International Economic Review

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Gender Criteria Gap in Evaluation: Role of Perceived Intentions and Outcomes

Nisvan Erkal[1] | Lata Gangadharan[2] | Boon Han Koh[3]

[1]Department of Economics, University of Melbourne, Parkville, Victoria, Australia | [2]Department of Economics, Monash University, Clayton, Victoria, Australia | [3]Department of Economics, University of Exeter, Exeter, UK

Correspondence: Lata Gangadharan (lata.gangadharan@monash.edu)

Received: 16 August 2024 | **Revised:** 27 March 2026 | **Accepted:** 2 April 2026

Keywords: biases in beliefs | discrimination | evaluation criteria | gender criteria gap | laboratory experiments | outcome bias | social preferences

ABSTRACT

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 (perceived intentions) and make discretionary payments. We find that while payments to male leaders are determined by both outcomes and perceived intentions, 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 if evaluators hold them in high regard.

JEL Classification: C92, D91, J71

1 | Introduction

Subjective performance measures are frequently used by organizations as part of their performance evaluation. These measures may complement or even replace objective performance measures because it is difficult to capture all aspects of employees' contributions through explicit contracts and quantifiable metrics (Baker et al. 1994; Prendergast 1999). However, subjective assessments may induce investment in influence activities by employees (see, e.g., de Janvry et al. 2023; Milgrom and Roberts 1988) or lead to biased evaluations by evaluators, such as those based on gender (e.g., Bohren et al. 2019).

An obvious channel through which gender biases can enter the evaluation process is the beliefs evaluators form about employees' actions or abilities since evaluators often cannot observe the full determinants of outcomes. In this paper, we introduce and provide evidence for a different channel through which gender biases can arise, the *gender criteria gap*, where different criteria are used to evaluate men and women.

We consider a leadership setting where leaders assume responsibility for the outcomes of others (Edelson et al. 2018; Ertac and Gurdal 2012). Specifically, leaders determine the payoffs of a group of individuals through the actions they choose. Group members evaluate the leaders after observing the outcomes which are jointly determined by unobservable actions and luck. Outcomes serve as signals and can assist evaluators to update their beliefs about the actions taken. Evaluators then make remuneration decisions and determine the discretionary payments received by the leaders. Such discretionary payments, which can take the form of bonuses or pay increments, are common features of remuneration packages offered by many organizations. Since the actions taken by the leaders are not observable, we collect data on the evaluators' beliefs, which reflect their perceptions of the leaders' intentions. We assume that leaders can be rewarded (or

penalized) for both their perceived intentions and/or the resulting outcomes. The latter can arise when evaluators find it challenging to assess employees solely based on the merits of the actions taken without being influenced by the outcomes of those actions, a phenomenon referred to as outcome bias.[1]

We analyze to what extent evaluators base their payment decisions on their beliefs about the actions taken by their leaders and to what extent their evaluation is also influenced by the observed outcome itself. If beliefs or outcomes play differential roles in the evaluation of (i.e., discretionary payments received by) male and female leaders, this suggests that male and female leaders are subjected to different evaluation criteria. We define this as the gender criteria gap.

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 leaders' 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 the evaluators' initial beliefs (prior beliefs) about the leaders' actions before observing any outcomes and their updated beliefs (posterior beliefs) after outcomes are realized, how these beliefs compare to the Bayesian benchmark, how they are mapped to discretionary payments, and importantly, how all these vary with gender. This approach is important for establishing the drivers of gender biases 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 while the other two are appointed as members who act as evaluators. The leader's investment decision is implemented for the group. The outcome for the group depends on both the leader's choice, which is unobservable to the evaluators, and luck. 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. The leader's gender is revealed to the group and members 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 leader.

To illustrate the type of leadership paradigm motivating our design, consider politicians 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 depends not only on competency but also on prosocial preferences.[3] The importance of prosocial preferences for leadership outcomes is what we emphasize in our experimental design.[4]

Leaders' investment choices are thus determined by both their altruism toward other group members and the discretionary payments they anticipate to receive. Evaluators, in turn, may recognize that both of these factors shape leaders' decisions and take them into account when determining the discretionary payments. We conjecture that, first, gender discrimination in the discretionary payments may arise due to gender differences in beliefs. For example, 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.[5] Alternatively, gender discrimination in the discretionary payments may be driven by gender differences in evaluation criteria, where evaluators place different weight or emphasis on the leader's outcomes (outcome bias) and/or their beliefs while determining their payments. For example, if evaluators are socially conditioned to think that, irrespective of perceived intentions, more favorable outcomes from men are worthy of higher financial remuneration than those from women, then discretionary payments may be higher for male leaders.

Our results reveal gender differences in discretionary payments. Strikingly, we do not detect any gender differences in evaluators' beliefs about leaders' investment choices. Hence, we show that gender biases in beliefs are not the channel through which gender differences emerge in discretionary payments. However, we detect a *gender criteria gap*. Specifically, while male leaders' discretionary payments are determined by both outcomes and evaluators' beliefs of the leaders' investment choices, female leaders' discretionary payments are predominantly determined by outcomes. Although we do not find a systematic gender difference in the weight placed on outcomes, there is a significant gender difference in the weight placed on beliefs. Hence, different criteria are used in the evaluation of male and female leaders.

Our results thus offer a new perspective on how beliefs contribute to gender biases in the evaluation of leaders. By documenting the presence of a gender criteria gap, our findings imply that policies designed to counteract gender discrimination should not only target biases in beliefs, but also the differential criteria that evaluators may use in their subjective evaluation of men and women. 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. As we discuss in Section 5, our findings underscore the importance of providing clearer guidance to evaluators on the criteria to be used in the performance evaluation process.

Our research contributes to three strands of the literature. First, we relate to the research on gender biases in performance evaluation in diverse domains, such as teaching (e.g., Boring 2017; MacNell et al. 2015; Mengel et al. 2019), music (e.g., Goldin and Rouse 2000), science and medicine (e.g., Jensen et al. 2018; Régner et al. 2019; Sarsons 2022), leadership (e.g., Eckel et al. forthcoming; Erkal et al. 2023; Grossman et al. 2019), financial management (e.g., Egan et al. 2022), retail (e.g., Benson et al. 2026), academia (e.g., Eberhardt et al. 2023), and group work (e.g., Sarsons et al. 2021).[6] Some studies (e.g., Boring 2017; Eberhardt et al. 2023; MacNell et al. 2015; Mengel et al. 2019) suggest that men and women are assessed differently against a predetermined set of criteria.

Distinguishing between belief-based and preference-based discrimination, Bohren et al. (2019) and Coffman et al. (2021) find that discrimination against women tends to be the former rather than the latter.[7] Barron et al. (2024) distinguish between explicit and implicit belief-based discrimination, while Albrecht et al. (2013), Campos-Mercade and Mengel (2023), and Erkal et al. (2023) examine biases in belief updating. While our study also investigates gender differences in beliefs, our main focus is on whether and how evaluators use their beliefs when determining the compensations made to men and women. We contribute to this literature by developing and testing the concept of the gender criteria gap. We do this with a focus on beliefs about intentions and show that perceived intentions is not a criterion used in the evaluation of women.

The second strand of literature we contribute to is the emerging research on beliefs as predictors of behavior. Although economic models, in general, assume that beliefs are drivers of decisions, evidence suggests that the link is often attenuated (e.g., Coibion et al. 2022; Costa-Gomes and Weizsäcker 2008; Giglio et al. 2021). Moreover, information interventions designed to correct mistaken beliefs seem to have modest effects on decisions (Haaland et al. 2023). Yang (2025) shows evidence that this may be due to difficulties individuals face in translating their beliefs into decisions. We contribute to this literature by studying this in the context of discrimination and showing that gender may interact with this translation of beliefs to decision making.

Finally, our study connects to the literature on outcome bias in the compensation of decision makers and advances this literature by providing the first evidence on whether outcome bias depends on the gender of the evaluated agent. 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 et al. (2013) and Brownback and Kuhn (2019) document that individuals' judgment about others' actions 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. Unlike previous studies, intentions are not directly observed in our context and hence, evaluators form beliefs about the leaders' intentions, which we refer to as perceived intentions. We show that there is a gender difference in the role perceived intentions and outcomes play in evaluators' reciprocal decisions. In contrast to previous findings, perceived intentions do not play a role in the case of female leaders. Reciprocity toward female leaders is shaped by outcomes only.

2 | Experimental Design

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

2.1 | 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 be assigned to be evaluators (referred to as "members" in the experiment). All participants are instructed 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.

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 are informed that leaders' investment decisions are not observable, but their outcomes are. Hence, outcomes serve as noisy signals about the leaders' decisions.

Participants complete five investment tasks with different parameterizations as shown in Table 1. The costs and probabilities of success remain constant across all five tasks, while the returns from the investments vary, leading to different payoffs for the leader and the evaluators. We use five investment tasks for two reasons. Our first reason is to test the robustness of our results since the leader's investment decisions may be sensitive to the payoff structure. For instance, Erkal, Gangadharan, and Koh (2022) find that leaders are averse to making the self-interested investment choice if evaluators receive zero payoff in case of failure. Our second reason is to collect multiple decisions from each leader and evaluator to increase the statistical power for our analysis.[10] In all five tasks, 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.

2.2 | 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 evaluators following the approach described in Bordalo et al. (2019). The experimenter calls out each group separately by their group number and asks the group's participants to raise their hands. The experimenter then announces the last three digits of the ID number of the group's leader,[11] and the leader is asked to respond by saying "here."

Our protocol enables the leader's gender to be discreetly revealed to evaluators through their voice. Specifically, participants are seated in individual cubicles with high partitions, ensuring that they are unable to see one another but can hear their leader's

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

Note: 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) + $(0.75 \times 150 + 0.25 \times 0)$ (expected return from Investment X) = 212.5 ECU, while the expected net earnings to each evaluator is given by $(0.75 \times 150 + 0.25 \times 0)$ (expected return from Investment X) = 112.5 ECU.

voice. The greeting is restricted to a single word "here" as in Bordalo et al. (2019) to limit the amount of information that may be conveyed by the leader's voice beyond gender (such as their ethnicity). We ask all group members to raise their hands when their group number is called to avoid drawing any obvious attention to the leader's announcement. Finally, from this point onward in the experiment, whenever there is a reference to the leader, evaluators see on their computer screens the pronoun corresponding to their leader's gender (see Appendix Figure C1 for examples of decision screens for evaluators who are matched with a female leader). Hence, information about the leader's gender is first introduced to the evaluators using the leader's voice, and it is further reinforced throughout the experiment using gender-specific pronouns. Using this protocol, differences in evaluations can be attributed to differences in the leader's gender.[12]

2.3 | Leadership Task — Stage 2: Elicitation of Beliefs and Discretionary Payments

After the groups and roles are revealed, for each investment task, we elicit evaluators' beliefs about the investment choices of the leaders. We regard these beliefs as the perceived intentions of the leader. Evaluators report two sets of beliefs about their 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 (Erkal et al. 2020; Hossain and Okui 2013).

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 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.

2.4 | 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 each session. In the majority of cases (82%), each leader was matched with one male evaluator and one female evaluator. While we reveal the gender of the leader to the group using the leader's voice, the gender composition of the evaluators in the group is not revealed.

Participants received printed instructions for 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 order in which the five investment tasks were presented to participants was 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.[13] Participants earned 41.09 AUD on average.

We collected data from 591 participants in two waves, with the first wave in 2019 (351 participants) and the second wave in 2025 (240 participants). The session sizes were large, with most sessions having 24 participants (eight groups). All 591 participants made investment decisions as leaders in the experiment before the roles were revealed. After the roles were revealed, 394 evaluators reported their beliefs and made discretionary payment decisions for 197 leaders.[14]

Given our sample size, our power calculations suggest that we are able to detect a 0.129 standard deviation difference between female and male leaders in payoff adjustments and prior beliefs. Simulations using data from Erkal, Gangadharan, and Koh (2022) reveal that we are able to detect a gender difference of 0.2–0.25 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 4.2.1). To determine our statistical power for detecting the gender criteria gap, we performed power calculations using simulations based on the data from the first wave ($N = 233$ evaluators). The combined dataset provides us with a statistical power of 0.95 to detect the gender criteria gap observed in the first wave at the 5% level, and a power of at least 0.85 to detect the same gender criteria gap at the 1% level.

3 | Conceptual Framework

In this section, we describe what factors may influence leaders' and evaluators' decisions in the context of our experiment. This discussion is based on a simple model presented in Appendix B and will guide our analysis in the next section.

In our experiment, each leader makes an unobservable discrete investment choice of X or Y on behalf of a group of N players. We assume the leaders' investment choices to be determined by both their type (i.e., the weight they put on others' payoff relative to their own payoff) and the discretionary payments they expect to receive. If the expected discretionary payments are sufficiently high, then even a purely self-interested leader will find it optimal to choose Investment X. However, for lower expected discretionary payments, there exists a leader type who is indifferent between choosing Investment X and Investment Y. Leaders with a higher concern for others' payoffs than this threshold type will choose Investment X.

The parameterization of the investment tasks in the experiment is such that a self-interested leader will never choose Investment X if they expect zero discretionary payments, and they will always choose Investment X if they expect the maximum payment for a high outcome and the lowest possible payment for a low outcome. Hence, for intermediate levels of expected discretionary payments, which is what we observe in our data, whether a leader will choose Investment X will depend on their relative concern for others' payoff.

To form beliefs about the choices made by the leaders, evaluators form second-order beliefs on the leaders' expectations of discretionary payments. Given these beliefs, they can solve for the leader type who is indifferent between choosing Investment X and Investment Y. Hence, conceptually, when evaluators state their beliefs on the likelihood that the leader chooses Investment

X, they are stating their beliefs about the leader having sufficiently high social preferences. These beliefs, which we refer to as perceived intentions, may be different between male and female leaders if evaluators expect them to differ both in terms of the distributions from which they draw their types and in terms of their expectations on discretionary payments.

After observing the outcome Q, evaluators update their prior beliefs and decide 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, the 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 regard them as reciprocal actions.

We consider a model of reciprocity where reciprocal actions are potentially determined by both (perceived) intentions (as represented by the evaluators' posterior beliefs) and outcomes.[15] 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., Brownback and Kuhn 2019; Gurdal et al. 2013). Hence, in this framework, outcomes 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.

Different evaluators may put different weights on the perceived intentions versus the realized outcome. As an extreme case, if two evaluators care about intentions only and if they have the same posterior beliefs, then they will choose the same discretionary payment irrespective of the outcome. On the other hand, if two evaluators care about outcomes only and if they observe different outcomes, then they will choose different discretionary payments even if their posterior beliefs are the same.

In this framework, gender differences in discretionary payments may stem from gender differences in beliefs or perceived intentions, gender differences in the direct weight put on outcomes (outcome bias), and gender differences in the weight put on perceived intentions. Evaluators use different criteria in the determination of the payments for male and female leaders if they put different weights on outcomes and different weights on beliefs depending on the leader's gender. We refer to the gender difference we may observe in the criteria used as the *gender criteria gap*.

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 outcomes and beliefs about the leaders' intentions?
4. If so, are there gender differences in the weights evaluators put on outcomes versus perceived intentions? 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?

4 | Results

We start our analysis by first examining whether there are gender differences in discretionary payments in Section 4.1 (Research Question 1). Then, in Section 4.2, we examine the channels driving the differences in discretionary payments, specifically whether there are gender differences in beliefs (Research Question 2) and in the weights placed by evaluators on perceived intentions versus outcomes (Research Questions 3 and 4). Finally, we examine the robustness of our results in Section 4.3.

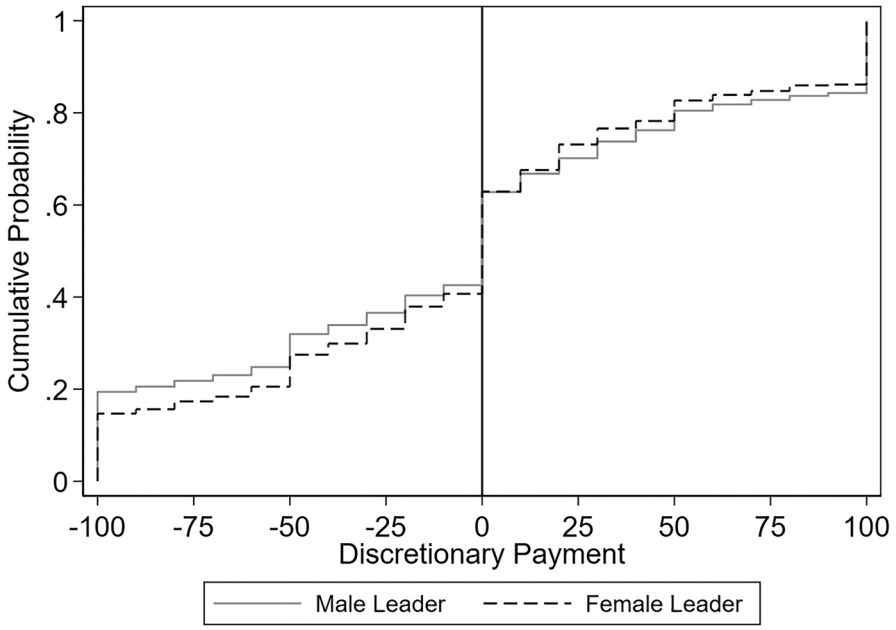
4.1 | Evaluators' Discretionary Payments

We first examine the overall payments made to female and male leaders (Research Question 1). Overall, 78.8% of leaders receive either a positive discretionary payment (i.e., bonus) or a negative discretionary payment (i.e., penalty) from evaluators, with male and female leaders being equally likely to not receive any discretionary payments (20.2% of male leaders and 22.2% of female leaders; p-value = 0.502).[16] However, panel (a) of Figure 1 shows that there is a statistically significant gender difference in the distribution of discretionary payments (Kolmogorov–Smirnov test: p-value = 0.017). Specifically, penalties given to male leaders are higher than those given to female leaders. On the other hand, male leaders receive higher bonuses than female leaders. In panel (b), we observe that this is driven by female leaders being less likely to receive the maximum penalty (i.e., −100 ECU) or the maximum bonus (i.e., +100 ECU) as compared to male leaders. We summarize as follows.

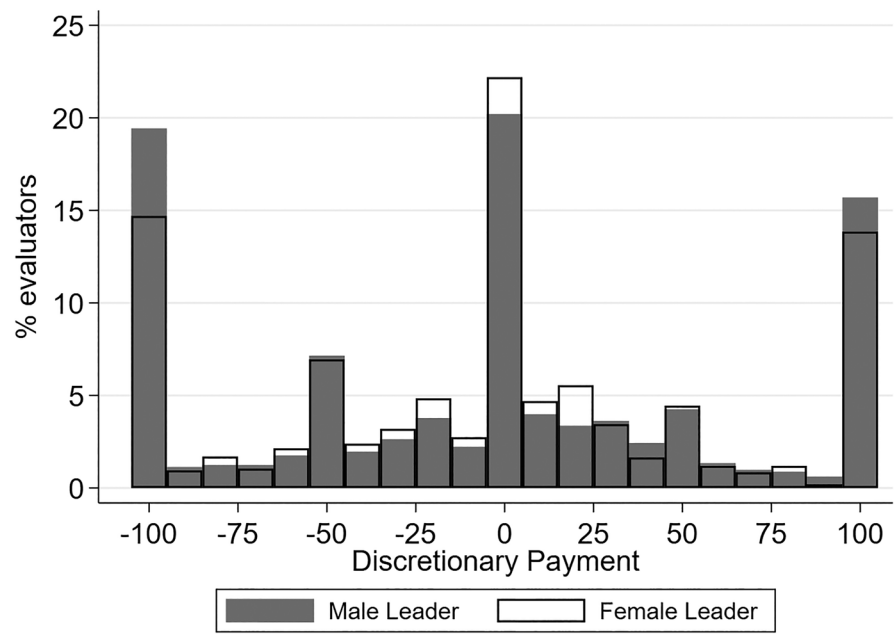
Result 1. *Male and female leaders are equally likely to receive a discretionary payment. However, the distribution of discretionary payments differs between male and female leaders.*

We next investigate whether leaders' expectations of the discretionary payments are in line with the actual discretionary payments they receive. Panel (a) of Figure 2 presents the empirical cumulative distributions of female and male leaders' expectations regarding these payments. The panel reveals that male and female leaders expect different levels of discretionary payments (Kolmogorov–Smirnov test: p-value < 0.001). For bonuses, we observe that, consistent with evidence on the "kernel of truth" (e.g., see Bordalo et al. 2019), the gender difference in leaders' expectations shown in panel (a) is in the same direction but much larger than the gender difference observed in Figure 1. For penalties, the gender differences are reversed. While female leaders receive lower penalties than male leaders, they *expect* to receive higher penalties than male leaders.

Panels (b) and (c) of Figure 2 compare the actual discretionary payments that leaders receive (solid lines) with their expectations of discretionary payments (dashed lines). We find that the distributions of expectations of both female leaders (panel b) and



(a) Empirical cumulative distributions of discretionary payments

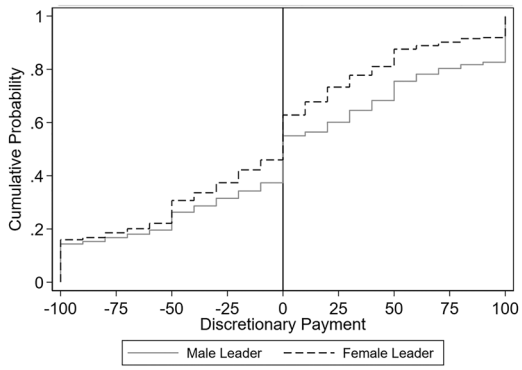


(b) Histograms of discretionary payments

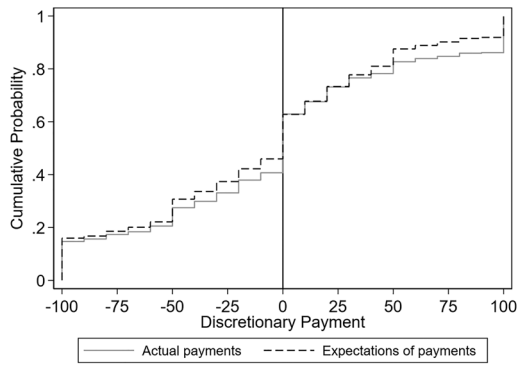
FIGURE 1 | Distributions of discretionary payments, by the leader's gender. *Note:* Panel (a) presents the empirical cumulative distribution of discretionary payments for male leaders (solid line) and female leaders (dashed line). The distributions are statistically significantly different between genders (Kolmogorov–Smirnov test: p-value = 0.017). Panel (b) presents the distribution of payments as a histogram and reveals that much of this difference is driven by female leaders being less likely than male leaders to receive the maximum penalty (−100 ECU) or the maximum bonus (+100 ECU).

male leaders (panel c) differ significantly from the distributions of actual discretionary payments received (Kolmogorov–Smirnov tests: p-values = 0.024 and <0.001, respectively).

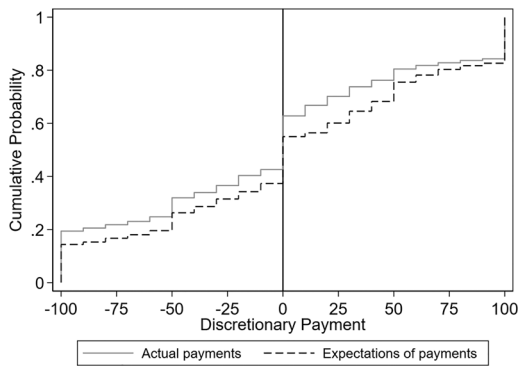
More importantly, male and female leaders' expectations depart from the actual payments in different ways. Specifically, female leaders tend to be pessimistic, expecting higher penalties and lower bonuses than what they receive. On the other hand, male leaders are optimistic in that they expect lower penalties and higher bonuses than what they receive from the evaluators. Our findings suggest that leaders' expectations about discretionary payments may be influenced by gender differences in confidence



(a) Male versus female leaders (expectations about payments)



(b) Expectations versus actual (female leaders)



(c) Expectations versus actual (male leaders)

FIGURE 2 | Leaders' expectations of evaluators' discretionary payment decisions versus evaluators' actual discretionary payment decisions, by the leader's gender. *Note:* Panel (a) presents the empirical cumulative distribution of leaders' beliefs about the discretionary payments that they will receive, separately for male leaders (solid line) and female leaders (dashed line). The distributions are statistically significantly different between gender (Kolmogorov–Smirnov test: p-value < 0.001). Panels (b) and (c) compare the actual discretionary payments that leaders receive (solid line) with their expectations of discretionary payments (dashed line). The distributions of expectations differ from the distributions of actual payments received for both female leaders (Kolmogorov–Smirnov test: p-value = 0.024) and male leaders (Kolmogorov–Smirnov test: p-value < 0.001).

or by societal norms about the role gender plays in shaping discretionary payments.

4.2 | Channels Driving Gender Differences in Discretionary Payments

In this section, we explore potential channels driving discretionary payment decisions. We first outline our estimation strategy in Section 4.2.1, and we report our results in Sections 4.2.2 and 4.2.3.

4.2.1 | Estimation Strategy

We start by investigating whether evaluators' prior beliefs and updating behavior differ between male and female leaders (Research Question 2). In all our analyses, belief is a variable that takes an integer value in the range [0, 100], where a higher belief implies that the evaluator thinks the leader is more likely to have chosen Investment X.

To examine gender differences in evaluators' prior beliefs, we estimate the following equation:

$$\mu_j = \beta_0 + \beta_1 \text{Female}_L + \beta_2 \text{Investment X}_j + \boldsymbol{\beta X} + \varepsilon_j, \quad (1)$$

where μ_j is evaluator j's prior belief that the leader has chosen Investment X, Female_L indicates whether the leader is female, Investment X_j is an indicator variable for whether evaluator j chooses Investment X in Stage 1, $\boldsymbol{X}$ represents a vector of covariates for task parameters and the leader's characteristics, and ε_j captures non-systematic errors. β_2 allows us to examine whether evaluators' own investment choice as leaders influences their prior beliefs, that is, whether a consensus effect is present (Erkal, Gangadharan, and Koh 2022). Our primary outcome variable of interest is β_1, which tests whether there is any gender difference in evaluators' prior beliefs.

To examine gender differences in posterior beliefs, we consider the following econometric specification:

$$\begin{aligned}\text{logit}\left(\sigma_j(X|Q)\right) &= \delta \text{logit}\left(\mu_j\right) + \gamma_H I(Q = Q_H) \cdot \text{logit}(p) \\ &+ \gamma_L I(Q = Q_L) \cdot \text{logit}(1-p) + \varepsilon_j, \end{aligned} \quad (2)$$

where $\text{logit}z = \log(\frac{z}{1-z})$,[17] $I(\cdot)$ is an indicator function for the observed outcome (Q) of the investment, $\sigma_j(X|Q)$ and μ_j represent evaluator j's reported posterior beliefs (given Q) and prior beliefs, respectively, and ε_j captures nonsystematic errors. p and $1-p$ denote the probability of a high outcome under Investments X and Y, respectively. In our experiment, $p = 0.75$.

The specification in (2) nests the theoretical Bayesian benchmark as a special case with $\delta = \gamma_H = \gamma_L = 1$, and any deviation in the estimated parameters from 1 is interpreted as non-Bayesian updating behavior. Hence, this specification allows for a comprehensive analysis of systematic biases in belief updating.[18]

Deviations from the Bayesian benchmark may arise from base-rate neglect and/or an under- or over-responsiveness to signals, independent of the evaluators' priors. The main parameters of

interest are γ_H and γ_L, which represent the evaluator's response to a signal of high outcome (i.e., investment succeeds) and low outcome (i.e., investment fails), respectively, when updating their beliefs. $\gamma_H < 1$ ($\gamma_L < 1$) 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.[19] We estimate Equation (2) using ordinary least squares (OLS) and compare the coefficients for male and female leaders to analyze whether evaluators exhibit gender biases when updating their beliefs about the leader's investment choices.

Next, we evaluate whether evaluators' discretionary payments are shaped by perceived intentions and/or the leaders' outcomes (Research Question 3), and whether there are gender differences in the weights that evaluators place on perceived intentions versus outcomes (Research Question 4). To do this, we estimate the following equation:

$$\Delta_j = \beta_0 + \beta_1 \sigma_j + \beta_2 I(Q = Q_H) + \boldsymbol{\beta X} + \varepsilon_j, \tag{3}$$

where Δ_j is the discretionary payment given by evaluator j, $I(\cdot)$ and σ_j are the leader's outcome and evaluator's posterior belief defined in a similar way as in Equation (2), $\boldsymbol{X}$ represents a vector of covariates for task parameters and the leader's characteristics, and ε_j represents non-systematic errors.

The coefficients of interest are β_1 and β_2, which capture the weights evaluators place on perceived intentions and the leader's outcomes, respectively, in determining discretionary payments. Note that our analysis in Equation (2) allows us to evaluate whether the differences in discretionary payments are driven by differences in the posterior beliefs (σ_j) that enter Equation (3). Importantly, independent of whether there are gender differences in beliefs, β_1 in Equation (3) allows us to evaluate whether and how these beliefs affect decisions about discretionary payments. β_2, on the other hand, allows us to capture the direct effect of the leader's outcome over and beyond its impact via beliefs; that is, it allows us to measure outcome bias. According to the informativeness principle in contract theory (Bolton and Dewatripont 2005), outcomes should not matter in determining discretionary payments over and beyond their role as signals for belief updating. Hence, β_2 measures whether an evaluator overweighs a signal relative to its informational content.

Since we are interested in whether these weights differ based on the leader's gender, we compare β_1 and β_2 between male and female leaders. To do this, we run two separate regressions (one for male leaders and one for female leaders) using seemingly unrelated estimation. We then conduct a Wald test to examine the equality of coefficients (separately for high outcome and posterior belief) across the two models, clustering standard errors at the evaluator level.[20]

We present all our analysis taking into account evaluators who do not update their beliefs or who update their beliefs inconsistently during the experiment. Specifically, panels (a) and (b) of Appendix Figure C2 present the distribution of evaluators who do not update their beliefs (i.e., have posterior beliefs equal to prior beliefs) or update their beliefs inconsistently (i.e., have posterior beliefs in the opposite direction to that predicted by Bayes' rule), respectively. We classify an evaluator as a non-updater if all their posterior beliefs are equal to their prior beliefs for all five rounds of the investment task (25 evaluators, 6.4% of the sample), and as an inconsistent updater if 25% or more of their posterior beliefs are in the opposite direction to that predicted by Bayes' rule (78 evaluators, 19.9% of sample).[21]

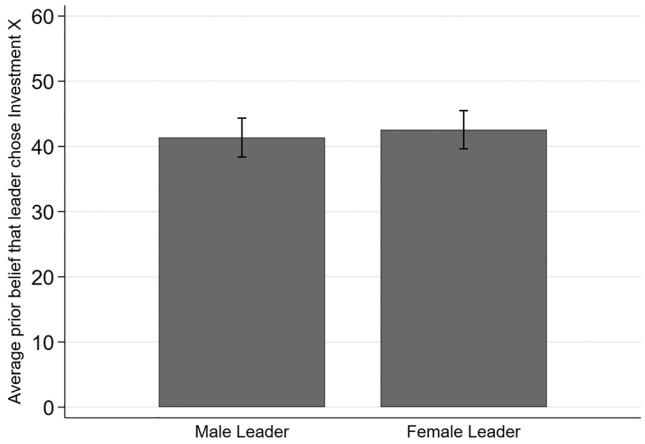


FIGURE 3 | Evaluators' prior belief that the leader has chosen Investment X, by the leader's gender. *Note:* The figure presents evaluators' mean prior belief that the leader has chosen Investment X, separately for male and female leaders. There are no statistically significant differences in evaluators' average prior beliefs about the investment choices made by male and female leaders (p-value = 0.570). Error bars represent 95% confidence intervals accounting for standard errors clustered at the participant level.

One concern is that non-updaters and inconsistent updaters may not be paying attention to the experiment or have a lower understanding of the decision-making environment.[22] Consequently, we also present our results by excluding these evaluators as their inclusion may lead to biased conclusions, since the empirical test of the relationship between the evaluators' discretionary payments and their beliefs relies on the assumption that the evaluators' belief reports reflect their true beliefs.[23]

4.2.2 | Are There Gender Differences in Evaluators' Beliefs?

We first examine whether there exist gender differences in evaluators' prior and posterior beliefs about the leader that can explain the gender difference in discretionary payment decisions.

Figure 3 presents evaluators' prior beliefs that the leader has chosen Investment X.[24] The figure 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 = 0.570). OLS estimates of Equation (1) reported in Table 2 yield similar conclusions. In column (1) of the table, we control for the evaluators' investment decisions as leaders, the difference in investment returns between a successful and a failed investment, and whether the investment provides a return of zero in case of failure. 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,

TABLE 2 | OLS regressions of evaluators' prior belief that the leader has chosen Investment X.

	Dependent variable: Prior belief			
Variables	**(1)**	**(2)**	**(3)**	**(4)**
Female leader	0.115	0.155	−0.481	−0.257
	(1.736)	(1.731)	(1.742)	(1.970)
Chose Investment X as leader	22.010***	20.915***	20.320***	20.663***
	(1.616)	(1.583)	(1.586)	(1.770)
High Return—Low Return	0.074***	0.076***	0.082***	0.088***
	(0.014)	(0.014)	(0.015)	(0.016)
Zero return if investment fails	−0.080	0.004	0.374	0.173
	(1.169)	(1.169)	(1.232)	(1.417)
Inconsistent or non-updater		−8.243***		
		(2.060)		
Constant	11.567***	12.309*	11.255	9.701
	(3.895)	(6.664)	(6.863)	(7.961)
Control for task order	✓	✓	✓	✓
Control for Wave 2 data	✓	✓	✓	✓
Control for beliefs about leader's DG behavior	✓	✓	✓	✓
Individual controls		✓	✓	✓
Excluding non-updaters			✓	✓
Excluding inconsistent updaters				✓
Observations	1965	1965	1840	1450
# participants (clusters)	393	393	368	290
R^2	0.210	0.228	0.190	0.210

Note: Robust standard errors clustered at the participant level in parentheses. Investment X refers to the costlier investment option for the leader but yields a higher success probability. Wave 2 data are the second wave of data collection conducted in 2025. We classify an evaluator as a non-updater if all their posterior beliefs are equal to their prior beliefs for all five rounds of the investment task, and as an inconsistent updater if 25% or more of their posterior beliefs are in the opposite direction to that predicted by Bayes' rule. In columns (2)–(5), 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.

***$p < 0.01$, **$p < 0.05$, *$p < 0.10$.

previous experience with economics experiments, CRT score, and whether they are classified as either an inconsistent updater or a non-updater. Finally, we demonstrate the robustness of our results with the exclusion of non-updater (column (3)), and of both non-updaters and inconsistent updaters (column (4)).[25]

Turning to evaluators' updating behavior, Table 3 presents OLS regression estimates of Equation (2) separately by the leader's gender. In the full sample (panel (a)), 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 = 0.690 and 0.456, respectively). Importantly, the results are robust to the exclusion of evaluators classified as non-updaters and inconsistent updaters (panels (b) and (c)).[26,27]

In sum, we find no gender differences in both prior beliefs and updating behavior. Consequently, we observe no statistically significant gender differences in evaluators' posterior beliefs given both a high and a low outcome (t-tests: p-values = 0.636 and 0.314, respectively). We summarize this result as follows.

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

Even though evaluators expect male and female leaders to be equally likely to choose high investment, they may have different beliefs about their motivations. Hence, a full understanding requires data on their beliefs about leaders' motivations. As discussed in Section 3, leaders' investment decisions are driven by their altruism and their beliefs about the discretionary payments they will receive. This is consistent with the analysis presented in Appendix D. While we do not have data on evaluators' beliefs about leaders' expectations of discretionary payments, we use evaluators' beliefs about the leader's behavior in the dictator game as a proxy for their beliefs about the leader's altruism. Appendix Figure C4 reveals that evaluators anticipate female leaders to transfer higher amounts to their matched partner

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**
(a) Full sample			
δ: Logit (prior belief)	0.577***	0.578***	0.985
	(0.060)	(0.060)	
γ_H: High outcome × logit (p)	0.779**	0.839	0.690
	(0.099)	(0.111)	
γ_L: Low outcome × logit $(1-p)$	0.701***	0.823	0.456
	(0.113)	(0.120)	
$\gamma_H - \gamma_L$	0.079	0.015	
	(0.150)	(0.142)	
Observations	2000	1930	
# participants (clusters)	200	193	
R^2	0.415	0.428	
(b) Excluding non-updaters			
δ: Logit (prior belief)	0.469***	0.460***	0.914
	(0.059)	(0.057)	
γ_H: High outcome × logit (p)	0.811*	0.940	0.407
	(0.099)	(0.120)	
γ_L: Low outcome × logit $(1-p)$	0.764**	0.842	0.641
	(0.111)	(0.126)	
$\gamma_H - \gamma_L$	0.047	0.098	
	(0.140)	(0.155)	
Observations	1880	1800	
# participants (clusters)	188	180	
R^2	0.322	0.325	
(c) Excluding non-updaters and inconsistent updaters			
δ: Logit (prior belief)	0.568***	0.517***	0.583
	(0.062)	(0.068)	
γ_H: High outcome × logit (p)	0.951	1.100	0.390
	(0.109)	(0.136)	
γ_L: Low outcome × logit $(1-p)$	1.040	1.049	0.960
	(0.113)	(0.136)	
$\gamma_H - \gamma_L$	−0.089	0.052	
	(0.138)	(0.170)	
Observations	1450	1450	
# participants (clusters)	145	145	
R^2	0.469	0.397	

Note: Robust standard errors clustered at the participant level in parentheses. Investment X refers to the costlier investment option for the leader but yields a higher success probability. We classify an evaluator as a non-updater if all their posterior beliefs are equal to their prior beliefs for all five rounds of the investment task, and as an inconsistent updater if 25% or more of their posterior beliefs are in the opposite direction to that predicted by Bayes' rule. 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(High Outcome) + I (Low Outcome) = 1, there is no constant term in the regression.

***$p < 0.01$, **$p < 0.05$, *$p < 0.10$. Null hypothesis is coefficient = 1.

in the dictator game than male leaders, although this result is marginally statistically significant (Kolmogorov–Smirnov test: p-value = 0.076). Hence, there is suggestive evidence that evaluators anticipate female leaders to be more altruistic than male leaders.[28] Given Result 2, our conjecture is that evaluators believe male leaders anticipate higher discretionary payments than female leaders.[29] That is, if evaluators expect female leaders to be more prosocial *and* male leaders to have higher expectations of discretionary payments, then these two opposing effects may explain why we do not see gender differences in their prior beliefs about the leaders' investment decisions. Moreover, the lack of a gender difference in posterior beliefs suggests that evaluators do not respond to signals from male and female leaders differently.

4.2.3 | Are There Gender Differences in the Criteria Used to Determine Discretionary Payments?

Despite not observing any gender differences in beliefs (Result 2), we observe gender differences in the discretionary payments received by leaders (Result 1). To investigate this issue further, we turn to the drivers of evaluators' discretionary payment decisions.

Figure 4 presents fitted line plots of evaluators' discretionary payments against their posterior beliefs separately for female leaders and male leaders. The lines are based on estimates of OLS regressions of evaluators' discretionary payments using Equation (3), as presented in Table 4, along with 95% confidence intervals.[30]

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 < 0.001), evaluators' posterior beliefs do not play a role in shaping payments to female leaders (column (1): p-value = 0.131). The difference between female and male leaders in the estimated impact of evaluators' posterior beliefs on discretionary payments is statistically significant (p-value = 0.044). Importantly, the gender difference in the coefficients on posterior beliefs is economically large. An increase in evaluators' beliefs leads to an increase in discretionary payments, that is, 202% higher for male leaders than for female leaders.[31] We also observe that the gender criteria gap in posterior beliefs persists (and becomes even stronger in magnitude and statistical significance) when we exclude non-updaters (columns (3) and (4)) and both non-updaters and inconsistent updaters (columns (5) and (6)).

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. The direct impact of outcomes on discretionary payments is slightly higher for female leaders than for male leaders. Nonetheless, the gender difference in the emphasis on outcomes is smaller than that on beliefs. For female leaders, keeping beliefs constant, achieving a high outcome leads to an increase in discretionary payments, that is, only 23% higher than that for male leaders.[32] While this difference is not statistically significant in the full sample (column (1) vs. column (2): p-value = 0.151), it is statistically significant with the exclusion of non-updaters (column (3) vs. column (4): p-value = 0.089) and the exclusion of both non-updaters and inconsistent updaters (column (5) vs. column (6): p-value = 0.044).[33] Hence, there is weak evidence that the outcome bias is larger for female leaders than for male leaders, with this result holding only when we exclude both non-updaters and inconsistent updaters.

We also investigate whether outcomes play a larger role than beliefs in driving discretionary payment decisions by computing the ratio of the coefficients on outcomes versus beliefs. These are reported in Table 4. In the main sample, this ratio is 386.4 for female leaders (column (1)) and 104.3 for male leaders (column (2)). This implies that for female leaders, evaluators have to shift their beliefs by 386 percentage points in order to have the same direct effect that a high outcome has on discretionary payments. However, for male leaders, the corresponding shift in evaluators' posterior beliefs is much lower, at 104 percentage points. We observe similar patterns even with the exclusion of non-updater and inconsistent updaters.[34]

We summarize our results in the following way.

Result 3. *There exists a gender criteria gap in the determination of discretionary payments. This is driven by different weights evaluators place on perceived intentions of the leaders.*

The gender criteria gap can thus potentially explain why there are differences in the discretionary payments 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 receive lower penalties than male leaders when beliefs are low, male leaders end up with higher bonuses than female leaders when beliefs are high. 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 18% of evaluators award a bonus for a low outcome, and this is more likely for male leaders (20%) than for female leaders (15%) (Fisher's exact test: p-value = 0.004).[35,36]

4.3 | Further Analyses

In this section, we evaluate the robustness of the gender criteria gap by considering: (i) evaluators who are unable to correctly predict their leaders' gender at the end of the experiment; and (ii) the evaluator's gender. Our core findings remain qualitatively unchanged when we analyze the data accounting for these factors.

First, not all evaluators correctly predict their leader's gender in the postexperimental questionnaire. For the 90.1% of evaluators who make accurate predictions, the accuracy rate does not depend on the gender of the leader or the evaluator (p-values = 0.606 and 0.451, respectively). In our main analysis, we categorize the data according to the evaluator's predictions of the leader's gender. For robustness, in this section, we restrict our analysis of the determinants of discretionary payments to those evaluators who correctly predict the leader's gender. As shown in Appendix Table C6, Result 3 remains robust to the exclusion of evaluators who incorrectly predict the leader's gender.

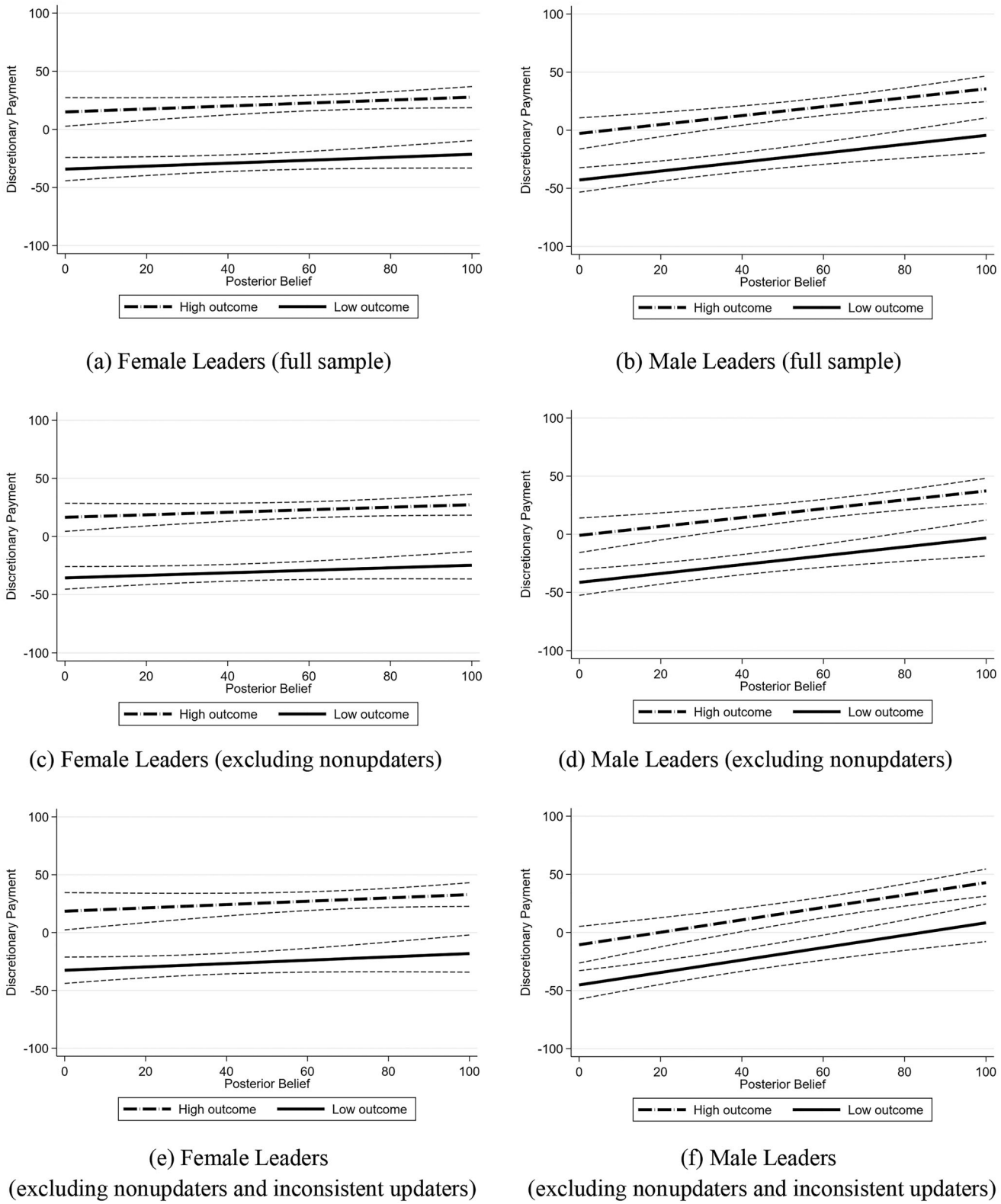


FIGURE 4 | Fitted line of discretionary payments against evaluators' posterior belief that the leader has chosen Investment X and leader's outcomes, by the leader's gender. *Note:* The figure presents fitted line plots of evaluators' discretionary payments against their posterior beliefs separately for female leaders (left three panels) and male leaders (right three panels). The lines are based on estimates of OLS regressions of evaluators' discretionary payments using Equation (3), as presented in Table 4. The dashed lines above and below each fitted line represent 95% confidence intervals.

TABLE 4 | OLS regressions of discretionary payments, by the leader's gender.

	Dependent variable: Discretionary payments								
Variables	**Female leader (1)**	**Male leader (2)**	***p*-Value (1) vs. (2)**	**Female leader (3)**	**Male leader (4)**	***p*-Value (3) vs. (4)**	**Female leader (5)**	**Male leader (6)**	***p*-value (5) vs. (6)**
High outcome	49.192***	40.033***	0.151	52.099***	40.479***	0.089	51.016***	34.562***	0.044
	(4.142)	(4.890)		(4.341)	(5.336)		(5.852)	(5.760)	
Posterior belief	0.127	0.384***	0.044	0.109	0.382***	0.034	0.144	0.535***	0.008
	(0.084)	(0.097)		(0.080)	(0.102)		(0.104)	(0.104)	
Inconsistent or non-updater	−14.789**	−6.081							
	(6.482)	(7.760)							
% endowment transferred in DG	0.388**	0.452**		0.278	0.345*		0.253	0.305	
	(0.192)	(0.203)		(0.206)	(0.197)		(0.237)	(0.222)	
High Return—Low Return	0.005	−0.021		0.010	−0.024		0.014	−0.039*	
	(0.013)	(0.016)		(0.013)	(0.018)		(0.015)	(0.020)	
Zero return if investment fails	−2.273*	−0.717		−2.079	−0.464		−1.436	0.001	
	(1.289)	(1.628)		(1.336)	(1.716)		(1.604)	(2.035)	
Constant	−34.098**	−34.685***		−40.140***	−33.354**		−37.835**	−24.820*	
	(13.179)	(12.464)		(13.873)	(13.505)		(14.670)	(14.842)	
Outcome/Belief: Coefficient	386.4	104.3		479.7	106.0		353.7	64.6	
Test of high outcome = 100 × Belief: *p*-value[a]	0.001	0.897		<0.001	0.867		0.014	0.182	
Control for task order	✓	✓		✓	✓		✓	✓	
Control for Wave 2 data	✓	✓		✓	✓		✓	✓	
Control for beliefs about leader's DG behavior	✓	✓		✓	✓		✓	✓	
Excluding non-updaters				✓	✓		✓	✓	
Excluding inconsistent updaters							✓	✓	
Observations	2000	1930		1880	1800		1450	1450	
# participants (clusters)	200	193		188	180		145	145	
R^2	0.255	0.222		0.250	0.216		0.247	0.230	

Note: Robust standard errors clustered at the participant level in parentheses. Wave 2 data are the second wave of data collection conducted in 2025. We classify an evaluator as a non-updater if all their posterior beliefs are equal to their prior beliefs for all five rounds of the investment task, and as an inconsistent updater if 25% or more of their posterior beliefs are in the opposite direction to that predicted by Bayes' rule.

[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). ***$p < 0.01$, **$p < 0.05$, *$p < 0.10$.

Next, we further investigate Result 3 by separating the analysis based on the evaluator's gender (Appendix Table C7). 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, or they may even treat female leaders less favorably (e.g., Arvate et al. 2018; Derks et al. 2016).

The analysis we present here is exploratory due to the smaller number of observations we have resulting in potentially lower statistical power. We observe evidence that the gender criteria gap is exhibited by female evaluators. First, columns (2) and (4) of panel (a) reveal that the payments made to male leaders, by both male and female evaluators, are increasing in their posterior beliefs (p-values = 0.005 and 0.004, respectively). On the other hand, columns (1) and (3) of panel (a) reveal that the discretionary payments made to female leaders do not depend on their posterior beliefs for female evaluators, but they do for male evaluators (columns (1) and (3): p-values = 0.302 and 0.012, respectively). Importantly, the gender difference in the weight placed on beliefs in the determination of discretionary payments is statistically significant for female evaluators (p-value = 0.006) but not for male evaluators (p-value = 0.719). This result is robust to the exclusion of non-updaters (panel (b)) and both non-updater and inconsistent updaters (panel (c)).

5 | Discussion

Although subjective performance measures are often used in performance evaluation, they are prone to various biases. We examine whether different criteria are used in the evaluation of male and female leaders. In our setting, outcomes are determined by a combination of the leaders' choices and luck, and costly investment choices 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 outcomes.

We find significant differences in the distribution of discretionary payments for male and female leaders. Evaluators choose steeper discretionary payments for male leaders than for female leaders, with male leaders being more likely to receive both the maximum penalty and the maximum bonus than female leaders. Moreover, there is a gender difference in leaders' expectations of discretionary payments, with male leaders being optimistic and female leaders being pessimistic about the payments they will receive.

Investigating the channels driving discretionary payments, our analysis shows that the gender difference in these payments cannot be explained by gender differences in perceived intentions. Instead, we find that it stems from a gender criteria gap. Specifically, while male leaders' discretionary payments are determined by both outcomes and perceived intentions, female leaders' 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 their actions; however, in the case of female leaders, the deviation is such that evaluators do not use their beliefs at all in determining discretionary payments.

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 et al. (2026) 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 the interests of the group.

A potential reason for the gender criteria gap is that evaluators believe male and female leaders are driven by different motivations in the decisions they make, possibly shaped by social norms. This may cause evaluators to emphasize different criteria when evaluating men and women. A more complete understanding of this conjecture would require additional data on evaluators' beliefs about leaders' motivations. Another conjecture is that evaluators are more uncertain in their beliefs regarding women than in their beliefs about men. This may arise, for instance, if leadership in general is seen to be a gender-incongruent domain for women. Higher uncertainty in evaluators' beliefs about women could result in less emphasis being placed on those beliefs in the evaluation process. The potential drivers of the gender criteria gap provide avenues for future research.

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 high discretionary payments, but men can receive high discretionary payments 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. While outcomes are determined by both luck and the actions taken, beliefs about the actions taken by female leaders are disregarded. This distinction is consequential because when evaluators rely on outcomes alone, actions are not interpreted through the lens of effort or intent, and female leaders are fully exposed to outcome risk. The disproportionate emphasis given to luck (rather than beliefs) has the potential to distort choices in risky environments and can perpetuate gender gaps. On the other hand, focusing entirely on beliefs about intentions, which are subjective, can also be problematic if the beliefs are not correctly formed. While the optimal mix of objective and subjective incentive structure may depend on organizational objectives, our results imply that employers should take measures to apply the same criteria across genders.

Our findings further our understanding of the factors that may be driving the observed gender gaps in performance pay. For example, if outcomes play too large a role in performance evaluation, then having a review system where evaluators are invited to reflect on and record their beliefs before outcomes are realized may be helpful. Our analysis also reveals that this gender criteria gap is exhibited primarily by female evaluators. This suggests that while increasing the representation of women on hiring committees or company boards may have benefits due to,

for example, the role-model effect, it may not necessarily mitigate other biases that may be present in evaluations.

Our findings point toward 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 perceived intentions versus outcomes may be a source of discrimination, with outcomes playing a disproportionately larger role than beliefs in case of women. This type of gender bias is distinct from biased beliefs and leads to a new channel through which discrimination can occur. In showing this new channel, we consider a setting where 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 is important to examine if the gender criteria gap emerges and leads to gender discrimination in environments where other leadership qualities also play a role in shaping outcomes.

Acknowledgments

We would like to thank the Editor, three anonymous reviewers, Billur Aksoy, Kai Barron, Puja Bhattacharya, Andy Brownback, Pol Campos-Mercade, Samuel Engel, 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, University of Konstanz, Masaryk University, Harvard University, Centre for Leadership at the University of Exeter, and participants at various conferences for their comments and feedback. We gratefully acknowledge funding from the Australian Research Council (DP210102183). Previous version of this paper was circulated under the title "Discrimination in Evaluation Criteria: The Role of Beliefs versus Outcomes." This study has been approved by the Business and Economics Human Ethics Advisory Group at the University of Melbourne (ID 1544873). For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising. Experimental software and replication files are available at: https://github.com/boonhankoh/EGK-gender_criteria_gap.

Open access publishing facilitated by Monash University, as part of the Wiley - Monash University agreement via the Council of Australasian University Librarians

Data Availability Statement

The data that support the findings of this study are openly available in GitHub at https://github.com/boonhankoh/EGK-gender_criteria_gap.

Endnotes

[1] According to 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 "informativeness 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).

[2] In addition to leadership, prosocial motivation plays an important role in many other jobs in the economy (Besley and Ghatak 2018; Bowles and Polanía-Reyes 2012). 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 (Alt and Lassen 2003; Lederman et al. 2005; Persson et al. 1997).

[4] While prosocial preference is an important aspect of leadership, it represents only one dimension of the multifaceted roles that leaders typically assume. Other dimensions of leadership, such as fostering coordination, managing group dynamics, and guiding collective decision making (e.g., De Paola et al. 2022; Gangadharan et al. 2016, 2019; Grossman et al. 2019; Karpowitz et al. 2024; Reuben and Timko 2018; Roy and Houser 2024), are also important but not the focus of our study.

[5] Such beliefs may arise as gender stereotypes suggest that prosocial actions are expected of women to a larger extent (see, e.g., Aguiar et al. 2008; Brañas-Garza et al. 2018; Heilman and Chen 2005; Solnick 2001). In their survey papers, Croson and Gneezy (2009), Niederle (2016), and Bilén et al. (2021) report that in practice, the relative prosociality of women is a context-dependent phenomenon.

[6] Other explanations for gender gaps in the labor market include gender differences in preferences (see, e.g., Croson and Gneezy 2009), and institutional factors (see, e.g., Erkal, Gangadharan, and Xiao 2022; Hernandez-Arenaz and Iriberri 2019; Recalde and Vesterlund 2023).

[7] Individuals may also be influenced by gender stereotypes that shape both prior and updated beliefs about their *own* ability (e.g., Bordalo et al. 2019; Coffman 2014; Coffman et al. 2024).

[8] Instructions can be found in Appendix A.

[9] 10 ECU = 1 AUD.

[10] As an additional robustness check, we examine whether our results are driven by participants' behavior in any specific investment task, by reestimating our main analyses excluding one investment task at a time. Our results remain robust.

[11] The ID number was issued to each participant prior to the start of the experiment. See discussion of the experimental procedures in Section 2.4.

[12] Evaluators are asked to predict the leader's gender and ethnicity in the post-experimental questionnaire. 90.1% of them predict their leader's gender correctly, which is comparable to the rate found in Bordalo et al. (2019). The accuracy rate is independent of the leader's or the evaluator's gender (p-values = 0.606 and 0.451, respectively). In our analysis, we categorize the data according to the evaluators' predictions of the leader's gender. In Section 4.3, we show that our main conclusions remain unchanged if we exclude evaluators who are unable to predict their leader's gender correctly in the post-experimental questionnaire.

[13] We conducted text analysis of participants' free-form text responses to the post-experimental survey question: "*Were there any parts of the experiment that were not clear to you? If yes, please explain briefly*." A research assistant, who was unaware of both the purpose of the study and our results, coded these responses. Of the 590 participants, only 4.6% (27) indicated that they were confused by the payment schemes used in the study. Nonetheless, participants who indicated that they were confused by the payment scheme answered a similar proportion of comprehension questions correctly on the first attempt as compared to the rest of the sample (p-value = 0.317). Taken together, we find only a small proportion of participants reporting that they were confused by the payment scheme used in the study, and this group of participants did not perform worse on the comprehension questions than the rest of the sample.

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

[15] See Falk and Fischbacher (2006). Unlike Falk and Fischbacher (2006), in the simple model we present in Appendix B, we assume that leaders have private information about their types (which determine their altruism). The evaluators' perception of the leaders' underlying intention is determined by the evaluators' belief about the leaders' type.

[16] Unless otherwise stated, for all tests reported in the text, we report p-values of t-tests with standard errors clustered at the participant level.

[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. As robustness, we consider the same estimation where we: (i) drop observations where evaluators state 0 or 100 as either their prior belief or posterior belief; (ii) replace the logit of 0 and 100 with the logit of 0.1 and 99.9 as an approximation; and (iii) replace the logit of 0 and 100 with the logit of 0.0001 and 99.999 as an approximation. Our results on the gender difference in the attribution of leaders' high and low outcomes remain robust in all three specifications.

[18] Moreover, because the theoretical relationship between posterior beliefs and prior beliefs (i.e., Bayes' rule) is non-linear, a common approach used in the literature is to estimate a log-linear transformation of Bayes' rule as defined in (2). See, for example, Grether (1980), Barron (2021), Coutts (2019), and Möbius et al. (2022) for similar estimation approaches. Benjamin (2019) provides a review.

[19] On the other hand, δ represents the emphasis 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.

[20] We also consider an alternative specification with a regression of the pooled data that include interaction terms of the leader's gender. The estimates of the interaction terms, reported in Appendix Table C1, provide similar conclusions.

[21] These proportions are in line with what has been previously found in the literature (Barron 2021; Coutts 2019; Erkal, Gangadharan, and Koh 2022; Erkal et al. 2023; Möbius et al. 2022).

[22] Both non-updaters and inconsistent updaters answer fewer comprehension questions correctly on the first attempt as compared to the rest of the sample (75.3% for non-updaters; 75.6% for inconsistent updaters; 80.5% for rest of sample; Wilcoxon rank-sum tests: p-values = 0.091 and 0.020, respectively). In addition, non-updaters spent significantly less time on both the prior and posterior belief decision screens than the rest of the sample (Wilcoxon rank-sum tests: p-values = 0.038 and 0.002, respectively), although we do observe that inconsistent updaters spent a bit more time on the posterior belief decision screen (Wilcoxon rank-sum test: p-value = 0.043). Finally, inconsistent updaters perform significantly worse in the CRT than the rest of the sample (Wilcoxon rank-sum test: p-value = 0.017).

[23] Appendix Figure C3 presents distributions of discretionary payments by the leader's gender with the exclusion of non-updaters and the exclusion of both non-updaters and inconsistent updaters. We observe similar patterns in discretionary payment decisions to those observed in Figure 1 (Result 1).

[24] Error bars in all figures represent 95% confidence intervals accounting for standard errors clustered at the participant level.

[25] Further examining prior beliefs separately by the evaluator's gender, we do not find any statistically significant gender differences in the prior beliefs held by female and male evaluators (p-values = 0.236 and 0.696, respectively).

[26] The results hold even when the analysis is conducted separately by the leaders' and evaluators' gender (Appendix Table C2). In contrast to this finding, in Erkal et al. (2023), where evaluators cannot make discretionary payment decisions, the authors find gender differences in the attribution of low outcomes but not in the attribution of high outcomes. Our results show that these gender differences in attribution cease to exist in the presence of discretionary payments.

[27] Columns (1) and (2) of Table 3 also reveal that, in the full sample, evaluators are more likely to attribute both high and low outcomes of female leaders to luck than a Bayesian (tests of $\gamma_H = 1$ and $\gamma_L = 1$ in column (1): p-values = 0.026 and 0.009, respectively). 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 $\gamma_L = 1$ in column (2): p-values = 0.147 and 0.142, respectively). This result disappears once we exclude inconsistent updaters (panel (c)). The estimates in Appendix Table C2 suggest that this is largely driven by the behavior of female evaluators. Nonetheless, in all cases, the estimates are close in magnitude and not statistically significantly different between male and female leaders, both overall and for female evaluators.

[28] We also observe that female leaders are more prosocial than male leaders in their dictator game decisions (p-value = 0.036).

[29] This belief would be consistent with leaders' expectations as reported in panel (a) of Figure 2.

[30] In Appendix Figure C5, we also add bubble plots of the discretionary payments separately for high outcomes (gray bubbles) and low outcomes (white bubbles), where the size of each bubble is proportional to the number of observations.

[31] A 10-percentage point increase in evaluators' beliefs leads to an increase in discretionary payments worth 3.84 ECU for male leaders, while this increase is only worth 1.27 ECU for female leaders.

[32] Controlling for beliefs, the discretionary payments of male and female leaders increase by 40.0 ECU and 49.2 ECU, respectively, on average, with a high outcome.

[33] Appendix Table C3 presents estimates with the inclusion of individual controls, where we obtain similar conclusions. In addition, Appendix Table C4 shows that our main conclusions do not change when we consider evaluators' prior beliefs instead of their posterior beliefs as a determinant of discretionary payments. We observe that the gender difference in the weight placed on outcomes is no longer statistically significant when we exclude non-updaters (p-value = 0.510) and when we exclude both non-updaters and inconsistent updaters (p-value = 0.550). It is important to note, however, that the estimated coefficient of high outcome in this analysis cannot be interpreted as outcome bias, since it captures both the indirect effect of outcome (through posterior beliefs) and its direct effect.

[34] Another thought experiment that we can do is to evaluate whether the change in discretionary payments made to leaders is higher for a change in outcome (from low to high) or a change in posterior belief (from a belief of 0 to 100 percentage points that the leader has chosen Investment X). The p-values of this Wald test are also reported in Table 4. 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 = 0.897). This result holds even with the exclusion of non-updaters and inconsistent updaters.

[35] On the other hand, it is also possible for male leaders to receive penalties for high outcomes if the evaluators have sufficiently low beliefs about the male leaders' decisions. For instance, 18% of evaluators impose a penalty for a high outcome, and this is higher for male leaders (21%) than female leaders (16%) (Fisher's exact test: p-value = 0.002).

[36] Another question we can ask is whether the outcomes and beliefs play different roles in determining whether leaders receive positive or negative discretionary payments. Appendix Table C5 presents OLS regressions of whether male and female leaders receive a positive discretionary payment (columns (1) and (2)) or a negative discretionary payment (columns (3) and (4)). The estimates reveal that outcomes matter more for female leaders than for male leaders both when determining whether they receive a positive payment (p-value = 0.058) or a negative payment (p-value = 0.013). Evaluators' beliefs play a role in the determination of both positive and negative payments for male leaders (p-values = 0.022 and <0.001 in columns (2) and (4), respectively), but not for female leaders (p-values = 0.113 and 0.472

in columns (1) and (3), respectively). The gender difference in the importance of beliefs in determining positive payments is small in magnitude and not statistically significant (p-value = 0.484).

References

Aguiar, F., P. Brañas-Garza, R. Cobo-Reyes, N. Jimenez, and L. M. Miller. 2008. "Are Women Expected to be More Generous?" *Experimental Economics* 12, no. 1: 93–98.

Albrecht, K., E. von Essen, J. Parys, and N. Szech. 2013. "Updating, Self-Confidence, and Discrimination." *European Economic Review* 60: 144–169.

Alt, J. E., and D. D. Lassen. 2003. "The Political Economy of Institutions and Corruption in American States." *Journal of Theoretical Politics* 15, no. 3: 341–365.

Arvate, P. R., G. W. Galilea, and I. Todescat. 2018. "The Queen Bee: A Myth? The Effect of Top-Level Female Leadership on Subordinate Females." *Leadership Quarterly* 29, no. 5: 533–548.

Baker, G., R. Gibbons, and K. J. Murphy. 1994. "Subjective Performance Measures in Optimal Incentive Contracts." *Quarterly Journal of Economics* 109, no. 4: 1125–1156.

Baron, J., and J. C. Hershey. 1988. "Outcome Bias in Decision Evaluation." *Journal of Personality and Social Psychology* 54, no. 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., R. Ditlmann, S. Gehrig, and S. Schweighofer-Kodritsch. 2024. "Explicit and Implicit Belief-Based Gender Discrimination: A Hiring Experiment." *Management Science* 71, no. 2: 1600–1622.

Benjamin, D. J. 2019. "Errors in Probabilistic Reasoning and Judgment biases." In *Handbook of Behavioral Economics: Applications and Foundations*, edited by B. D. Bernheim, S. DellaVigna, and D. Laibson, Vol. 2, 69–186. Elsevier.

Benson, A., D. Li, and K. Shue. 2026. "'Potential' and the Gender Promotion Gap." *American Economic Review* 116, no. 2: 375–417.

Bertrand, M., and S. Mullainathan. 2001. "Are CEOs Rewarded for Luck? The Ones Without Principals Are." *Quarterly Journal of Economics* 116, no. 3: 901–932.

Besley, T., and M. Ghatak. 2018. "Prosocial Motivation and Incentives." *Annual Review of Economics* 10: 411–438.

Bilén, D., A. Dreber, and M. Johannesson. 2021. "Are Women More Generous Than Men? A Meta-Analysis." *Journal of the Economic Science Association* 7: 1–18.

Bohren, J. A., A. Imas, and M. Rosenberg. 2019. "The Dynamics of Discrimination: Theory and Evidence." *American Economic Review* 109, no. 10: 3395–3436.

Bolton, P., and M. Dewatripont. 2005. *Contract Theory*. MIT Press.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer. 2019. "Beliefs About Gender." *American Economic Review* 109, no. 3: 739–773.

Boring, A. 2017. "Gender Biases in Student Evaluations of Teaching." *Journal of Public Economics* 145: 27–41.

Bowles, S., and S. Polanía-Reyes. 2012. "Economic Incentives and Social Preferences: Substitutes or Complements?" *Journal of Economic Literature* 50, no. 2: 368–425.

Brañas-Garza, P., V. Capraro, and E. Rascón-Ramírez. 2018. "Gender Differences in Altruism on Mechanical Turk: Expectations and Actual Behaviour." *Economics Letters* 170: 19–23.

Brownback, A., and M. A. Kuhn. 2019. "Understanding Outcome Bias." *Games and Economic Behavior* 117: 342–360.

Campos-Mercade, P., and F. Mengel. 2023. "Non-Bayesian Statistical Discrimination." *Management Science* 70, no. 4: 2549–2567.

Charness, G. 2004. "Attribution and Reciprocity in an Experimental Labor Market." *Journal of Labor Economics* 22, no. 3: 665–688.

Charness, G., and D. I. Levine. 2007. "Intention and Stochastic Outcomes: An Experimental Study." *Economic Journal* 117, no. 522: 1051–1072.

Coffman, K. 2014. "Evidence on Self-Stereotyping and the Contribution of Ideas." *Quarterly Journal of Economics* 129, no. 4: 1625–1660.

Coffman, K., M. Collis, and L. Kulkarni. 2024. "Stereotypes and Belief Updating." *Journal of the European Economic Association* 22, no. 3: 1011–1054.

Coffman, K., C. L. Exley, and M. Niederle. 2021. "The Role of Beliefs in Driving Gender Discrimination." *Management Science* 67, no. 5: 3551–3569.

Coibion, O., Y. Gorodnichenko, and M. Weber. 2022. "Monetary Policy Communications and Their Effects on Household Inflation Expectations." *Journal of Political Economy* 130, no. 6: 1537–1584.

Costa-Gomes, M. A., and G. Weizsäcker. 2008. "Stated Beliefs and Play in Normal-Form Games." *Review of Economic Studies* 75, no. 3: 729–762.

Coutts, A. 2019. "Good News and Bad News Are Still News: Experimental Evidence on Belief Updating." *Experimental Economics* 22, no. 2: 369–395.

Croson, R., and U. Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature* 47, no. 2: 448–474.

de Janvry, A., G. He, E. Sadoulet, S. Wang, and Q. Zhang. 2023. "Subjective Performance Evaluation, Influence Activities, and Bureaucratic Work Behavior: Evidence From China." *American Economic Review* 113, no. 3: 766–799.

De Paola, M., F. Gioia, and V. Scoppa. 2022. "Female Leadership: Effectiveness and Perception." *Journal of Economic Behavior & Organization* 201: 134–162.

Derks, B., C. Van Laar, and N. Ellemers. 2016. "The Queen Bee Phenomenon: Why Women Leaders Distance Themselves From Junior Women." *Leadership Quarterly* 27, no. 3: 456–469.

Eberhardt, M., G. Facchini, and V. Rueda. 2023. "Gender Differences in Reference Letters: Evidence From the Economics Job Market." *Economic Journal* 133, no. 655: 2676–2708.

Eckel, C., N. Erkal, L. Gangadharan, and P. J. Grossman. Forthcoming. "Gender and Leadership: Role of Beliefs." In *Handbook of Experimental Gender Economics*, edited by M. Cubel and C. Schwieren. Edward Elgar Publishing.

Edelson, M. G., R. Polania, C. C. Ruff, E. Fehr, and T. A. Hare. 2018. "Computational and Neurobiological Foundations of Leadership Decisions." *Science* 361, no. 6401: eaat0036.

Egan, M. L., G. Matvos, and A. Seru. 2022. "When Harry Fired Sally: The Double Standard in Punishing Misconduct." *Journal of Political Economy* 130, no. 5: 1184–1248.

Erkal, N., L. Gangadharan, and B. H. Koh. 2020. "Replication: Belief Elicitation With Quadratic and Binarized Scoring Rules." *Journal of Economic Psychology* 81: 102315.

Erkal, N., L. Gangadharan, and B. H. Koh. 2022. "By Chance or By Choice? Biased Attribution of Others' Outcomes When Social Preferences Matter." *Experimental Economics* 25, no. 2: 413–443.

Erkal, N., L. Gangadharan, and B. H. Koh. 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., L. Gangadharan, and E. Xiao. 2022. "Leadership Selection: Can Changing the Default Break the Glass Ceiling?" *Leadership Quarterly* 33, no. 2: 101563.

Ertac, S., and M. Y. Gurdal. 2012. "Deciding to Decide: Gender, Leadership and Risk-taking in Groups." *Journal of Economic Behavior & Organization* 83, no. 1: 24–30.

Falk, A., and U. Fischbacher. 2006. "A Theory of Reciprocity." *Games and Economic Behavior* 54, no. 2: 293–315.

Fischbacher, U. z-T. 2007. "Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics* 10, no. 2: 171–178.

Gangadharan, L., T. Jain, P. Maitra, and J. Vecci. 2016. "Social Identity and Governance: The Behavioral Response to Female Leaders." *European Economic Review* 90: 302–325.

Gangadharan, L., T. Jain, P. Maitra, and J. Vecci. 2019. "Female Leaders and Their Response to the Social Environment." *Journal of Economic Behavior and Organization* 164: 256–272.

Gauriot, R., and L. Page. 2019. "Fooled by Performance Randomness: Overrewarding Luck." *Review of Economics and Statistics* 101, no. 4: 658–666.

Giglio, S., M. Maggiori, J. Stroebel, and S. Utkus. 2021. "Five Facts About Beliefs and Portfolios." *American Economic Review* 111, no. 5: 1481–1522.

Goldin, C., and C. Rouse. 2000. "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians." *American Economic Review* 90, no. 4: 715–741.

Greiner, B. 2015. "Subject Pool Recruitment Procedures: Organizing Experiments With ORSEE." *Journal of the Economic Science Association* 1, no. 1: 114–125.

Grether, D. M. 1980. "Bayes Rule as a Descriptive Model: The Representativeness Heuristic." *Quarterly Journal of Economics* 95, no. 3: 537–557.

Grossman, P. J., C. Eckel, M. Komai, and W. Zhan. 2019. "It Pays to be a Man: Rewards for Leaders in a Coordination Game." *Journal of Economic Behavior and Organization* 161: 197–215.

Gurdal, M. Y., J. B. Miller, and A. Rustichini. 2013. "Why Blame?" *Journal of Political Economy* 121, no. 6: 1205–1247.

Haaland, I., C. Roth, and J. Wohlfart. 2023. "Designing Information Provision Experiments." *Journal of Economic Literature* 61, no. 1: 3–40.

Heilman, M. E., and J. J. Chen. 2005. "Same Behavior, Different Consequences: Reactions to Men's and Women's Altruistic Citizenship Behavior." *Journal of Applied Psychology* 90, no. 3: 431–441.

Hernandez-Arenaz, I., and N. Iriberri. 2019. "A Review of Gender Differences in Negotiation." In *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press.

Hossain, T., and R. Okui. 2013. "The Binarized Scoring Rule." *Review of Economic Studies* 80, no. 3: 984–1001.

Jensen, K., B. Kovacs, and O. Sorenson. 2018. "Gender Differences in Obtaining and Maintaining Patent Rights." *Nature Biotechnology* 36, no. 4: 307–309.

Karpowitz, C. F., S. D. O'Connell, J. Preece, and O. Stoddard. 2024. "Strength in Numbers? Gender Composition, Leadership, and Women's Influence in Teams." *Journal of Political Economy* 132, no. 9: 3077–3114.

Lederman, D., N. V. Loayza, and R. R. Soares. 2005. "Accountability and Corruption: Political Institutions Matter." *Economics and Politics* 17, no. 1: 1–35.

MacNell, L., A. Driscoll, and A. N. Hunt. 2015. "What's in a Name: Exposing Gender Bias in Student Ratings of Teaching." *Innovative Higher Education* 40, no. 4: 291–303.

Mengel, F., J. Sauermann, and U. Zölitz. 2019. "Gender Bias in Teaching Evaluations." *Journal of the European Economic Association* 17, no. 2: 535–566.

Milgrom, P., and J. Roberts. 1988. "An Economic Approach to Influence Activities in Organizations." *American Journal of Sociology* 94: S154–S179.

Möbius, M. M., M. Niederle, P. Niehaus, and T. S. Rosenblat. 2022. "Managing Self-Confidence." *Management Science* 68, no. 11: 7793–7817.

Niederle, M. G. 2016. "Gender." In *The Handbook of Experimental Economics*, edited by J. H. Kagel and A. E. Roth, Vol. 2, 481–562. Princeton University Press.

Persson, T., G. Roland, and G. Tabellini. 1997. "Separation of Powers and Political Accountability." *Quarterly Journal of Economics* 112, no. 4: 1163–1202.

Player, A., G. Randsley de Moura, A. C. Leite, D. Abrams, and F. Tresh. 2019. "Overlooked Leadership Potential: The Preference for Leadership Potential in Job Candidates Who Are Men vs. Women." *Frontiers in Psychology* 10: 755.

Prendergast, C. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature* 37, no. 1: 7–63.

Recalde, M. P., and L. Vesterlund. 2023. "Gender Differences in Negotiation: Can Interventions Reduce the Gap?" *Annual Review of Economics* 15, no. 1: 633–657.

Régner, I., C. Thinus-Blanc, A. Netter, T. Schmader, and P. Huguet. 2019. "Committees With Implicit Biases Promote Fewer Women When They Do Not Believe Gender Bias Exists." *Nature Human Behaviour* 3, no. 11: 1171–1179.

Reuben, E., and K. Timko. 2018. "On the Effectiveness of Elected Male and Female Leaders and Team Coordination." *Journal of the Economic Science Association* 4, no. 2: 123–135.

Roy, M., and D. Houser. 2024. "Identity, Leadership, and Cooperation: An Experimental Analysis." *European Economic Review* 165: 104741.

Sarsons, H. 2022. "Interpreting Signals in the Labor Market: Evidence From Medical Referrals." Working Paper.

Sarsons, H., K. Gërxhani, E. Reuben, and A. Schram. 2021. "Gender Differences in Recognition for Group Work." *Journal of Political Economy* 129, no. 1: 101–147.

Solnick, S. J. 2001. "Gender Differences in the Ultimatum Game." *Economic Inquiry* 39, no. 2: 189–200.

Wolfers, J. 2007. "Are Voters Rational? Evidence From Gubernatorial Elections." Working Paper.

Yang, J. 2025. "On the Decision-Relevance of Subjective Beliefs." Working Paper.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Supplementary Appendix A: Experimental Instructions. **Supplementary Appendix B**: Conceptual Framework. **Supplementary Appendix C**: Additional Tables and Figures. **Supplementary Appendix D**: Investigation of leaders' investment decisions.