Discrimination from Below: Experimental Evidence from Ethiopia *

Shibiru Ayalew Shanthi Manian Ketki Sheth

Globally, women are underrepresented in leadership positions. A potential explanation is that gender discrimination by subordinates reduces the effectiveness of female leadership. Using a lab-in-the-field experiment in Ethiopia, we test whether leader gender affects the way subjects respond to leadership. We find subjects are ten percent less likely to follow the *same* advice from a female leader than an *otherwise identical* male leader. Subjects also give lower evaluations to hypothetical female managerial candidates. However, we find that ability information reverses discrimination. When leaders are presented as highly trained and competent, subjects are *more* likely to follow advice from women than men. This pattern suggests that beliefs about men and women's ability (i.e., statistical discrimination) play an important role in driving this discriminatory behavior. Our results show that gender discrimination affects adherence to leadership, and signals of ability may be an important tool for gender equity policies aimed at increasing female representation.

JEL Codes: O1, J7

Keywords: gender; discrimination; advice; lab in the field; leadership

*We are grateful to the East Africa Social Science Translation (EASST), administered by the Center for Effective Global Action (CEGA), for financial support, and to Adama Science and Technology University for supporting our study, sharing data, and the staff which provided invaluable assistance with implementing the study design. We also thank Prashant Bharadwaj, Monica Capra, Edward Miguel, Karthik Muralidharan, Aurelie Ouss, Siqi Pan, Lise Vesterlund, Sevgi Yuksel, and various seminar participants for helpful suggestions and comments. This study was preregistered at the AEA RCT Registry (AEARCTR-0002304). No third party had the right to review this paper prior to its circulation.
Ayalew: Adama Science and Technology University (shibekoo84@gmail.com), Manian (corresponding author): Washington State University (shanthi.manian@wsu.edu), Sheth: University of California Merced (ksheth@ucmerced.edu).

1 Introduction

Improving gender equity in leadership positions is a key priority of the global development agenda.[1] Globally, women remain underrepresented in leadership roles, with the largest gender gaps concentrated in low-income countries. For example, women hold just 17 percent of board directorships in the world's 200 largest companies (African Development Bank, 2015). In this paper, we explore a potential explanation: that gender discrimination makes subordinates less likely to adhere to female leadership. Successful performance in leadership often depends on how well others adhere to one's advice and direction. If women face such discrimination from below by subordinates, then female leaders may appear less effective than male leaders despite taking identical actions.

Using a novel lab-in-the-field experiment in Ethiopia, we study whether individuals follow advice differently when they are randomly assigned to a male versus female leader. Strikingly, although the female and male leaders are otherwise identical, subjects respond differently to the same guidance when provided by a woman rather than a man. When given no information on their leader's ability, subjects are 10 percent less likely to follow a female leader's guidance. As a result, these female-led subjects earn fewer total points. That is, female leadership appears less effective despite being no different from male leadership, and subjects are made worse off by discounting female leadership.[2]

We then estimate whether information about the leader's ability can mitigate discrimination by subordinates. We inform a random subset of subjects that their leader is of high ability. This information has significantly higher returns for female leaders, and this differential response is large enough to *reverse* the gender gap. Among those who were informed that their leader was of high ability, subjects were *more* likely to follow the guidance of a female leader, making male leaders appear less effective despite identical behavior. Together,

[1] See, for example, Sustainable Development Goal 5.5: "ensure women's full and effective participation and equal opportunities for leadership at all levels of decision-making in political, economic, and public life."

[2] In addition to leader gender, subjects receive information about their leader's age range and occupation. All information other than gender is the same for male and female leaders.

the results of the experiment show that subordinates can be less likely to follow female leadership due to gender discrimination. However, a signal of high ability can be sufficient to reduce or even reverse such gender discrimination.

This pattern of results also provides suggestive evidence of the mechanisms underlying this discrimination. The fact that information about the leader's ability reverses the gender gap—rather than simply reducing it—implies that differing beliefs about the underlying ability of male and female leaders are contributing to gender discrimination.[3] That is, our results suggest that a simple dislike of following female leaders cannot fully account for this discrimination, and that statistical discrimination plays an important role in the results.[4] The reversal suggests that the same information about leader ability differentially changes beliefs about male versus female ability.

There are two key strengths to the study design. First, we conduct a framed lab-in-the-field experiment with a sample of full-time employees. In general, discrimination from below is difficult to identify using correspondence or audit studies because it requires varying the behaviors and characteristics of those in a position of relative seniority. By conducting a lab experiment we are able to overcome this constraint and provide clean identification of discrimination towards others in senior positions. Furthermore, we document these results in a unique sample of highly educated, high-skilled employees at a large Ethiopian university. The individuals in our sample work in a hierarchical management setting and face daily decisions about following directions from those in more senior positions. Second, we supplement our lab-in-the-field experiment by asking subjects to evaluate a resume for a hypothetical senior management position in which the candidate gender is randomly assigned. Subjects gave lower evaluations to female candidates for the position, providing additional evidence

[3]We do not claim that these beliefs are necessarily accurate reflections of differences between men and women; for the remainder of the paper, we use the convention of referring to any discrimination based on beliefs about the underlying groups, accurate or not, as statistical discrimination.

[4]It is difficult to completely disentangle statistical discrimination and taste-based discrimination because in equilibrium, taste-based discrimination and statistical discrimination may be mutually reinforcing. Average difference in male and female leadership ability may emerge because of taste-based discrimination or gender norms that prevent women from obtaining human capital. As a result, statistical discrimination may emerge that, in turn, reinforces these gender norms.

that subjects discriminate based on gender when evaluating leadership positions.

This paper has three main contributions. First, to the best of our knowledge, we are among the first to show that gender discrimination affects adherence to leadership. Leaders will be less effective if subordinates are unwilling to listen to them, and we show that leader gender directly affects subjects' decision to follow leadership even when that decision is costly. Thus, our results show that gender discrimination from subordinates may make one leader appear less effective than another, despite both having taken identical actions. This highlights a potential mechanism underlying gender gaps in leadership positions.

This finding builds on several important studies documenting differential responsiveness to female versus male leaders, advisers, and experts, particularly in low-income countries (Gangadharan et al., 2016; Grossman et al., 2019). This recent evidence includes female manager trainees in Bangladeshi garment factories being seen as less effective, female-owned businesses in Ghana receiving fewer customers, and farmers perceiving female agricultural trainers in Malawi as less knowledgeable despite the female trainers being equally effective at diffusing new technologies (Macchiavello et al., 2015; Hardy and Kagy, 2018; BenYishay et al., 2020).[5] This literature documents a consistent differential response to women, but leaves open the question of whether it is driven by gender discrimination. In these natural settings, men and women often differ on a number of characteristics. And, even when men and women are observably similar, subtle differences in communication style, confidence, and risk preferences can drive gender gaps in adherence to leadership.[6]

In our study, we build on this literature by offering explicit identification of discrimination based on gender. We observe all information presented to subjects and experimentally

[5]In high-income countries, female university professors receive lower teaching evaluations (Mengel, Sauermann and Zölitz, 2019; Boring, 2017) and female experts are more likely to be punished for random negative shocks (Egan, Matvos and Seru, 2017; Landsman, 2018; Sarsons, 2017). Sarsons (2017) also shows that male experts are more likely to be rewarded for positive shocks, and that this implies that signals are interpreted differently for men and women.

[6]A significant literature documents average differences by gender in communication style, confidence, and risk preferences (see Niederle (2017) for a review). In the United States, Manian and Sheth (2020) document no gender discrimination in adherence to a leader's advice, but show that female leaders preferred using less assertive language which does reduce adherence to advice.

assign leader gender. We limit the interaction between subjects and leaders to written communications, and pre-scripted messages are used to ensure that leader gender is the only difference between the two groups. Thus, we capture how individuals respond to gender itself, as opposed to correlates of gender.[7] Our lab-in-the-field results support the notion that gender discrimination contributes to the gaps documented in field experiments; likewise, the field experiments highlight the external validity and real-world consequences of our findings.

Our second contribution is to show that providing information about female leaders' ability can alleviate gender discrimination. Since the reduced adherence to female leaders documented in the literature can be driven by multiple factors, it remains an open question whether information on ability can help close such gender gaps. We find that it can. Information about ability raised subjects' adherence to female leaders by 12 percentage points, while it had no detectable effect for male leaders. The pattern we observe, in which ability information reverses the gender gap in adherence, implies that subjects make inferences about leader ability based on leader gender (i.e., it is consistent with statistical discrimination as a dominant driver of the results).

By showing that information can reduce discrimination, we contribute to the literature on anti-discrimination interventions at the individual level, which has previously focused primarily on encouraging contact between groups rather than information interventions.[8]

[7]Psychologists have used lab experiments to study discrimination toward female leaders in a variety of settings (See Eagly (2013) for a review). However, this literature focuses on high-income countries and generally does not involve real stakes. Low-income countries differ from high-income countries in two important ways. First, gender gaps in a variety of human capital outcomes (e.g., educational attainment) are larger in low-income countries (Duflo, 2012; Jayachandran, 2015). Such gender gaps could generate different beliefs about male and female ability. Second, a diverse set of cultural factors affects gender norms around the world, and poverty can increase reliance on such traditional norms (Jayachandran, 2015). This suggests that both beliefs and preferences will differ in low-income versus high-income country contexts. This in turn implies different patterns of discrimination, since beliefs and preferences underpin theories of discrimination. Introducing real stakes in experiments allows for the opportunity to see how decisions are made when they have a consequence to the subject, as we expect in most real world settings.

[8]The notion that intergroup contact may reduce discrimination is known as the "contact hypothesis." See Paluck, Green and Green (2019) for a review of interventions based on the contact hypothesis. Dahl, Kotsadam and Rooth (2020) show that male members of the military rate women's performance more highly after being assigned to a mixed-gender team. Lowe (2020) and Mousa (2020) also provide recent evidence that contact-based interventions are effective for ethnic discrimination. Contact-based interventions may work because they provide information: interacting with members of a different group changes beliefs about their ability. However, to the best of our knowledge, this mechanism has not been directly tested.

Our finding that ability information reduces discrimination at the individual level reinforces evidence from the United States that interventions that signal ability can reduce gender and racial gaps in the labor market. For example, education and occupational licensing have been shown to have higher returns for black men and women (Arcidiacono, Bayer and Hizmo, 2010; Blair and Chung, 2020). Our results suggest that signals of ability, such as credentials, may be an important tool for gender equity policies directed at increasing female representation.

Our third contribution is the finding that a signal of ability can reverse gender discrimination outside a dynamic context. Bohren, Imas and Rosenberg (2019) show that an ability signal can reverse a gender gap in evaluations because evaluators account for discrimination faced in obtaining the ability signal. A key difference between this paper and Bohren, Imas and Rosenberg (2019) is that our experiment has no dynamic component: subjects have no reason to believe that it would be more difficult for women to obtain the ability signal in our experiment. This suggests a broader phenomenon in which subjects respond particularly favorably to women of high ability, perhaps due to a general environment in which women commonly face barriers to attaining skills or accolades. Importantly, such reversals indicate that positive discrimination in favor of high-ability women does not preclude the existence of discrimination against women in other contexts.

The rest of the paper proceeds as follows. Section 2 provides details on the experimental design of our study. In Section 3 and 4, we present our findings and supporting evidence. Section 5 discusses potential mechanisms and policy implications of the results, and Section 6 concludes.

2 Study Design

We conducted our study in Adama, Ethiopia, with a sample of full-time administrative employees at Adama Science and Technology University (ASTU) that hold a BA or higher.

Our primary results are based on an experiment we conducted in a subsample of these employees. We constructed the sample ourselves through local recruitment at the university. The subjects are high-skilled administrative employees of an institution, and are unlikely to have participated as subjects in prior research. We supplement the experimental results with data from a survey experiment and institutional human resources data on the universe of ASTU administrative employees.

2.1 Context

Ethiopia generally performs poorly on global indicators of gender inequality. For example, in the World Economic Forum's 2016 Global Gender Gap Report, Ethiopia ranked 109 of 144. This low rank was driven by their rank on sub-indices related to education and labor market outcomes: they ranked 106 on "Economic participation and opportunity" and 132 on educational attainment.

Adama Science and Technology University (ASTU) is an elite public university located about 100 km from the capital, Addis Ababa. To provide context for the potential beliefs of subjects in our sample, we use institutional human resources data to describe the characteristics of ASTU administrative employees (Table I). Educational attainment among employees is high: on average, employees completed 12 years of education, which corresponds to secondary school completion. In contrast, in the Ethiopian population more broadly, 48.3 percent of women and 45.7 percent of men are out of secondary school (World Bank, 2017). Nearly 30 percent of the sample has a BA or higher, while the gross tertiary enrollment ratio in Ethiopia is just 8 percent (World Bank, 2017). Turnover among administrative employees at ASTU is low: average job tenure is 8 years. We observe significant differences in job tenure and salary by gender: women have been with the institution longer, but are paid less on average.

Female administrative employees have significantly fewer years of education: they are 37 percent less likely to hold a Bachelors degree and 75 percent less likely to hold a Masters

Table I: Summary Statistics: Adama Science and Technology Administrative Employees

	(1) Total	(2) Male	(3) Female	(4) Diff.
Female	0.56			
	(0.50)			
Tenure	8.00	7.61	8.31	-0.71*
	(5.55)	(5.95)	(5.20)	
Years of education	12.87	13.04	12.73	0.31*
	(3.01)	(3.23)	(2.83)	
BA or higher	0.30	0.38	0.23	0.14***
	(0.46)	(0.48)	(0.42)	
MA or higher	0.02	0.04	0.01	0.03***
	(0.15)	(0.20)	(0.09)	
Salary	2354.62	2629.83	2135.97	493.85***
	(1536.24)	(1878.60)	(1151.46)	
Salary|BA or higher	3613.11	3681.16	3525.79	155.37
	(1624.55)	(1769.13)	(4161.84)	
Observations	1685	746	939	1685

Standard deviations in parentheses. Summary statistics are shown for the entire set of administrative employees at Adama Science and Technology University. Female is an indicator for the subject being female, Tenure is the number of years the subject has been employed by the University, Years of education are based on the subject's highest education level completed, BA or higher is an indicator for whether the subject holds a Bachelors degree, MA or higher is an indicator for whether the subject holds a Masters degree, and salary is the subject's monthly salary reported in Ethiopian Birr. Salary|BA or higher is the salary conditional on the sample who hold a BA or higher. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

degree. The salary gap we observe on average disappears when limiting attention to those with advanced degrees. Thus, we find that women are less likely to have obtained an advanced degree, a credible signal of ability, but that its differential return is higher for women.

2.2 Leadership Game: Lab-in-the-Field Experiment

2.2.1 Sample

Using a list of employees provided by the human resource department, we contacted all ASTU administrative employees with a BA or higher, and implemented the experiment until we reached 150 female subjects and 150 male subjects (see Table III below for summary statistics on this sample). Thus, relative to all university administrative employees, those in the experiment were more educated, had higher salaries, and were balanced on gender. Within this sample, there is no salary or tenure difference across subject gender, though women have fewer years of education than men even conditional on obtaining a bachelors degree.

Unlike in the United States, recruitment of subjects in this lab-in-the-field experiment was not routine: there was no systematic recruitment pool or reliable method to recruit subjects in advance of the experiment. Instead, enumerators would go to the unit at which the employee worked to recruit the subject to participate within the next few days, with most subjects participating on the same day they were informed of the experiment.

Subjects were informed that they were participating in "an experiment in the economics of decision making," and were not informed of the hypotheses regarding gender and ability.

2.2.2 Overview of design

The basic setup of the experiment is that subjects are randomly assigned to either a male or female "leader," are asked to complete two games, and are told that the role of the leader is to provide assistance in the second game. The subject never sees the leader, and interaction between the leader and subject is limited to written messages that are from the leader and

identical across all subjects. In this way, we are able to hold the leader's behavior constant across male and female leaders. The subject is given some information about their leader: their leader's gender, as well as their leader's age range, and that their leader works in a similar position at a different university. All information other than gender is the same for male and female leaders. The experiment is designed to quantify how the leader's gender affects the likelihood of subjects following the guidance provided by their leaders, and test whether providing information that their leader is highly able mitigates any gender gap.

The leaders were real individuals who played the games as described to the subjects a week prior at another university in Ethiopia. Unlike the subjects in the primary study, the leaders were given extensive training on how to play each game. We recruited a sample of eight potential leaders and selected the two top performing leaders, one male and one female, to be assigned to subjects. To hold behavior constant, the leaders played ahead of time, and we selected one male and one female leader who played in the same way and had the same outcomes to be matched to subjects. Leaders received a bonus based on the average performance of the team members assigned to them. Subjects were told that their leader's compensation was partly based on how well the subject performed on the game. Analysis on the sample of recruited leaders is not possible because only eight individuals were recruited to be potential leaders. The purpose of using real individuals as leaders, as opposed to describing fictitious leaders with these same characteristics, was to avoid deceiving our subjects.

To prime our subjects to consider leadership, we framed the experiment by referring to the person providing advice as a team leader. Enumerators explicitly referred to the "leader," using the relevant word in Amharic, throughout the experiment. Though our results on advice giving may be broader than just leadership settings, we maintain the "leader" descriptor, instead of "advisor," because of this framing. In addition, we recognize that a leader's role is more than just providing advice; however, by focusing on one aspect of leadership, we are able to causally estimate the effect of gender on advice following, holding

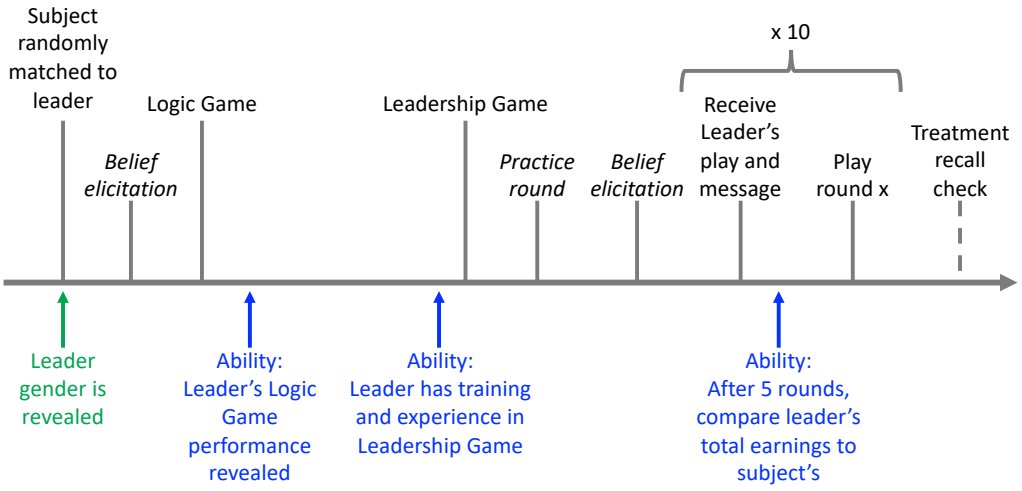

Figure I: Timeline of Leadership Game

all other aspects of leadership constant.

The experiment consists of two parts. The first part is a logic game, the Tower of Hanoi; we refer to this as the Logic Game. The second part is a game adapted from Cooper and Kagel (2005) in which subjects receive advice; we refer to this as the Leadership Game. The primary purpose of the first game, the Logic Game, is to serve as an input to the high ability signal treatment. The primary purpose of the second game, the Leadership Game, is to measure whether subjects follow their leader's directions. Figure I provides an overview of the experiment.

In the Logic Game, subjects are asked to solve the Tower of Hanoi (see Appendix Figure A.1 for details of the puzzle and Appendix Figure B for compensation schedule). How well a person solves the Logic Game is measured by the number of moves required, where fewer moves are better. Prior to playing, we asked subjects how many moves they thought *they* would require to solve the puzzle, how many moves they thought *their leader* would require to solve the puzzle, and finally how many moves they thought their leader guessed *they*

would require to solve the puzzle. These questions were specified in our preanalysis plan. However, the responses to these questions were bunched at the minimum number of moves and were highly skewed to the right, and therefore were ineffective tools for precisely eliciting beliefs. We observe no statistically significant difference across treatment assignments or across female and male subjects; also, mean differences for all three measures by subject gender and randomly assigned leader gender are less than one move. These results can be found in Appendix D.

The second component, the Leadership Game, was a game adapted from Cooper and Kagel (2005). We selected this game because it has a clear correct answer, but the correct answer is difficult to guess. We intentionally chose a complex game in order to create a clear and important role for leader advice. In this two-player game, nature first selects Player 1's type (A or B with 50 percent probability). Player 1 moves first. Player 2 then responds after seeing what Player 1 has selected, but without knowing Player 1's type. The payoff structure is shown in Figure II.[9]

The key insight is that for a Player 1 Type B, the optimal play is 5. The logic is as follows. A naive Player 1 Type B will select 3, observing that conditional on Player 2's selection, 3 always provides the highest payoff. But a Player 1 Type B can be "strategic" by selecting 5. If he selects 5, he can signal his type, because 5 is strictly dominated for Type A. If Player 2 knows that Player 1 is Type B, Player 2 is better off playing "Out" (Figure II). A similar logic applies to playing 4.

The leader advises the subject to play strategically and select 5 in this game. Because we are interested in how subjects respond to such advice, we assigned all subjects to be Player 1 Type B, and Player 2 was played by a computer. We programmed a mobile phone app to draw from the actual distribution of Player 2 responses by university students in Cooper and Kagel (2005). To make this clear to the subjects, they were told that the computer did not know whether they were Type A or Type B. In addition, we included the following

[9]The original game by Cooper and Kagel had 7 possible plays for Player 1 to select. We adapted the game to exclude the extreme options, leaving only 5 possible plays.

Player 1

Type A				Type B			
A's choice	In	Out		B's choice	In	Our	*Expected Payoff (not shown)*
1	168	444		1	276	568	*299*
2	150	426		2	330	606	*395*
3	132	426		3	352	628	*466*
4	56	182		4	334	610	*525*
5	-188	-38		5	316	592	*573*

Player 2 (Computer)

Computer's choice	Type A	Type B
In	500	200
Out	250	250

Figure II: The Leadership Game Payoffs (colors and expected payoffs not shown to subjects)

statement: "Though you are playing a computer, the computer has been programmed to mimic how real life university students have played this game, and so the computer does not always respond in the same way to a given number."

After learning the directions of the Leadership Game and completing comprehension questions, the subjects played a "practice round." In the practice round, the subjects selected the number they planned on playing prior to getting any advice from their leader. They did not see how the computer responded to this selection. Subjects were then asked what they believed was the probability of receiving each possible payoff in their first round, and the probability of their leader receiving each possible payoff in their leader's first round. Using these two questions, we calculate the subject's belief of the expected point value for him/herself and their leader in the first round of the Leadership Game. Our expectation was for subjects to report non-zero probabilities on only two of the options when eliciting beliefs of their own payoff (as the subject selects which number they will play), but the majority of subjects did include positive probabilities on more than two possible payoffs. Thus, these belief elicitation measures have a relatively large variance due to significant measurement

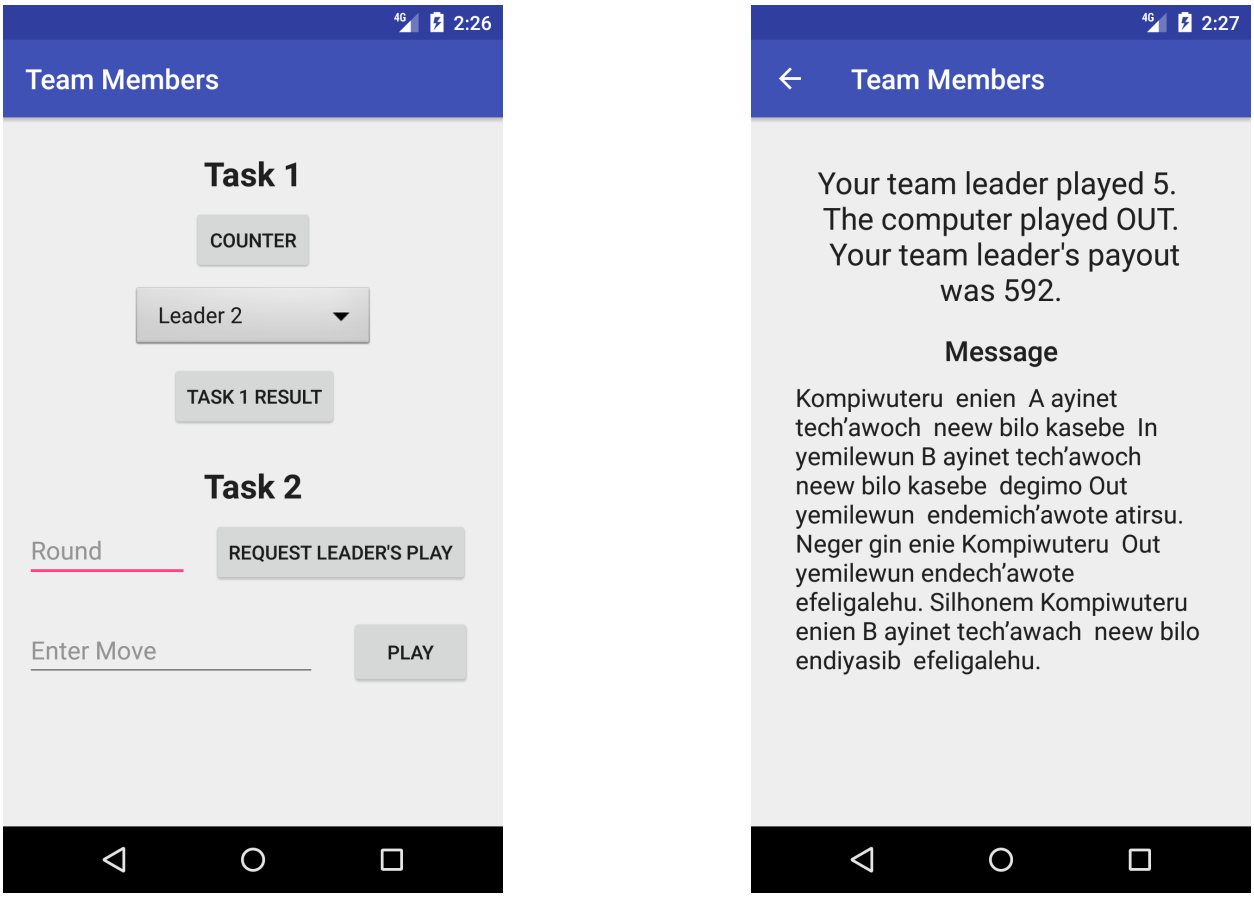

Figure III: Leader result and messages as shown to subjects
Note: the translated messages are shown in Appendix C.

error.

The subject then played 10 rounds on the Leadership Game. Prior to each round, the subject observed how their assigned leader played for that given round and the points the leader received. The leaders always selected 5 and received 592 points. In addition, subjects were told that the leader could send them messages. To control the content of the messages, messages were pre-written and leaders simply chose whether or not to send the messages to the subjects. All leaders chose to send the messages. The messages were displayed on an Android app by the enumerator (Figure III), and became increasingly informative over the rounds of the Leadership Game. The enumerator recorded the leader's play and outcome for each round on a piece of paper in front of the subject. The messages are provided in Appendix C.

Table II: Experimental Treatments

Male leader & Control	Female leader & Control
Male leader & Ability signal	Female leader & Ability signal

We completed the experiment in a span of 6 days. Options in the Leadership Game were relabeled for Day 5 and Day 6, such that Player 1 selected from two different sets of letters for Day 5 and 6, and the computer responded with "left/right" and "up/down".[10]

2.2.3 Experimental Treatments

We implemented a cross-cutting randomization of two treatments: leader gender and information on the leader being of high ability. As shown in Table II, subjects were randomly assigned to one of four groups: Male leader with no information on ability (control); female leader with no information on ability; male leader with a signal of high ability; or female leader with a signal of high ability.[11]

Leader Gender

Subjects were randomly assigned to either the male leader or the female leader. As discussed above, the information provided to the subjects about how the leader played is identical for both leaders, and subjects do not personally interact with their leaders. This ensures that the leaders were identical to each other, except for gender. In addition to explicitly telling the subjects the gender of their leader, we provided gendered pseudonyms for the leader, mentioned 23 times in the enumerator's script.[12] We also relied on the gendered grammatical structure of the local language, Amharic, to make the leader's gender salient. To confirm that subjects were aware of their leader's gender, we asked subjects a series of questions at

[10]Results are robust to including day fixed effects and we observe no consistent differential pattern of choices for subjects playing later in the study.

[11]We randomized leader gender and then independently randomized the ability treatment, so the subjects are not perfectly evenly distributed across treatments; i.e., a given treatment arm has 23 to 28 percent of subjects, as opposed to 25 percent.

[12]Subjects were informed that the name was a pseudonym to protect the privacy of their leader.

the end of the experiment on the characteristics of their leader, including gender, on the last two days of the experiment. 95 percent recalled the correct gender of their leader.

Leader Ability

We cross-randomized subjects to receive information on their leader being of high ability. This high ability treatment consisted of three components. First, after the "Tower of Hanoi" Logic Game, the enumerator informed the subject that their leader solved the Logic Game in 15 moves, and noted how many moves fewer this was than their own performance. In all cases, the leaders' number of moves was equal to or lower than the subjects' number of moves. Second, in the introduction to the Leadership Game, subjects were explicitly told that unlike themselves, the leader had already played the game and was an experienced player. And third, after 5 rounds of playing the Leadership Game, the enumerator totalled the points earned by the leader versus the subject to highlight the (expected) point advantage by their leader. From the subject's perspective, both the number of moves in the Logic Game and the total points mid-way through the Leadership Game are continuous measures of ability.

2.2.4 Subject Understanding

For our results to be valid, subjects must understand the gender of their leader, understand that earning more points increased their compensation, and know how to follow the advice provided.

In our validity exercises, we show that leader gender, our key independent variable of interest, was known almost unanimously among subjects (95 percent). We also observe that subjects play the game and select a number consistent with maximizing compensation. If subjects did not understand or care to maximize compensation, we should expect a uniform distribution among the numbers selected. But even from the first play of the game, we observe significant differences in the numbers selected, showing that subjects were trying to maximize their compensation. In the practice round of the game, prior to any advice

provided, 3, the naive selection that appears to provide the highest compensation, is the most common number selected (32 percent). Similarly, 1, the selection that provides the lowest compensation, is the least chosen (8 percent).

Finally, we require that the subjects understand that the advice was for them to select the number 5 in the game. It is highly unlikely that subjects did not understand the directions provided in the advice: the game does not advance until a number is selected, subjects are given detailed instructions on how to play the game, including comprehension questions and playing a practice round, subjects are paired one-to-one with the enumerator, and the advice on which number to select is given in simple terms in the local language. It may be the case that some subjects did not understand the *reasoning* behind the advice—indeed, the game is purposefully complex—but this is not required for our experimental design to be valid. In addition, given the randomized selection of subjects into treatment status, the distribution of subject understanding should be balanced across treatment status. Even if there is variation in subjects understanding the game and the underlying logic, the enumerators directly tell subjects how to implement the advice and there is near complete accuracy in their belief of whether that advice is coming from a female or a male. Indeed, contexts in which there is a lack of clarity on the reasoning behind advice are common and relevant to understanding what conditions change the likelihood of following advice.

2.2.5 Validity of randomization

Subjects were assigned a treatment once they arrived for the experiment. The randomization was stratified by subject gender.[13]

Table III confirms the validity of our randomization by comparing subject characteristics across treatment arms. Using information on the subjects provided by the human resources

[13]We generated a random ordering for treatment assignments to be assigned as subjects arrived. For the last two days of the experiment, we re-randomized using a blocked randomization in groups of four, due to concerns of not meeting our recruitment targets (although we were ultimately successful in meeting the target). In all analyses, we account for differing randomization probabilities using inverse probability weights. Analyses based on the exclusion of weights do not statistically differ from the main results reported in the paper.

Table III: Randomization balance

	(1) Fem. subject	(2) ln(Salary)	(3) Level	(4) Years Ed.	(5) MA or higher	(6) Job tenure
Female leader only (F)	0.0173	-0.0213	-0.145	0.00175	0.00848	238.2
	(0.0817)	(0.0634)	(0.446)	(0.0813)	(0.0401)	(328.3)
Ability signal only (A)	-0.0189	-0.00813	0.151	0.0556	0.0354	71.63
	(0.0803)	(0.0597)	(0.424)	(0.0865)	(0.0427)	(335.7)
Female leader × Ability (FA)	-0.0383	-0.00636	-0.149	0.117	0.0587	-276.9
	(0.0840)	(0.0610)	(0.420)	(0.100)	(0.0494)	(342.2)
Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	304	304	304	304	304	304
p-val: F = A	0.649	0.839	0.510	0.535	0.535	0.586
p-val: A = FA	0.812	0.977	0.481	0.554	0.650	0.268
p-val: F = FA	0.503	0.821	0.994	0.251	0.312	0.0959
Sample Mean	0.484	8.092	13.45	16.17	0.0822	3020.7

Robust standard errors in parentheses. All dependent variables refer to subject characteristics taken from institutional data. Fem. subject is an indicator for the being female, ln(Salary) is the log of annual salary, Level refers to internal categorization of the seniority and skill of a position, Years Ed. is the number of years of education reported, MA or higher is an indicator of whether the subject holds a Masters degree or higher, and Job tenure is the number of days of employment with the university. Day FE are fixed effects referring to the day the subject participated in the experiment. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

department, we confirm that subject characteristics are balanced across the four treatment groups. We also confirm pairwise balance in the bottom three rows of Table III.

In addition to the subjects' characteristics, we may be concerned that the pseudonyms we used to connote gender also contained information on other important characteristics (e.g., ethnicity, age). In Ethiopia, there are significant differences in ethnicity (Amhara and Oromic are the two dominant ethnicities) and religion (Orthodox Christianity and Islam are dominant). The pseudonyms assigned to leaders were selected from a listing exercise conducted for another study in an Amharic region of Ethiopia (Ahmed and Mcintosh, 2017).[14] We use 193 unique names and no name is used for more than five subjects to reduce the

[14]We therefore oversample Oromic names in our selection.

Table IV: Pseudonym balance

	(1) Amhara	(2) Oromo	(3) Age	(4) Grade	(5) Orthodox
Female leader only (F)	-0.0188	-0.00914	0.670	0.219	-0.0220
	(0.0554)	(0.0708)	(2.365)	(0.263)	(0.0700)
Ability signal only (A)	-0.0537	-0.0104	-0.932	0.145	-0.0689
	(0.0568)	(0.0697)	(2.278)	(0.227)	(0.0665)
Female leader × Ability (FA)	-0.0265	0.00721	-0.409	0.160	-0.0477
	(0.0597)	(0.0754)	(2.517)	(0.270)	(0.0712)
Day FE	Yes	Yes	Yes	Yes	Yes
Observations	304	304	304	304	304
p-val: F = A	0.544	0.985	0.444	0.781	0.466
p-val: A = FA	0.658	0.807	0.816	0.956	0.743
p-val: F = FA	0.900	0.826	0.648	0.848	0.700

Robust standard errors in parentheses. Pseudonym characteristics are assigned based on the characteristics of actual individuals with a given name, drawn from a listing exercise conducted for another study in Ethiopia. The ethnicities, Amhara and Oromo, and religion, Orthodox Christian, are equal to 1 if there was at least one individual with the relevant characteristic. Age and grade represent the average age and educational attainment of all individuals with a given name. Day FE are fixed effects referring to the day the subject participated in the experiment. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table V: Leader "error" balance

	(1) Error	(2) Error
Female leader only (F)	0.00943	0.00643
	(0.0187)	(0.0174)
Ability signal only (A)	0.00202	-0.00126
	(0.0182)	(0.0162)
Female leader × Ability (FA)	-0.0118	-0.00627
	(0.0187)	(0.0186)
Day FE	Yes	Yes
Round FE	Yes	Yes
Play FE	No	Yes
Observations	3344	3339
p-val: F = A	0.681	0.620
p-val: A = FA	0.443	0.771
p-val: F = FA	0.252	0.483

Standard errors in parentheses, clustered at subject level. Error is an indicator of whether the computer played "IN" in response to the subject playing strategically (i.e., 4 or 5) or if the Computer played "OUT" in response to the subject playing 2 or 3. Day FE are fixed effects referring to the day the subject participated in the experiment. Round FE are fixed effects referring to the ten rounds of the game. Play FE are fixed effects referring to the number played by the subject. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

concern of characteristics associated with a name being correlated with treatment status. The listing exercise had also collected information on the following basic demographic information on characteristics of the person with the given name: ethnicity, religion, age, and grade completed. Table IV confirms that the characteristics associated with the pseudonym assigned to each subject are balanced across treatment arms.

A final concern is that due to the randomized responses by the computer, leader ability could appear different across treatments despite holding leader behavior constant. Subjects may perceive their leader as less able if they do not follow their leader's advice and happen to obtain a higher payoff in a given round than the leader, or if they follow their leader's advice but happen to receive a low payoff. Table V shows that these "errors" are balanced

across treatments both unconditionally (Column 1) and conditional on the subject's play (Column 2), confirming that differential error rates are not driving our results.

2.2.6 Estimating Equations

Our primary research question is whether subjects differentially adhere to women versus men in leadership positions. Thus, we estimate whether subjects were less likely to follow their leader's advice to play strategically in the Leadership Game as a function of their leader's (randomly assigned) gender. We additionally estimate whether information indicating the leader is of high ability will have higher returns to female leaders and mitigate such gender gaps. We do so by estimating the following equation using a linear probability model:

$$R_{ir} = \alpha + \beta_1 * FL_i + \beta_2 * Ability_i + \beta_3 FL * Ability_i + \epsilon_{ir} \quad (1)$$

where R is an indicator for playing strategically (defined as playing 4 or 5) for subject i in round r (of 10 rounds).[15] FL is an indicator for being randomly assigned a female leader, *Ability* is an indicator for being randomly assigned to information about the leader's high ability, and $FL * Ability$ is the interaction of the two indicators. We additionally include an indicator of whether the individual chose to play strategically in their practice round selection, subject characteristics listed in Table III, day fixed effects (i.e., the six days of the experiment), and round fixed effects (i.e., the 10 rounds of the game) to increase precision of our estimates and to directly control for changes we made on the latter days of the experiment. Standard errors are clustered at the individual level, corresponding to the level of randomization (Bertrand and Mullainathan, 2004; McKenzie, 2012).

β_1 is the difference in following the advice provided by a female leader, relative to a male leader, when no information on high ability is provided. β_2 is the average marginal effect of

[15]We use an indicator for playing 4 or 5 based on our pre-specified outcome of interest in our pre-analysis plan, following the earlier work of Cooper and Kagel (2005). Results when using an indicator for selecting 5 only as the dependent variable are reported in the Appendix, and do not statistically differ from the main results reported in the paper.

providing the high ability signal for male leaders. β_3 is the differential return to the signal of high ability for female leaders (i.e., the relative difference in the change in following advice for female leaders relative to male leaders as a response to the high ability signal). If β_3 is positive, this indicates that the returns to the high ability signal are larger for female leaders. This would suggest that a signal of high ability can mitigate discrimination against female leaders. An additional parameter of interest is $\beta_1 + \beta_3$, the gender gap in following the leader conditional on receiving a signal of high ability. If $\beta_1 + \beta_3 > 0$ and $\beta_1 < 0$, this represents a reversal of the gender gap.

3 Leadership Game Results

Table VI shows results for our primary outcome: whether subjects follow the leader's advice and play strategically in the Leadership Game. The coefficients correspond to those in estimating equation (1). We show results for the first round of the Leadership Game (Column 1), the first half of the Leadership Game (Column 2), and for all rounds of the Leadership Game (Column 3).

Discrimination from below, no ability information

We find that in the absence of information on ability, subjects were 6 percentage points less likely to follow the advice of a female leader (Column 3, β_1). The coefficient estimate on β_1 is remarkably stable across rounds. The magnitude of this effect represents a 10 to 13 percent reduction in adherence relative to a male leader's recommendation.

This discrimination against female leaders is costly. For subjects who did not receive the ability signal, having a female leader reduced total points earned by .34 standard deviations, which is statistically significant at the 5 percent level.

Table VI: Results: Following the Leader's Advice

Dependent Variable:	Strategic Play		
	(1)	(2)	(3)
	Round 1	Rounds 1-5	All Rounds
(β_1) Fem. Leader	-0.0502	-0.0822**	-0.0604*
	(0.0810)	(0.0391)	(0.0344)
(β_2) Ability	-0.0361	-0.0443	-0.00234
	(0.0783)	(0.0393)	(0.0343)
(β_3) Fem. leader × Ability	0.295***	0.154***	0.123***
	(0.112)	(0.0542)	(0.0472)
Covariates	X	X	X
Day FE	X	X	X
Round FE		X	X
Practice round	X	X	X
Observations	301	1505	3010
Control group mean	0.479	0.614	0.618
$\beta_1 + \beta_3$	0.245***	0.0722*	0.0624*
P-val.: $\beta_1 + \beta_3$	0.00153	0.0571	0.0569

Standard errors in parentheses, clustered at subject level. Strategic play is defined as playing 4 or 5. Practice Round is an indicator for whether the subject played strategically in a practice round prior to any advice from the leader. Covariates are subject's gender, ln(salary), level of employment, years of education, an indicator for having a masters degree, and tenure. Day FE are fixed effects referring to the day the subject participated in the experiment. Round FE are fixed effects referring to the ten rounds of the game. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Returns to ability information

Providing information on the leader's ability substantially increases adherence to advice from female leaders. The coefficient β_3 is large and significant, which means that the return to the signal of high ability is higher for female leaders than for male leaders. In the first round of the game, the signal of high ability differentially increased the likelihood of the female leader's advice being followed by 29.5 percentage points—a 69 percent increase. Across all rounds of the game, the differential return to the ability signal for female leaders was 12.3 percentage points (a 22 percent increase). The larger coefficient in the first round of the game (Column 1) may reflect the fact that subjects would have been most uncertain about the quality of the advice in earlier rounds. This large return to ability signals for female leaders (β_3) diminishes as the game proceeds, suggesting that the signal becomes less important as subjects observe objective evidence that the leader is providing good advice.[16] Thus, the signal of ability significantly increased the effectiveness of female leadership.

In contrast, information on ability did not change subjects' likelihood of following advice from male leaders (β_2). This suggests that subjects expected men to be of high ability, and so the signal did not substantially change subjects' beliefs about the ability of their male leader.

Discrimination from below, with ability information

The last two rows of Table VI test for discrimination from below among subjects who received the high ability signal. Interestingly, conditional on being given a signal of their leader being of high ability, subjects are *more* likely to follow the directions provided by female leaders ($\beta_1 + \beta_3 > 0$). The gender gap ranges from 24.5 percentage points in the first round (Column 1) to 6.2 percentage points across all rounds (Column 3).

[16] We do not present later rounds in isolation because early round decisions influence later round decisions, and early decisions are a function of treatment status. Using later rounds alone as a dependent variable thus raises concerns about endogeneity.

Robustness

Our results are robust to a number of changes in the specification: results do not statistically differ when excluding covariates, using a probit model, redefining the dependent variable as selecting 5 only, and excluding any round (see Appendix A.5 and Appendix A.6). Our results are also robust to determining statistical significance using randomization inference.[17] We estimate our results separately for male and female subjects in Appendix Table A.7. The results are similar across subject genders.[18] If anything, the reversal of discrimination may be somewhat stronger among female subjects.

We also provide results on subject expectations. We measured subjects' beliefs on how well their leader would perform in the Leadership Game prior to playing. We estimate our main estimating equation, Equation 1, with beliefs as a dependent variable in Table VII. The magnitudes and signs of the coefficients align with the main results on following leader advice (Table VI). Female leaders are expected to perform worse than male leaders when no information is provided on ability: their expected performance is 7.43 fewer points. However, the ability signal to female leaders increases expected performance by 18.46 points. Thus, when leaders are presented as high-ability, subjects expect female leaders' to earn 11.04 *more* points than male leaders. Our estimates have large standard errors—as noted earlier, the responses to the belief elicitation exercise suggest that it was difficult for subjects to understand. Despite such measurement error, the pattern in beliefs mirrors what we observe in the Leadership Game.

[17] Using 1,000 draws, the p-value is similar to our primary specification. For β_1 the p-value for the one-sided test is 0.04 and two-sided is 0.082. For β_3, the p-value is 0.009 for the one-sided test and 0.015 for the two-sided test. For $\beta_1 + \beta_3$, the p-value is 0.04 for the one-sided test and 0.09 for the two-sided test.

[18] Estimating a single model that interacts the subject's gender with treatment also does not yield statistical differences by subject gender.

Table VII: Beliefs about leaders

Dependent Variable:	Beliefs on leader's performance
	(1)
(β_1) Fem. Leader	-7.425
	(9.051)
(β_2) Ability	4.064
	(9.381)
(β_3) Fem. leader × Ability	18.46
	(13.30)
Covariates	X
Day FE	X
Observations	300

Robust standard errors in parentheses. Dependent variable refers to the expected points earned in the Leadership Game by the leader, based on the subject's reported probability of the leader receiving each possible outcome. Covariates are subject's gender, ln(salary), level of employment, years of education, an indicator for having a masters degree, and tenure. Day FE are fixed effects referring to the day the subject participated in the experiment. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4 Supporting Evidence: Resume Evaluation Experiment

We provide additional evidence of gender discrimination toward management positions from a resume evaluation experiment that we implemented the week after the Leadership Game. We provided subjects with a job description for a senior management position, then asked subjects to evaluate a hypothetical candidate for that position. The gender of that candidate was randomly determined. Though it is hypothetical, this exercise complements the Leadership Game as a more closely linked exercise to a real-world labor market decision (i.e., evaluation of a job candidate).

4.1 Resume Evaluation Design

Respondents were asked to review a complete resume for a senior management position. It is customary to note the gender of the candidate on resumes in Ethiopia; therefore, names were not used and the gender was listed directly on the resume. An example is shown in Figure

IV. To ensure the salience of candidate gender, we implemented a "comprehension" test before asking subjects to evaluate the resume. The test asked subjects a series of questions about the resume, include candidate gender. 95 percent of subjects correctly identified the candidate's gender. Subjects were randomly assigned one of two possible resumes that were designed to be comparable in quality. To guard against social desirability bias, we compare evaluations across subjects only; that is, in the analysis sample, subjects are not directly comparing a male and a female candidate.[19]

After reviewing the resume and completing the comprehension questions, subjects evaluated the potential candidate on an increasing Likert scale from 1 to 5 on competence, likeability, and willingness to hire. They additionally suggested a salary to be offered to the candidate.[20]

Because of uncertainty in scheduling survey interviews with subjects, we again randomized the treatment assignment by creating a random ordering in groups of four (two resume versions * two candidate gender) for each enumerator, and then had the enumerator provide the resume indicated by the ordering of the list when interviewing subjects.[21] We success-

[19]In the experiment, after evaluating the resume, subjects were given a second resume of the opposite gender and asked to compare the two candidates directly. Our original analysis plan specified comparing evaluations within subjects, but we find evidence that providing a second resume to our subjects revealed that gender was a key component of interest, and subjects responded accordingly. Averaging across all subjects, we find that relative to the first resume, the second resume was rated more positively if it was a female candidate and more negatively if it was a male candidate. The estimations testing for social desirability bias, along with estimation specified in the preanalysis plan, are shown in Appendix Table A.8. Importantly, when subjects were given the initial resume to evaluate, they were not told that a second resume would follow. In addition, even if subjects had known beforehand that the purpose of the resume evaluation was gender, the results from the second resume suggest that social desirability bias would have resulted in female resumes being evaluated more positively, causing our estimates to be a lower bound of gender discrimination against women.

[20]The exact questions were as follows: 1."I will first ask you about the competency of the candidate. By competency, I mean for you to evaluate the candidate based on how well you think he will perform on the requirements of the job. Based on the resume, is his competency: poor, fair, good, very good, or excellent?" 2. "I will now ask you about the likeability of the candidate. By likeability, I mean for you to evaluate the candidate based on how well you think he will get along with his colleagues, including the employees he will directly supervise. Based on the resume, is his likeability: poor, fair, good, very good, or excellent?" 3. "I will now ask you about how willing you would be to hire the candidate for the position. Based on the resume, would you be very unwilling, slightly unwilling, neither unwilling or willing, slightly willing, or very willing to hire him?" 4. "If this job candidate were hired, what monthly salary would you offer him, in Ethiopian birr?"

[21]We find 6 subjects for which the assigned treatment resume differs from the enumerator's recorded resume for the subject. All analysis uses assigned treatment resume.

I. Personal Information

Name: ----------
Sex: [Randomly Determined: Female/Male]
Birthdate: 21/07/1984

Personal Summary:

I am an outgoing, ambitious, and confident individual, whose passion for the HR sector is equally matched by my experience in it. For the previous 6 years, my primary role at ----- has been to provide HR support, guidance, advice, and services to all company staff. This has taught me to translate corporate goals into human resource development programs, as well as given me extensive knowledge of HR administration, principles, practices, and laws. I have experience sourcing candidates, overseeing hiring processes, and resolving employee relations issues. This has given me experience interacting with many different types of people and I have developed strong interpersonal skills for resolving conflicts. I am always looking for ways to improve systems in human resources, consistently complete tasks to their natural end, work well under pressure and deadlines, and adapt to changing environments.

II. Work Experience

Title: Employee and Labor Relations Consultant in Human Resources
Period of employment: 2010 - Present

Figure IV: Resume Evaluation Experiment: Example Resume

fully followed up with 74 percent of the experimental subjects who complete the resume evaluation component in its entirety.[22,23] Table VIII confirms the validity of our randomization by documenting that subject characteristics were balanced across the randomly assigned candidate gender.

The resume evaluation provides an additional test of gender discrimination towards po-

[22]An additional 12.8 percent also participated in the resume evaluation, but chose to not respond to at least one of the evaluation questions, primarily the salary offer. We observe the same pattern for the difference in evaluation of a female resume on the remaining evaluation questions for which these subjects do provide a response. Attrition was not due to lack of consent or desire to participate, but rather driven by the difficulty in finding the same subjects by the enumerators. Because we implemented the survey over the summer, many employees were on leave. In general, subjects we were successful in following up with were paid less and had lower level positions in university. We do not observe differences in the lab experiment results when limiting the sample to those who completed the resume experiment.

[23]Prior to arrival in Ethiopia, we expected to implement the resume evaluation with 600 subjects. However, due to difficulties in recruitment and implementation by enumerators, we decided to limit the resume evaluation to just those subjects that participated in the experimental game. This decision was made prior to any data collection for the resume evaluation, and no other subjects were asked to evaluate the resumes.

Table VIII: Resume Experiment Balance

	(1) Fem. subject	(2) ln(Salary)	(3) Level	(4) Years Ed.	(5) MA or higher	(6) Job tenure
Female Resume	0.0213	-0.0256	-0.181	0.0189	0.00517	412.0
	(0.0671)	(0.0493)	(0.355)	(0.0712)	(0.0354)	(264.9)
Observations	225	225	225	225	225	225

Robust standard errors in parentheses. Female Resume is an indicator for whether the subject reviewed a resume for a female candidate. Regressions additionally control for Resume Type, which refers to which of two resume versions the subject reviewed. All dependent variables refer to subject characteristics taken from institutional data. Fem. subject is an indicator for the being female, ln(Salary) is the log of annual salary, Level refers to internal categorization of the seniority and skill of a position, Years Ed. is the number of years of education reported, MA or higher is an indicator of whether the subject holds a Masters degree or higher, and Job tenure is the number of days of employment with the university. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

tential managers. We test for this using the following linear regression model:

$$Outcome_i = \alpha + \gamma_1 * FC_i + \epsilon_i \tag{2}$$

where *Outcome* is competence, likeability, hireability, or salary offer (in logs); FC is an indicator of whether the resume was randomly assigned to be a female candidate; and i represents subject. The coefficient of interest is γ_1, the difference in how subjects evaluated female candidates relative to male candidates. We additionally include *ResumeType* as a control for which of the two "candidate" resume was given, and use robust standard errors.

4.2 Resume Evaluation Results

Table IX shows the effect of candidate gender on resume evaluation scores. On all measures, female candidates were evaluated more poorly than male candidates. Female candidates were rated less competent, less likeable and less likely to be hired. They were offered a 12 percent lower salary. Only the latter result is statistically significant, though the pattern of lower

Table IX: Resume Evaluation Results

	(1) Competence	(2) Likeability	(3) Likelihood of Hire	(4) Log Salary Offer
Female Resume	-0.0933	-0.0337	-0.172	-0.115**
	(0.122)	(0.111)	(0.140)	(0.0534)
Observations	225	225	225	225

Robust standard errors in parentheses. Competence, Likeability, and Likelihood to Hire were asked using a Likert Scale, increasing from 1 to 5. Log Salary Offer is the log of the salary the subject suggested as an offer to the candidate in Birr. Female Resume is an indicator for the resume belonging to a randomly assigned female candidate. Regression specifications include the resume version, and subject's gender, ln(salary), level of employment, years of education, an indicator for having a masters degree, and tenure as covariates. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

evaluation is consistent across all measures.[24] These results provide additional evidence that employees discriminate based on gender in evaluating those in senior leadership positions. The results are consistent with the gender discrimination we observe in the Leadership Game in the absence of high ability information.[25]

This exercise differs from typical correspondence studies in that our sample is not involved with human resources or hiring decisions. We interpret our results as suggestive survey evidence on how subjects may generally view managers. One possible interpretation of the results is that differences in salary offered reflect differences in expectations of the candidate's outside option. However, there is no gender wage gap among staff who hold advanced degrees at the university, suggesting that the results are more consistent with gender discrimination.

5 Discussion

Our results provide evidence of discrimination towards leaders based on their gender. In addition, we show that providing information about leaders' high ability can reverse dis-

[24] 39 additional subjects participated in the resume experiment, but did not provide a salary estimate for the resume they reviewed. Among these subjects, competency and likelihood of hiring are rated lower for women, but likeability is rated higher.

[25] We do not observe statistically significant differences by subject gender (see Appendix).

crimination in favor of women. These results raise the question of *why* we observe this discrimination. Existing literature largely categorizes discrimination into either "taste-based" discrimination based on preferences, or "statistical" discrimination based on beliefs. In Appendix E, we provide a conceptual framework exploring the implications of taste-based and statistical discrimination in our context. We summarize these implications here.

In our context, we can conceive of taste-based discrimination as a fixed utility cost of obeying a female leader, regardless of her ability. This utility cost might reflect a general dislike of taking directions from women, or an unwillingness to follow female leaders' advice due to gender norms that female leadership is inappropriate. Taste-based discrimination is costly to the subject, because they are less likely to obey a woman despite facing the same expected payoff from following a woman or a man.

In contrast, we can think of statistical discrimination as any difference in a subject's expected payoff based on leader gender. This might be due to correct inferences based on subject gender; for example, it may be that women face barriers to education and thus are less skilled on average. It could also be due to incorrect beliefs about the competence of different genders on average. Under statistical discrimination, subjects expect that a male or female leader is providing lower quality advice, and so they would be less likely to heed the leader's directions simply because the expected payoff to doing so is lower. If these beliefs are correct (i.e., there are true differences in ability between men and women), statistical discrimination could improve subjects' payoffs because leader gender indeed provides relevant information.

In the absence of ability information, both taste-based and statistical discrimination are consistent with reduced adherence to female leadership. However, our finding that gender discrimination *reverses* conditional on providing information about the leader's high ability suggests a more important role for statistical discrimination. If workers preferred to disobey women regardless of their ability, then we should expect that gender gaps diminish with information about high ability, but never reverse. That is, if subjects simply disliked following

women's advice, then they should always be weakly less likely to follow female leaders. In contrast, this reversal can occur if, conditional on knowing that the leader is of high ability, leader gender affects subjects' beliefs about their expected payoff. Indeed, the results on beliefs about the leader in Table VII are consistent with statistical discrimination in that subjects' beliefs about the leader's performance differ by the leader's gender. This is consistent with the idea that leader gender changes expected payoffs.

Moreover, the finding that the ability treatment reverses discrimination suggests that the same signal results in different inferences about men's versus women's underlying ability. There could be many reasons to believe that the same signal has different implications for men's versus women's underlying ability. One example where a reversal of gender discrimination can occur is if subjects believe that women are less able on average, but that female ability has an entirely different distribution in which the variance of female ability is much larger. Then, the same signal can imply higher ability for a female leader than a male leader because the female leader must be more of an outlier in her potential ability distribution.

Another important example of signals having different implications by gender is when subjects believe that the ability signal is more difficult for women to obtain. Bohren, Imas and Rosenberg (2019) provide evidence that gender discrimination in obtaining a signal can result in the signal's interpretation differing by gender. Indeed, an ability signal would have different implications by gender if there is any barrier to obtaining such a signal that differentially affects women. For example, if women are less likely to obtain advanced degrees, then conditional on observing an advanced degree, a person may reasonably infer higher ability for a woman with an advanced degree than when observing the same advanced degree for a man.

Note that such statistical discrimination can occur because gender norms or societal preferences drive the initial barrier to obtaining the signal (e.g., women have less access to education due to gender norms). Indeed, in equilibrium, taste-based discrimination and statistical discrimination may be mutually reinforcing. Taste-based discrimination can generate

average differences in male and female ability. As a result, statistical discrimination may emerge that, in turn, reinforces gender norms around occupational specialization.[26]

Discrimination from below is a potential explanation for the under-representation of women in senior management in developing countries. An implication of our findings is that discrimination from below can generate disparate promotion probabilities for men versus women even when an employer is unbiased. The pattern of discrimination we observe suggests that women who are promoted would be positively selected. If discrimination from below reduces the performance of teams led by women in more junior levels of leadership, women are less likely to achieve the performance standard required for promotion. Women who exceed the standard despite discrimination are then more likely to be qualified than their male counterparts.[27]

Thus, discrimination from below can generate both under-representation of women in leadership positions, and positive selection of female leaders in more senior positions. Conditional on obtaining a high enough senior position, female leaders may then see a reduction or even a reversal in discrimination from below. This insight can help reconcile, for example, the large gender disparities for the median woman in South Asia with the fact that the four largest South Asian countries have all had a female head of government.[28] In addition to highlighting the importance of conducting studies on discrimination in various settings, our findings help reconcile why discrimination and gender inequities on average may not translate to similar patterns of inequities among the elite.

Discrimination from below can also result in a self-reinforcing belief that women are less qualified to hold leadership positions. That is, if subordinates believe that women are less effective leaders and therefore do not follow their leadership, this will make female leadership less effective. This is turn justifies the initial belief. Such discrimination from below suggests

[26]We thank an anonymous referee for highlighting this point.

[27]This implication follows from the seminal model of Coate and Loury (1993), in which an employer maximizes her payoff by setting a minimum standard and promoting those who exceed the minimum standard.

[28]Sen, Amartya. "More Than 100 Million Women Are Missing." The New York Review of Books, December 20, 1990.

that even if a woman alters her leadership style or increases her human capital, she may still fall short of her male counterparts. In addition, if subordinates discount valuable advice from female leaders, it may lower their own welfare.

Our results suggest that signals of ability, such as certifications and credentials, may mitigate discrimination from below and disrupt self-reinforcing beliefs about female leaders. Indeed, such ability signals have been important in mitigating gender and racial gaps in other contexts. Key examples can be found in the literature on gender and racial wage gaps in the United States. For example, Dougherty (2005) and Arcidiacono, Bayer and Hizmo (2010) find that higher education degrees reduce gender and racial wage gaps respectively. Blair and Chung (2020) show that occupational licenses, particularly licenses with permanent felony bans, have higher wage returns for black men relative to white men. Tray (1982) shows that veteran status functions as an ability signal, and black veterans experience a higher wage premium than white veterans. The value of such real-world ability signals in developing countries is an important area for future research.

6 Conclusion

This paper studies how leader gender influences the way individuals respond to leadership. We find evidence for gender discrimination in the decision to follow directions from leaders. While we use a leadership framing, our results highlight discriminatory concerns in advice-giving contexts more generally. Discrimination from below can generate gender disparities in any position in which successful performance requires individuals to follow one's advice or direction. Our results further raise concerns about how best to evaluate leaders and highlight a tension between gender equity and successful performance. Performance metrics based on subordinate or client responsiveness may be problematic in reaching equity goals. It also suggests that simply providing opportunities to "sit at the table" may not be sufficient to overcome gender disparities.

We find that a credible signal of high ability has significantly larger returns for female leaders than for male leaders. When we do not provide information about leader ability, we observe discrimination against female leaders. However, the gender gap in following the leader is not only mitigated, but reversed, when the leader is presented as highly trained and competent. Thus, a simple informational intervention describing leader ability is sufficient to eliminate discrimination against female leaders. Our results imply that subjects are using gender as a proxy for quality of the advice, suggesting that statistical discrimination plays a role in subjects' decision to follow the leader's advice.

It is worth noting that our analysis is focused on the real-stakes outcome of following the leader's advice, and we do not collect subjects' evaluation of their leaders on other characteristics. In future research, it would be worthwhile to study how gender discrimination and ability signals affect subjective evaluations of leaders.

Our results suggest that providing women with credible signals of their ability may improve their performance by reducing discrimination from below. It also follows that gender equity efforts should not be limited to only those who hire and evaluate employees; changing the beliefs of *all* employees with respect to manager gender is important for improving gender equity. In future research, it would be useful to identify and test methods of communicating ability information to a broad audience.

References

African Development Bank. 2015. "Where are the women: inclusive boardrooms in Africa's top listed companies?" https://www.afdb.org/en/documents/document/where-are-the-women-inclusive-boardrooms-in-africas-top-listed-companies-53810/.

Ahmed, Shukri, and Craig Mcintosh. 2017. "The Impact of Commercial Rainfall Index Insurance: Experimental Evidence From Ethiopia." https://gps.ucsd.edu/_files/faculty/mcintosh/mcintosh_paper_ams-ethiopia-impact.pdf.

Arcidiacono, Peter, Patrick Bayer, and Aurel Hizmo. 2010. "Beyond Signaling and Human Capital: Education and the Revelation of Ability." *American Economic Journal: Applied Economics*, 2(4): 76–104.

Becker, Gary Stanley. 1957. *The economics of discrimination.* Chicago:Univ. of Chicago Press.

BenYishay, Ariel, Maria Jones, Florence Kondylis, and Ahmed Mushfiq Mobarak. 2020. "Gender Gaps in Technology Diffusion." *Journal of Development Economics*, 143.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, 94(4): 991–1013.

Blair, Peter Q, and Bobby W Chung. 2020. "Job Market Signaling through Occupational Licensing." *NBER Working Paper Series*, 24791.

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg. 2019. "The Dynamics of Discrimination: Theory and Evidence." *American Economic Review*, 109.

Boring, Anne. 2017. "Gender biases in student evaluations of teaching." *Journal of Public Economics*, 145: 27–41.

Coate, Stephen, and Glenn C. Loury. 1993. "Will Affirmative-Action Policies Eliminate Negative Stereotypes?" *American Economic Review*, 83(5): 1220–1240.

Cooper, David J., and John H. Kagel. 2005. "Are two heads better than one? Team versus individual play in signaling games." *American Economic Review*, 95(3): 477–509.

Dahl, Gordon B, Andreas Kotsadam, and Dan-Olof Rooth. 2020. "Does Integration Change Gender Attitudes? The Effect of Randomly Assigning Women to Traditionally Male Teams*." *The Quarterly Journal of Economics.*

Dougherty, Christopher. 2005. "Why Are the Returns to Schooling Higher for Women than for Men?" *The Journal of Human Resources*, 40(4): 969–988.

Duflo, Esther. 2012. "Women Empowerment and Economic Development." *Journal of Economic Literature*, 50(4): 1051–1079.

Eagly, Alice H. 2013. "Women as Leaders: Leadership Style Versus Leaders' Values and Attitudes." In *Gender and work: Challenging conventional wisdom*. Harvard Business School Press. https://www.hbs.edu/faculty/conferences/2013-w50-research-symposium/Documents/eagly.pdf.

Egan, Mark L, Gregor Matvos, and Amit Seru. 2017. "When Harry Fired Sally: The Double Standard in Punishing Misconduct." *NBER Working Paper Series*, 23242.

Gangadharan, Lata, Tarun Jain, Pushkar Maitra, and Joseph Vecci. 2016. "Social identity and governance: The behavioral response to female leaders." *European Economic Review*, 90: 302–325.

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 and Organization*, 161.

Hardy, Morgan, and Gisella Kagy. 2018. "It's Getting Crowded in Here: Experimental Evidence of Demand Constraints." https://www.dropbox.com/s/kdz1or4r0404k9w.

Jayachandran, Seema. 2015. "The Roots of Gender Inequality in Developing Countries." *Annual Review of Economics*, 7(1): 63–88.

Landsman, Rachel. 2018. "Gender Differences in Executive Departure." https://drive.google.com/file/d/0B4Gus2bxOyznZXM3TVlfTzZ2TWM/view.

Lowe, Matthew. 2020. "Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration." https://osf.io/pxsj4/.

Macchiavello, Rocco, Andreas Menzel, Atonu Rabbani, and Christopher Woodruff. 2015. "Challenges of Change: An Experiment Training Women to Manage in the Bangladeshi Garment Sector." *Centre for Competitive Advantage in the Global Economy, University of Warwick Working Paper Series*, 256.

Manian, Shanthi, and Ketki Sheth. 2020. "Follow my Lead: Assertive Cheap Talk and the Gender Gap." https://drive.google.com/file/d/1l_gkQwmNXaIF3Iqq0FRnr5qgktYYWeWt/view?usp=sharing.

McKenzie, David. 2012. "Beyond baseline and follow-up: The case for more T in experiments." *Journal of Development Economics*, 99(2): 210–221.

Mengel, Friederike, Jan Sauermann, and Ulf Zölitz. 2019. "Gender Bias in Teaching Evaluations." *Journal of the European Economic Association*, 17(2).

Mousa, Salma. 2020. "Building social cohesion between Christians and Muslims through soccer in post-ISIS Iraq." *Science*, 369(6505).

Niederle, Muriel. 2017. "Gender." In *The Handbook of Experimental Economics*. Vol. 2. Princeton University Press.

Paluck, Elizabeth Levy, Seth A. Green, and Donald P. Green. 2019. "The contact hypothesis re-evaluated." *Behavioural Public Policy*, 3(2): 129–158.

Sarsons, Heather. 2017. "Interpreting Signals in the Labor Market: Evidence from Medical Referrals." https://drive.google.com/file/d/1bDV1Tqhl6SX2CtM6Sf1c95PF5eloDJtr/view.

Tray, Dennis De. 1982. "Veteran Status as a Screening Device." *The American Economic Review*, 72(1): 133–142.

World Bank. 2017. "World Development Indicators." https://datacatalog.worldbank.org/dataset/world-development-indicators.

For Online Publication

A Tower of Hanoi

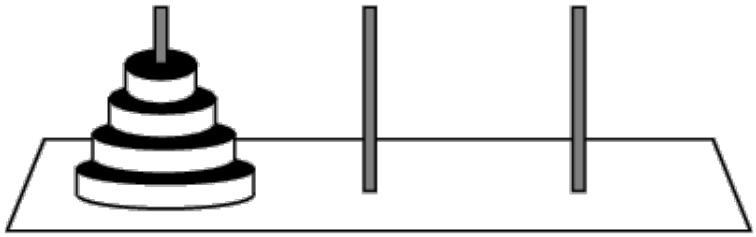

Figure A.1: Tower of Hanoi

Subjects are asked to move the tower from one pole to another. They can only move one disk at a time, and a larger disk cannot be placed on a smaller disk. The subject is asked to solve the Tower using four disks and told that the minimum moves are 15.

B Subject Compensation Schedule

Enumerator ID_____ Subject Number_____

Payout Schedules Provided to Subject:

Payout Schedule for Game 1: (*Show each of these as different tables at the relevant time*.)

Number of Moves – Number of Guessed Moves			Number of Moves to Solve	
0	$1.7		15	$2.00
1	$1.65		16	$1.94
2	$1.6		17	$1.88
3	$1.55		18	$1.82
4	$1.5		19	$1.76
5	$1.45		20	$1.70
6	$1.4		21	$1.64
7	$1.35		22	$1.58
8	$1.3		23	$1.52
9	$1.25		24	$1.46
10	$1.2		25	$1.40
11	$1.15		26	$1.34
12	$1.1		27	$1.28
13	$1.05		28	$1.22
14 or more, or failed to solve the puzzle.	$1		29 or more, or failed to solve the puzzle.	$1.16

Payout Schedule for Game 2:

Type A				Type B		
A's choice	Computer: In	Computer: Out		B's choice	Computer: In	Computer: Out
1	168	444		1	276	568
2	150	426		2	330	606
3	132	408		3	352	628
4	56	182		4	334	610
5	-188	-38		5	316	592

Conversion rate: 100 Points = 1 USD (e.g., 568 = 5.68)

The computer makes its decisions to try to get the maximum points possible. The computer receives points in the following way:

Computer Decides:	Type A	Type B
In	500	200
Out	250	250

Figure A.2: Subject Compensation Schedule

C Messages Sent by Leaders

- Round 3: When I play 5, the Computer guesses I am Type B and so plays Out.
- Round 4: When I play 5, the Computer guesses I am Type B and so plays Out. Remember, my payment is based on how well you play the game - Trust me, you and I will both make more if you play 5.
- Rounds 5 and 6: Remember, the computer wants to play In when it thinks I'm Type A and Out when it thinks I'm Type B. But I want the computer to play Out. So I need to make the computer think I am Type B.
- Round 7: Remember, the computer wants to play In when it thinks I'm Type A and Out when it thinks I'm Type B. But I want the computer to play Out. So I need to make the computer think I am Type B. When I play 5, the computer thinks I must be Type B, because Type A is always better off on another number even if the Computer chooses In.
- Round 8: Remember, the computer wants to play In when it thinks I'm Type A and Out when it thinks I'm Type B. But I want the computer to play Out. So I need to make the computer think I am Type B. When I play 5, the computer thinks I must be Type B, because Type A is always better off on another number even if the Computer chooses In.This is why I want you to Play 5, so we can both earn more.
- Rounds 9 and 10: Remember, the computer wants to play In when it thinks I'm Type A and Out when it thinks I'm Type B. But I want the computer to play Out. So I need to make the computer think I am Type B. When I play 5, the computer thinks I must be Type B, because Type A is always better off on another number even if the Computer chooses In. If I play 3, then the Computer cannot tell if I am A or B and so will assume half the time it is better to Play In - that means that on average, I earn less when Playing 3 because half the time I earn 352. But when I play 5, most times the Computer chooses Out and I earn 592. So on average, I earn more when I play 5 because it signals to the computer that I must not be Type A and so the computer can get more points if it plays Out.

D Prespecified Estimations, Robustness, and Heterogeneity by Subject Gender

Table A.1: Self Confidence in Performance on Games by Subject Gender

Dependent Variable:	Beliefs on own performance	
	(1)	(2)
	Game 1 (Tower)	Game 2 (Signaling)
Female Subject	-0.0226	3.340
	(0.456)	(6.391)
Constant	17.02***	467.7***
	(0.923)	(11.81)
Day FE	X	X
Observations	304	303

Robust standard errors in parentheses. Column 1 refers to the number of predicted moves for the subject to move the tower. Column 2 refers to the expected points earned in Game 2, based on the self-reported probability of receiving each possible outcome. Day FE are fixed effects referring to the day the subject participated in the experiment. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.2: Confidence in Leader Performance

Dependent Variable:	Beliefs on leader's performance	
	(1)	(2)
	Game 1 (Tower)	Game 2 (Signaling)
(β_1) Fem. Leader	-0.171	-5.812
	(0.403)	(9.056)
(β_2) Ability		6.362
		(9.527)
(β_3) Fem. leader × Ability		14.39
		(12.98)
Day FE	X	X
Observations	304	301

Robust standard errors in parentheses. Column 1 refers to the number of predicted moves for the leader to move the tower. Column 2 refers to the expected points earned in Game 2 by the leader, based on the subject's reported probability of the leader receiving each possible outcome. Day FE are fixed effects referring to the day the subject participated in the experiment. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.3: Confidence in Leader Performance by Subject Gender

Dependent Variable:	Beliefs on leader's performance	
	(1)	(2)
	Game 1 (Tower)	Game 2 (Signaling)
Fem. Leader	-0.534	8.680
	(0.516)	(8.899)
Female Subject	-0.840	15.21
	(0.549)	(9.225)
Fem. leader × Fem. Subject	0.742	-15.18
	(0.819)	(12.85)
Day FE	X	X
Observations	304	301

Robust standard errors in parentheses. Column 1 refers to the number of predicted moves for the leader to move the tower. Column 2 refers to the expected points earned in Game 2 by the leader, based on the subject's reported probability of the leader receiving each possible outcome, and includes an indicator for belonging to the ability treatment arm as additional covariate. Day FE are fixed effects referring to the day the subject participated in the experiment. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.4: Beliefs on Tower of Hanoi

Dependent Variable:	Perceived		Perceived - Expected	Expected
	(1)	(2)	(3)	(4)
Fem. Leader	-1.013	-0.726	0.612	-0.401
	(0.684)	(0.531)	(0.554)	(0.604)
Female Subject	-1.204*	-1.013*	0.937	-0.332
	(0.665)	(0.543)	(0.576)	(0.660)
Fem. leader × Fem. Subject	1.173	0.823	-0.684	0.478
	(0.955)	(0.706)	(0.748)	(0.912)
Leader beliefs first				0.100
				(0.615)
Leader beliefs first × Fem. subj.				0.123
				(0.913)
Day FE	X	X	X	X
Observations	304	304	304	304

Robust standard errors in parentheses. Column 1 and 2 refers to the number of moves the subject reports as the leader's expected performance of the subject, Column 3 refers to the difference in the leader's expected performance of the subject relative to the subject's own expectations of his/her performance, Column 4 refers to the subject's own expectations of his/her performance. Column 2 includes expectations of one's own performance as an additional covariate. Leader beliefs first is an indicator for whether the subject was first asked about the leader's performance rather than his/her own performance. Day FE are fixed effects referring to the day the subject participated in the experiment. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.5: Leadership Game Results

	All Rounds				
	(1)	(2)	(3)	(4)	(5)
Dependent Variable:	SP	SP	SP	SP	Play 5
(β_1) Fem. Leader	-0.0604*	-0.0590*	-0.0518	-0.0605*	-0.0668*
	(0.0344)	(0.0352)	(0.0360)	(0.0349)	(0.0399)
(β_2) Ability	-0.00234	-0.00301	-0.00590	0.00762	0.00813
	(0.0343)	(0.0350)	(0.0362)	(0.0350)	(0.0405)
(β_3) Fem. leader × Ability	0.123***	0.115**	0.115**	0.115**	0.0978*
	(0.0472)	(0.0479)	(0.0491)	(0.0481)	(0.0559)
Covariates	X		X	X	X
Day FE	X	X	X		X
Round FE	X	X	X		X
Probit Specification				X	
Practice round	X	X		X	X
Observations	3010	3020	3030	3010	3010
Control group mean	0.618	0.618	0.618	0.618	0.374
$\beta_1 + \beta_3$	0.0624	0.0561	0.0633	0.0550	0.0310
P-val.: $\beta_1 + \beta_3$	0.0569	0.0891	0.0586	0.0970	0.434

Standard errors in parentheses, clustered at subject level. SP refers to strategic play (i.e., subject selecting 4 or 5); Play 5 refers to subjecting selecting 5. Covariates include subject's gender, ln(salary), level of employment, years of education, an indicator for having a masters degree, and tenure. Day FE are fixed effects referring to the day the subject participated in the experiment. Round FE are fixed effects referring to the ten rounds of the game. Practice Round is an indicator for whether the subject played strategically in a practice round prior to any advice from the leader. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.6: Excluding Rounds

Dependent Variable:	Strategic Play									
	(1) R1	(2) R2	(3) R3	(4) R4	(5) R5	(6) R6	(7) R7	(8) R8	(9) R9	(10) R10
(β_1) Fem. Leader	-0.0616*	-0.0581	-0.0578	-0.0591*	-0.0535	-0.0657*	-0.0657*	-0.0699**	-0.0567	-0.0562
	(0.0346)	(0.0367)	(0.0358)	(0.0355)	(0.0346)	(0.0360)	(0.0351)	(0.0345)	(0.0353)	(0.0347)
(β_2) Ability	0.00141	0.00838	-0.00293	-0.00501	0.00974	-0.0160	-0.00786	-0.0129	-0.00235	0.00409
	(0.0358)	(0.0350)	(0.0358)	(0.0351)	(0.0346)	(0.0346)	(0.0355)	(0.0352)	(0.0350)	(0.0341)
(β_3) Fem. leader × Ability	0.104**	0.113**	0.130***	0.131***	0.119**	0.137***	0.126***	0.136***	0.121**	0.112**
	(0.0486)	(0.0498)	(0.0491)	(0.0485)	(0.0476)	(0.0482)	(0.0485)	(0.0482)	(0.0487)	(0.0470)
Covariates	X	X	X	X	X	X	X	X	X	X
Day FE	X	X	X	X	X	X	X	X	X	X
Round FE	X	X	X	X	X	X	X	X	X	X
Practice round	X	X	X	X	X	X	X	X	X	X
Observations	2709	2709	2709	2709	2709	2709	2709	2709	2709	2709
Control group mean	0.660	0.651	0.649	0.634	0.632	0.646	0.645	0.642	0.634	0.629
$\beta_1 + \beta_3$	0.0421	0.0550	0.0724 **	0.0716 **	0.0657 **	0.0708 *	0.0605 *	0.0662 *	0.0645 *	0.0556 *
P-val.: $\beta_1 + \beta_3$	0.225	0.107	0.0332	0.0322	0.0476	0.0296	0.0730	0.0515	0.0586	0.0865

Each column excludes the round indicated in the column header. Standard errors in parentheses, clustered at subject level. Strategic play is defined as playing 4 or 5. Practice Round is an indicator for whether the subject played strategically in a practice round prior to any advice from the leader. Covariates include subject's gender, ln(salary), level of employment, years of education, an indicator for having a masters degree, and tenure. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.7: Leadership Game: Results by subject gender

Dependent Variable:	Strategic Play		
	(1) All subjects	(2) Male Subjects	(3) Female Subjects
(β_1) Fem. Leader	-0.0590*	-0.0683	-0.0600
	(0.0352)	(0.0488)	(0.0530)
(β_2) Ability	-0.00301	0.0107	-0.0144
	(0.0350)	(0.0517)	(0.0481)
(β_3) Fem. leader × Ability	0.115**	0.0979	0.135**
	(0.0479)	(0.0682)	(0.0683)
Day FE	X	X	X
Round FE	X	X	X
Practice round	X	X	X
Observations	3020	1560	1460
Control group mean	0.618	0.618	0.618
$\beta_1 + \beta_3$	0.0561	0.0296	0.0751
P-val.: $\beta_1 + \beta_3$	0.0891	0.540	0.0885

Standard errors in parentheses, clustered at subject level. Strategic play is defined as playing 4 or 5. Practice Round is an indicator for whether the subject played strategically in a practice round prior to any advice from the leader. Covariates are subject's gender, ln(salary), level of employment, years of education, an indicator for having a masters degree, and tenure. Day FE are fixed effects referring to the day the subject participated in the experiment. Round FE are fixed effects referring to the ten rounds of the game. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.8: Resume Evaluation Results: Social Desireability Bias

	(1) Competence	(2) Likeability	(3) Likelihood of Hire	(4) Log Salary Offer
Panel A: Social Desireability Bias				
Female Resume	-0.0729	-0.0283	-0.149	-0.123**
	(0.119)	(0.108)	(0.143)	(0.0521)
Reviewed Second	-0.0142	-0.0381	-0.147	-0.113**
	(0.119)	(0.115)	(0.141)	(0.0491)
Female × Reviewed Second	0.237	0.142	0.402*	0.227**
	(0.211)	(0.193)	(0.242)	(0.0993)
Panel B: Female Resume Evaluation				
Female Resume	0.0457	0.0425	0.0496	-0.0121
	(0.0607)	(0.0589)	(0.0704)	(0.0147)
Observations	450	450	445	441

Standard errors are clustered at the subject level and are in parentheses. Competence, Likeability, and Likelihood to Hire were asked using a Likert Scale, increasing from 1 to 5. Log Salary Offer is the log of the salary the subject suggested as an offer to the candidate in Birr. Female Resume is an indicator for the resume belonging to a randomly assigned female candidate. Reviewed Second is an indicator for whether the candidate was reviewed second. All regressions include the version of the resume and the ordering of the resumes as covariates. We restrict results to the sample used in the primary specification in the table for consistency; additional reductions in the number of observations are due to individuals who did not respond on the second resume. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.9: Resume Evaluation by Subject Gender

	(1) Competence	(2) Likeability	(3) Likelihood of Hire	(4) Log Salary Offer
Female Resume	-0.196	-0.0344	-0.237	-0.0737
	(0.166)	(0.149)	(0.217)	(0.0672)
Female Subject	-0.119	-0.0325	-0.129	-0.0735
	(0.162)	(0.155)	(0.185)	(0.0701)
Female Resume × Female Subject	0.240	0.0125	0.169	-0.0943
	(0.238)	(0.217)	(0.287)	(0.102)
Observations	225	225	225	225

Robust standard errors in parentheses. Competence, Likeability, and Likelihood to Hire were asked using a Likert Scale, increasing from 1 to 5. Log Salary Offer is the log of the salary the subject suggested as an offer to the candidate in Birr. Female Resume is an indicator for the resume belonging to a randomly assigned female candidate. Female subject is an indicator for the subject being female. Regression specifications include the resume version as a covariate. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

E Conceptual Framework

In this section, we provide a simplified conceptual framework, incorporating both taste-based and statistical discrimination, to show how our experimental results shed light on the importance of these two sources of discrimination. We consider a person's decision to follow the advice of either a male or a female leader. We assume that both the male and female leader have equal underlying ability θ. However, we allow both the mean and variance of ability in the population to vary by gender $g \in \{m, f\}$, so $\theta \sim N(\bar{\theta}_g, \sigma_g^2)$.[29] Mirroring our experiment, we focus on female and male leaders of high ability, so $\theta \geq \bar{\theta}_g$ for all g. The subject does not observe the leader's ability.

In the experiment, we study discrimination when the subject has no information about the leader except gender, and discrimination when the subject gets a signal indicating that the leader is of high ability. We consider these two cases in turn.

[29]Given large differences in educational attainment between men and women in Ethiopia, for example, it may make sense to assume that mean ability is higher among men, and ability among women exhibits higher variance.

No ability signal

Suppose first that the subject has no information about the leader except gender. Thus, the subject forms a belief $E(\theta|g)$ and chooses her action based on that belief. If she chooses to follow the leader's advice, she receives payoffs according to a continuous and increasing function $f(E(\theta|g))$. We also allow the subject's utility from following the advice to depend directly on the leader's gender, as in a model of "taste-based" discrimination (Becker, 1957). Therefore, suppose the subject has the utility function $u(g, f(E(\theta|g))) = f(\bar{\theta}_g) - c_g$, where c is the "taste-based" cost associated with following each gender. We standardize the utility of not following the leader to 0. The subject will then follow the leader's advice if the expected payoff from following the leader exceeds the taste-based cost of following the leader's directions:

$$f(\bar{\theta}_g) > c_g$$

Discrimination occurs when subjects are strictly less likely to follow the advice of a female leader than a male leader of equal ability.

In the absence of any other information about the leader, it is straightforward to see that both taste-based discrimination and statistical discrimination toward women reduce the share of subjects following the female leader relative to the male leader.[30] If there is taste-based discrimination against women ($c_f > c_m$), then the expected payoff from following the leader must be higher for the female leader than the male leader, to compensate for the distaste. If there is statistical discrimination against women (i.e., $\bar{\theta}_f < \bar{\theta}_m$), subjects are less likely to follow the female leader because the expected payoff from doing so is simply lower.

[30]We note that discrimination could also occur when statistical discrimination is positive (i.e., $\bar{\theta}_f > \bar{\theta_m}$), but taste-based discrimination is severe enough to outweigh the added benefit of following the female leader. Here, our intention is not to rule out the possibility of positive discrimination, but rather to focus on which mechanism can generate the empirical observation that subjects are less likely to follow female leaders.

The role of the ability signal

We now consider the possibility of introducing additional information about leader ability. Let s be a noisy but unbiased signal of ability: $s = \theta + u$, where u is independent of θ and is normally distributed with mean zero: $u \sim N(0, \eta^2)$.[31] Note that for a male and female leader of equal ability, the distribution of s is the same for them both. We assume Bayesian updating and obtain:

$$E(\theta|s, g) = \lambda_g \bar{\theta}_g + (1 - \lambda_g)s$$

where $\lambda_g = \frac{\eta^2}{\eta^2 + \sigma_g^2}$.

In other words, when there is a signal of ability, subjects form beliefs by taking a weighted average of the prior and the signal. The weights depend on the relative noise of the prior versus the ability signal: if the prior is noisier, the ability signal will be given more weight, whereas if the ability signal is noisier, the prior will be given more weight.

Comparing no signal with a high ability signal

We now consider how a high ability signal affects discrimination. We compare the gender gap in following the leader when the subjects receives no signal, versus when the subject receives a high ability signal, which is the primary empirical comparison in our experiment. Specifically, we consider the case where the subject discriminates against the female leader in the no-signal condition, and explore the circumstances that can generate a *reversal* when there is a high ability signal — that is, the subject discriminates in favor of the female leader in the high ability signal condition.

After observing a signal of high ability, subjects are weakly more likely to follow both male and female leaders relative to the no-signal case. If $s \geq \bar{\theta}_g$ for all g, then $E(\theta|s, g) \geq E(\theta|g)$

[31] In the experiment, subjects actually receive three signals: the leader's (continuous) score on an initial game played by subjects; a statement that the leader has training and experience in the primary experimental game, where the leader provides advice; and the leader's (continuous) score halfway through this primary game. We model a single continuous signal for clarity.

and the expected payoff from following the leader increases.

We consider two cases: when there is taste-based discrimination only, or when there is statistical discrimination only. When there is taste-based discrimination only, we have $c_f \geq c_m$ for all subjects, but beliefs about ability are identically distributed. In this case, the condition for following the leader is $f(E(\theta|s)) > c_m$ if the leader is male and $f(E(\theta|s)) > c_f$ if the leader is female.

Under only taste-based discrimination, $c_f > c_m$, the signal of high ability can reduce, but cannot reverse the gender gap in following the leader. A high ability signal increases the expected payoff from following the leader, so it makes discrimination more costly. However, because the expected payoff is independent of leader gender, any given expected payoff is weakly more likely to exceed the distaste for following a male leader than a female leader. Thus, under taste-based discrimination, the share following the female leader can never exceed the share following the male leader.

This implies that if a signal of high ability reverses the gender gap in following the leader, this must be due to a reversal of beliefs relative to priors. Holding taste preferences constant, any reduction in the gender gaps in beliefs will translate into a corresponding reduction in discrimination from below. Therefore, we now consider the case where there is statistical discrimination only ($c_f = c_m$), and focus on beliefs for the remainder of this section.

We return to our assumption that the priors on ability may vary by gender. With no signal of high ability, the gender gap in beliefs is simply $\bar{\theta}_m - \bar{\theta}_f$, which is assumed to be positive. A reversal occurs when $\bar{\theta}_m - \bar{\theta}_f > 0$, but the ability signal makes the gender gap in beliefs negative. In the case of a high ability signal, the gender gap in beliefs is:

$$E(\theta|s,m) - E(\theta|s,f) = \lambda_m \bar{\theta}_m - \lambda_f \bar{\theta}_f + (\lambda_f - \lambda_m)s$$

If the prior is that male leaders have higher mean ability, $\bar{\theta}_m > \bar{\theta}_f$, but similar variances, $\sigma_m^2 = \sigma_f^2$ then a signal of high ability will reduce, but not reverse, the gender gap. The

gender gap will be negative only if the variance of female ability is large relative to male ability, so that much more weight is placed on the signal for female leaders:

$$\frac{\lambda_f}{\lambda_m} < \frac{s - \bar{\theta}_m}{s - \bar{\theta}_f}$$

In the special case of $s = \bar{\theta}_m$,[32] that is, the signal indicates that the leader is of average male ability, differences in prior variances in ability also cannot reverse the gender gap. The gender gap can reverse if the shape of the belief distribution in entirely different for male and female leaders (i.e., at least one distribution is non-normal). Additionally, the gender gap can reverse if the signal itself is being interpreted differently for each gender, so that subjects infer different mean ability from the same signal for each gender. A simple example of a signal that will not be gender neutral is when subjects assume that for the same level of ability, a female leader will produce, on average, a lower signal than men (e.g., $s = \theta - \gamma_g + u, where \gamma_f > \gamma_m = 0$). There could be many reasons for the belief that women will produce a lower signal given equal underlying ability, such as social norms that make it more difficult for women to obtain such signals.

Summary of empirical predictions

The conceptual framework suggests the following empirical predictions:

1. If there is either taste-based or statistical discrimination from below, subjects will be less likely to follow the advice of a female leader than an otherwise identical male leader in a no-signal treatment.

2. If there is either taste-based or statistical discrimination from below, the gender gap in following the leader is reduced in a treatment where subjects observe a signal that the leader is of high ability.

[32] We focus on this special case because our results suggest that the signal of high ability in our experiment indicated average male ability, i.e., $s = \bar{\theta}_m$.

3. A reversal in the gender gap from the no-signal treatment to the high signal treatment indicates that discrimination is driven by beliefs. A high ability signal cannot reverse the gender gap in following the leader under standard assumptions of taste-based discrimination.