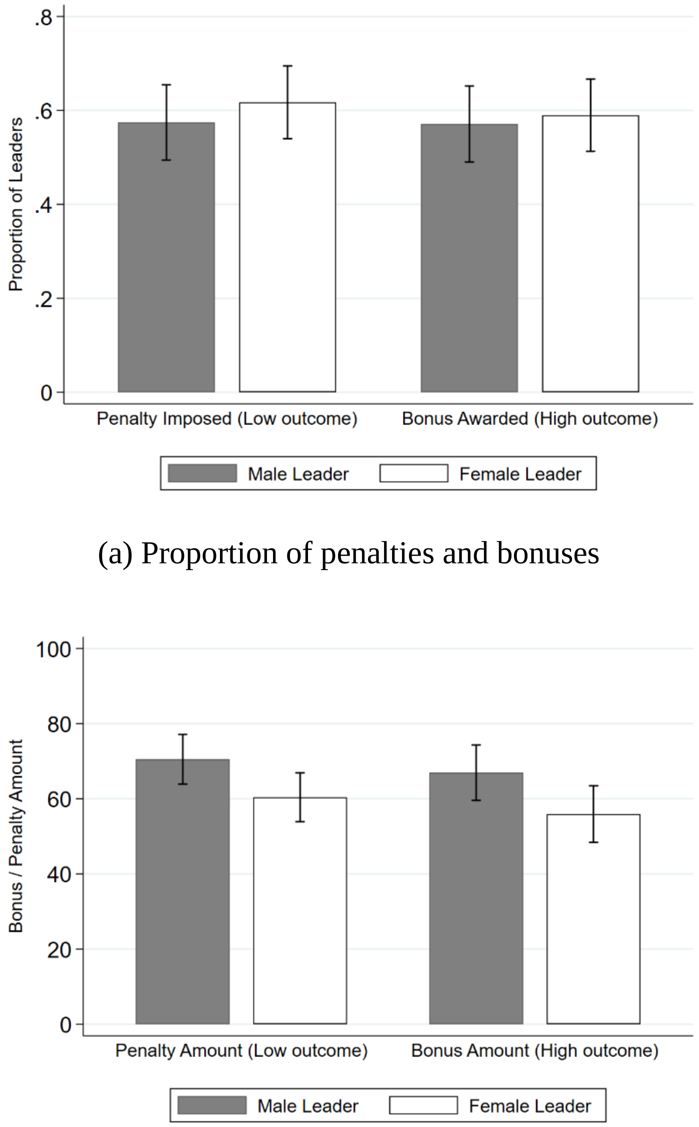


Figure 1: Evaluators' discretionary payments, by outcome and the leader's gender

Note: Error bars represent 95% confidence intervals accounting for standard errors clustered at the participant level.



(a) Proportion of penalties and bonuses

(b) Average penalty and bonus amounts

Figure 2: 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. Error bars represent 95% confidence intervals accounting for standard errors clustered at the participant level.

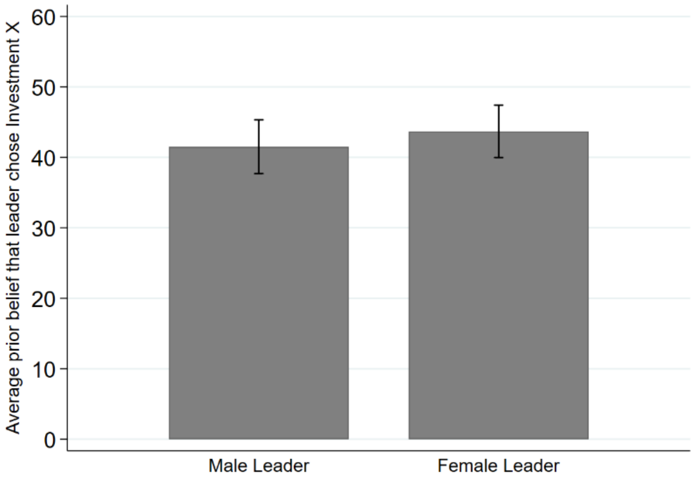
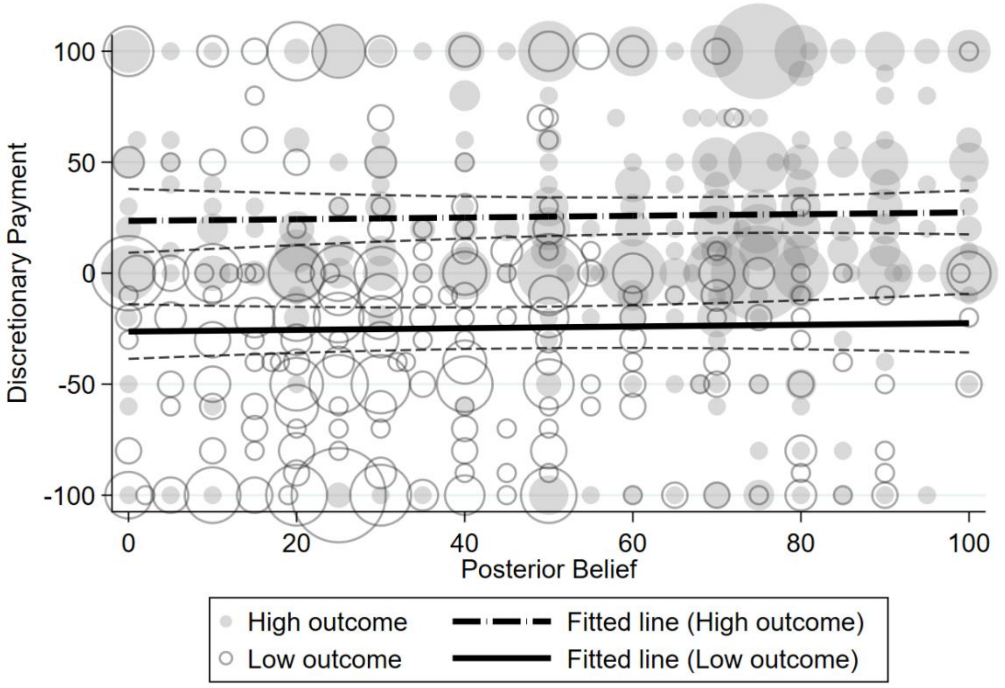
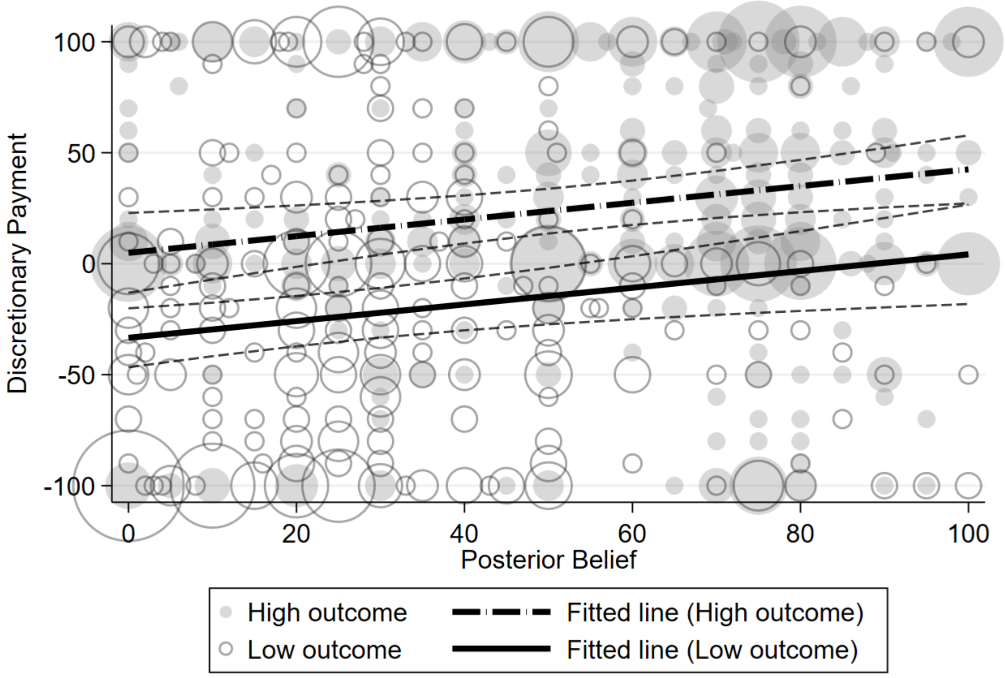


Figure 3: Evaluators' prior belief that the leader has chosen Investment X, by the leader's gender

Note: Error bars represent 95% confidence intervals accounting for standard errors clustered at the participant level.



(a) Female Leaders



(b) Male Leaders

Figure 4: Discretionary payments against evaluators' posterior belief that the leader has chosen Investment X and leader's outcomes

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

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 different 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.

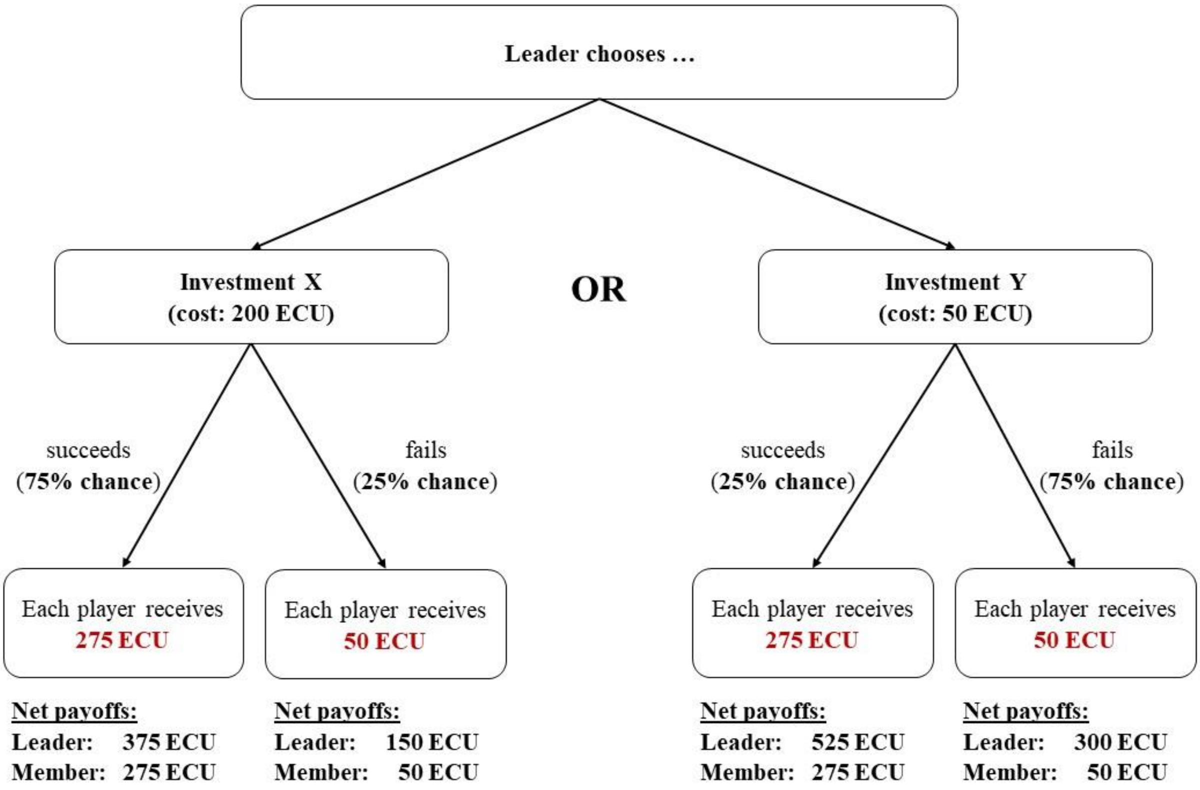


Figure 1: Investment Options in Part 1 (Example of a Task)

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.

> Question 1
>
> **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.

> Question 2(a)
>
> **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?
>
> Question 2(b)
>
> **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. A higher number means that you think your Leader is more likely to have chosen Investment X.

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. <u>Note that your earnings will not be affected by your decisions to increase or decrease your Leader's payoff.</u>

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)

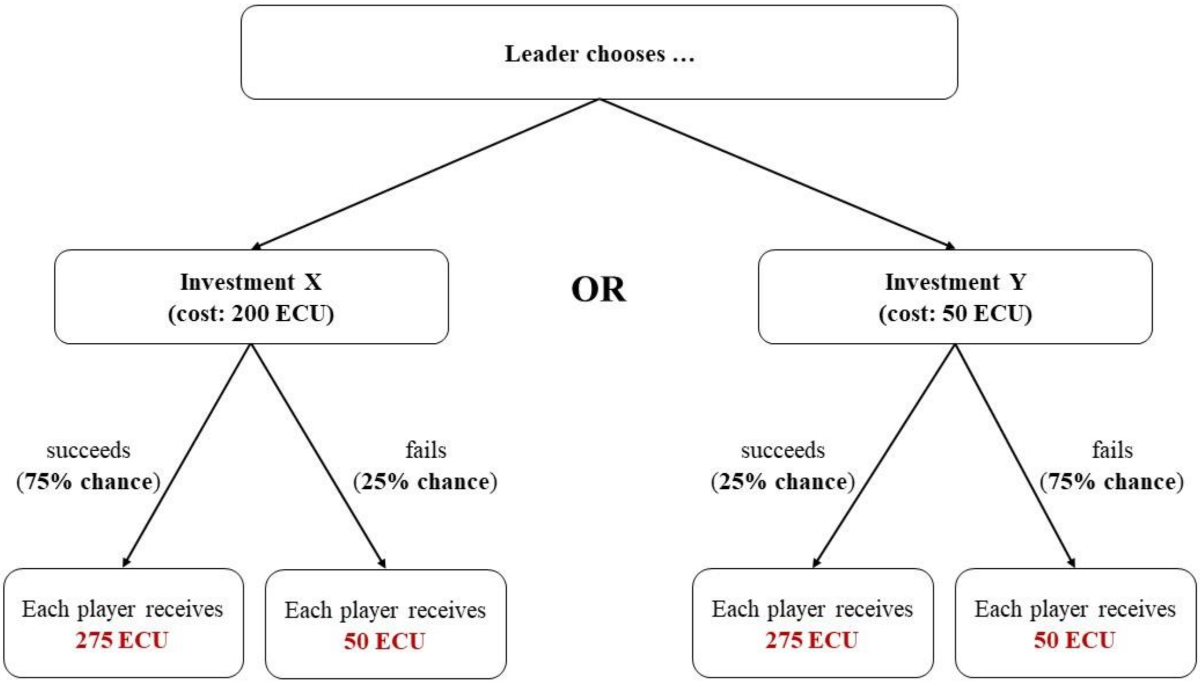
6. If I am a Member, then I will be paid for:
   (a) my Leader's investment decision in Stage 1 only.
   (b) my prediction of my Leader's investment decision in Stage 2 only.
   (c) both my Leader's investment decision in Stage 1 AND my prediction of his/her decision in Stage 2.
   (d) either my Leader's 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 fails. 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 succeeds. 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 Characterization of Equilibrium

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

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

(ii) $\Delta(\cdot,\cdot): \{Q_L, Q_H\} \rightarrow \mathbb{R}$, 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 investment choice by comparing the expected utility from choosing X with the expected utility from choosing Y. The expected utility from choosing X 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 Y can be written in a similar way. The optimal investment choice partitions the interval $[0,1]$ into two sets:

$$T_L = \{\alpha \in [0,1]: i = X\} \text{ and } T_H = \{\alpha \in [0,1]: i = Y\}$$

Let α^* denote the type who is indifferent between choosing X and Y. 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

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

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.

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

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

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

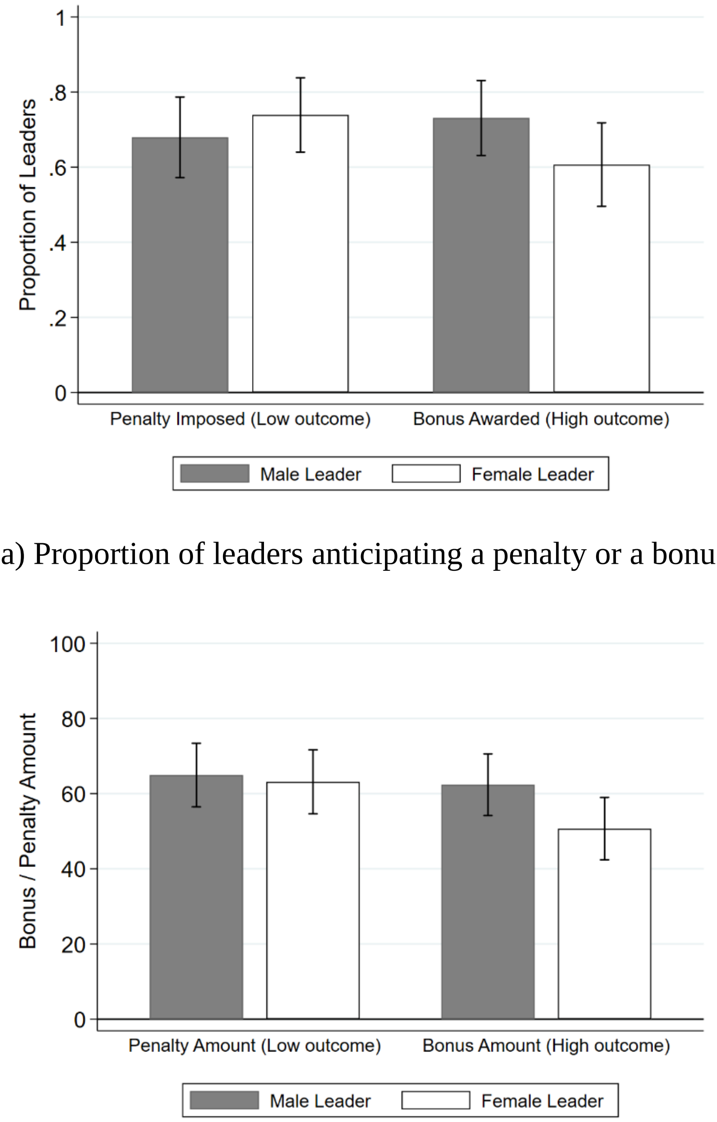
Appendix C Additional Analyses

Table C1: Hurdle model estimations of penalty and bonus decisions conditional on outcomes

	Penalty decisions (Low outcome)		Bonus decisions (High outcome)	
Variables	Proportion imposed (1)	Average amount (2)	Proportion awarded (3)	Average amount (4)
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
# participants (clusters)	233	170	233	173
R-squared		0.123		0.172

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

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.



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

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

Figure C1: 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. Error bars represent 95% confidence intervals accounting for standard errors clustered at the participant level.

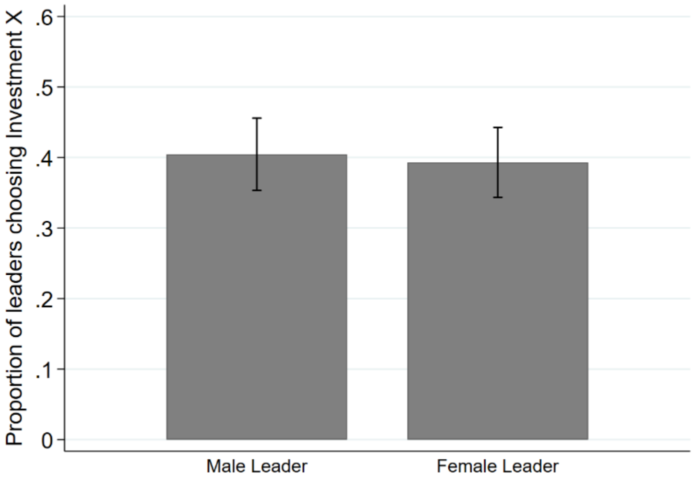


Figure C2: Proportion of participants choosing Investment X as leaders, by the leader’s gender

Note: Error bars represent 95% confidence intervals accounting for standard errors clustered at the participant level.

Table C2: Probit regressions of leaders' investment choice

	Dependent variable: = 1 if leader chooses high investment	
Variables	(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
# participants (clusters)	350	350

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

DG: Dictator Game; RT: Risk Task.

Note that all participants make investment decisions as leaders in the experiment. 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 C3: OLS regressions of evaluators' posterior belief that the leader has chosen Investment X, by both the leader's and evaluator's gender

	Dependent variable: Logit (posterior belief)					
	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
δ: Logit (prior belief)	0.562	0.484	0.561	0.516	0.551	0.817
	(0.081)	(0.106)		(0.095)	(0.118)	
γ_H: High outcome × logit (p)	0.795	0.902	0.717	0.776	0.704	0.725
	(0.174)	(0.241)		(0.135)	(0.155)	
γ_L: Low outcome × logit $(1-p)$	0.955	1.149	0.487	0.593	0.904	0.266
	(0.169)	(0.224)		(0.167)	(0.224)	
$\gamma_H - \gamma_L$	-0.160	-0.247		0.183	-0.200	
	(0.221)	(0.221)		(0.218)	(0.237)	
Observations	560	590		600	580	
# participants (clusters)	56	59		60	58	
R-squared	0.457	0.388		0.390	0.418	

Investment X refers to the costlier investment option for the leader but yields a higher success probability. 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 $I(\text{High Outcome}) + I(\text{Low Outcome}) = 1$, there is no constant term in the regression.

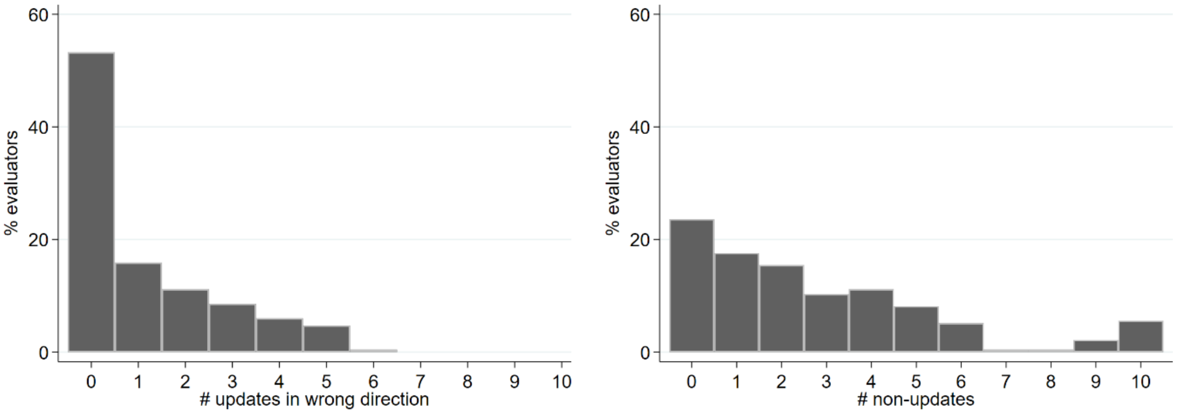
Appendix D Analyses using Restricted Sample

As shown in Figure D1, some evaluators in our sample 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.

Specifically, we classify an evaluator as an inconsistent updater if they have 25% or more of their posterior beliefs in the opposite direction to that predicted by Bayes' rule, and as a non-updater if all their posterior beliefs are equal to their prior beliefs. We find that evaluators classified as either inconsistent updaters or non-updaters answer more comprehension questions incorrectly on the first attempt on average than the rest of the sample (Wilcoxon rank-sum test: p-value = 0.083), suggesting that these evaluators may have a lower understanding of the instructions during the experiment.

Hence, in this section, we exclude evaluators if they are classified as either an inconsistent updater or a non-updater. 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 on belief updating (Coutts, 2019; Barron, 2021; Erkal, Gangadharan, and Koh, 2022; Möbius et al., 2022).[23] Our main conclusions remain broadly unchanged in the restricted sample, with the exception of the gender difference in penalties for low outcomes which is no longer statistically significant.

[23] Barron, K. (2021). Belief updating: Does the 'good-news, bad-news' asymmetry extend to purely financial domains? *Experimental Economics,* 24:31-58.
Coutts, A. (2019). Good news and bad news are still news: Experimental evidence on belief updating. *Experimental Economics,* 22(2):369-395.
Erkal, N., Gangadharan, L., Koh, B.H. (2023). Do women receive less blame than men? Attribution of outcomes in a prosocial setting. *Journal of Economic Behavior & Organization,* 210:441-452.
Möbius, M.M., Niederle, M., Niehaus, P., Rosenblat, T.S. (2022). Managing self-confidence. *Management Science*.



(a) Inconsistent updates

(b) Non-updates

Figure D1: Distribution of inconsistent and non-updates by evaluators

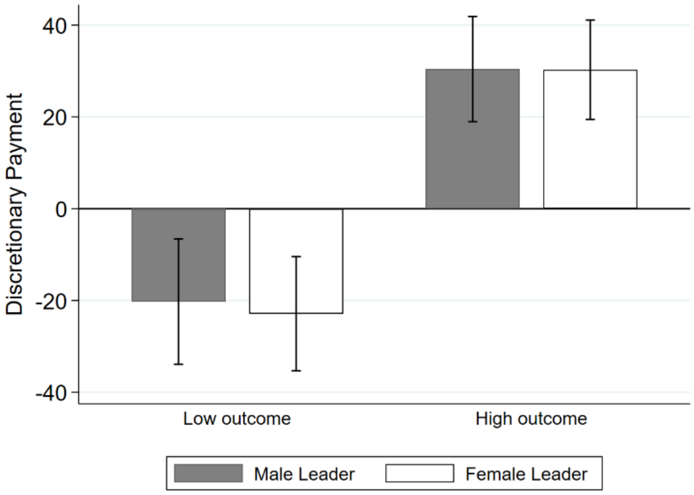
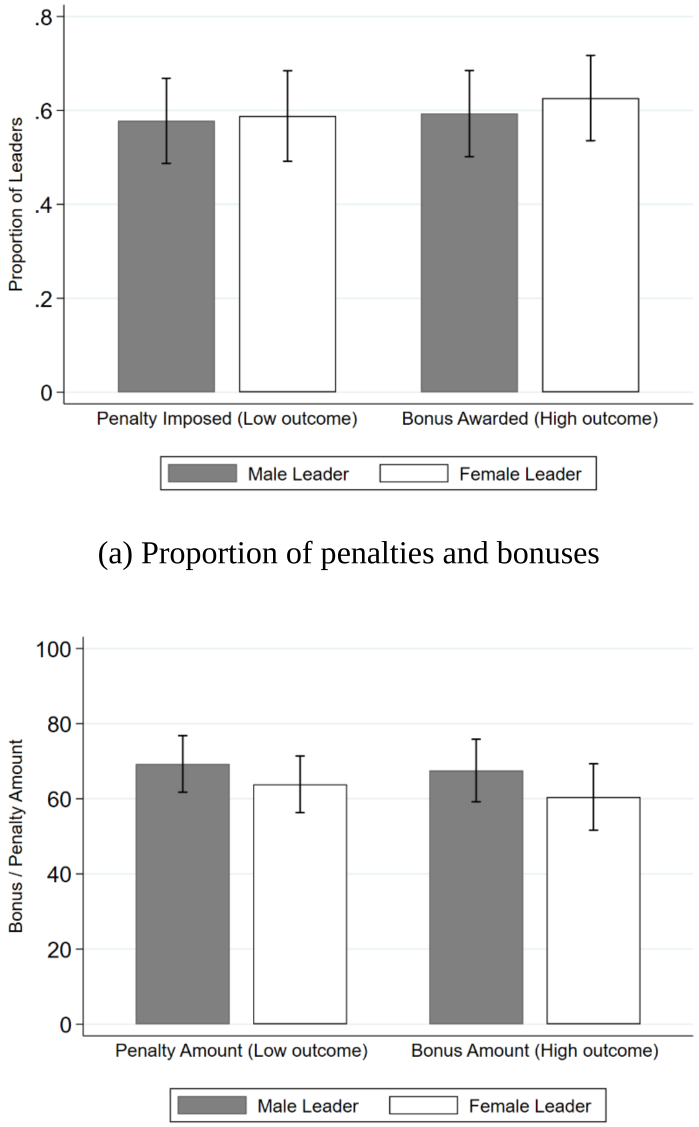


Figure D2: Evaluators' discretionary payments, by outcome and the leader's gender (restricted sample)

Note: Error bars represent 95% confidence intervals accounting for standard errors clustered at the participant level. The figure excludes participants classified as inconsistent or non-updaters.



(a) Proportion of penalties and bonuses

(b) Average penalty and bonus amounts

Figure D3: Evaluators' penalty and bonus decisions, by outcome and the leader's gender (restricted sample)

Note: The average penalty and bonus amounts in panel (b) are computed conditional on a penalty being imposed and a bonus being awarded, respectively. Error bars represent 95% confidence intervals accounting for standard errors clustered at the participant level. The figures exclude participants classified as inconsistent or non-updaters.

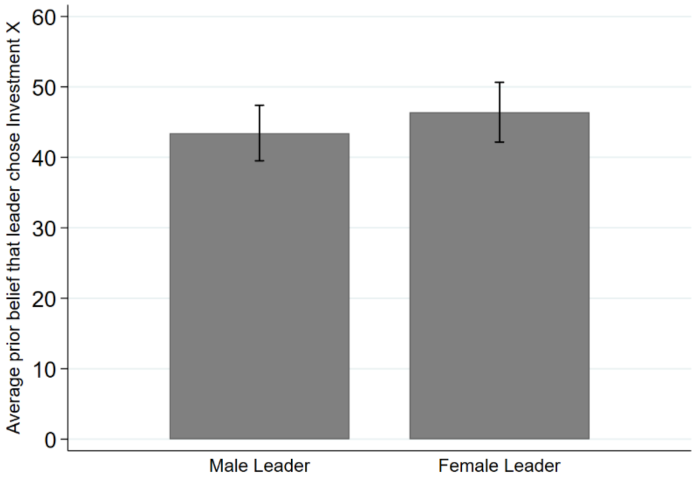
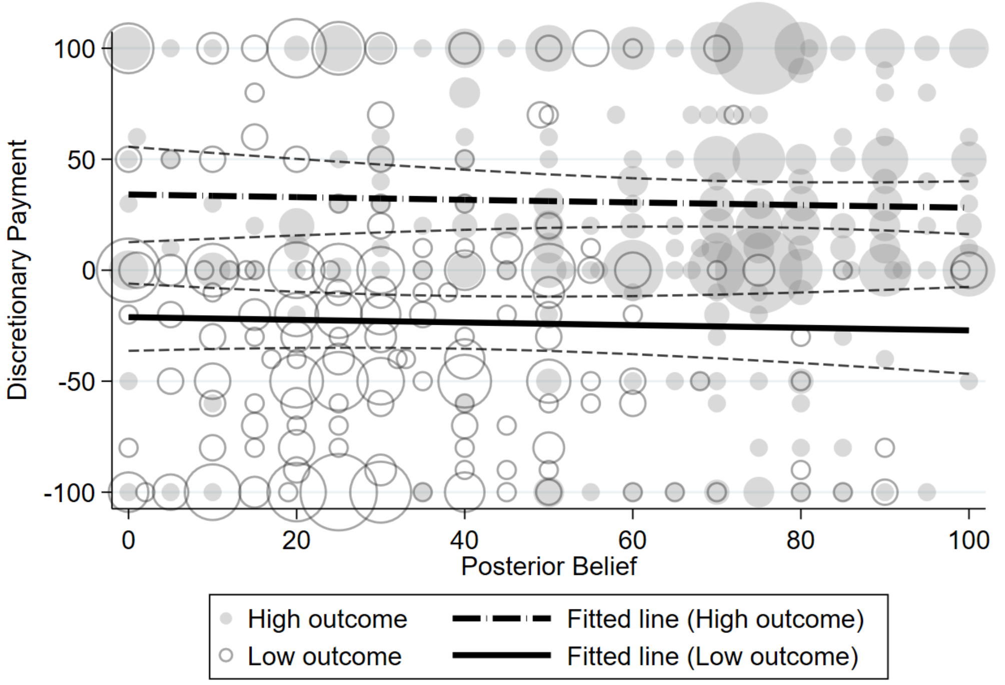
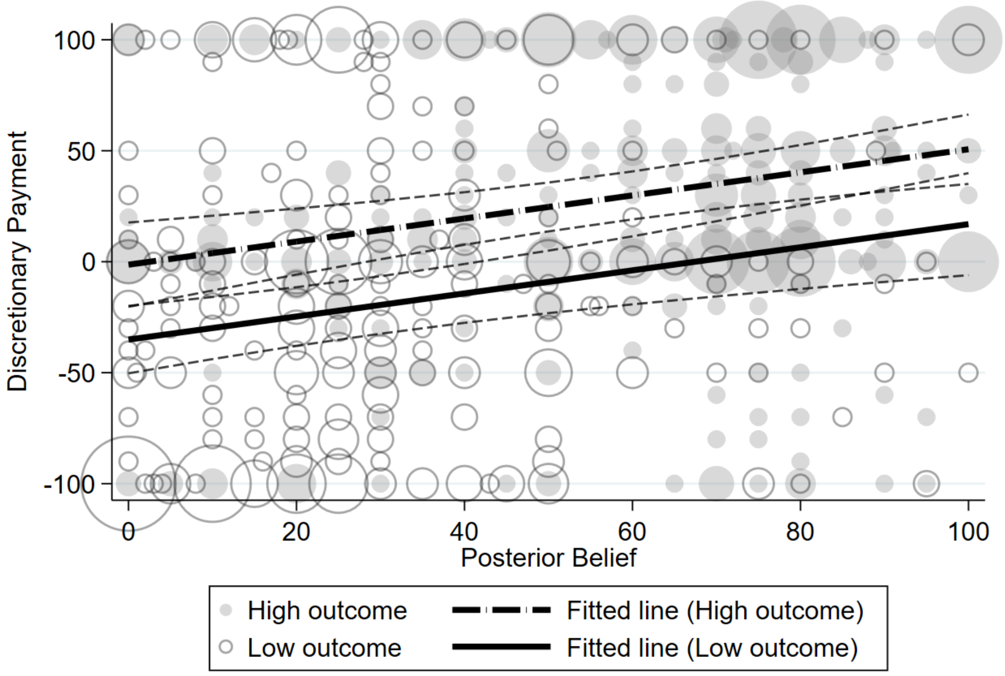


Figure D4: Evaluators' prior belief that the leader has chosen Investment X, by the leader's gender (restricted sample)

Note: Error bars represent 95% confidence intervals accounting for standard errors clustered at the participant level. The figure excludes participants classified as inconsistent or non-updaters.



(a) Female Leaders



(b) Male Leaders

Figure D5: Discretionary payments against evaluators' posterior belief that the leader has chosen Investment X and leader's outcomes (restricted sample)

Note: Dashed lines above and below each fitted line represent 95% confidence intervals. The figures exclude participants classified as inconsistent or non-updaters.

Table D1: Hurdle model estimations of penalty and bonus decisions conditional on outcomes (restricted sample)

	Penalty decisions (Low outcome)		Bonus decisions (High outcome)	
	Proportion imposed	Average amount	Proportion awarded	Average amount
Variables	(1)	(2)	(3)	(4)
(a) Pooled				
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
# participants (clusters)	174	124	174	133
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.

In column (2), the gender difference in the average penalty sent to leaders for a low outcome reported is not statistically significant (p-value = 0.256).

Table D2: OLS regressions of evaluators' prior belief that the leader has chosen Investment X (restricted sample)

	Dependent variable: Prior belief	
Variables	(1)	(2)
Female leader	2.622	2.153
	(2.466)	(2.509)
Chose Investment X as leader	21.167	21.261
	(2.287)	(2.228)
# risky choices in RT	0.723	0.673
	(0.785)	(0.734)
High Return – Low Return	0.090	0.090
	(0.023)	(0.023)
Zero return if investment fails	0.869	0.859
	(1.964)	(1.970)
Constant	17.359	18.674
	(6.186)	(10.040)
Individual controls	N	Y
Control for task order	Y	Y
Observations	870	870
# participants (clusters)	174	174
R-squared	0.187	0.198

Investment X refers to the costlier investment option for the leader but yields a higher success probability. RT stands for the Risk Task. Robust standard errors clustered at the participant level in parentheses. This analysis excludes participants classified as inconsistent or non-updaters. 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 D3: OLS regressions of evaluators' posterior belief that leader has chosen Investment X by the leader's gender (restricted sample)

	Dependent variable: Logit(posterior belief)		
	Female Leader	Male Leader	p-value
Variables	(1)	(2)	(1) vs. (2)
δ : Logit(prior belief)	0.602	0.442	0.119
	(0.050)	(0.089)	
γ_H : High outcome × logit(p)	0.986	1.061	0.720
	(0.114)	(0.173)	
γ_L : Low outcome × logit($1 - p$)	1.088	1.256	0.434
	(0.125)	(0.174)	
$\gamma_H - \gamma_L$	-0.102	-0.195	
	(0.160)	(0.181)	
Observations	840	900	
# participants (clusters)	84	90	
R-squared	0.530	0.369	

Investment X refers to the costlier investment option for the leader but yields a higher success probability. 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 $I(\text{High Outcome}) + I(\text{Low Outcome}) = 1$, there is no constant term in the regression.

In columns (1) and (2), evaluators are no different from a Bayesian in their attribution of both high and low outcomes of female leaders (p-values = 0.906 and 0.484, respectively).

Table D4: OLS regressions of discretionary payments, (i) by the leader's gender and (ii) by the leader's and evaluator's gender (restricted sample)

	Dependent variable: Discretionary payments								
	Pooled			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)
High outcome	55.279	33.758	0.048	52.455	33.007	0.143	57.345	36.906	0.260
	(7.815)	(7.732)		(9.149)	(9.975)		(13.632)	(12.588)	
Posterior belief	-0.060	0.520	0.002	-0.125	0.422	0.026	0.051	0.570	0.076
	(0.126)	(0.136)		(0.148)	(0.202)		(0.237)	(0.182)	
Constant	-71.525	-76.654		-58.023	-98.791		-87.662	-60.945	
	(30.223)	(23.701)		(33.474)	(33.701)		(56.911)	(39.743)	
Test of High outcome = 100 × Belief [a]									
p-value	0.001	0.331		0.004	0.738		0.134	0.448	
Individual controls	Y	Y		Y	Y		Y	Y	
Control for task order	Y	Y		Y	Y		Y	Y	
Observations	840	900		450	500		390	400	
# participants (clusters)	84	90		45	50		39	40	
R-squared	0.231	0.219		0.236	0.199		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.

[a] This tests the null hypothesis that the coefficient on high 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 low outcome to a high outcome), and (ii) with respect to a given change in belief (from a belief that the leader has chosen Investment Y with certainty to a belief that the leader has chosen Investment X with certainty).

The direct impact of outcomes on discretionary payments is statistically significantly different between female leaders and male leaders in column (1) (p-value = 0.048). Moreover, the gender criteria gap is statistically significant for both female and male evaluators (columns 2 and 3: p-values = 0.026 and 0.076, respectively).