Are women less effective leaders than men? Evidence from experiments using coordination games

Lea Heursen[a], Eva Ranehill[b] and Roberto A. Weber[c]

[a] Department of Economics, Ludwig-Maximilians-Universität Munich
[b] Department of Economics University of Lund and University of Gothenburg
[c] Department of Economics, University of Zurich

November 21, 2022[*]

Abstract

We study whether one reason behind female underrepresentation in leadership is that female leaders are less effective at coordinating followers' actions. Two experiments using coordination games investigate whether female leaders are less successful than males in persuading followers to coordinate on efficient equilibria. In these settings, successful coordination hinges on higher-order beliefs about the leader's capacity to convince followers to pursue desired actions, making beliefs that women are less effective leaders potentially self-confirming. We find no evidence that such bias impacts actual leadership performance, precisely estimating the absence of a gender leadership gap. We further show that this result is surprising given experts' priors.

Keywords: gender; coordination games; leadership; experiment
JEL codes: D23, C72, C92, J1

[*] We are greatly thankful to Amanda Chuan, Marina Gertsberg, Pavitra Govindan, Christina Rott and Silvia Saccardo who gave helpful comments on earlier drafts, as well as to participants at several conferences and seminars for helpful comments and suggestions. We thank the Swiss National Science Foundation (100010_149451), Jan Wallander and Tom Hedelius Foundation (Handelsbankens Forskningsstiftelser Grant P2010-0133:1) and the Deutsche Forschungsgemeinschaft through CRC TRR 190 (project number 280092119) for generous financial support. We are also grateful to Philipp Grossman and Ernesto Reuben for sharing their data with us. Simon Grässli provided excellent research assistance. The research reported in this paper was approved by the Research Ethics Board of the Department of Economics at the University of Zurich.

1. Introduction

Economic research has devoted considerable attention to the persisting underrepresentation of women in leadership positions, studying the explanatory power of factors such as gender gaps in human capital acquisition, parental investment, economic preferences and discrimination (Bertrand and Hallock 2001; Gobillon et al. 2015; Blau and Kahn 2017; Goldin and Rouse 2000; Kleven et al. 2019; Buser et al. 2014; Preece and Stoddard 2015). In this study, we propose and test another explanatory factor—that women may be placed in leadership roles at lower rates than men simply because they are less effective leaders.

We focus on gender differences in one essential function of leadership—the ability to persuade groups of followers to pursue a common vision and coordinate on a course of action promoted by the leader (Kotter 2001; Dionne et al. 2004; Cooper et al., 2020). To address this question, we implement two laboratory experiments in which the leader's task is to persuade a group of followers to coordinate in implementing the leader's direction.[1] In both experiments, the leader sends free-form statements to all followers in a group, before the followers individually and simultaneously select actions. In the coordination games we employ, following the leader's direction is beneficial to followers only to the extent that other followers do so as well. Hence, both followers' first-order beliefs about the value of the leader's directive *and* their higher-order beliefs over whether others will be persuaded can matter for the leader's capacity to successfully motivate followers' actions.

Our studies are designed to test two different mechanisms that may result in a gender gap in leader persuasiveness. First, male and female leaders' directives may differ objectively in terms of, for example, quality or style. In our experiments, that would mean that, *ceteris paribus,* male and female leaders make different types of statements, and that these differences influence the degree to which their directives enhance coordination.

Second, even if male and female leaders are equally capable in issuing directives, gendered perceptions may impact how persuasive directives from male versus female leaders are perceived to be. The perception that female leaders are less effective than male leaders is widespread. For example, according to the last wave of the World Values Survey (2010-2014), 48.5% of respondents worldwide agree with the statement that men make

[1] The experimental setting allows us to randomly assign individuals to groups, and to the role of leader, and thus causally estimate the effectiveness of different types of leaders. Moreover, we can exogenously vary whether followers know a leader's gender, holding all other aspects of the decision environment constant, to causally explore the impact of gender stereotypes about effective leadership.

better political leaders than women.[2] Previous research in psychology and sociology indicates that attitudes towards female leaders are often more negative than those towards male leaders (e.g., Rudman and Kilianski 2000; Rudman and Glick 2001; Eagly and Karau 2002) and that women may be seen as less legitimate leaders and encounter more resistance to their authority (Ridgeway 2001). Characteristics associated with femininity may differ from those typically associated with effective leadership (see, e.g., Koenig et al. 2011) and gender stereotypes about higher male competence and agency may lead to men having greater influence than women (Carli 2017). In economic research, a handful of studies (e.g., Macchiavello et al. 2020; Ayalew, Manian and Sheth 2021; Born, Ranehill, and Sandberg 2022) document less positive perceptions of female leaders; we review this literature in more detail in the next section.

The null hypothesis for our studies (H0) is that men and women are equally persuasive as leaders—i.e., that they achieve, on average, the same group outcomes in our two experiments. Against this null hypothesis, we test two alternative hypotheses, based on the above two mechanisms underlying a potential gender difference. First, if female leaders' directives are worse at generating support from followers, we should observe female leaders to be less persuasive leaders independently of whether followers observe their gender (H1a). Second, if actual leadership quality does not differ, but performance differences arise only due to self-confirming gendered perceptions of leadership quality, then we expect to see a difference only when the leader's gender is known to followers (H1b).

Our first experiment uses a variant of the "turnaround game," first introduced by Brandts and Cooper (2006) and widely employed in recent research on leader effectiveness. The game involves a group of followers who play a repeated weak-link coordination game where each follower independently selects a level of investment for a group project. Investments exhibit complementarities—the return from the group project is based on the *minimum* amount invested by any follower and investments higher than the minimum yield no benefit. This feature leads most groups to coordinate on the least efficient equilibrium after a few rounds. At this stage, we introduce a leader who is incentivized to move the group to more efficient equilibria; that is, to use their directives to elicit investment from *all* followers. Previous literature indicates that statements made by leaders, as well as leaders' characteristics, can matter in this environment and influence the extent to which groups

[2] The corresponding share of respondents indicating a preference for men as business executives is 43%. However, respondents could only agree or disagree with the statement and the question does not allow identifying the share of respondents who would agree with the reversed statement.

coordinate on efficient equilibria (Brandts and Cooper 2007; Brandts et al. 2015; Bhalotra et al. 2021).

In these settings, we find that male and female leaders are equally effective at issuing directives that coordinate followers and increase their investments in the group project. This holds both when the gender of the leader is visible to followers and when it is not, even though followers, on average, hold stereotypical views on gender and leadership, as measured both using an implicit association test and more explicit attitudinal measures.

The absence of a gender gap in our first study is consistent with two other concurrent studies, also employing weak-link coordination games to evaluate the effectiveness of male and female leadership (Reuben and Timko 2018; Grossman et al. 2019). However, as we demonstrate in Section 3.3, neither these studies nor our first experiment, which employs a considerably larger sample size, can rule out sizable gender effects. Further, while weak-link coordination games may be useful for studying leadership for many reasons (Foss 2001; Weber et al. 2001; Brandts and Cooper 2007; Brandts et al. 2015), they may not be ideal for identifying differences across leader categories. In such games, followers always gain by adhering to a leaders' advice if others do so. The structure of the game means that followers should pursue a leader's recommended action whenever they believe that the probability others will do so is above a threshold determined by the game's payoffs. Thus, the weak-link game will only detect a gap in effectiveness between two leaders if the expected follower response probabilities lie on opposite sides of this threshold.[3] For many game parameters and leader characteristics, different groups of leaders may be on the same side of the threshold and thus equally persuasive, making weak-link games ineffective for detecting some differences in leader persuasiveness. We formally illustrate this point in Appendix B1.

Based on these considerations, we design a second experiment—which is the primary focus of this paper—providing a more precise test of the relative persuasiveness of male and female leaders. In this experiment, two followers play a pure-matching coordination game in which they simultaneously decide which of two competing leaders to follow. Leaders are incentivized to persuade as many pairs of followers as possible to coordinate on their preferred action through a written statement shown to the followers. As in Experiment 1, we explore the impact of varying leader gender and its visibility. Hence, both experiments

[3] For example, in a lab-in-the-field study in India, Bhalotra et al. (2021) show that a leader's religious identity mediates their effectiveness in a weak-link game, indicating that such identity may produce beliefs about leaders that lie on opposite sides of the threshold necessary to induce follower coordination.

are based on coordination games and leverage second-order beliefs to explore the gender gap in leader persuasiveness. However, since in Experiment 2 followers only care about picking the same leader, even very small differences in leaders' actual or perceived persuasiveness are likely to have big effects on which leader is followed.

The data from Experiment 2 show that the leaders' statements are attended to by followers and positively impact coordination rates. That is, leaders are persuasive and their persuasiveness is heterogeneous. However, despite such heterogeneity, men and women are followed at nearly identical rates, independently of whether their gender is visible. Thus, the more sensitive test of our hypotheses in Experiment 2 again yields no indication of a gender gap in leader effectiveness and instead a precisely estimated null effect.

We further document that this null effect is surprising to researchers in the field of gender economics. Before revealing the results of Experiment 2, we elicited predictions regarding differences in leader persuasiveness by gender, for both the gender blind and the gender visible conditions, from participants at a workshop. On average, these researchers expected a large difference when leader gender was visible, in contrast with our results. These researchers also expected no gender differences when followers did not observe leader gender, which is consistent with our results.

We began our research expecting to document that biased beliefs about the persuasiveness of male and female leaders can generate real differences in their respective success in coordinating follower teams. However—surprisingly to us and, apparently, to other researchers in economics—we find little evidence of such leadership gender gaps. Our results thus serve as a counterpoint to arguments that women should not lead in environments where buy-in from followers is critical merely because the perception that they are less effective can be self-reinforcing.

We make several contributions to the existing literature. To our knowledge, only one previous study—a thesis chapter by Timko (2017)—investigates gender gaps in leader effectiveness in coordination games, using a design that separates the effects of leader quality from follower responses.[4] Further, our second experiment introduces a novel game that is designed to detect even the slightest systematic difference in the persuasiveness of two groups of leaders. We show theoretically and empirically that this experiment is a useful tool for systematically studying what makes some leaders more persuasive in settings

[4] While similar to our paper in its approach, Timko (2017) is also relatively underpowered, comprising a total of 15 groups across 4 conditions (leader gender X leader gender visibility).

requiring coordinated responses. Moreover, as illustrated in Section 3.3, our estimates of no gender gap in leadership effectiveness are considerably more precise than previous ones. Finally, by documenting the priors of a relevant research community and establishing results that are substantially different from these priors, our main finding provides valuable information allowing us to correct possibly mistaken beliefs about a gender gap in leader effectiveness and its source.

The rest of this paper is organized as follows. The next section provides a review of related literature other than the closely related work by Reuben and Timko (2018) and Grossman et al (2019) that we discuss above. Section 3 presents a brief overview of our first experiment, while Section 4 provides a more in-depth account of our second, and more novel, experiment. Section 5 concludes and discusses important open questions.

2. Related Literature

Several academic fields have studied the effectiveness of male and female leadership. Meta-analyses of research in organizational behavior, based largely on subjective evaluations of leaders' abilities and effectiveness, find that the gender stereotype of the work context mediates the extent to which men and women are perceived as equally effective leaders (Eagly et al. 1995; Eagly and Karau 2002; Paustian-Underdahl et al. 2014). Generally, men are perceived as relatively more effective the more male dominated the organization and the more male stereotyped the role.

Gipson et al. (2017) present an overview of the management research covering studies that investigate differences in organizational outcomes under male and female leadership. This mainly correlational line of work finds no consistent impact of the gender composition of high-level management positions, such as the CEO, on, for example, financial outcomes or investments in CSR activities.

A growing body of studies in economics—often using natural experiments—investigates whether male and female leaders generate different collective outcomes, finding mixed results. For example, Chattopadhyay and Duflo (2004) make use of the random political reservations for women at the village level in India and find that female village leaders are more reactive to the priorities of female constituents. However, exploring the impact of gender quotas in candidate lists in local Spanish elections, Campa and Bagues (2017) find no impact of an increased share of women, neither with respect to the size nor the composition of public spending. Another example from the corporate sector is presented by Matsa and Miller (2013), who find companies that increased the share of female board

members in response to the Norwegian gender quota to have higher labor costs and lower operating profits. However, these results are contradicted in Eckbo et al. (2016) who find no differences when extending the sample period.

More closely related to the question explored here—whether male and female leaders are more effective at influencing the behavior of followers—is a recent strand of research using experimental methods to evaluate reactions to male and female leadership and advice. One strand of this research explores the impact of male and female leaders primarily in developing countries, finding, e.g., that female trainee managers in the Bangladesh garment industry are initially seen as less effective (Macchiavello et al. 2020), that female agricultural trainers in Malawi are perceived as less knowledgeable (BenYishay et al. 2020) and that female leaders elicit less support among administrative employees at an Ethiopian university (Ayalew, Manian, and Sheth 2021). Further, in field and lab experiments using student samples, Gloor et al. (2020), De Paola et al. (2018) and Chakraborty and Serra (2019) find that female leaders are evaluated more negatively and experience more backlash than male leaders, and that this effect is most pronounced in male dominated groups. Born et al. (2022) document, in a laboratory experiment, that groups selecting leaders discriminate against female candidates.[5] Other studies report mixed results on the effectiveness of male and female leaders in experimental public good games. In a field experiment in India, Gangadharan et al. (2016) find that male villagers contribute less when their group has a female leader, whereas Grossman et al. (2015) observe no difference in contributions under male and female leadership in the laboratory. Finally, Brandts and Rott (2021) implement a laboratory experiment and find no impact of advisor gender on behavior of advisees in a setting where advisors provide guidance on whether to enter a tournament.

3. Experiment 1

To focus on our more novel second study, we provide only minimal details on Experiment 1 here, deferring additional details to Section 1.1 in the online appendix.

[5] A growing body of research in economics finds women to be less recognized for their expertise than men, which provides an additional indication that leading roles may be more challenging for women than men (Grunspan et al. 2016; Boring 2017; Mengel et al. 2019; Sarsons et al.2021; Shurchkov and Geen 2019). This effect seems to be particularly strong in stereotypically male fields of knowledge (Bohren et al. 2019; Bordalo et al. 2019).

3.1 Experiment Design

The turnaround game employed in our first experiment is based on a weak-link coordination game (Van Huyck et al. 1990; Weber et al. 2001) in which a leader must induce a group with a history of coordination failure to coordinate on a more efficient equilibrium. Participants were randomly assigned to the role of a "CEO" or an "employee" and played 16 periods of the weak-link game in stable "firms" of 5 employees and a CEO.[6] In each period, employees independently decided how to allocate a total of 40 hours of work time between a "safe" and a "risky" project. The payoff of the risky project was potentially higher than that of the safe project but depended on the minimum number of hours allocated to that project by any employee in the firm and a randomly chosen rate of return (for payoff tables see Appendix Table A1).

During periods 1-6 the CEO merely observed the group outcomes. From period 7 onwards, we introduced two changes. First, CEOs were given the opportunity to direct their groups to a more efficient equilibrium through a free-form message sent to group members at the onset of each period. Second—following previous studies—we increased the randomly chosen rate of return.[7] The CEO payment during Part 1 consisted of the average group member payoff. In Part 2, the CEO payoff increased in the minimum investment in the risky project, incentivizing leaders to move the group to a more efficient equilibrium.

The experimental conditions in Part 2 varied the gender of the CEO (Female vs. Male) and whether the picture of the CEO was shown (Visible vs. Blind) in a 2x2 design.[8] Sessions were gender balanced and comprised 4 groups, each randomized to one of the 4 conditions.

Finally, before receiving their payment, all participants completed the Gender and Authority Measure (GAM) (Rudman and Kilianski 2000) to measure explicit preferences for male versus female authorities and the Implicit Association Test (IAT) (Greenwald et al. 1998; Greenwald et al. 2003), to measure the strength of participants' implicit association

[6] We used these labels, as well as the introduction of a risky rate of return, to strengthen associations with male-dominated contexts and thereby strengthen potential bias toward male leaders.

[7] The rate of return of either $r = 5$ or $r = 6$ in Part 1 was chosen to make efficient coordination in Part 1 challenging. The increase in the rate of return to $r = 8$ or $r = 10$ in the second part follows previous literature (Brandts and Cooper 2007; Brandts, Cooper and Weber 2015) and gives leaders a better chance of overcoming the coordination problem. Apart from the change in the rate of return employee payoffs were calculated similarly throughout the experiment.

[8] Photographs of each participant were taken at the beginning of a session (as described in the invitation email). Participants were informed that the CEO's photo would be visible either to all employees in a firm, or for none of them, and that the CEO would not know whether the photograph was shown (to prevent this knowledge from systematically affecting leaders' communication). There is debate on the best way to communicate gender in the laboratory. While photos also convey other characteristics than gender, we chose to disclose gender by photos as an innocuous way to do so. For example, it does not require participants to choose gendered names or avatars. Photos, contrary to avatars, are also a natural part of professional settings.

of leadership with male or female gender (see Section 1.2 of the online appendix for more information on these two measures).

We conducted the experiment in English at the Laboratory for Experimental and Behavioral Economics at the University of Zurich, with a total of 600 participants. Table 1 lists the number of participants in the different conditions. We oversampled groups in which the gender of the leader was visible to followers.

Table 1. Session Overview and Number of Observations (Experiment 1)

Treatment	Groups	Leaders	Followers (men, women)
Female Leader Gender Visible	30	30	150 (79,71)
Male Leader Gender Visible	30	30	150 (75,75)
Female Leader Gender Blind	20	20	100 (56, 44)
Male Leader Gender Blind	20	20	100 (44, 56)
Total	100	100	500

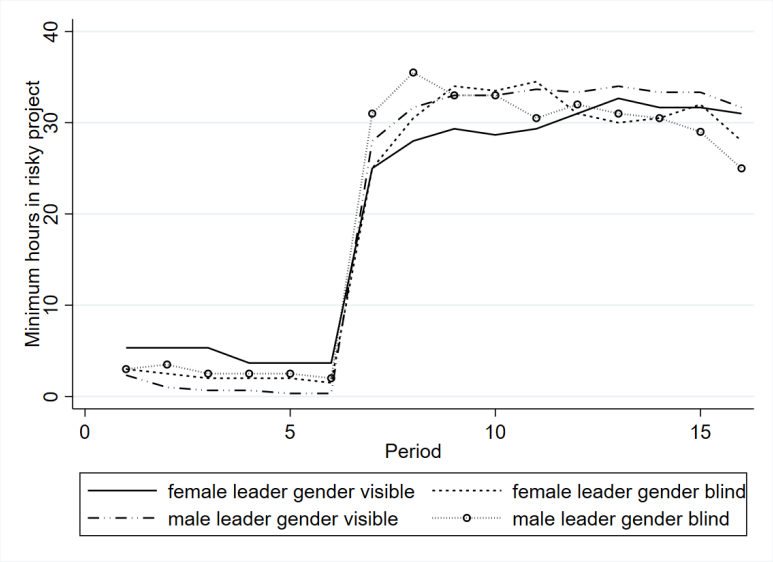
3.2 Results

Among participants in the role of employee, we find a moderate but significant tendency to associate leadership with maleness and to prefer male authorities ($p < 0.001$ of two-sided t-tests against H0 of no bias for both the IAT and the GAM; see Appendix Figure A1 for more detail). However, despite such bias in favor of male leaders, we find only small and statistically insignificant differences in the outcomes of male- and female-led groups.

Figure 1 presents our main measure of leadership effectiveness—the average minimum number of hours invested in in the risky project in a group—across periods.[9] In period 7, the first period with active leaders, the small gender gaps in leadership effectiveness are not statistically significant (two-sided t-tests of equality of means; gender visible: $p=0.407$, gender invisible: $p=0.136$). Across periods 7-16, the average effect of having a female rather than male leader is 0.15 fewer hours invested in the risky project in the gender-blind condition ($p=0.967$) and 2.7 in the gender-visible condition (p-value=0.419). Regression analysis employing various specifications and controls confirm these results (see Tables A2 and A3 in the appendix).

[9] We find no significant differences in Part 1, before leaders became active. A Kruskal-Wallis test fails to reject the null hypothesis that the distributions of group outcomes are similar across conditions both over all of Part 1 ($p=0.177$) and in the final period of Part 1 ($p=0.083$).

Figure 1. Average Minimum Group Investment by Condition (Experiment 1)



3.3 Discussion of Experiment 1

Our results suggest that, contrary to our motivating hypothesis, female leaders are no less effective at inducing change to an efficient, but risky, equilibrium in a weak-link coordination game. We were surprised by these results, though as we note earlier, they concord with findings from similar concurrent studies (Reuben and Timko, 2018; Grossman et al, 2019). However, while our study uses a considerably larger sample size than these studies, the estimates from all three studies nevertheless have limited precision and cannot rule out sizable gender gaps.[10]

Further, following the observation of a null result we realized that weak-link and turnaround games, while widely used to study leadership, may not be ideal for detecting discrimination favoring one kind of leader over another. In these games, followers' always benefit if they coordinate on following a leader, meaning that they have an economic interest not to discriminate against *any* leader. Moreover, given the payoff structure of the game, a follower is incentivized to follow the leader's recommendation whenever she thinks the likelihood that all others will do so is above a specific threshold. Thus, differences in the effectiveness of leaders, even if substantive, are likely only detected if these likelihoods lie on opposite sides of this threshold. Given the relatively high rates of leader effectiveness

[10] We demonstrate this point in Figure A2, which shows the estimated gender gaps in effectiveness in the gender visible condition of our Experiment 1 and those in the studies by Reuben and Timko (2018) and Grossman et al. (2019), which all employ weak-link coordination games. In all three cases, men are more effective at obtaining efficient coordination, but the effects are not statistically different from zero.

observed, the above concerns provide a potential interpretation of our and others' observation of no gender gap in leader effectiveness.

To make the above argument more concrete, Appendix B1 presents a simple model of behavior in coordination games when followers are encouraged to select an action by a leader with a given level of persuasiveness. This analysis illustrates the above limitations to using weak-link coordination games to study discrimination against different types of leaders. It also suggests an alternative design as a more precise test of the relative persuasiveness of male and female leaders. Instead of having a group of followers decide whether to follow a single leader, each group must coordinate on following *one of two* leaders. In this case, the followers' expected payoffs are higher by following any leader perceived as even slightly more persuasive. Thus, even a very slight tendency to believe that male leaders will receive more support than female leaders from other followers should result in a clear preference for male leaders.

4. Experiment 2

Motivated by the above considerations, we designed a novel study involving a pure-matching coordination game in which followers must coordinate on following the recommendations from one of two leaders. We focus on cases in which leaders differ with respect to their gender.

4.1 Design of Experiment 2

Participants were assigned the role of either a CEO of a small start-up or an Investor. CEOs were incentivized to persuade followers to invest in their firm. To obtain CEOs' directives, initial sessions comprised only CEOs, who used these sessions to craft a message to Investors to attract their investment. Subsequent sessions consisted only of Investors.

This second experiment retained features of the first experiment, such as an underlying coordination game and the importance of second order beliefs about leader effectiveness. It also retained the professional context and riskiness of payoffs, intended to, if anything, strengthen associations with male stereotypes. Finally, as in Experiment 1, half of all Investor pairs also saw the portraits of the two CEOs above their messages.

The Leader Competition Game

Across several rounds, pairs of Investors, anonymous to each other, played the coordination game described below, framed as an investment task. They saw the messages of two competing CEOs displayed in random order on their screens and had to choose,

privately and simultaneously, in which of the two firms to invest. Investors played a pure-matching coordination game in which they could earn money only if they both chose to invest in the same firm. A CEO earned money in a round only if their message persuaded both Investors to invest in his or her firm.

Below we describe the two kinds of sessions in detail.

The CEO Sessions

When arriving at a session, participants in the role of CEO received full instructions for the game. This included a description of their role as CEO and the role that would subsequently be performed by Investors. CEOs were informed that they could think of their role as being the CEO of a small start-up firm. The first and main task for each CEO was to draft a message to the Investors to convince them to invest in that CEO's firm. In the instructions, we aimed to further strengthen associations with male stereotyped contexts by stressing the competitive aspect of the game—e.g., by referring to the other CEO with whom they would be paired as the CEO of a "competing firm." The CEOs were further told that the Investors would see their message and that of the CEO of the competing firm when making their decisions in each round. In addition, the computer would randomly determine whether to show the CEOs' photos along with their messages, and if photos were shown, the photos of both CEOs would be shown. This way, all CEOs crafted their message under the assumption that at least sometimes, their photo would be shown with it. Because of this, CEOs' beliefs about the effect that their gender has on their ability to persuade Investors are constant across our visibility treatments.

Each CEO had 40 minutes to write a message and was restricted to use between 60 and 700 characters (about 10-100 words). We asked all CEOs to refrain from using personal identifying information or offensive language. Two of these messages are reproduced in Figure 3. Overall, participants drafted their message with great care.[11]

While drafting their messages, CEOs were called, one at a time, to have a portrait photograph taken. The photos were taken by an experienced photographer in a different room with professional photo equipment. Participants were instructed to pose like they would in a photo for their *curriculum vitae*. The photographer ensured that the portraits were very similar in terms of composition and facial expression.

[11] The average length of a message is 451 characters (about 64 words) and the overall rate of misspelled words in all messages is very low, at 0.014, in an environment without an automatic spell checker. This likely reflects the participants' efforts at carefully crafting their messages.

We chose to reveal leader gender through portrait photographs because we think portraits are more likely to trigger implicit biases than, for example, names or avatars, while at the same time being less conducive to revealing our gender research focus. Moreover, portraits are a natural part of professional settings. Portraits could pose a challenge to our identification strategy if portrait characteristics other than gender impacted leader persuasiveness and these characteristics were correlated with gender in our sample. To mitigate this concern, we hired independent coders who rated participant portraits on many characteristics that we subsequently analyze.

At the end of the session, after writing their messages, the CEOs were asked to choose 3 out of 6 possible "products" that their firm would develop and sell. The products were represented by 6 letters shown on the screen (e.g., "Product X"), and CEOs simply clicked on three products to make their choice. Before the participation of the Investors, we randomly drew 1 of the 6 products to be the "successful" product for this study. This choice was relevant for the payoff of Investors who earned extra money from coordinating their investments in firms where the CEO had also selected a successful product. While this part of the study is simply a random lottery, we implemented it to add an element of risk. A risky environment could, if anything, strengthen the association of leadership with maleness.

Participants in the role of CEO received a show-up fee of CHF 15. In addition to the show-up fee, CEOs earned CHF 2 for games in which both Investors invested in the CEO's firm over that of a competing CEO. CEOs knew that their message would be used in several games and that they would be paid according to the behavior of Investors in a randomly selected subset of 24 of these games. This performance-based part of their payment was mailed to the CEOs after the completion of the Investor sessions. The final payoff of a CEO was thus the show-up fee plus the sum of the payments across 24 games in which the CEO's message was used, potentially ranging from CHF 15 to 63.

The Investor Sessions

After all CEOs participated in the study, we invited a new set of participants to take part as Investors. As with the CEOs, subjects in the role of Investor first went through the full instructions for the study, which explained the role of the CEOs who had participated earlier, as well as the role of the Investors.

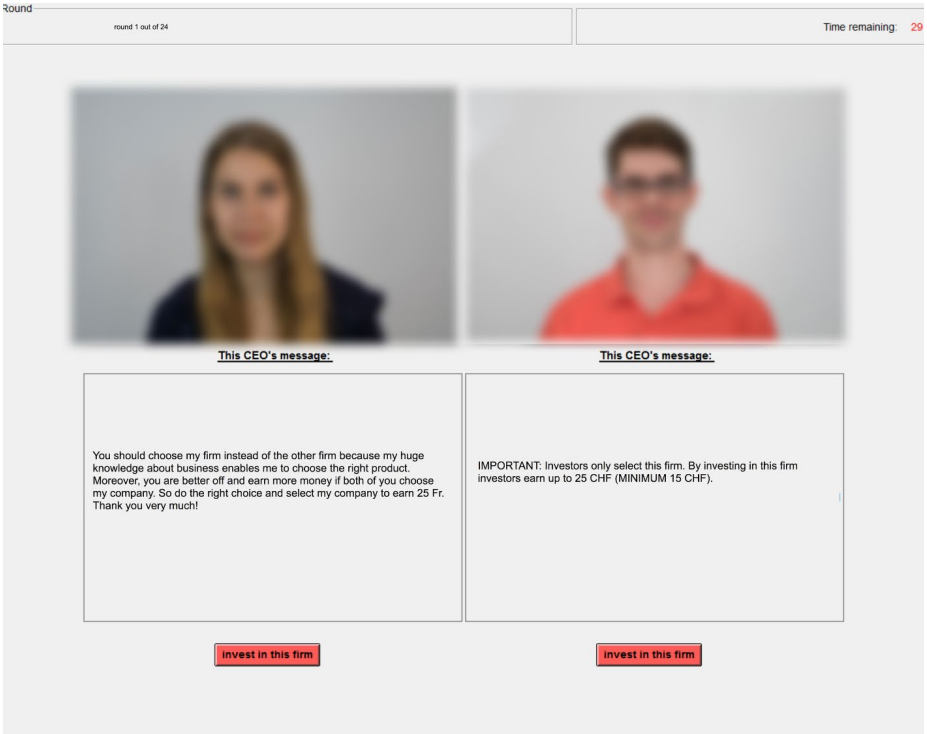
Investors then played 24 rounds of the coordination game described above. In each round, Investors were randomly assigned a new pair of CEOs. An Investor never

encountered the same CEO more than once. Investors were randomly re-matched in each round and could be matched with the same Investor more than once. However, Investors never knew the identity of their matched counterpart, and did not know whether they had met a particular Investor before. Investors received no feedback on any choices that were made until the end of the session and, therefore, did not find out whether coordination in any specific period was successful until then.

For each Investor, one of the 24 games was randomly selected for payment. If the Investors did not coordinate in this game both Investors earned 0. If they coordinated, but the CEO had chosen an unsuccessful product, they earned 15 CHF. Finally, if they coordinated, and the CEO had chosen a successful product, they earned 25 CHF. These incentives were paid in addition to a 15 CHF show-up fee. Investors saw an overview of their earnings at the end of the session, including which of the 24 rounds was selected for payment and whether coordination was successful in this round.

Investors made their investment decisions by clicking on one of the two competing firms presented on their screen. Figure 3 provides an example of the decision interface for the case in which the leaders' portraits are visible. Investors were told that the order in which CEOs were displayed on their individual screen was random and that they could therefore not use screen position to coordinate.

Figure 3. Decision Screen for Investors in Leader Competition Game



Notes: Faces of participants are blurred here to preserve their anonymity. Whenever pictures were displayed to participants in the experiment, they were not blurred.

In a final questionnaire, we collected the same individual measures regarding explicit and implicit gender stereotypes on leadership as in Experiment 1, the Implicit Association Test and the Gender and Authority Measure (see Section 1.2 of the online appendix for more details). The questionnaire also contained basic demographic questions, e.g., about age and gender.

Experimental Conditions

We varied whether the CEOs matched with Investors in a game were male or female and whether leader gender was visible to Investors. We did not implement a condition without leadership, that is, with pictures of CEOs only without their statements, since our research question is about the persuasiveness of leaders' directives and not about characteristics of portraits that facilitate coordination. The experimental conditions varied within-subject for Investors, with different treatment configurations across games. We stratified randomization of the CEO pairs such that, in expectation, 2/3 of matches would involve one male and one female CEO. In the experiment, 61.2% (N=940) of all investment games involved a male and a female CEO, 19.7% (N=302) involved two female CEOs, and 19.1% (N=294) involved two male CEOs. We included same-gender pairs to produce a natural mix of leader pairs and decrease the likelihood that participants guessed our research question.

Table 2. Participants and Leader Competition Games per Experimental Condition

Number of Participants			
		CEO	*Investor*
Gender	female	48	61
	male	48	67
Total		**96**	**128**
Leader Competition Games per Experimental Condition			
		Gender visible	*Gender invisible*
Gender of Leaders	mixed-gender	470	470
	2 females	151	151
	2 males	147	147
Total		**768**	**768**

Table 2 presents an overview of the experimental conditions implemented. In exactly half of all investment games (768 of 1536), the portraits of the two CEOs were displayed

along with their messages in each period. This balance was achieved by letting two sets of Investor pairs play the investment game with each randomly formed CEO pair. One of these Investor pairs had access to the CEOs' portraits and messages, while the other pair only saw messages. Thus, for each CEO pair, we have one game with photographs and one without.

Experiment Procedures

Experiment 2 was conducted in English at the Laboratory for Experimental and Behavioral Economics at the University of Zurich. In total, 224 participants from the University of Zurich and the Swiss Federal Institute of Technology in Zurich took part.

The CEO sessions were programmed in Qualtrics, and we collected the data for 96 participants in the role of CEO, half of them women, on two days. To facilitate taking the portrait photographs, sets of four participants were recruited to arrive in 15-minute intervals. Upon arrival, participants read the instructions available as a printout on their work terminal. Participants could only start drafting the message to Investors once they had correctly answered several comprehension questions.

About two weeks after the CEO sessions, we conducted a total of four Investor sessions with 32 participants per session. All 128 participants listened to a recording of the instructions before the experiment started, and the experiment started only once all participants had correctly answered a set of comprehension questions. The Investor sessions were programmed in Ztree (Fischbacher 2007).

The instructions for CEOs and Investors are provided in the online appendix (Sections 2.2-2.3) The sessions for the CEOs and the Investors lasted, respectively, 1 and 1.25 hours. CEOs earned an average of 44.8 CHF and Investors earned 30.8 CHF.

Coding of CEO Messages and Portrait Characteristics

After collecting the data, we hired 10 coders from the same subject pool as the participants to code the messages sent by the CEOs and rate characteristics of their pictures. Half of the coders, all native English speakers, categorized the content of the messages according to 15 pre-determined categories and provided their best guess about the CEO's gender. The categories asked about the message content (such as whether the message mentioned risk or the competing company) as well as tone (see Table A6 in the appendix for a list of the categories and summary statistics). To these variables we add the length of the CEO message. The remaining 5 coders assessed the portraits on 14 pre-determined characteristics such as whether the person in the picture looks competent, likeable, or

trustworthy. In addition, we asked the coders whether they would choose the person as a leader, and whether they knew the person (see appendix Table A6 for an overview of the categories and summary statistics). We collected these data to explore whether there are attributes of the messages or the portraits other than gender that differ between the women and men in our sample of leaders.

4.2 Results

Establishing Priors: Expert Prediction Task

Before presenting results, we provide estimates of the prior beliefs held by a set of experts regarding the outcomes in Experiment 2. DellaVigna et al. (2019) argue that collecting predictions before communicating experimental results may mitigate hindsight and publication bias and help assess the novelty and information value of empirical results.

In October 2018, before we shared the results from this study, we asked participants at a workshop on gender and economics to predict the results. We distributed a one-page handout summarizing the experiment (shown in the online appendix Section 2.4) to the workshop participants. They had about 30 minutes to make a prediction regarding what percent of the time followers who faced two competing leaders of opposite sex would invest in the male leader's firm when gender was (i) not known and (ii) known. Participation was voluntary and incentivized. For each prediction, we selected the person whose guess was closest in absolute value to the actual percentage; that person received 50 Euros. Almost every workshop participant present took part, yielding 54 answers.

The average prediction for the gender-blind condition is that male CEOs are followed to the same extent as females, in 49 percent of the cases. However, for the gender visible condition, the average prediction is that 63 percent of followers will follow the male CEO. Thus, researchers expected male and female CEOs to write messages that were equally effective but expected followers to rely on gender stereotypes about leadership to coordinate their investments. Put differently, this sample of researchers expects our experiment to reject the null hypothesis in favor of H1b, that performance differences of female and male leaders arise due to self-confirming perceptions of leadership quality.

Do Messages and Photographs Improve Coordination Rates?

Table 3 presents the empirical coordination rates in our experiment, by treatment. If followers were to simply randomize over the two investment options, the expected coordination rate would be 0.5. The average coordination rate is 0.592 in the gender-blind

condition and 0.642 when gender is visible. Both outcomes are significantly more frequent than the 0.5 benchmark ($p<0.001$ for both conditions, test of proportions, n=768 per condition). These results indicate that leaders are instrumental for successful coordination and that Investors pay attention to the leaders when making investment decisions. Moreover, the coordination rate in the gender-visible condition is significantly higher than when gender is blind ($p=0.005$, test of proportions, n=1556),[12] providing evidence that followers rely on characteristics in the pictures to coordinate, over and above the message content.

We simulated the rate of successful coordination in our experiment under the assumption that followers play a mixed strategy that puts 50% probability on each investment option. We simulated 100,000 such experiments, each with 768 one-shot coordination games. Not a single simulated experiment had an average coordination rate that was equally far from the 0.5 benchmark as the effect sizes we observe in our data, i.e. 0.59 when gender is blind and 0.64 when gender is visible. These simulations corroborate that the leader competition game measures some characteristics of leaders and their messages that influence their persuasiveness in coordinating followers.

Table 3. Rate of Successful Coordination by Treatment

Experimental condition		Average rate of coordination	p-value (difference from 0.5)	Obs.
Gender blind	mixed-gender leaders	0.602	0.0000	470
	2 female leaders	0.610	0.0072	151
	2 male leaders	0.544	0.2836	147
	Total	0.592	0.0000	768
Gender visible (messages + pictures)	mixed-gender leaders	0.626	0.0000	470
	2 female leaders	0.722	0.0000	151
	2 male leaders	0.612	0.0065	147
	total	0.642	0.0000	768

Are Followers Biased in Favor of Male Leadership?

Before we explore whether leaders' gender impacts their ability to obtain follower support, we study followers' implicit and explicit biases regarding gender and leadership. Similarly to Experiment 1, the average score obtained in the IAT and the GAM indicate a slight bias in favor of male leaders ($p <0.007$ of two-sided t-tests against H0 of no bias; see Appendix Figure A1 for more detail), with 40% of the followers revealing a moderate to

[12] Note that the underlying random leader pairings are the same in these two samples.

strong bias in implicitly associating leadership more with men than with women (see Figure A1 in the appendix).

Are there Gender Differences in Leader Persuasiveness?

To test our main hypotheses, we only consider data from mixed-gender pairings, pitting male and female leaders against one another. In the gender-blind condition, followers chose to invest in the project promoted by the male CEO 50.0% of the time. The corresponding number for the gender visible condition is 50.4%. This supports our null hypothesis of no leader gender gap in persuasiveness against either alternative hypothesis of greater persuasiveness by male leaders. A test of proportions, with high levels of statistical power to reject the 0.5 benchmark,[13] confirms that neither difference is statistically significant (p=0.948 (gender blind), p=0.794 (gender visible), n=940 in both conditions). Hence, Experiment 2 reveals a precise estimate of no substantial gender gap in leadership persuasiveness—with or without gender being visible.

As an alternative test of H1b, we exploit the fact that for each pairing of leaders we observe how two pairs of followers—one pair in the gender blind and one in the gender visible condition—respond to those leaders' directives (see Table A4 in the Appendix). In 30.4% of such cases, a female leader was more persuasive in the gender visible condition compared to the gender-blind condition, in 30.7% of the cases the male leader was more persuasive with gender visible, and in 38.9% of the cases there was no change.

Complementary regression analysis predicting the probability that a leader is followed is presented in Table 4. Column 1 replicates the null result from our non-parametric analysis, regressing the likelihood of being followed on an indicator for female leader, gender visibility and their interaction. Noteworthy is the very low R-squared, indicating that none of the variance in followers' choices is explained by this set of control variables. Estimates presented in column 2 show no indication that female or male followers respond differently to women as leaders, regardless of whether the leader's gender is visible.

[13] We calculated the power of a two-sided one-sample test of proportions to reject the null hypothesis that women are followed at the same rate as men, at the 0.05 level of statistical significance or higher, given 940 observed decisions (see Figure A3 in the appendix). As an example, for the average gap in following rates predicted by research economists (37 percent), this test has a power very close to 1.

Table 4. Predicting Leader Persuasion

	OLS Regressions Predicting Likelihood of Being Followed		
	(1)	(2)	(3)
Female leader	0.00213	-0.0223	-0.0668
	(0.0358)	(0.0432)	(0.0377)
Picture	0.00532	-0.00916	-0.0117
	(0.0201)	(0.0306)	(0.0295)
Female leader X Picture	-0.0106	0.0183	0.0233
	(0.0286)	(0.0442)	(0.0429)
Female follower		-0.0257	-0.0253
		(0.0326)	(0.0313)
Female leader X Female follower		0.0514	0.0556
		(0.0472)	(0.0459)
Female follower X Picture		0.0306	0.0359
		(0.0438)	(0.0401)
Female leader X Picture X Female follower		-0.0611	-0.0717
		(0.0670)	(0.0626)
Message length			0.311***
			(0.0597)
Mentions investor			0.0197
			(0.0290)
Mentions product			-0.0384
			(0.0286)
Sustainability or Social Responsibility			0.117***
			(0.0305)
Assertive			0.0256
			(0.0180)
Constant	0.499***	0.511***	0.307***
	(0.0282)	(0.0313)	(0.0448)
R^2	0.000	0.000	0.045
N	3760	3760	3760

Notes: OLS regression with the dependent variable equaling 1 if the leader's recommendation is followed and 0 otherwise. *Female leader* indicates that a leader is a woman, *picture* indicates that the leader's picture was visible to followers. *Female follower* indicates that the follower is a woman. Among the elicited message characteristics, we included the variables *Message length*, *Mentions investor*, *Mentions product*, *Sustainability or social responsibility* and *Assertive* in Specification 3. These message characteristics were included in the analysis since they exhibit at least a fair amount of intercoder agreement and (marginally) significantly correlate with leader persuasiveness (see Figure 6 for these results and the notes to Figure 5 for a description of these categories). Standard errors in parentheses (clustered at the leader level). *Significant at the 5% level,** at the 1% level, *** at the 0.1% level.

We can also further investigate our research question by studying the heterogeneous effectiveness of individual male and female leaders. For each of the 96 leaders in our sample, we estimate that leader's persuasiveness using the empirical rate at which they are followed when they compete against a leader of the opposite sex. Figure 4 displays the

empirical cumulative frequencies obtained separately for male and female leaders in the gender blind (panel a) and gender visible (panel b) conditions. For both genders, we observe substantial heterogeneity in leader persuasiveness. However, in both cases the distributions of male and female leaders look similar, and we find no evidence that they are drawn from different populations (Wilcoxon-Mann-Whitney tests p>0.84).

Figure 4. Rate at Which Leaders in Mixed-Gender Pairs are Followed by Treatment: Cumulative Frequency

a. Gender Blind *b. Gender Visible*

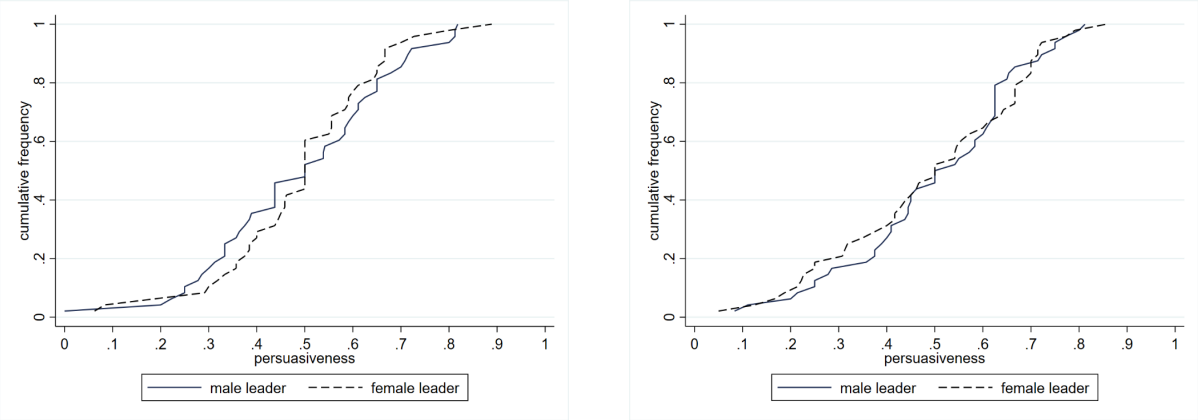
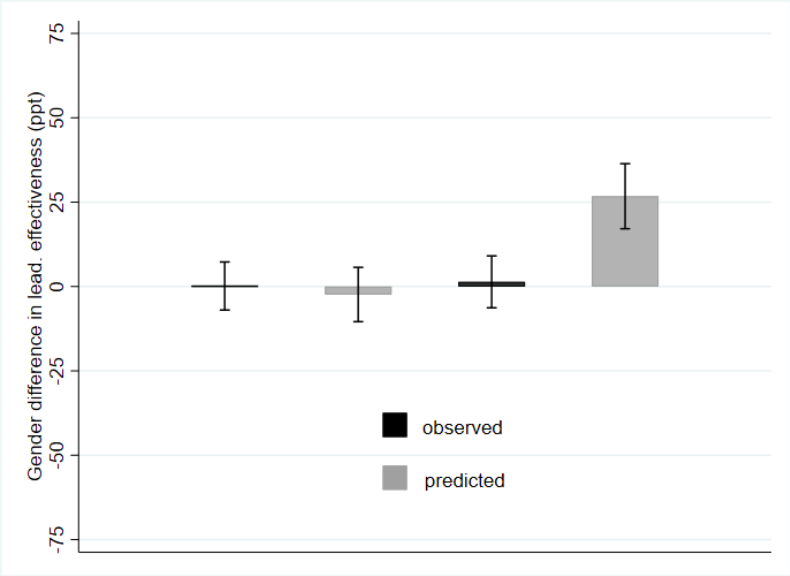


Figure 4 illustrates the precision with which we estimate no gender gap in leader effectiveness. The figure shows the mean difference in the persuasiveness of individual male and female leaders. The vertical axis ranges from -100 percentage points (corresponding to female leaders always being chosen over male leaders) to 100 percentage points (corresponding to male leaders always being chosen over female leaders). In the gender-visible condition, the average gender gap is 1.38 percentage points with a 95-% confidence interval of [-6.31, 9.07]; in the gender-blind condition it is 0.14 points with the corresponding confidence interval of [-6.98, 7.27].[14]

We contrast this with the predicted gender gap in persuasiveness based on our sample of researchers. The expert predictions expect the null hypothesis to hold in the gender-blind comparison, with an average predicted persuasiveness gap of -2.38, but expect a gap of 26.8 when gender is visible (i.e., 63.4 – 36.6). Clearly, our data reject this predicted effect size for the gender visible condition.

[14] These confidence intervals are based on the t-distribution, treating each of the 96 leaders as an independent observation.

Figure 4. Observed and Predicted Gender Gaps in Leader Effectiveness

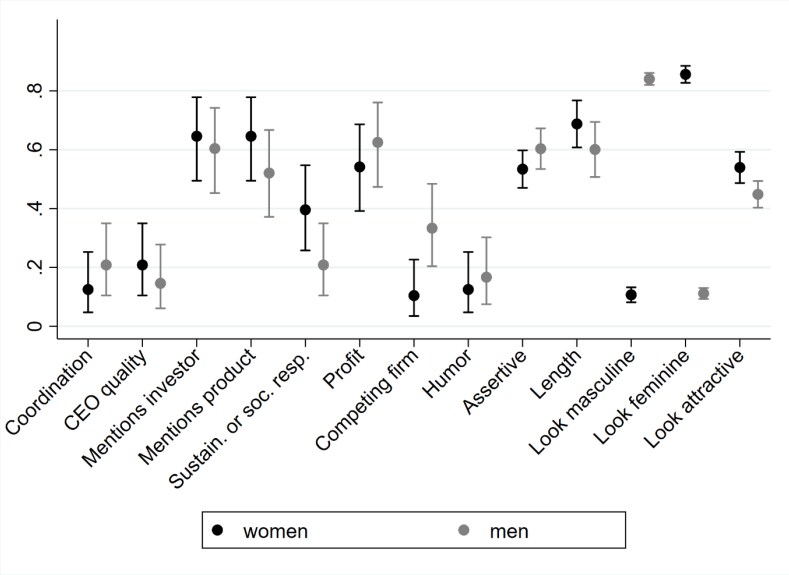


Notes: The black bars depict the average observed and the grey bars the average predicted gender gap in leadership effectiveness. Gender difference in leader effectiveness is measured as the percentage point difference in the average success of male and female leaders, with spikes representing 95-% confidence intervals. Success ranges from 0% to 100% and corresponds to the rate at which a leader is followed when matched with a competing leader of the opposite sex.

Gender Differences in CEOs' Messages and Portraits

Although we find no evidence of gender differences in the effectiveness of leaders, we can use our data in an exploratory manner to study whether male and female leaders send different kinds of messages or whether their portraits convey different characteristics. For this purpose, 5 independent raters, all native English speakers from the same student pool as our study participants, classified the content of the messages into 15 categories; 5 different students were further hired to rate the portraits of our leaders along 15 dimensions. Coders' agreement is measured through Kippendorff's alpha, which is suitable for binary and ordinal data such as the Likert scale ratings we used. Tables D2 and D3 of the Online Appendix provide an overview of all message and portrait categories, together with summary statistics for the male and female leaders, separately, and the alpha statistic for intercoder agreement. For the subsequent exploratory analysis, we retain those characteristics for which we observe at least a fair amount of intercoder agreement (an alpha statistic of approximately 0.4 or higher), which holds for 9 messages and 3 picture characteristics. We also include a measure of message length, coded as the number of characters in a leader's message as a fraction of the longest message

Figure 5. Message and Portrait Characteristics by CEO Gender



Notes: Dots display means or proportions and spikes the associated 95-% confidence intervals. Variables Coordination-Humor: binary variables, variables Assertive-Look attractive ordinal variables scaled to 0/1 where higher numbers above the neutral outcome 0.5 mean more agreement. All characteristics were identified by 5 independent coders and aggregated at the leader-level by taking the mean of all ratings (for Likert-scale ratings) or the median of all classifications (for binary classifications). *Coordination* indicates whether a message mentions coordination. *CEO quality* indicates whether a message mentions the CEO or qualities of the CEO. *Mentions investor* indicates whether a message mentions the investors. *Mentions product* indicates whether a message mentions a product or provides details on the products offered by the firm. *Sustainability or social responsibility* indicates whether a message mentions sustainability or social responsibility. *Profit* indicates whether a message mentions profit. *Competing firm* indicates whether a message mentions the competing firm. *Humor* indicates whether a message uses humor. *Assertive* is the extent to which a message is written in a forceful tone. *Length* refers to the number of characters in the message, measured as a fraction of the longest message. *Look masculine*, *Look feminine* and *Look attractive* are aggregated from ratings of the extent to which the person in the picture looks masculine, feminine or attractive.

Figure 5 presents the frequency with which the coders categorized the messages of male and female CEOs according to the pre-specified dimensions. The main observation is that messages written by male and female CEOs are generally perceived (by coders blind to the leader's gender) as similar. Messages of female CEOs mention sustainability and corporate social responsibility more often, and the competing firm less often, than male CEOs. Further, messages by female CEOs are rated to be written in a somewhat less assertive tone, on average, than messages by male CEOs. However, these differences are not substantial when considering the exploratory nature of this analysis.

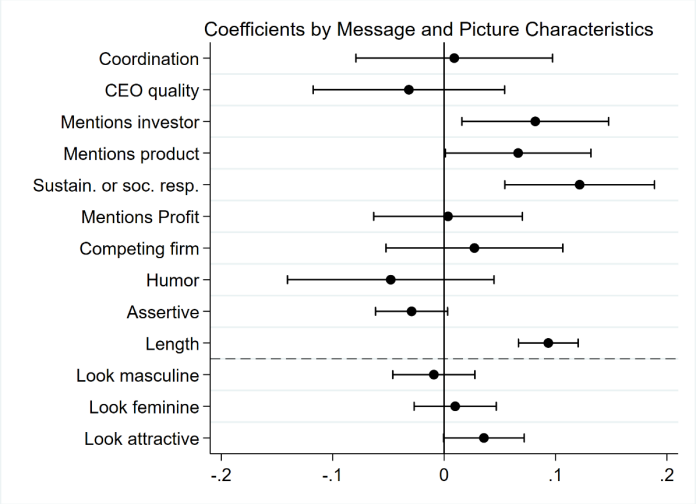
Turning to the classification of leader portraits, we generally find intercoder agreement to be very low. As expected, pictures of male (female) CEOs are perceived as much more masculine (feminine). Pictures of female CEOs are also rated as somewhat more attractive than pictures of male CEOs, although our measure of intercoder agreement at just below 0.4 indicates a high degree of subjectivity.

These data also allow for an exploratory analysis of the relationship between message and portrait characteristics and leader persuasiveness. Specifically, we investigate how such characteristics influence the empirical rate at which a leader is followed whenever their message is used or their portrait is shown in our study, not only in mixed gender pairings. Figure 6 presents the coefficients from separate OLS regressions using a leader's overall persuasiveness as the dependent variable, using each different message and portrait characteristic as explanatory variables. Messages that mention sustainability or social responsibility are more effective, as are messages that mention the investor or the product that is offered. Messages that are judged to be more assertive in tone appear to be somewhat less effective. This analysis provides suggestive evidence that certain message properties make them more effective, though, as we observed earlier, there are few differences in the extent to which male and female leaders send different types of messages. We find little indication that the physical characteristics of the CEOs—as evaluated by the coders—have strong relationships with the leaders' effectiveness. Particularly noteworthy is the observation that being perceived as masculine or feminine has little impact on persuasiveness.

To investigate whether leader or message characteristics influence gender gaps in leader persuasiveness, we add those message characteristics that are (marginally) significantly correlated with leader persuasiveness as control variables to the models in Table 4, in which we predict the likelihood of being followed as a function of leader gender and leader gender visibility (see column 3 of Table 4). This analysis uses only observations in which leaders compete against someone of the other sex.. We find no evidence of a gender gap in leaders' ability to persuade holding constant message characteristics that matter individually, since men and women write similar messages, on average.[15]

[15] The picture characteristic *look attractive* also marginally significantly correlates with leader persuasiveness. When we additionally control for this characteristic in a model that predicts the likelihood of being followed as a function of leader gender, leader follower and the relevant message characteristics but only for the sample of follower that saw pictures, the inference on the indicator "female leader" does not change.

Figure 6: Message and Picture Characteristics Predicting Persuasiveness



Note: Coefficients and 95-% confidence intervals for message and picture characteristics from OLS regressions of leader persuasiveness (rate at which a leader is followed across all leader competition games) on each characteristic separately. For picture characteristics, the sample is restricted to leader competition games in which leader gender was visible to followers (estimates below the dotted reference line). *Length of message* is the fraction of the longest message and transformed to a z-score. All other individual message and picture characteristics were identified by 5 independent coders and aggregated at the leader-level by taking the mean of all ratings (for Likert-scale ratings) or the median of all classifications (for binary classifications). The aggregated variables that are continuous (those constructed from Likert-scale ratings) are transformed to z-scores. See notes to Figure 5 for a description of these categories.

4.3 Discussion of Experiment 2

We designed Experiment 2 to detect even slight differences in the persuasiveness of male and female leaders. Compared to Experiment 1, we also removed any incentives for leaders to adapt their messages based on their beliefs about followers' response. However, despite this stronger test, Experiment 2 replicates the null results obtained in Experiment 1 and in earlier research, although with considerably greater precision. The results from the gender-blind condition indicate that male and female leaders communicate equally persuasively, and we find no indication that stereotypes impact followers' coordination decisions in the gender visible condition. Our survey of experts' expectations establishes that this null result is surprising, relative to the priors of researchers with expertise in gender economics. This lack of a gender effect also occurs despite clear evidence of heterogeneous leader ability influencing coordination outcomes.

5. Conclusion

We present results from two experimental studies designed to test the hypothesis that female leaders are less persuasive than men, and hence less effective at motivating coordinated responses by followers. In both experiments, leaders send messages that direct

followers toward specific actions in simultaneous-move coordination games. We add contextual features to the games to enhance the perception of a stereotypically male domain. In order to test two channels through which differential leader persuasiveness may operate, we experimentally vary the gender of the leader and whether the gender is observed by followers.

In our first experiment, followers play a weak-link coordination game in which a safe investment option guarantees a fixed amount of money, independent of the actions of other followers. Leaders recommend to followers an alternative action that entails greater risk but also a potentially high return. Thus, leader effectiveness involves obtaining coordinated responses by followers on a risky action. In this context, we observe that women are equally effective as men at convincing followers to take more risk, regardless of whether followers know the leader's gender. However, we also show that this paradigm, also employed by other researchers to investigate similar questions, may be a suboptimal instrument for detecting differences in persuasiveness between different groups of leaders.

With our Experiment 2, we introduce a paradigm for uncovering characteristics of leaders and their directives that make some leaders more persuasive than others. We use this paradigm to test whether the *gender* of the person issuing a directive impacts its credibility, but many other applications are possible, for example, varying age or ethnicity of those who attempt to persuade others. In Experiment 2, two leaders give competing directives to followers. Followers simply want to coordinate on the same action in a simultaneous-move pure matching coordination game but are, otherwise, indifferent as to whose advice to follow. In this decision environment, even slight differences in the persuasiveness of female and male leaders should cause followers' behavior in the experiment to reveal a gender gap in leadership effectiveness. However, we again find no evidence in support of the hypothesis of differential effectiveness of male and female leaders. On the contrary, we estimate that female leaders who are in direct competition with male leaders are followed at an empirical rate very close to 50%, regardless of whether their gender is known to followers. This is remarkable because the overall rate at which followers achieve coordination in the leader competition game is 59% when followers see messages and 64% when followers see messages and pictures, indicating that leader characteristics matter substantially. In other words, while we find that the leader competition game is a useful instrument to measure leadership effectiveness, we find no evidence whatsoever that gender mediates differences in effectiveness.

We also document that this precisely estimated zero result is surprising to a set of research economists with expertise in the field of gender economics. DellaVigna and Pope (2018) find that the average forecast of expert researchers typically gets the level of an effect as well as the direction right. In contrast, we document that expert researchers anticipate a substantial gender gap in leader persuasiveness in our condition of Experiment 2 with visible leader photos, but that the actual data reveal no effect. In a recent paper on non-significant results in empirical economics, Abadie (2020) makes a compelling argument that failures to reject a point null hypothesis should be given a much higher *scientific* significance by the intellectual community, particularly when these rejections are surprising. Given the expert predictions, our results are clearly surprising and change the prior beliefs held by the research community, including our own. A similar argument is put forward by Bertrand (2020) who writes that the finding of no gender difference should be considered as valuable as a finding of a gender difference, since these results can help to correct gender stereotypical thinking.

We hope that our results will contribute to updating widely held stereotypes on gender and leadership effectiveness. This is important, as mistaken beliefs may contribute to the low rate at which women are promoted into positions of leadership for which the ability to motivate and direct followers is key to good performance. For example, if the panel of experts we surveyed regarding Experiment 2 were consulting a board attempting to select a leader to represent their "firm" in the condition of that experiment in which leader gender is visible, they might presumably select male leaders who they expect to perform better. In this case, it could take a long time for observational data to generate disconfirming evidence to correct such stereotypes.[16]

Our results provide an indication that the source of the gender gap in top leadership positions cannot be justified by a general tendency for women to be inherently less persuasive at directing coordinated action by followers. In our studies, there is nothing that women do that makes them less effective, and we find no evidence that a perception that they might be less effective yields self-confirming outcomes in which female leaders are less able to influence follower behavior. Of course, this does not mean that such equilibrium discrimination cannot happen outside the laboratory. Indeed, it might be more likely to occur in contexts where perceptions of competence and ability play a larger role than in our studies

[16] Relatedly, recent work by Bursztyn et al. (2020) finds direct evidence that mistaken beliefs—in their case about social norms regarding female labor supply—contribute to sustaining a gender gap in labor market outcomes.

(see, e.g., Bohren et al. 2019), and where existing biases that women are less capable or knowledgeable might lead directives from female leaders to be less effective. In this regard, our study can be viewed as a useful starting point, demonstrating that a general perception of lower female leader efficacy alone cannot produce such effects.

References

Abadie, Alberto. 2020. "Statistical Nonsignificance in Empirical Economics." *American Economic Review: Insights* 2 (2): 193–208. https://doi.org/10.1257/aeri.20190252.

Ayalew, Shibiru, Shanthi Manian, and Ketki Sheth. 2021. "Discrimination from below: Experimental Evidence from Ethiopia." *Journal of Development Economics* 151 (June): 102653. https://doi.org/10.1016/j.jdeveco.2021.102653.

BenYishay, Ariel, Maria Jones, Florence Kondylis, and Ahmed Mushfiq Mobarak. 2020. "Gender Gaps in Technology Diffusion." *Journal of Development Economics* 143 (March): 102380. https://doi.org/10.1016/j.jdeveco.2019.102380.

Bertrand, Marianne. 2020. "Gender in the Twenty-First Century." *AEA Papers and Proceedings 110* (May): 1–24. https://doi.org/10.1257/pandp.20201126.

Bertrand, Marianne, and Kevin F Hallock. 2001. "The Gender Gap in Top Corporate Jobs." *Industrial and Labor Relations Review* 55 (1): 3–21. https://doi.org/10.2307/2696183.

Bhalotra Sonia, Irma Clots-Figueras, Lakshmi Iyer, Joseph Vecci.2021."Leader Identity and Coordination." *The Review of Economics and Statistics;* doi: https://doi.org/10.1162/rest_a_01040..

Blau, Francine D., and Lawrence M. Kahn. 2017. "The Gender Wage Gap: Extent, Trends, and Explanations." *Journal of Economic Literature* 55 (3): 789–865. https://doi.org/10.1257/jel.20160995.

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg. 2019. "The Dynamics of Discrimination: Theory and Evidence." *American Economic Review* 109 (10): 3395–3436. https://doi.org/10.1257/aer.20171829.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2019. "Beliefs about Gender." *American Economic Review* 109 (3): 739–73. https://doi.org/10.1257/aer.20170007.

Boring, Anne. 2017. "Gender Biases in Student Evaluations of Teaching." *Journal of Public Economics* 145 (January): 27–41. https://doi.org/10.1016/j.jpubeco.2016.11.006.

Born, Andreas, Eva Ranehill, and Anna Sandberg. "Gender and Willingness to Lead: Does the Gender Composition of Teams Matter?" *The Review of Economics and Statistics* 2022; 104 (2): 259–275.

Brandts, Jordi, and David J. Cooper. 2006. "A Change Would Do You Good .... An Experimental Study on How to Overcome Coordination Failure in Organizations." *American Economic Review* 96 (3): 669–93. https://doi.org/10.1257/aer.96.3.669.

———. 2007. "It's What You Say, Not What You Pay: An Experimental Study of Manager-Employee Relationships in Overcoming Coordination Failure." *Journal of the European Economic Association* 5 (6): 1223–68. https://doi.org/10.1162/JEEA.2007.5.6.1223.

Brandts, Jordi, David J. Cooper, and Roberto A. Weber. 2015. "Legitimacy, Communication, and Leadership in the Turnaround Game." *Management Science* 61 (11): 2627–45. https://doi.org/10.1287/mnsc.2014.2021.

Brandts, Jordi, and Christina Rott. 2021. "Advice from Women and Men and Selection into Competition." *Journal of Economic Psychology* 82: 102333. https://doi.org/10.1016/j.joep.2020.102333.

Bursztyn, Leonardo, Alessandra L Gonzalez, and David Yanagizawa-Drott. in press. "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia." *American Economic Review*.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek. 2014. "Gender, Competitiveness, and Career Choices." *The Quarterly Journal of Economics* 129 (3): 1409–47. https://doi.org/10.1093/qje/qju009.

Campa, Pamela, and Manuel Bagues. 2017. "Can Gender Quotas in Candidate Lists Empower Women? Evidence from a Regression Discontinuity Design." 2017–06. Working Papers. Department of Economics, University of Calgary. https://ideas.repec.org/p/clg/wpaper/2017-06.html.

Carli, Linda L. 2017. "Social Influence and Gender." In *The Oxford Handbook of Social Influence*, edited by Stephen G. Harkins, Kipling D. Williams, and Jerry Burger. Oxford University Press.

Chakraborty, Priyanka, and Danila Serra. 2019. "Gender Diferences in Top Leadership Roles: Does Worker Backlash Matter?" http://faculty.smu.edu/dserra/ChakrabortySerraLeadershipFeb2019.pdf.

Chattopadhyay, Raghabendra, and Esther Duflo. 2004. "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India." *Econometrica* 72 (5): 1409–43. https://doi.org/10.1111/j.1468-0262.2004.00539.x.

Cooper, David J., John R. Hamman, and Roberto A. Weber. 2020. "Fool Me Once: An Experiment on Credibility and Leadership." *The Economic Journal*, 130(631): 2105-2133. https://doi.org/10.1093/ej/ueaa059.

De Paola, Maria, Francesca Gioia, and Vincenzo Scoppa. 2018. "Teamwork, Leadership and Gender." *IZA Discussion Paper Series*, no. 11861.

DellaVigna, Stefano, and Devin Pope. 2018. "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy* 126 (6): 2410–56. https://doi.org/10.1086/699976.

DellaVigna, Stefano, Devin Pope, and Eva Vivalt. 2019. "Predict Science to Improve Science." *Science* 366 (6464): 428–29. https://doi.org/10.1126/science.aaz1704.

Dionne, Shelley D, Francis J Yammarino, Leanne E Atwater, and William D Spangler. 2004. "Transformational Leadership and Team Performance." *Journal of Organizational Change Management* 17 (2): 177–93. https://doi.org/10.1108/09534810410530601.

Eagly, Alice H., and Steven J. Karau. 2002. "Role Congruity Theory of Prejudice toward Female Leaders." *Psychological Review* 109 (3): 573–98. https://doi.org/10.1037//0033-295X.109.3.573.

Eagly, Alice H., Steven J. Karau, and Mona G. Makhijani. 1995. "Gender and the Effectiveness of Leaders: A Meta-Analysis." *Psychological Bulletin* 117 (1): 125–45. https://doi.org/10.1037/0033-2909.117.1.125.

Eckbo, B. Espen, Knut Nygaard, and Karin S. Thorburn. 2016. "How Costly Is Forced Gender-Balancing of Corporate Boards?" SSRN Scholarly Paper ID 2746786. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2746786.

Fischbacher, Urs. 2007. "Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics* 10 (2): 171–78. https://doi.org/10.1007/s10683-006-9159-4.

Foss, Nicolai. 2001. "Leadership, Beliefs and Coordination: An Explorative Discussion." *Industrial and Corporate Change* 10 (2): 357–88.

Gangadharan, Lata, Tarun Jain, Pushkar Maitra, and Joseph Vecci. 2016. "Social Identity and Governance: The Behavioral Response to Female Leaders." *European Economic Review* 90 (November): 302–25. https://doi.org/10.1016/j.euroecorev.2016.01.003.

Gipson, Asha N., Danielle L. Pfaff, David B. Mendelsohn, Lauren T. Catenacci, and W. Warner Burke. 2017. "Women and Leadership: Selection, Development, Leadership Style, and Performance." *The Journal of Applied Behavioral Science* 53 (1): 32–65. https://doi.org/10.1177/0021886316687247.

Gloor, Jamie L., Manuela Morf, Samantha Paustian-Underdahl, and Uschi Backes-Gellner. 2020. "Fix the Game, Not the Dame: Restoring Equity in Leadership Evaluations." *Journal of Business Ethics* 161 (3): 497–511. https://doi.org/10.1007/s10551-018-3861-y.

Gobillon, Laurent, Dominique Meurs, and Sébastien Roux. 2015. "Estimating Gender Differences in Access to Jobs." *Journal of Labor Economics* 33 (2): 317–63. https://doi.org/10.1086/678495.

Goldin, Claudia, and Cecilia Rouse. 2000. "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians." *American Economic Review* 90 (4): 715–41. https://doi.org/10.1257/aer.90.4.715.

Greenwald, A. G., D. E. McGhee, and J. L. Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74 (6): 1464–80.

Greenwald, Anthony G., Brian A. Nosek, and Mahzarin R. Banaji. 2003. "Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm." *Journal of Personality and Social Psychology* 85 (2): 197–216. https://doi.org/10.1037/0022-3514.85.2.197.

Grossman, Philip J., Catherine Eckel, Mana Komai, and Wei Zhan. 2019. "It Pays to Be a Man: Rewards for Leaders in a Coordination Game." *Journal of Economic Behavior & Organization* 161 (May): 197–215. https://doi.org/10.1016/j.jebo.2019.04.002.

Grossman, Philip J., Mana Komai, and James E. Jensen. 2015. "Leadership and gender in groups: An experiment." *Canadian Journal of Economics/Revue canadienne d'économique* 48 (1): 368–88. https://doi.org/10.1111/caje.12123.

Grunspan, Daniel Z., Sarah L. Eddy, Sara E. Brownell, Benjamin L. Wiggins, Alison J. Crowe, and Steven M. Goodreau. 2016. "Males Under-Estimate Academic Performance of Their Female Peers in Undergraduate Biology Classrooms." *PLOS ONE* 11 (2): e0148405. https://doi.org/10.1371/journal.pone.0148405.

Kleven, Henrik, Camille Landais, and Jakob Egholt Søgaard. 2019. "Children and Gender Inequality: Evidence from Denmark." *American Economic Journal: Applied Economics* 11 (4): 181–209. https://doi.org/10.1257/app.20180010.

Koenig, Anne M., Alice H. Eagly, Abigail A. Mitchell, and Tiina Ristikari. 2011. "Are Leader Stereotypes Masculine? A Meta-Analysis of Three Research Paradigms." *Psychological Bulletin* 137 (4): 616–42. https://doi.org/10.1037/a0023557.

Kotter, John P. 2001. "What Leaders Really Do." *Harvard Business Review*, 2001. https://hbr.org/2001/12/what-leaders-really-do.

Macchiavello, Rocco, Andreas Menzel, Atonu Rabbani, and Christopher Woodruff. 2020. "Challenges of Change: An Experiment Promoting Women to Managerial Roles in the Bangladeshi Garment Sector." Working Paper 27606. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w27606.

Matsa, David A, and Amalia R Miller. 2013. "A Female Style in Corporate Leadership? Evidence from Quotas." *American Economic Journal: Applied Economics* 5 (3): 136–69. https://doi.org/10.1257/app.5.3.136.

Mengel, Friederike, Jan Sauermann, and Ulf Zölitz. 2019. "Gender Bias in Teaching Evaluations." *Journal of the European Economic Association* 17 (2): 535–66. https://doi.org/10.1093/jeea/jvx057.

Northouse, Peter G. 2012. *Leadership: Theory and Practice*. 6th ed. Los Angeles: Sage publications.

Paustian-Underdahl, Samantha C., Lisa Slattery Walker, and David J. Woehr. 2014. "Gender and Perceptions of Leadership Effectiveness: A Meta-Analysis of Contextual Moderators." *Journal of Applied Psychology* 99 (6): 1129–45. https://doi.org/10.1037/a0036751.

Preece, Jessica, and Olga Stoddard. 2015. "Why Women Don't Run: Experimental Evidence on Gender Differences in Political Competition Aversion." *Journal of Economic Behavior & Organization* 117 (C): 296–308.

Reuben, Ernesto, and Krisztina Timko. 2018. "On the Effectiveness of Elected Male and Female Leaders and Team Coordination." *Journal of the Economic Science Association* 4 (2): 123–35. https://doi.org/10.1007/s40881-018-0056-3.

Ridgeway, Cecilia L. 2001. "Gender, Status, and Leadership." *Journal of Social Issues* 57 (4): 637–55. https://doi.org/10.1111/0022-4537.00233.

Rudman, Laurie A., and Peter Glick. 2001. "Prescriptive Gender Stereotypes and Backlash Toward Agentic Women." *Journal of Social Issues* 57 (4): 743–62. https://doi.org/10.1111/0022-4537.00239.

Rudman, Laurie A., and Stephen E. Kilianski. 2000. "Implicit and Explicit Attitudes Toward Female Authority." *Personality and Social Psychology Bulletin* 26 (11): 1315–28. https://doi.org/10.1177/0146167200263001.

Sarsons, Heather, Klarita Gërxhani, Ernesto Reuben, and Arthur Schram. 2021. "Gender Differences in Recognition for Group Work." *Journal of Political Economy* 129 (1): 101–47. https://doi.org/10.1086/711401.

Schelling, Thomas C. 1969. "Models of Segregation." *The American Economic Review* 59 (2): 488–93.

Shurchkov, Olga, and Alexandra V. M. van Geen. 2019. "Why Female Decision-Makers Shy Away from Promoting Competition." *Kyklos* 72 (2): 297–331. https://doi.org/10.1111/kykl.12202.

Timko, Krisztina. 2017. "The Selection Process and Not Gender Matters for Effective Leadership." *PhD Thesis*.

Van Huyck, John B., Raymond C. Battalio, and Richard O. Beil. 1990. "Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure." *The American Economic Review* 80 (1): 234–48.

Weber, Roberto, Colin Camerer, Yuval Rottenstreich, and Marc Knez. 2001. "The Illusion of Leadership: Misattribution of Cause in Coordination Games." *Organization Science* 12 (5): 582–98. https://doi.org/10.1287/orsc.12.5.582.10090.

Appendix A – Additional analysis

The Appendix contains the following additional figures:

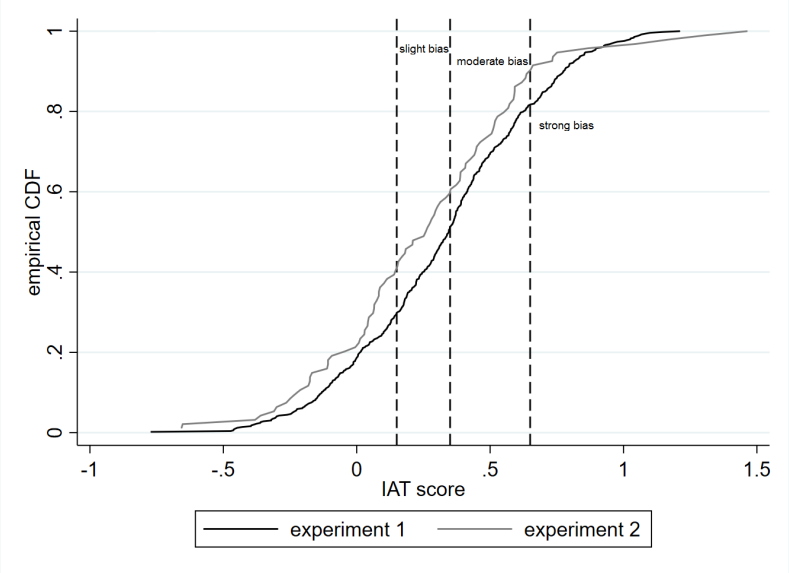
1. Experiments 1 and 2: Empirical Distribution of IAT and GAM Scores in Sample of Followers
2. Observed Gender Leader Effectiveness Gaps in Experiment 1 and 2 Related Studies
3. Experiment 2: Power Curve for Experiment 2

The Appendix contains the following additional tables

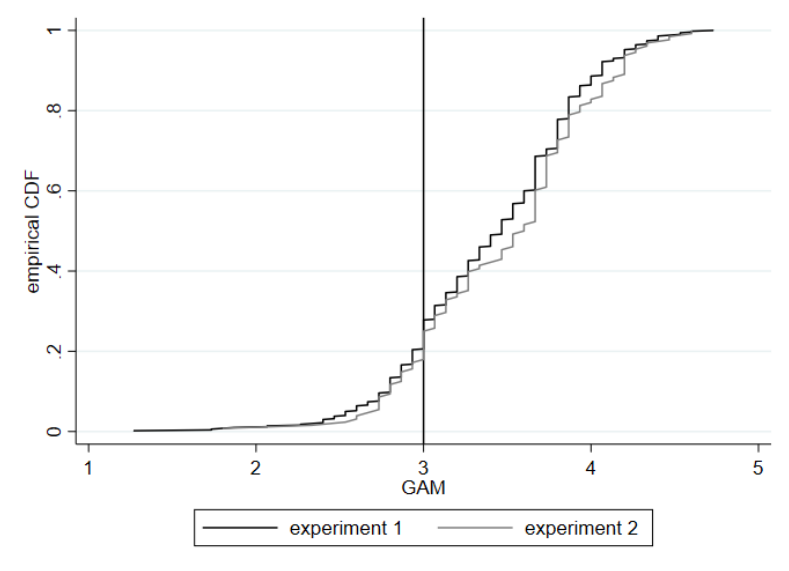
1. Experiment 1: Employee and CEO Payoffs in Parts 1 and 2
2. Experiment 1: Results of OLS regressions with group-level outcomes
3. Experiment 1: Results of OLS regressions with individual follower behavior as the outcome variable
4. Experiment 2: Follower behavior in response to the same leader pair in a round by gender visibility

Figure A1. Gender Stereotypes and Attitudes toward Female Leadership in Follower Samples

Panel a) IAT score

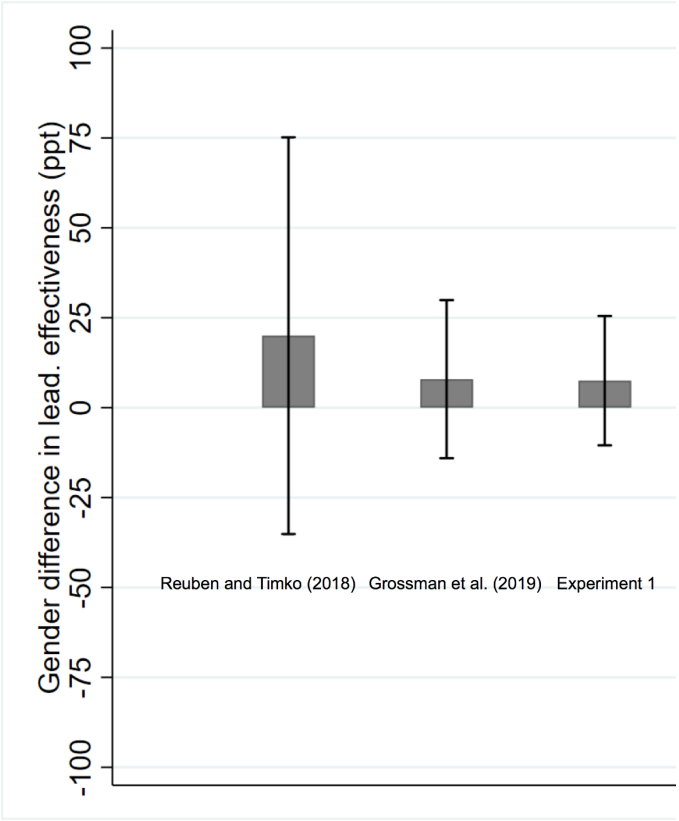


Panel b) GAM score



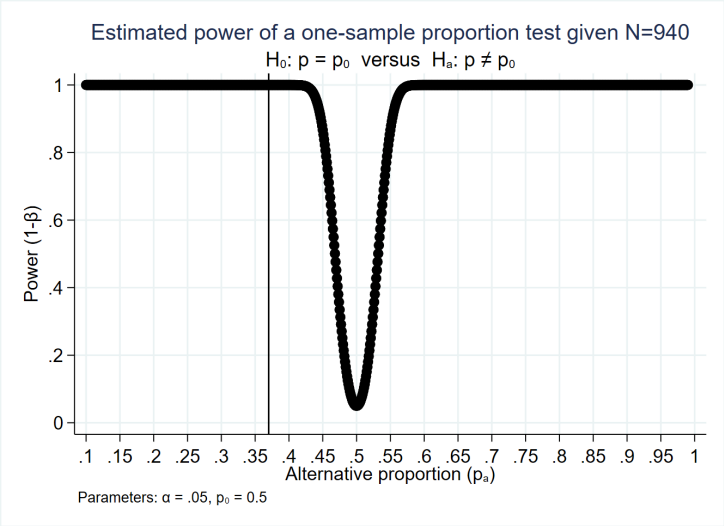
Notes: Panel (a): The IAT measures the implicit tendency to associate leadership more strongly with men than with women. It ranges from -2 to 2 and positive numbers above 0.15 are generally considered as indicating a bias in associating leadership more strongly with men than with women. The higher the number, the stronger the bias, with values above 0.35 indicating a modest bias and numbers above 0.65 indicating a strong bias in favor of men. Panel (b): The GAM ranges from 1-5 and measures an explicit preference for male versus female authorities. Scores above 3 indicate a preference in this direction, with higher scores indicating a stronger preference. The mean IAT score of participants who played the role of followers in Experiment 1 is 0.33 and 0.26 in Experiment 2. The mean GAM scores are 3.41 and 3.5 in these samples. Section 1.2 of the online appendix provides more details on the IAT and how we implemented it and the GAM.

Figure A2. Observed Gender Gap in Leader Effectiveness in Experiment 1 and Related Studies (Gender Visible Conditions)



Notes: The gender difference in leader effectiveness is measured as the percentage point difference between the average success of a male leader minus that of a female leader in an experiment. Success ranges from 0% to 100% and is the outcome a leader achieved relative to what would be maximal success in an experiment. The bars show the average observed gender gap in leader effectiveness in gender visible conditions, along with 95-% confidence intervals. The unit of observation is the leader, and we use data from the first period after leaders were introduced. While similar in spirit, Reuben and Timko (2018) and Grossman et al. (2019) differ from our study design along several dimension. Reuben and Timko (2018), for example, have additional treatment arms that vary procedures for selecting leaders. To compare effect sizes across studies, we selected the experimental conditions that are the most similar across the three studies. Section 1.5 of the online appendix provides more details on the comparisons.

Figure A3. Power Curve for Experiment 2



Notes: Figure depicts the power of our study, using a two-sided one-sample test of proportion to reject the null hypothesis that women are followed at the same rate as men, that is, at a rate of 50%, at the 0.05 level of statistical significance or higher. The power is plotted as a function of alternative hypotheses, that is, alternative proportions at which women could be followed in the leader competition game whenever they are pitted against a male leader. The reference line is at the average rate of female following that the sample of research economists predicted for the gender visible condition.

Table A1. Employee and CEO Payoffs in Parts 1 and 2 (Experiment 1)

Part 1 (Periods 1-6)

Hours spent on the risky project	Minimum hours spent by others on the risky project					Minimum hours spent by others on the risky project				
	0	10	20	30	40	0	10	20	30	40
0	200	200	200	200	200	200	200	200	200	200
10	150	200	200	200	200	150	210	210	210	210
20	100	150	200	200	200	100	160	220	220	220
30	50	100	150	200	200	50	110	170	230	230
40	0	50	100	150	200	0	60	120	180	240
CEO Payoff	average earnings of employees in same firm									
	Rate of return = 5					Rate of return = 6				

Part 2 (Periods 7-16)

Hours spent on the risky project	Minimum hours spent by others on the risky project					Minimum hours spent by others on the risky project				
	0	10	20	30	40	0	10	20	30	40
0	200	200	200	200	200	200	200	200	200	200
10	150	230	230	230	230	150	250	250	250	250
20	100	180	260	260	260	100	200	300	300	300
30	50	130	210	290	290	50	150	250	350	350
40	0	80	160	240	320	0	100	200	300	400
	Minimum hours spent on the risky project in firm									
	0	10	20	30	40	0	10	20	30	40
CEO Payoff	150	230	310	390	470	150	250	350	450	550
	Rate of return = 8					Rate of return = 10				

Table A2. Impact of Gender Leader on Group Performance (Experiment 1)

	OLS regressions predicting group performance (minimum hours invested in risky option)				
	(1) First period of Part 2	(2) First period of Part 2	(3) Average for all of Part 2	(4) All periods in Part 2	(5) All periods (panel)
Female leader	-5.739 (4.165)	-0.017 (3.525)	-0.091 (3.851)	-0.091 (3.535)	-0.028 (2.113)
Leader visibility	-2.044 (3.831)	0.329 (3.080)	1.666 (3.544)	1.666 (3.615)	0.240 (2.257)
Female leader X Leader visibility	1.175 (5.457)	-1.215 (4.579)	-2.928 (5.046)	-2.928 (4.931)	-0.0312 (3.010)
Part 1 group performance	0.521* (0.270)	0.091 (0.243)	0.118 (0.250)	0.118 (0.204)	
Leader recommendation (number of hours)		0.778*** (0.107)			
Constant	29.70*** (3.020)	2.287 (4.751)	30.76*** (2.793)	30.76*** (2.780)	3.510** (1.536)
N (groups)	100	82	100	1000 (100)	1600 (100)
R^2	0.0284	0.3924	0.0032	0.0072	0.5945

Notes: Female leader is an indicator for whether the leader l of a group is female and *Leader visibility* indicates whether the portrait was shown to followers. *Female leader X Leader visibility* interacts these two variables. *Group performance in Part 1* is the highest group equilibrium played—the maximum of the minimum hours a group invested in the risky project—over the last three periods of Part 1. For robustness, we also attempt to control for group outcomes in Part 1 in different ways—e. g. the group hours in the risky project in period 6—which does not change our results. *Leader recommendation (number of hours)* records the specific numeric investment recommendation the leader gave to followers, as coded by independent raters, whenever such a specific request was made. In later rounds, there are fewer and selected instances in which leaders gave numeric investment recommendations since many participants stopped doing this once their group was coordinated at 40 hours investment in the risky project. This is why we only control for investment recommendations when we predict minimum hours in risky project in period 7. Not all leaders gave numeric recommendations in period 7 so our sample for this analysis is a subset of the full sample. In the panel specification, we control for period fixed effects. Standard errors in parentheses (clustered at the group level in models 4 and 5) *Significant at the 10% level,** at the 5% level, *** at the 1% level.

Table A3. Impact of Gender Leader on Followers' Initial Individual Investments (Experiment 1)

	OLS regressions predicting hours invested in risky project by follower i in period 7			
			Male followers	***Female followers***
	(1)	(2)	(3)	(4)
Female Leader	-3.669 (2.366)	0.944 (1.259)	-3.795 (2.305)	-4.152 (2.889)
Leader Visibility	-2.852 (2.310)	-0.670 (1.001)	-4.559** (2.305)	-1.340 (2.754)
Female Leader X Leader Visibility	2.814 (3.259)	-0.335 (1.920)	4.951 (3.259)	1.190 (3.887)
Group Performance Part 1	0.263 (0.105)	0.009 (0.056)	0.142 (0.109)	0.407 (0.134)
Leader Recommendation (number of hours)		0.710*** (0.082)		
Constant	35.743*** (1.581)	9.913 (3.295)	37.112*** (1.413)	34.590*** (2.063)
N followers	500	410	254	246
R^2	0.0322	0.549	0.0319	0.0492

Notes: Female leader is an indicator for whether the leader l of a group is female and *leader visibility* indicates whether the portrait was shown to followers. *Female leader X leader visibility* interacts these two terms. *Group performance in Part 1* is the highest group equilibrium played—the maximum of the minimum hours a group invested in the risky project —over the last three periods of Part 1.*Leader recommendation (number of hours)* records the specific numeric investment recommendation the leader gave to followers, whenever such a specific request was made. Standard errors in parentheses (clustered at the group level). *Significant at the 10% level,** at the 5% level, *** at the 1% level.

Table A4. Follower Behavior in Response to the Same Leader Pair in a Round by Gender Visibility (Experiment 2)

		Leader visible		
		Both choose M	One F one M	Both choose F
Leader invisible	Both Choose M	12.7	12.1	5.1
	One F one M	13.0	13.6	13.2
	Both Choose F	6.0	11.7	12.6

Notes: The table displays the outcomes of follower behavior in response to the same leader pair of mixed gender by gender visibility. We observe the same 470 leader pairings once when their gender was visible to followers and once when it was not. Entries in the diagonal of that table represent instances in which the behavior of followers was the same, irrespective of whether the gender of the leaders was known. Cells shaded in light gray indicate instances in which gender-visibility brought about a more female following. White cells indicate instances in which followers followed the male leader more, when the gender of the two leaders was visible.

Appendix B1. Leader Persuasiveness in Coordination Games

In our first experiment, we model leadership as the ability to persuade all followers to collectively change behavior and move from an investment level of zero units to a recommended positive amount invested in the risky project. Successful leaders can influence equilibrium selection in this coordination game. In the following paragraphs, we introduce a simple theoretical framework that captures how leaders mediate equilibrium selection in one-shot simultaneous move coordination games through an individual trait that we refer to as "leader persuasiveness." This notion is applied to two types of coordination games. The analysis provides insights into why we designed and implemented a second experiment.

1.1. Weak-Link Game

The first coordination game that we consider is a simplified version of the weak-link game of Experiment 1. In its simplest representation, the weak-link game is the following version of a Stag Hunt coordination game. Followers $i \in \{1,2,3,\dots,n\}$ with $n \geq 2$ choose simultaneously between two actions. They can either invest a positive amount in a group project, $I_G = 1$, or not invest, $I_G = 0$. Followers all have the same payoff function. They receive some guaranteed amount $\pi_0 > 0$ when they do not invest in the group project ($I_G = 0$). This amount is independent of the action of other followers. When they choose to invest ($I_G = 1$) in the group project, followers receive a payoff equal to π_G whenever all others invest, and 0 otherwise, with $\pi_G > \pi_0$. Figure B1 illustrates the payoff table for followers for the case of two players.

Before all followers make their choice, a leader l can send a written statement r_l to all followers, in which she advises the followers on whether to invest in the group project. All followers read r_l before they simultaneously decide whether to invest in the group project, and that is common knowledge. The leader receives a payoff equal to π_G when all followers invest in the group project and receives a payoff of 0 otherwise.

This game has two pure-strategy equilibria in which all followers choose the same action. From the perspective of the individual follower who decides on which strategy to play, the two strategies differ in the risk and return profiles.

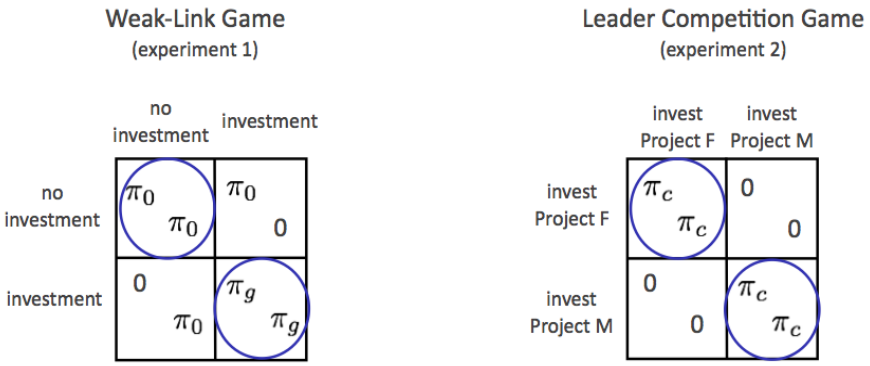
1.2. Leader Competition Game

The design of Experiment 2 builds on a simple pure-matching coordination game. Followers $i \in \{1,2,3,\dots,n\}$ with $n \geq 2$ choose simultaneously between two actions. They

can either play action F or action M. Followers receive a payoff of $\pi_C > 0$ when they all play the same action, which can be either F or M, and a payoff of 0 if their actions do not match. See Figure B1 for an illustration of follower payoffs for the case of two players. Before all followers make their choice, two leaders l^1, l^2 can each send a written statement r_{l^1}, r_{l^2}. In the statement, a leader advises the followers on which action they should play. All followers read the two statements r_{l^1}, r_{l^2} before they simultaneously decide whether to play F or M, and that is common knowledge. One leader is a woman and the other leader is a man. The leaders have opposing preferences. The female leader obtains a payoff of $\pi_C > 0$ when *all* followers choose action F and a payoff of 0 otherwise. The male leader obtains a payoff of $\pi_C > 0$ when *all* followers choose action M and, otherwise, a payoff of 0.

The second-stage game between the followers has two pure-strategy equilibria in which all followers choose the same action, as well as mixed-strategy equilibria involving randomization. From the perspective of the individual follower who decides on which strategy to play, the two strategies are equivalent in terms of their risk and return profiles.

Figure B1. Follower Payoffs in the Weak-Link Game and the Leader-Competition Game when N=2



Notes: Blue circled cells indicate that the outcome is a pure-strategy Nash Equilibrium.

1.3. Leader Persuasiveness and Leader Social Credibility

We posit that the followers' perception of leadership quality is instrumental in selecting the equilibrium that followers will play.

Definition 1. *A leader's persuasiveness* $\alpha(r, k) \in [0,1]$ *is a follower's subjective perception of the degree to which the leader's statement is convincing*

The leader's persuasiveness is a function of the statement r that the leader writes to convince followers to play a specific strategy in a coordination game. Moreover, we

conceptualize that the leader's persuasiveness may also be a function of observable leader characteristics k, for example, the leader's gender, ethnicity, height, eye color, voice tone. Gender is the leader characteristic that randomly varies in the modeled experiments and because of this, we use $k \in \{f, m\}$ as a summary of all relevant leader characteristics. In our leader-blind conditions, we eliminate the possibility that k influences persuasiveness, meaning that any gender difference must come from differences in male and female leaders' distributions of message quality, r. The gender-visible conditions, on the other hand, allow for differences in persuasiveness to arise solely because of different leader characteristics, k.

In simultaneous move coordination games, a follower wants to take the action that she believes others are likely to take because of the strong complementarity in actions. Because of this, a follower forms a belief $p \in [0,1]$ about how likely it is that all others are going to take the advice of a leader instead of ignoring it. We conceptualize a leader's "social credibility," $p(\alpha)$, as the subjective probability that a follower assigns to all others in the group playing the equilibrium that is promoted by the leader.[17]

Definition 2. *A leader's social credibility* $p(\alpha) \in [0,1]$ *is the follower's prior belief about how likely it is that all others in the group will be convinced by the leader's advice instead of ignoring it.*

We conceptualize that this prior belief $p(\alpha)$ is a function of leader's persuasiveness $\alpha(r, k) \in [0,1]$. It is specific to either type of game and we make minimal and plausible assumptions about how a leader's persuasiveness $\alpha(r, k)$ maps into the leader's social credibility $p(\alpha)$.[18]

[17] It is common to advance prior beliefs about the relative likelihood of equilibria in theories of equilibrium selection. Harsanyi and Selten (1988), for example, conceptualize that players of a game understand that the outcome of the game will be a Nash equilibrium but that, initially, they do not know which one will be chosen, before they play the game. Through a specific way of reasoning about the relative likelihood that each equilibrium will be played, and player best-responding to these prior beliefs, the so called Harsanyi-Selten tracing procedure selects one equilibrium before the game is played.

[18] We impose that followers have a shared understanding of their leader's social credibility. This is a direct consequence of the modeling choice that all followers observe the same $\alpha(r, k)$. One way to justify it is that all followers read the same statement and that it seems plausible that followers have at least a similar view about the persuasiveness of that statement. For ease of exposition, instead of imposing some bounds on the degree to which the assessments $\alpha_i(r, k)$ may differ, we simply impose that $\alpha_i(r, k) = \alpha(r, k)$ for all followers i in a group.

> **Assumption 1 (A1)** In the Weak-Link Game, $p(0) = 0$, $p(1) = 1$ and $p'(\alpha) > 0$, in other words, $p(\,)$ is strictly monotonically increasing in α.

This simply means that the subjective probability that a follower assigns to the others in her group following the advice of the leader, instead of ignoring it, increases, as the leader's statement is perceived to be more personally convincing.

In the Leader-Competition Game, the leaders' social credibility is jointly determined, that is, each is a function of the persuasiveness of both leaders.

> **Assumption 2 (A2)** In the Leader-Competition Game, $p_1(\alpha^1, \alpha^2) > p_2(\alpha^1, \alpha^2)$ with $p_1(\alpha^1, \alpha^2) + p_2(\alpha^1, \alpha^2) = 1$ if $\alpha^1 > \alpha^2$ and $p_1(\alpha^1, \alpha^2) = 1 - p_2(\alpha^1, \alpha^2) = 0.5$ if $\alpha^1 = \alpha^2 \geq 0$.

In words, we assume that whenever a leader (l^1) is more persuasive than the other leader (l^2), a follower holds the prior belief that it is more likely that all others will choose the action promoted by l^1 than all playing the action promoted by l^2. Whenever the two leaders are exactly equally socially credible, a follower gives the same subjective probability to the two events that either all other followers will play M or that all other followers will play F.

When followers ignore the advice of their leader, they play the "babbling equilibrium" of the coordination game. This situation is strategically equivalent to the second-stage coordination game played without a pre-play communication phase. This means that the equilibrium that followers play in the absence of leadership is the same as the babbling equilibrium of the game.

> **Assumption 3 (A3)** When followers play the Weak-Link Game without the leader, followers play the equilibrium in which everyone invests zero units in the group project ($\boldsymbol{I_G^* = 0}$).

A3 reflects an important feature of the actual experimental context, in which groups converged to the play the risk-dominant equilibrium $\boldsymbol{I_G^* = 0}$ in a first part of the experiment without a leader. Moreover, based on earlier research (e.g., Van Huyck, Battalio, and Beil, 1990; Brands and Cooper, 2006), it seems likely that groups will converge to the risk-dominant but inefficient equilibrium in the absence of leadership in the weak-link game in our experiment.

> **Assumption 4 (A4)** When followers play the Leader-Competition Game without the leader, the behavior of followers is best described by the mixed strategy Nash equilibrium $\boldsymbol{q}^*$ of the simple coordination game in which each follower randomizes between the play of F and M with equal probability, e.g. $q_i^*(\frac{1}{2}, \frac{1}{2})$.

Empirically, **A4** must be satisfied in the aggregate; in anonymous experimental groups in which people interact once without communication they should coordinate around half of the time in this pure-matching coordination game. Nothing systematic can guide players' thinking in terms of which action is relatively more likely to be played by others.

> **Assumption 5 (A5)** Followers either take the advice of the leader at face value or they ignore it.

A5 rules out that followers do, for example, collectively the opposite of what the leader says. Empirically, A5 is the most realistic way to think about followers' reaction to the messages in a one-shot simultaneous move game in which they cannot communicate. Theoretically, it is in line with the notion of focal points in coordination games that Schelling (1969) introduced to recognize that humans often have a way to single out specific actions as more salient than others.

Under A1-A5 we can compare the weak-link coordination game and the leader competition game in their ability to uncover that women are less effective leaders than men. In the framework that we introduced, the effectiveness of a leader is linked to the leader's social credibility.

For illustration, consider a scenario in which all women have the same persuasiveness $\alpha(r, f)$ and all men have the same persuasiveness $\alpha(r, m)$ which implies that all female leaders have the same leader social credibility $p(\alpha_f)$ and all men have the same leader social credibility $p(\alpha_m)$.[19]

The experimenter never directly observes the leader's social credibility or the underlying leader persuasiveness. The experimenter only observes followers' behavior, that is, the equilibrium they play. From this, outside observers can judge whether a leader was effective. In the weak-link game, an effective leader gets followers to invest in the risky project. In the leader competition game, the effective leader is the one who persuades

[19] The key insights of this modeling framework extend to a stochastic environment in which we assume that leader persuasiveness is distributed according to the cumulative distribution functions $F_{r,f}()$ and $F_{r,m}()$. Men are, stochastically, more persuasive than women if $F_{r,m}()$ if $F_{r,m}()$ first order stochastically dominates $F_{r,f}()$.

followers to invest in the leader's preferred project. Our main interest is in how well each game performs in revealing differences in the social credibility of female and male leaders, given that outside observers only see the behavior of followers.

Proposition 1. *Assume A1-A5.*

- *In the Weak-Link Game female leaders are less effective than male leaders if and only if* $p(\alpha(r,f)) < \frac{\pi_0}{\pi_G} < p(\alpha(r,m))$.
- *In the Leader-Competition Game, female leaders are less effective than male leaders if and only if* $\alpha(r,f) < \alpha(r,m)$.

Proof. *See Appendix B2.*

In the statement of Proposition 1, we do not differentiate between the cases in which differences in persuasiveness arise through differences in the quality of messages sent by male versus female leaders (r) or followers' responses to leader characteristics (k). Which one, if any, causes women's differential effectiveness as leaders is ultimately an empirical question that we are disentangling in our two experiments.

Proposition 1 implies that the weak-link game can fail to indicate any differences in the effectiveness of male and female leaders, even if men are (substantially) more socially credible than women (i.e., if $p(\alpha(r,m))$ and $p(\alpha(r,f))$ are both less than or greater than $\frac{\pi_0}{\pi_G}$, even if $|p\big(\alpha(r,m)\big) - p(\alpha(r,f))|$ is large). In this game, all that is required in order for a follower to choose higher investment is that the follower believe it is sufficiently likely that others will do so. That is, there is a threshold belief about how likely others are to respond positively to a leader, and a belief below this threshold means that a follower should not increase her investment, while a higher belief means that the follower should. Thus, leaders are effective whenever their social credibility is above this threshold. It may be the case, therefore, that even when men and women differ in how credible they are believed to be, that they are both above or below the threshold, meaning that the weak-link game will not detect any differences in how effective men and women are to persuade others.

By contrast, Proposition 1 states that the leader competition game is guaranteed to show differences in leadership effectiveness whenever the statements of male leaders are perceived to be more persuasive than the statement by female leaders. This is the case because, in this game, any differences in the perceived persuasiveness will help followers to select which equilibrium to play.

Appendix B2 - Proofs

Lemma 1. *Assume A5. In the Weak-Link Game, the leader recommends* $I_G = 1$. *In the Leader-Competition Game, the female leader recommends* F *and the male leader recommends* M.

Proof. This follows directly from A5 under which it is weakly better for a leader to recommend that followers choose actions that are weakly better for the leader ■

Lemma 2. *Assume A1, A3, A5. A follower* i *in the Weak-Link Game follows the advice of the leader if and only if* $p(\alpha(r,k)) \geq \frac{\pi_0}{\pi_G}$.

Proof. Followers who ignore the advice of the leader play $I_G = 0$ by A3. The leader recommends the play of $I_G = 1$ by lemma 1. By the definition of leader social credibility, a follower observes (r,k) and assigns subjective probability $p\big(\alpha(\mathrm{r},\mathrm{k})\big)$ to the event that all other followers will take the advice of the leader and assigns probability $1 - p\big(\alpha(r,k)\big)$ to the event that all other followers ignore it. Her unique best-response to prior belief $p\big(\alpha(\mathrm{r},\mathrm{k})\big)$ is to ignore the advice and play $\mathrm{I_G} = 0$ when $p\big(\alpha(\mathrm{r},\mathrm{k})\big) < \frac{\pi_0}{\pi_G}$ and to take the advice of the leader and play $\mathrm{I_G} = 1$ when $\big(\alpha(\mathrm{r},\mathrm{k})\big) \geq \frac{\pi_0}{\pi_G}$. The primitives of the game are such that $\frac{\pi_0}{\pi_G} \in]0,1[$. By A1, strict monotonicity of $p(\cdot)$ in $\alpha(r,k)$ and $p(0) = 0, p(1) = 1$ this threshold value $p\big(\alpha(r,k)\big) = \frac{\pi_0}{\pi_G}$ is unique■

Lemma 3. Assume A2, A4, A5. A follower i in the Leader-Competition Game follows the advice of leader l^1 if and only if $\alpha^1(r,k) > \alpha^2(r,k) \geq 0$.

Proof. We will first show the *if part* of the proposition. By lemma 1, the two leaders recommend different actions. Label the strategy that leader l^1 promotes $L^1 \in \{\mathrm{F},\mathrm{M}\}$ and the strategy that the other leader l^2 promotes $\mathrm{L}^2 \in \{\mathrm{F},\mathrm{M}\}\backslash L^1$. Without-loss-of-generality, assume that leader l^1 is more persuasive, that is, $\alpha^1(r,k) > \alpha^2(r,k)$. By the definition of leader social credibility, follower i believes that with probability $p^1(\alpha^1,\alpha^2)$ everyone else in the group will follow the advice of leader l^1 and believes that with probability $1 - \mathrm{p}^1(\alpha^1,\alpha^2)$ the other followers will play the action recommended by leader l^2. The follower chooses L^1 since this is her unique response to her prior belief $p(\alpha^1(\mathrm{r},\mathrm{k}))$. To see this, note that her expected payoff from playing L^1 is strictly larger than her expected payoff from playing L^2:

$$E[\pi(L^1, p^1(\alpha^1, a^2))] = p^1(\alpha^1, \alpha^2)\,\pi_c > (1 - p^1(\alpha^1, \alpha^2))\pi_c = E[\pi(L^2, p^2(\alpha^1, \alpha^2))]$$

where the strict inequality follows from A2. This implies that she would never want to play L^2 or some mixed strategy $q(q_1, q_2)$ with $q_2 > 0$. In particular, the follower has no incentive to ignore the advice of the leader and play the babbling equilibrium of the game which is, by A4, the mixed strategy equilibrium with $q_i^*(\frac{1}{2}, \frac{1}{2})$. For the *only if* part, note that the only case the proof has not covered so far is the case in which $\alpha^1(r, k) = \alpha^2(r, k)$, since the assumption that l^1 is the more persuasive leader was without-loss-of-generality. By A2, followers play the mixed strategy equilibrium $q_i^*(\frac{1}{2}, \frac{1}{2})$ whenever the two leaders are exactly equally credible. This completes the proof of the lemma ■

Proposition 1. *Assume A1-A5. In the Weak-Link Game female leaders are less effective than male leaders if and only if* $p(\alpha(r, f)) < \frac{\pi_0}{\pi_G} < p(\alpha(r, m))$. *In the Leader-Competition Game, female leaders are less effective than male leaders if and only if* $\alpha(r, f) < \alpha(r, m)$.

Proof. We will first show the first part of the proposition. Consider the following cases $\frac{\pi_0}{\pi_G} \leq \mathrm{p}(\alpha(\mathrm{r}, \mathrm{f})) \leq \mathrm{p}(\alpha(\mathrm{r}, \mathrm{m}))$. By lemma 2 followers under female leadership will listen to her and play the equilibrium she recommends. Similarly, followers under male leadership will listen to him and play the equilibrium he recommends. Next consider the cases $p(\alpha(\mathrm{r}, \mathrm{f})) \leq \mathrm{p}(\alpha(\mathrm{r}, \mathrm{m})) \leq \frac{\pi_0}{\pi_G}$. By lemma 2 followers under either type of leadership will ignore the advice of the leader and play the babbling equilibrium. Lastly, if $p(\alpha(\mathrm{r}, \mathrm{m})) < \frac{\pi_0}{\pi_G} < \mathrm{p}(\alpha(\mathrm{r}, \mathrm{f}))$, it follows from lemma 2 that women will be more effective leaders than men, whose advice followers will ignore in equilibrium. This completes the proof of the first part of the proposition. The second part follows directly from lemma 3■