Discrimination from Below: Experimental Evidence on Female Leadership in Ethiopia

Shibiru Ayalew   Shanthi Manian   Ketki Sheth

February 17, 2018

Abstract

Globally, women are underrepresented in top management. We propose that this may result from discrimination from below: gender discrimination by subordinates can make a female leader less effective than an otherwise identical male leader. Using a novel "lab in the field" experiment in Ethiopia, we test whether leader gender affects the way subjects respond to leadership. We find evidence for discrimination: subjects are less likely to follow the advice of female leaders, and perform worse as a result. Using an information treatment, we then characterize this discrimination as belief-based rather than taste-based: when the leader is presented as highly trained and competent, the gender gap is reversed and subjects are *more* likely to follow women than men. The findings are consistent with a model of discrimination in which the same signal is interpreted differently for different genders, and which implies less discrimination at the "top" of the labor market. Consistent with this, we find no discrimination in a resume evaluation experiment for a senior management position. And, using a large sample of university administrative employees, we show that there is no gender wage gap among the highly educated.

1 Introduction

Globally, women are underrepresented in top management: for example, women hold just 17 percent of board directorships in the world's 200 largest companies (African Development Bank, 2015). In addition to equity considerations, these gaps suggest under-utilization of the productivity potential of the labor force. Existing explanations for these gaps have often focused on supply-side differences between male and female candidates (e.g., lower educational attainment or skill accumulation among women, differences in prefereences, or the notion that women are less likely to "lean in" and go after management positions (Niederle and Vesterlund, 2011; Sandberg and Scovell, 2013; African Development Bank, 2015)). In addition, a large literature documents discrimination from "above" in the hiring and promotion processes (Bertrand and Duflo, 2016). We propose a complementary explanation: that discrimination from "below"—gender discrimination by subordinates—can make a female leader appear less qualified than a male leader who is of equal ability ex-ante.

Success in management depends in large part on how well others adhere to the leader's advice and direction. Thus, even if women are equally skilled and have similar leadership styles, discrimination from below can reduce the performance of female-led teams. This can generate gender disparities in promotions to higher-level management even when male and female leaders are otherwise identical and, importantly, even when there is no discrimination in promotion decisions. This mechanism implies that policies that focus on altering female behavior or increasing human capital inputs to women will fall short of closing the gender gap in management. However, little well-identified evidence exists on whether individuals in the workforce respond differently to managers based solely on their gender. Evidence is particular scarce for developing countries, where the under-representation of women in senior management is even more severe.[1]

Using a novel laboratory experiment in a sample of 304 white-collar workers in Ethiopia, we study whether individuals respond differently when they are randomly assigned to a male versus female leader. Importantly, our design allows us to hold leader ability constant: there is no direct interaction between subjects and leaders, and pre-scripted messages are used to ensure that leader gender is the only difference between the two groups. We find that subjects are indeed 10 percent less likely to follow guidance provided by a female leader relative to a male leader. As a result, female-led subjects earn 3.5 percent fewer points.

The literature on discrimination distinguishes between "taste-based" discrimination, which is motivated by animus, and belief-based discrimination, where beliefs about a group are used to solve a signal extraction problem (Becker, 1957; Aigner and Cain, 1977; Guryan and Charles, 2013).[2] These different mechanisms suggest very different policy solutions. Therefore, to identify the source

[1]While women hold 17 percent of board directorships globally, the analogous figures in Africa, Asia and Latin America are 14 percent, 10 percent, and 6 percent respectively (African Development Bank, 2015).

[2]Belief-based discrimination is also often called "statistical" discrimination.

of discrimination, we cross-randomize an information treatment, where subjects are informed that the leader is highly trained and competent.

Interestingly, the gender gap is not only mitigated, but is actually reversed, when subjects receive information about leader ability. Moreover, the information treatment has no effect for male leaders: the likelihood that subjects follow male leaders is statistically indistinguishable with and without information on leader ability. These two facts allow us to characterize the discrimination as belief-based rather than taste-based: we show that a standard model where some individuals simply dislike female leaders cannot explain these findings. Moreover, in a model of belief-based discrimination, our results suggest that the same information about leader ability is interpreted differently for men versus women.

Reversals have previously been explained by dynamic models of discrimination, which could be one reason that signals would be interpreted differently for different groups (Fryer, 2007; Bohren, Imas and Rosenberg, 2017) We therefore consider the dynamic implications of discrimination from below. We develop a model based on Coate and Loury (1993) to show that because discrimination from below reduces the performance of female-led teams, women who nevertheless succeed in attaining management positions will be positively selected—that is, the underlying ability of an accomplished woman is higher than the underlying ability of an accomplished man. Then, conditional on making it to the "top" of the labor market, we may *not* observe discrimination against women.

We provide evidence consistent with this dynamic implication using a resume experiment for a senior management position. Subjects are asked to evaluate qualified candidates for a hypothetical position at their university along several dimensions (e.g., competence, likelihood of hiring, etc.) The candidate resumes listed a randomly assigned gender at the top, as is customary in Ethiopia. We find no differences in evaluations of candidates by gender along any dimension. As a second test of a dynamic model, we study gender wage gaps in our sample of 1,685 university administrative employees. We document an average gender wage gap of approximately 20 percent. At the same time, we document two empirical facts consistent with our model. First, women are much less likely to have a BA or higher than men. Second, we find no gender wage gap among highly educated employees (those with a BA or higher).

In a large literature on gender differences in labor market outcomes, this paper is one of the first, to our knowledge, to provide a well-identified estimate of gender discrimination from below.[3] We thus describe an understudied explanation for the persistent gender gap in senior-level management positions and provide a robust empirical test for its existence. While several papers have used similar methods to study discrimination from above in hiring and promotion decisions, evaluations, and credit or rental offers (Bertrand and Duflo, 2016), the evidence on discrimination from below has been more limited, in part because of the difficulty of randomly assigning leader gender while holding leadership style and ability constant in field settings.[4] Grossman et al. (2017) document discrimination toward female leaders in an incentivized coordination game; however, they do not distinguish between taste-based and belief-based discrimination. And while psychological research has documented differential responses to male and female leaders using hypothetical vignettes or trained actors (Eagly, 2013), we show that this discrimination persists when there are real stakes. Thus, our use of a "lab-in-the-field" experiment allows us to overcome many of the barriers in the existing literature and provide rigorous evidence for discrimination from below.

Secondly, we contribute to a growing literature on discrimination in low-income countries. Interestingly, most existing papers on female leadership in developing countries suggest that gender discrimination is more consistent with taste-based discrimination (or discrimination related to social norms), while our results are more consistent with belief-based discrimination (Beaman et al., 2009; Delavande and Zafar, 2013; Gangadharan et al., 2016). The large majority of the existing papers are based in South Asia.[5] Given substantial variation in gender norms and gender disparities around the world (Alesina, Giuliano and Nunn, 2013), we may not expect the magnitude or source of gender discrimination to be the same across different settings. And importantly, our finding that discrimination is at least partially belief-based has different policy implications. For example, many have suggested that more equitable gender attitudes tend to accompany the process of development (Duflo, 2012). However, if discrimination is belief-based, it may not be affected by changes in gender

[3]Blau and Kahn (2017) review the literature on gender differences in the labor market.

[4]A relatively new literature explores gender discrimination towards experts by using negative shocks ("mistakes") for identification (Egan, Matvos and Seru, 2017; Landsman, 2017; Sarsons et al., 2017). In addition to focusing on negative shocks, these settings are also qualitatively different in some ways from discrimination from below—for example, Sarsons et al. (2017) studies discrimination in general practitioner referrals to male versus female surgeons, which is more akin to a hiring decision.

[5]Several papers exploit India's political reservation system, which reserves seats for women in villages councils, as a natural experiment; however, only Beaman et al. (2009) and Gangadharan et al. (2016) explore implications for discrimination. Most papers explore the impacts of female leaders on outcomes; examples include Chattopadhyay and Duflo (2004); Beaman et al. (2012); Iyer et al. (2012); Kalsi (2017).

attitudes.

Our setting in Ethiopia may also be important to explaining the reversal of the gender gap that we observe in our experiment, as well as the lack of gender gaps among the highly educated in the resume experiment and institutional data. One reason that signals of ability may be interpreted differently as a function of gender is that it is more unusual for women to obtain those signals of ability in the first place. This model can thus 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.[6] Our results thus highlight the importance of conducting studies on discrimination in various settings, instead of assuming studies from high-income countries can be automatically applied to lower-income settings.

The rest of the paper proceeds as follows. In Section 2, we provide a theoretical framework to motivate our experiment. Section 3 provides details on the design of the three components of the study: the leadership game, resume evaluation experiment, and institutional data analysis. In Section 4, we present the results. Section 6 concludes and discusses policy implications of the results.

2 Theory

We study an employee's decision to follow the advice of either a male or a female manager. We assume that both the male and female manager 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 $\theta \sim N(\bar{\theta}_g, \sigma_g^2)$.[7] We further assume that both the female and the male manager are of high ability, so $\theta \geq \bar{\theta}_g$ for all g.

The employee does not observe the manager's ability. We first consider a base case in which the employee has no information about the manager except gender. Thus, the employee forms a belief $\tilde{E}(\theta|g)$ and chooses her action based on that belief. If she chooses to follow the manager's advice, she receives payoffs according to a continuous and increasing function $f(\tilde{E}(\theta|g))$. We also allow the employee's utility to depend directly on the manager's gender, as in a model of "taste-based" discrimination (Becker, 1957). Thus, the employee has utility function $u(g, f(\tilde{E}(\theta|g)))$. For

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

[7] 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 women exhibit higher variance

simplicity, we assume that utility is linear in payoffs, and that taste-based utility and utility from payoffs are additively separable. This yields $u(g, f(\tilde{E}(\theta|g))) = f(\tilde{E}(\theta|g)) - c(g)$, where c is the "taste-based" cost associated with following each gender. We standardize the expected utility of not following the manager to 0. The employee will then follow her manager's advice if the expected payoff from following the manager exceeds the taste-based cost of following the manager:

$$f(\tilde{E}(\theta|g)) > c(g)$$

Remark 1 *Employees are less likely to follow female managers if* $c(f) > c(m)$*, if* $\bar{\theta}_f < \bar{\theta}_m$*, or both.*

In the absence of any other information about the manager, both taste-based discrimination and statistical discrimination result in employees being less likely to follow the female manager relative to the male manager. If there is taste-based discrimination against women, then the expected payoff from following the manager must be higher for the female manager than the male manager to compensate for the distaste. If there is statistical discrimination against women (i.e., $\bar{\theta}_f < \bar{\theta}_m$), employees are less likely to follow the female manager because the expected payoff from doing so is simply lower.

The role of ability signals

We now consider the possibility of introducing additional information about manager 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: $u \sim N(0, \eta^2)$. Note that since the male and female manager are truly of equal ability, the distribution of s is the same for them both. We assume Bayesian updating and obtain:

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

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

In other words, when there is an additional signal of ability, employees 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.

Remark 2 *After observing a high signal, employees are weakly more likely to follow both male and female managers relative to the no-signal baseline.*

If $s \geq \bar{\theta}_g$ for all g, then $\tilde{E}(\theta|s,g) \geq \tilde{E}(\theta|g)$ and the expected payoff from following the manager increases.

We now consider the role of ability signals when there is taste-based discrimination only: $c(f) \geq c(m)$ for all employees, but beliefs about ability are identically distributed. In this case, the condition for following the manager is $f(\tilde{E}(\theta|s)) > c(m)$ if the manager is male and $f(\tilde{E}(\theta|s)) > c(f)$ if the manager is female.

Proposition 1 *Under only taste-based discrimination, ability signals cannot reverse the gender gap in following the manager.*

The ability signal increases the expected payoff from following the manager, so it makes discrimination more costly. However, if the expected payoff is independent of manager gender, any given expected payoff is weakly more likely to exceed the distaste for following a male manager than a female manager. Thus, under taste-based discrimination, the share following the female manager can never exceed the share following the male manager.

Proposition 1 above implies that if the ability signal reverses the gender gap in following the leader, this must be due to a reversal of beliefs relative to priors. In this section, we return to our initial assumption that the priors on ability may vary by gender. In this case, after observing a high signal of ability, the gender gap in beliefs is:

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

If the prior is that male managers have higher mean ability, the gender gap in beliefs will reverse if the variance of female ability is large relative to male ability, so that much more weight is placed on the signal for female managers:

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

However, suppose $s = \bar{\theta}_m$, that is, the signal indicates that the manager is of average male ability. The signal will have no effect of employees' response to a male manager, but will increase beliefs about the ability of a female manager.

Proposition 2 *A signal indicating that a female manager is equal to the average male manager can reduce, but cannot reverse, the gender gap in following the manager.*

In this case the gender gap in beliefs is $\lambda_f(\bar{\theta}_m - \bar{\theta}_f)$, which is weakly positive by assumption.

Signal differing by gender

A reversal in beliefs can be explained by a model in which employees interpret the same signal differently based on the gender of the manager. As a simple illustration of this point, let $s = \theta - \gamma_g + u$, for some constant γ_g, where $\gamma_m = 0$ and $\gamma_f > 0$. Therefore, for the same level of ability, the employee assumes that a female manager will produce, on average, a lower signal than men. Now we have:

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

Proposition 3 *If employees believe that the signal mean differs by gender, then it is possible for a signal $s = \bar{\theta}_m$ to reverse the baseline gender gap in beliefs about ability.*

For $s = \bar{\theta}_m$, the gender gap in beliefs is now $\tilde{E}(\theta|s, m) - \tilde{E}(\theta|s, f) = \lambda_f(s - \bar{\theta}_f) - (1 - \lambda_f)\gamma_f$. This can be negative if the penalty γ_f is large enough. Employees viewing the same signal from male and female managers will conclude that it indicates higher ability for the female manager, on average, and this may be enough to reverse the gap.

There may be several reasons that employees would interpret the same signal differently for male and female manager. One is gender stereotypes: employees may expect female managers to perform worse on math or logic problems, for example (Bordalo et al., 2016). In a labor market context, the canonical signal of ability is education. In the educational setting, this result is consistent with the dynamic model of discrimination described by Bohren, Imas and Rosenberg (2017), which is driven by barriers to entry in obtaining signals. In Ethiopia, as in many places around the world, barriers to entry for women in education are well documented. For example, the World Economic

Forum's 2016 Global Gender Gap Report ranked Ethiopia 132, out of 144 countries evaluated, for educational attainment.

Effect of discrimination from below on promotion

To conclude our theoretical treatment of discrimation from below, we adapt Coate and Loury (1993) to show that discrimination from below can generate disparate promotion probabilities for male versus female managers even when the employer is unbiased. The employer must decide whether to promote a manager to a higher level. We assume the employer's objective is to promote qualified managers; thus, employers receive a payoff of $x_q > 0$ if they promote a qualified manager and $-x_u < 0$ if they promote an unqualified manager. Employers do not observe whether managers are qualified, but they do observe the *performance* ϕ of the manager's team. Let $F_{i\in\{q,u\}}(\phi)$ denote the cumulative distribution function of ϕ for qualified and unqualified managers, respectively.

Because qualified managers improved the performance of their teams, we assume that $F_{q,g}(\phi) < F_{u,g}(\phi)$ for all ϕ and for all g. That is, the team performance of qualified managers first order stochastically dominates the team performance of unqualified managers for both men and women. In addition, we assume that employees are less likely to follow the good advice of qualified female managers due to discrimination. This reduces the performance of teams led by qualified female manager relative to teams led by qualified male managers. We assume $F_{q,m}(\phi) \leq F_{q,f}(\phi)$. In contrast, discrimination has no effect on the performance of teams led by unqualified managers (since the advice is unhelpful anyway). Thus $F_{u,m}(\phi) = F_{u,f}(\phi)$ for all ϕ.

Now suppose employers are unbiased and have a prior that the share π managers are qualified. After observing the team performance, they update to:

$$\xi(\pi, \phi) = \frac{\pi f_q(\phi)}{\pi f_q(\phi) + (1-\pi) f_u(\phi)}$$

As in Coate & Loury 1993, the employer's expected benefit from promoting any given manager is $\xi(\pi, \phi)x_q - (1 - \xi(\pi, \phi))x_u$. The employer maximizes her payoff by setting a minimum team performance standard $\underline{\phi} = \min\{\phi : \xi(\pi, \phi)x_q - (1 - \xi(\pi, \phi))x_u > 0\}$ and promoting managers whose teams exceed the minimum standard.

Proposition 4 *Even if the share of qualified managers is equal for men and women, discrimination* *from below will reduce the probability that female managers are promoted.*

By reducing the performance of the team, discrimination from below will reduce the probability that female-led teams exceed the minimum performance standard. Formally, women are promoted with probability $1 - [(1-\pi)F_{u,f} + \pi F_{q,f}]$ and men are promoted with probability $1 - [(1-\pi)F_{u,m} + \pi F_{q,m}]$. The difference in promotion probabilities is $(1-\pi)(F_{u,f} - F_{u,m}) + \pi(F_{q,f} - F_{q,m})$, which is strictly positive by assumption.

3 Study Design

We conducted the study in Adama, Ethiopia, in a sample of full-time administrative employees at Adama Science and Technology University (ASTU). Our primary results are based on an experiment we conducted in a subsample of these employees. We supplement the experimental results with data from a survey we conducted as well as institutional human resources data on the universe of ASTU administrative employees.

3.1 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-indexes related to education and labor market outcomes: they ranked 106 on "Economic participation and opportunity" and 132 on educational attainment. However, the country has instituted a number of affirmative action policies designed to reduce gender gaps. In 2016, as part of its annual Country Policy and Institutional Assessment (CPIA) exercise, the World Bank assigned Ethiopia a Gender Equality Rating of 3 on a scale of 1 (low) to 6.[8]

Adama Science and Technology University (ASTU) is an elite public university located about 100 km from the capital, Addis Ababa. Table 1 shows summary statistics for all administrative employees at ASTU, based on institutional data from the human resources department. Educational attainment in the sample is high: on average, employees completed 12 years of education, which

[8]The gender equality ranking assesses the extent to which the country has installed institutions and programs to enforce laws and policies that promote equal access for men and women in education, health, the economy, and protection under law.

Table 1: Summary Statistics

	(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)	
Observations	1685	746	939	1685

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard deviations in parentheses.

corresponds to secondary school completion. In contrast, in the Ethiopian population more broadly, 48.3 percent females and 45.7 percent males 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.

Women represent 56 percent of the sample, which suggests that they are over-represented in the sample, but only slightly. In 2012, women and men with an advanced education were almost equally likely to be in the labor force (although the female labor force participation rate is about 15 percentage points lower overall). We observe significant differences in job tenure by gender: women have been with the institution longer.

Importantly for the interpretation of our model, women in the sample have significantly fewer years of education and are 37 percent less likely to hold a BA. They are also 75 percent less likely to hold a Masters degree. Though we were unable to find comparable national statistics on education, this does mirror the general trend of gender gaps in education completion in Ethiopia. For example, in primary and secondary school, the gender parity index of school enrollment is 1. But for tertiary school, the gross enrollment gender parity index is .5 (World Bank, 2017).

3.2 Leadership Game

3.2.1 Sample

For the leadership game, we selected a subsample of the university administrative employees that hold a BA or higher.[9] Using a list of employees provided by the human resource department, we contacted all administrative employees with a BA or higher ($n = 500$), and implemented the game until we reached 150 female subjects and 150 male subjects. The game was ultimately implemented over 6 days. Most eligible subjects who did not participate (about 40 percent) were excluded because we were unable to locate them during the week of the study.

3.2.2 Overview of design

The leadership game was designed to test two questions, as previously described: whether individuals respond differentially to male and female leaders *holding all else constant*; and whether information about the leader's ability changes the gender gap. The basic setup of the game is that subjects are randomly assigned to either a male or female "leader", and the subject-leader pair completes several tasks. However, there is no direct interaction between the subject and the leader, allowing us to hold the leader's behavior constant across male and female leaders.[10] 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. In general, we are interested in the likelihood of subjects following the guidance provided by their leaders as a function of their leader's gender, and whether any gender gap can be mitigated by providing information about the leader's ability.

The game consists of two tasks: a logic game (Tower of Hanoi) and a signaling game adapted from Cooper and Kagel (2005). The primary purpose of the first task is to elicit beliefs on performance, and to serve as an input to the ability signal. The primary purpose of the second task is to measure whether subjects follow their leader's directions.

In the logic game, subjects are shown a Tower of Hanoi and are asked to move the tower from

[9] We restricted the game to highly educated employees because we wanted to focus on white collar workers, and because we believed that the game may be too complicated for subjects with low levels of education.

[10] The leaders were real individuals at another University who actually played the games as described to the subjects. To hold behavior constant, the leaders played ahead of time, and we only matched subjects to a subset of male and female leaders who played in the same way and had the same outcomes.

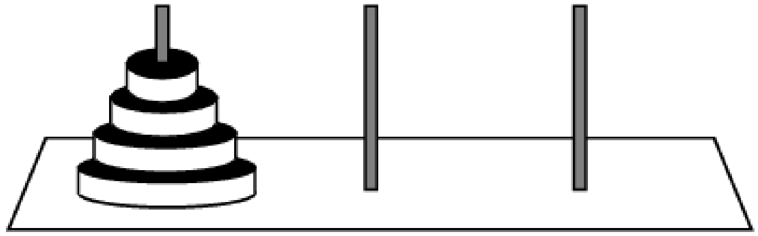

Figure 1: Tower of Hanoi

one pole to another (see Figure 1). They can only move one disk at a time, and a larger disk cannot be placed on a smaller disk. Subjects are shown the game and allowed to practice once with three disks. The subject is then told that they will be asked to solve the Tower using four disks and that the minimum moves are 15.

Prior to actually playing, we asked subjects how many moves they think *they* will require to move the tower, how many moves they think *the leader* will require to move the tower, and finally how many moves they think the leader guessed *they* would require to move the tower.[11] The subject is then given one chance to move the tower in a time limit of four minutes. The subject was paid as a decreasing function of number of moves required to move the tower and number of moves further from their guess (see Appendix Figure A for compensation schedule).

The second task was a signaling game adapted from Cooper and Kagel (2005). We selected this game because it has a clear correct answer that is difficult to guess. In this two-player game, nature first selects Player 1's type (A or B). Player 1 moves first, and Player 2 then responds after seeing what Player 1 has selected. The sequence of moves is shown in Figure 2 and the payoff structure is shown in Figure 3.[12]

The key insight is that for a Type B player 1, the optimal play is 5. The logic is as follows. A naive Type B Player 1 will select 3, observing that conditional on Player 2's selection, 3 always provides the highest payoff. But a Type B Player 1 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 3). A similar logic could be applied to playing 4.

We are interested in the strategic interaction between leaders and subjects. The research question

[11]The first two questions were randomly ordered for each subject

[12]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.

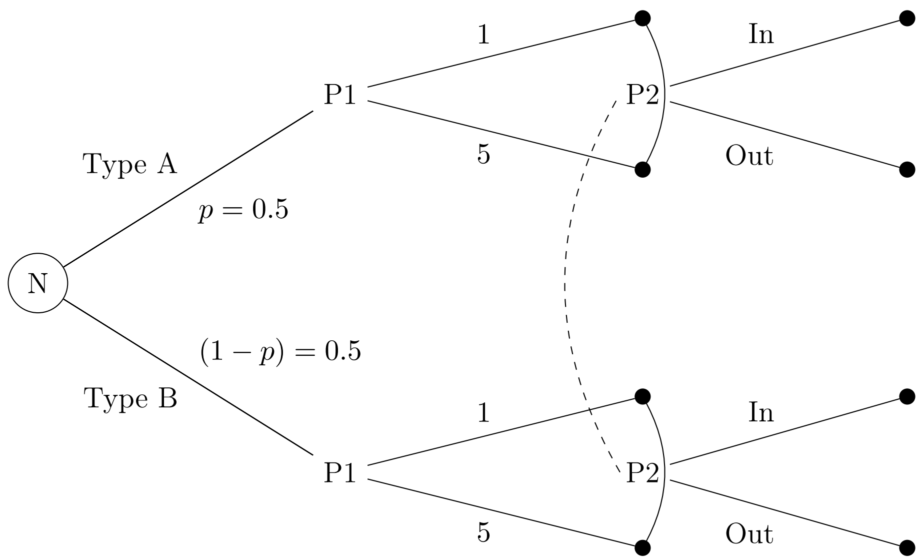

Figure 2: Simplified Game Tree for Game 2

Player 1:

Type A				Type B		
A	In	Out		B	In	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

Player 2:

	Type A	Type B
In	500	200
Out	250	250

Figure 3: Signaling Game Payoffs (colors not shown to subjects)

focuses on whether leader gender and ability change the likelihood that the subject will follow the leader's advice. We therefore assigned all subjects to be Player 1 Type B, and Player 2 was played by a computer. We programmed the 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 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 being introduced to the directions of the game, the subject was then asked to complete a "practice round" in which they selected which number they believed they would play, but the computer did not respond to this selection. Subjects were then asked what they believed was the probability of receiving each possible payoff in the first round. We expected non-zero probabilities on only two of the options (as the subject selects which number they will play), but the majority of subjects did include positive probabilities on more than two possible payoffs. The subject was then asked what they believed was the probability of the leader receiving each possible payoff in the first round. Using these two questions, we were able to calculated the subject's belief of the expected point value for him/herself and their leader.

The subject then played 10 rounds on the game. Prior to each round, the subject is able to observe how their assigned leader played for that given round.[13] In addition, subjects are told that the leader can 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.[14] The messages were displayed on an Android app by the enumerator (Figure 4). The enumerator additionally recorded the leader's play and outcome for each round on a piece of paper in front of the subject.

Figure 5 provides an overview of the experiment. We completed the game in a span of 6 days. Due to subjects discussing the game with colleagues, we relabeled the choices for Day 5 and Day 6. Specifically, Player 1 selected from two different sets of letters for Days 5 and 6, and the computer

[13]Leaders were selected at a different university a week prior. Unlike the subjects in the primary study, the leaders were given extensive training on how to play each task. We selected the two top performing leaders, one male and one female, to be assigned to subjects. Both of these leaders selected 5 for each round, and the Computer responded "Out" for every round. Leaders received a bonus based on the average performance of the team members assigned to them. Subjects were told that their leader's compensation is partly based on how well the subject performs on the task.

[14]All leaders chose to send the messages.

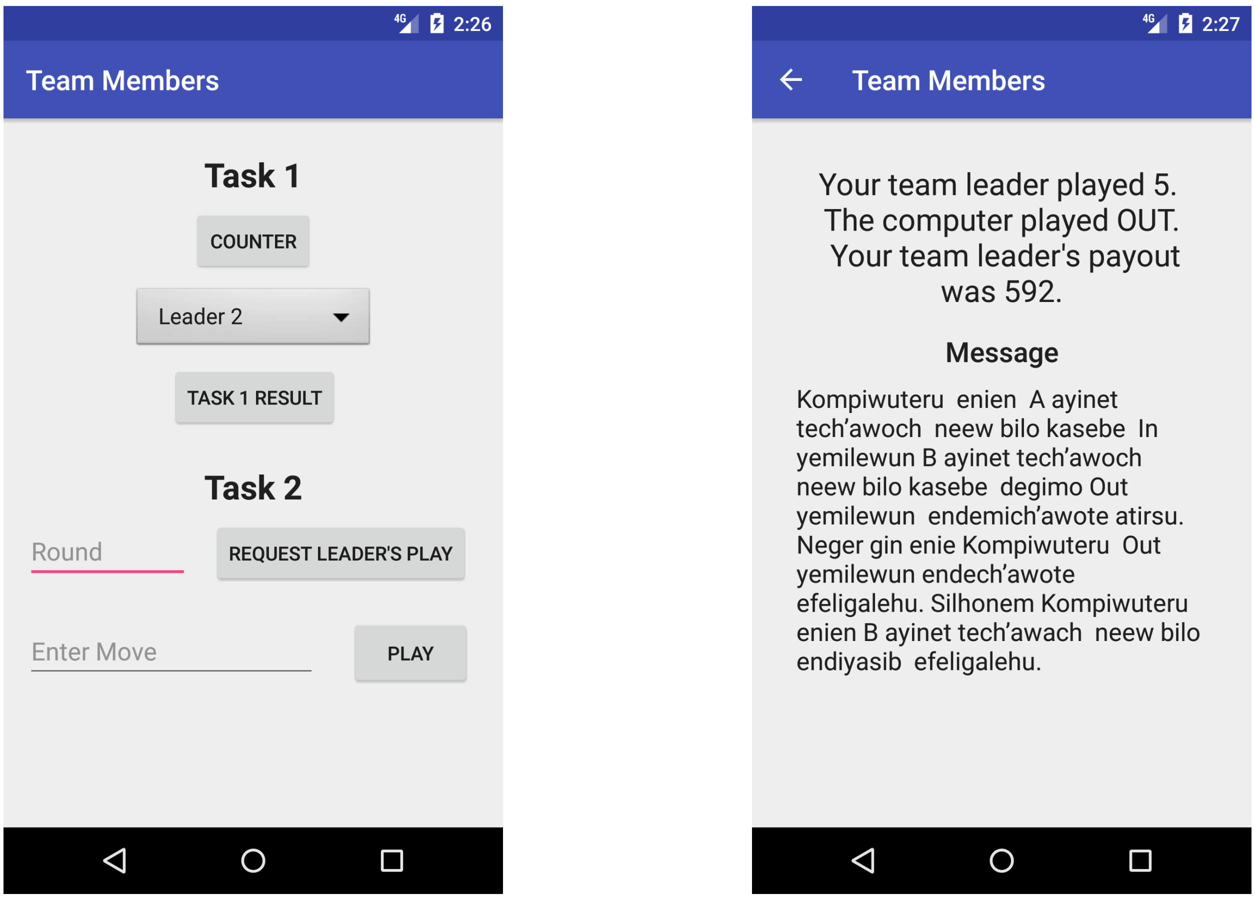

Figure 4: Leader result and messages as shown to subjects

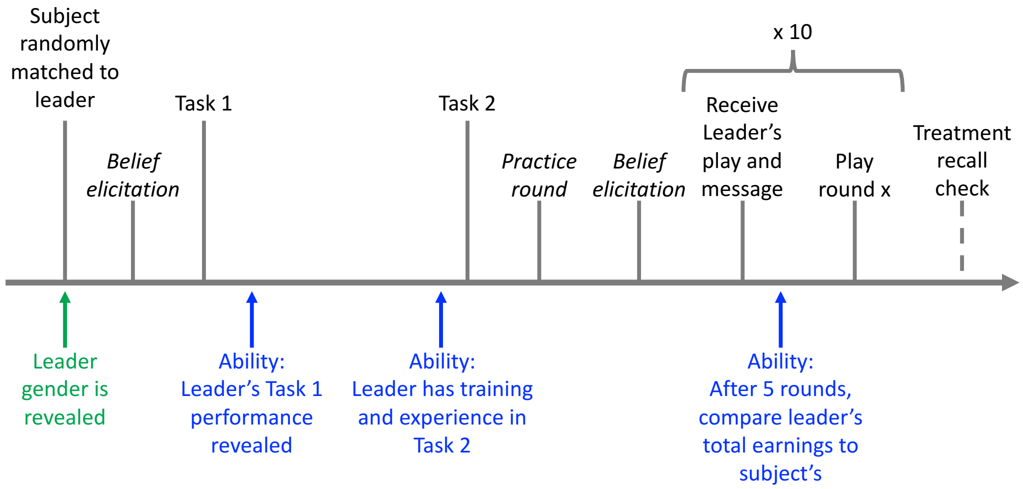

Figure 5: Timeline of Leadership Game

responded with "left/right" and "up/down."

3.2.3 Experimental Treatments

We implemented a cross cutting randomization of two treatments: leader gender and information on leader ability. Subjects were randomly assigned into one of four groups: Female leader with no information on ability, male leader with no information on ability, female leader with information on ability, and male leader with no information on ability.[15]

Leader Gender

Subjects were randomly assigned either the male leader or the female leader. Recall, the information provided to the subjects about how the leaders played are identical, 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 telling the subjects the gender of their leader, we provided gendered

[15] We randomized leader gender and then independently randomized the ability treatment, so the subjects are not perfectly evenly distributed across treatments. The distribution is as follows. Female leader with no information on ability: n = 78. Male leader with no information on ability: n = 71. Female leader with information on ability: n = 70. Male leader with no information on ability: n = 85.

pseudonyms for the leader (mentioned 23 times in the enumerator's script) and relied on the gendered language of 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 on the characteristics of their leader, including gender, on the last two days of the experiment. 95 percent recalled the correct gender of the leader.

Leader Ability

Subjects were randomly assigned whether or not they were given information on the leader's ability. The ability treatment consists of three components. First, after the "Tower of Hanoi" logic game, the enumerator informed the subject that the leader completed the task in the minimum number of moves, and noted how many moves fewer this was than their own performance. [16] Second, in the introduction to the second task, subjects were explicitly told that unlike themselves, the leader has already played the game and is an experienced player. Finally, after 5 rounds of play, the enumerator totalled the points earned by the leader versus the subject to highlight the (expected) point advantage by the leader.

3.2.4 Validity of randomization

Subjects were assigned a treatment once they arrived for the experiment. The randomization was stratified by subject gender. We had generated a random ordering of 150 treatment assignments per male and female subjects to be assigned as subjects arrived. For the last two days of the experiment, we rerandomized using a blocked randomization in groups of four, because we were concerned that we may not meet our recruitment targets (although we were ultimately successful in meeting the target). In all analyses, we account for differing randomization probabilities use inverse probability weights.

Table 2 confirms the validity of our randomization. Using information on the subjects provided by the human resources department, we confirm that subject characteristics are balanced across the four treatment groups using a linear regression of treatment assignment on each characteristic. We also confirm pairwise balance in the bottom three rows of Table 2. Column 1 confirms balance on

[16] Note that subjects were not informed of the extra practice and training that leaders received for the logic game, regardless of treatment assignment

Table 2: 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

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

subject, which is satisfied by construction because we stratified on subject gender. We also confirm balance on salary, job level, education, and tenure, none of which were used for stratification.

In additional to balance across subject characteristics, we may be concerned that the pseudonyms we used to connote gender also connoted other 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). To the extent that names connote information on ethnicity and religion, we want to confirm that our treatments are balanced across other information contained in the pseudonyms. 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). We therefore oversample Oromic names in our selection. 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 3 confirms that the characteristics associated with the pseudonym assigned to each subject in a given treatment are balanced across treatment arms.[17]

[17] The results in Table 2 and Table 3 are robust to the exclusion of day fixed effects.

Table 3: 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

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

A final concern is that due to the randomized responses of the computer, leader ability could appear different in different treatments despite holding leader behavior constant. Subjects may perceive the leader as less able if they do not follow the leader's advice and happen to obtain a higher payoff in a given round than the leader, or if they follow the leader's advice but happen to receive a low payoff. Table 4 shows that these "errors" are balanced across treatments both unconditionally (Column 1) and conditional on the subject's play (Column 2). This alleviates concerns that differential error rates could be driving our results.

3.2.5 Estimating Equations

Our primary research question is whether discrimination from below reduces the performance of female leaders. In the leadership game, this correspond to the hypothesis that subjects are less likely to follow the leader's advice to play strategically (defined as playing 4 or 5, following (Cooper and Kagel, 2005)). We additionally hypothesized that information indicating the leader is trained and competent mitigates and may reverse such gender gaps.

Table 4: Leader "error" balance

	(1) Error	(2) Error
Female leader only (F)	0.00622	0.00267
	(0.0183)	(0.0129)
Ability signal only (A)	0.0124	0.0127
	(0.0182)	(0.0123)
Female leader & Ability (FA)	0.0190	0.0113
	(0.0193)	(0.0138)
Day FE	Yes	Yes
Round FE	Yes	Yes
Play FE	No	Yes
Observations	3344	3339
p-val: F = A	0.730	0.420
p-val: A = FA	0.724	0.916
p-val: F = FA	0.500	0.536

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

To test these hypotheses we estimate the following equation using a linear regression model:

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

where R is one of two outcomes. Our primary outcome of interest is an indicator for playing strategically (i.e., selecting 4 or 5) for subject i in round r. We also study simply playing 5, since subjects may simply mimic the leader without understanding the logic behind their play. FL is an indicator for being randomly assigned a female leader, *Ability* is an indicator for being randomly assigned receipt of information on the leader's ability, and $FL * Ability$ is the interaction of the two indicators. [18] We additionally include an indicator of whether the practice round selection was equivalent to the outcome of interest, day fixed effects, and round fixed effects 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.

Based on our model, we hypothesized the following:

$\beta_1 < 0$: In the absence of information, directions provided by female leaders are less likely to be

[18] As previously described, we corrected for varying randomization probabilities using inverse probability weights. The exclusion of these weights does not qualitatively change the results.

followed relative to directions provided by male leaders.

$\beta_2 > 0$: Providing information on the leader's ability increases the likelihood of subject's following the leader's directions.

$\beta_3 > 0$: Providing information on the leader's ability increases the likelihood of a subject's following the female leader's directions more than male leaders.

We additionally hypothesized the standard prediction of information in discrimination for unbiased signals of ability:

$\beta_1 + \beta_3 <= 0$: Though information mitigates gender differentials, it will fall short of eliminating the gap completely. This sum represents the gender gap conditional on ability information (i.e., the probability of following the directions of a female leader with ability information relative to a male leader with ability information.) The reversal of the gender gap predicted by a model in which signals are interpreted differently for men and women corresponds to $\beta_1 + \beta_3 > 0$.

4 Results

4.1 Leadership Game

Table 5 shows our primary results from estimating equation (1).[19] Column 1 uses the outcome variable of strategic play (i.e., selecting 4 or 5), and Column 2 uses the outcome variable of playing 5. In both columns, we find that in the absence of ability, subjects with female leaders (and no information on ability) were 6 percentage points less likely to play in accordance with their leader's directions (see β_1). Relative to subjects with male leaders with no information on ability, this reflects a 10 percent reduction in adherence to the leader's recommendation.

Surprisingly, we find that information on ability had no effect for subjects with male leaders: subjects were equally likely to follow male leaders whether or not they were given information on the leader's experience or training (see β_2). This would suggest that for men, the signal did not provide information on ability that was greater than the expected group mean of men. In other words, the signal we provided of being capable of performing well on the tasks was already in line with the expectation of how average males would perform.

[19]The results are qualitatively similar when the practice round is excluded, but lose precision. Marginal effects and statistical significance are similar when using either probit or logit models.

Table 5: Leadership Game Results

Dependent Variable:	Strategic Play	Played 5	Total Points
	(1)	(2)	(3)
(β_1) Fem. Leader	-0.0590*	-0.0640	-179.8**
	(0.0352)	(0.0407)	(79.13)
(β_2) Ability	-0.00301	0.000942	49.36
	(0.0350)	(0.0409)	(78.99)
(β_3) Fem. leader × Ability	0.115**	0.0998*	101.0
	(0.0479)	(0.0562)	(113.9)
Day FE	X	X	X
Round FE	X	X	
Practice round	X	X	X
Observations	3020	3020	302
Control group mean	0.618	0.374	5099.2
$\beta_1 + \beta_3$	0.0561	0.0359	-78.84
P-val.: $\beta_1 + \beta_3$	0.0891	0.366	0.335

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. In Columns 1 and 2, standard errors are clustered at subject level. Strategic play is defined as playing 4 or 5. 5 is the highest expected value play, and the leader played 5 in every round.

However, the information on ability does have a large effect for subjects assigned to female leaders (see β_3). Interestingly, $\beta_1 + \beta_3 > 0$, which means that upon having information on ability, subjects were more likely to follow the directions provided by female leaders relative to male leaders. As shown in Section 2, if priors are normally distributed, this implies that the ability signal is interpreted differently for men and women, even though the information is identical.

The discrimination against female leaders in the absence of ability information is costly. Column 3 quantifies the effect of female leadership on the total number of points that subjects earned in the game. Column 3 shows that on average, the control group (male leader, no ability information) earned 5099 points, whereas the leader earned 5920 points in the game. Having a female leader reduced total points earned by 179.8, a reduction of 3.5 percent. The model in Section 2 thus predicts that an employer evaluating the female leader based on her team's performance would be less likely to conclude that she was qualified, and discrimination from below would reduce her probability of promotion to higher-level management.

Table 6 estimates our results separately for male and female subjects. Though less precise, the

Table 6: Heterogeneous Treatment Effects by Subject Gender

Dependent Variable:	Strategic Play		Played 5	
	(1) M. subjects	(2) F. subjects	(3) M. subjects	(4) F. subjects
(β_1) Fem. Leader	-0.0683	-0.0600	-0.0440	-0.0927*
	(0.0488)	(0.0530)	(0.0605)	(0.0554)
(β_2) Ability	0.0107	-0.0144	0.0828	-0.0804
	(0.0517)	(0.0481)	(0.0611)	(0.0520)
(β_3) Fem. leader × Ability	0.0979	0.135**	-0.00673	0.213***
	(0.0682)	(0.0683)	(0.0771)	(0.0799)
Day FE	X	X	X	X
Round FE	X	X	X	X
Practice round	X	X	X	X
Observations	1560	1460	1560	1460

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses, clustered at subject level. Strategic play is defined as playing 4 or 5. 5 is the highest expected value play, and the leader played 5 in every round.

estimates do suggest that the reversal of discrimination may be more driven by female subjects. If women have a greater understanding of the barriers females face to attain "signals of ability", then it is likely that females would be more likely to infer higher levels of ability for a given signal.

Our estimates of belief expectation on how well the leader will perform in Task 2 can act as a robustness check for our results. Unfortunately, the belief expectation exercises were difficult for subjects to understand and thus were likely very noisy estimates of belief. However, as Table 7 shows, the pattern of the magnitudes of the beliefs elicited for Task 2 directly align with the pattern of being likely to follow the leader's directions in Table 5. Female leaders (relative to male leaders) were assessed to perform more poorly (i.e., lower expected value) when no information was provided on ability, but were assessed to perform more effectively (relative to male leaders) when information on ability was provided. Our results lack statistical precision and thus cannot be differentiated from having no effect on expected value of performance, but the fact that they exhibit the same pattern as our primary results are suggestive of the robustness behind our results in Table 5.

Table 7: Beliefs about leaders

Dependent Variable:	Leader's performance	
	(1)	(2)
	Task 1	Task 2
Fem. leader (F)=1	1.304	-5.812
	(6.416)	(9.056)
Ability (A)=1	13.50**	6.362
	(6.466)	(9.527)
Fem. leader (F)=1 × Ability (A)=1		14.39
		(12.98)
Constant	456.2***	459.4***
	(12.78)	(13.00)
Day FE	X	X
Observations	301	301

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

5 Resume Experiment and Observational Data

5.1 Design

Upon completion of the experimental game, we implemented a resume evaluation experiment that began the following week. We successfully followed up with 86.8 percent of the experimental subjects.[20] 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. 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.[21] An example is shown in Figure 6.

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, indicating that they read the resumes carefully. However, to guard against social desirability bias,

[20]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.

[21]There were two model resume types, resulting in four possible resumes: female/type 1, male/type 1, female/type 2, male/type 2.

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 6: Resume Evaluation Experiment: Example Resume

Table 8: Resume Experiment Balance

	(1) Fem. subject	(2) Years of education	(3) ln(Salary)	(4) tenure$_{days}$	(5) Level
Female Resume	-0.0174	0.0620	-0.0400	401.7	-0.245
	(0.0618)	(0.0722)	(0.0454)	(246.5)	(0.324)
Resume Version	-0.0536	-0.0601	0.0219	-221.3	0.0767
	(0.0618)	(0.0722)	(0.0453)	(246.4)	(0.324)
Constant	0.528***	16.12***	8.078***	2994.0***	13.41***
	(0.0541)	(0.0631)	(0.0397)	(215.5)	(0.283)
Observations	264	264	264	264	264

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

we compare gender across subjects only; that is, in the analysis sample, subjects are not directly comparing a male and a female candidate.[22]

After reviewing each resume and completing the comprehension test, subjects evaluated the potential candidate on an increasing scale of 1 to 5 on competence, likeability, and willingness to hire. They additionally suggested a salary to be offered to the candidate.[23]

Because of uncertainty in scheduling survey interviews with subjects, we again randomized the treatment (which of the four resumes) by creating a random ordering in groups of four for each enumerator and then had them go in the order of their list when interviewing subjects. Table 8 confirms the validity of our randomization by documenting that subject characteristics were balanced across treatment arms.

[22]In the experiment, subjects were given a second resume of the opposite gender and asked to compare it; however, because of concerns about social desirability bias, evaluations of this second resume are excluded from this analysis. Importantly, when subjects were given the initial resume to evaluate, they were not told that a second resume would follow.

In evaluations of the second resume, it appears that subjects become aware that the two resumes differ on gender and anchored their responses to their first evaluation. The second resume is evaluated more favorably if the second candidate is female and less favorably if the second candidate is male. However, we remain confident that the experiment on the first resume is not affected by desirability bias.

[23]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?”

Table 9: Resume Evaluation Results

	(1) Competence	(2) Likeability	(3) Likelihood of Hire	(4) Log Salary
Female Resume	-0.000946	0.0392	-0.0870	-0.0400
	(0.127)	(0.113)	(0.155)	(0.0454)
Resume Version	0.246*	0.0336	-0.103	0.0219
	(0.127)	(0.113)	(0.155)	(0.0453)
Constant	3.466***	3.759***	4.121***	8.078***
	(0.111)	(0.0984)	(0.135)	(0.0397)
Observations	263	263	263	264

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

5.1.1 Estimating Equation

The resume evaluation allows us to test for an implication of discrimination from below: that due to positive selection of women into management positions, women with management experience will be positively selected, leading to no discrimination at the "top" of the labor market. We test for this using the following linear regression model:

$$Outcome_i = \alpha + \gamma_1 * FC_i + \gamma_2 * ResumeType_i + \epsilon_i \tag{2}$$

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, *ResumeType* is a control for which resume was given; and i represents subject.

5.2 Results

This section presents results from the resume evaluation experiment. As shown in Table 9, we surprisingly find no differential evaluation of resumes as a function of candidate gender. We do find that subjects are more likely to favor one type of resume, in particular with respect to competency, suggesting that subjects are paying attention to the quality of the resume when considering their responses. We additionally check for "resume comprehension" prior to asking the subject to evaluate the resume, confirming that candidates were aware of the gender of the candidate (95 percent correctly identified the candidate's resume). As an additional robustness check, we also show in

Table 10: Gender Wage Gap at Adama University

	(1) ln(Salary)	(2) ln(Salary)	(3) ln(Salary)
Female	-0.198***	-0.129***	-0.0861***
	(0.0234)	(0.0161)	(0.0197)
Tenure		0.0281***	0.0268***
		(0.00140)	(0.00168)
Years of education		0.0509***	0.0363***
		(0.00332)	(0.00402)
BA or higher		0.383***	0.337***
		(0.0262)	(0.0255)
MA or higher		0.395***	0.419***
		(0.0504)	(0.0647)
Constant	7.744***	6.701***	6.938***
	(0.0173)	(0.0403)	(0.281)
Work Unit FE	No	No	Yes
Observations	1685	1665	1665

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix tables 12 and 13 that there is no difference in extreme ratings of female v. male resumes. Female and male resumes are equally likely to receive both very low ratings and very high ratings. Thus, it is not the case that the lack of an average effect masks greater variance in evaluation of female or male resumes. Therefore, our lack of discrimination results suggests that though subjects were aware of the candidate's gender and were thoughtful about their responses, candidate gender had no impact on their evaluations.

We collect further evidence for the results of our model using administrative data from the human resources department. We begin by studying gender wage gaps in the entire set of administrative employees. In Table 10, Column 1, we show women earn about 19.8 percent less than men on average. This gap can be partially explained by job tenure and education (Column 2) and occupational sorting (Column 3), but the gap remains large and statistically significant at about 8.6 percent even after inclusion of these controls.

However, Table 11 shows that when we separate the sample by educational attainment, there is no gender wage gap among those with a BA or higher. Among those without a BA, the gender wage gap ranges from 13.4 to 18.8 percent, depending on controls (see β_1). But in each case, β_3 is

Table 11: No gender wage gap among the highly educated

	(1) ln(Salary)	(2) ln(Salary)	(3) ln(Salary)
(β_1) Female	-0.143***	-0.188***	-0.134***
	(0.0206)	(0.0174)	(0.0232)
(β_2) BA or higher	0.584***	0.278***	0.272***
	(0.0308)	(0.0328)	(0.0314)
(β_3) Female × BA or higher	0.123***	0.196***	0.127***
	(0.0436)	(0.0382)	(0.0397)
Other controls	No	Yes	Yes
Work Unit FE	No	No	Yes
Observations	1685	1665	1665
$\beta_1 + \beta_3$	-0.02	0.008	-.007
P-val.	0.613	0.819	0.830

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

positive and significant, and the sum $\beta_1 + \beta_3$ is small and statistically indistinguishable from zero, indicating no gender wage gap among the highly educated.

6 Conclusion

This paper uses a novel experimental design to study how leader gender influences the way individuals respond to leadership. We find a surprising pattern of results: while there is evidence for discrimination against female leaders when subjects have no other information about the leader, the gender gap reverses when the leader is presented as highly trained and competent. Conditional on signaling high ability, female leaders are *more* likely to be followed. Further, despite Ethiopia's poor performance on gender equity, and lower of levels of female educational attainment in general, we document a lack of discrimination in a resume evaluation experiment and no gender wage gap among the highly educated. This apparent contradiction—low levels of gender parity and education quality coupled with a lack of a gender gap among the elite—can be reconciled with a dynamic model of discrimination, in which the barriers to entry are higher for females, causing discrimination to disappear (or even reverse) at higher levels of educational attainment.

Our results in the experimental game, coupled with the results of our resume experiment and observational data, suggest that at higher levels of education and training, we may not find as much evidence of discrimination in outcomes. Importantly, however, this is not necessarily evidence of gender equality or lack of taste-based discrimination. Instead, selection into higher levels of education and training is different for women and men. If obtaining an advanced degree is harder for females, then conditional upon having an advanced degree, we may expect females to have greater ability than males. This model further suggests that as developing countries achieve gender parity in educational attainment, discrimination may begin to emerge at higher levels.

The discrimination we observe against female leadership in the absence of information is a potential explanation for why female representation in top management remains low globally despite large country-to-country variation in gender disparities in education and labor force participation. It also suggests that achieving gender parity in human capital accumulation is unlikely to solve underrepresentation of women in management.

References

African Development Bank. 2015. "Where are the women: inclusive boardrooms in Africa's top listed companies?" 1–119.

Ahmed, Shukri, and Craig Mcintosh. 2017. "the Impact of Commercial Rainfall Index Insurance: Experimental Evidence From Ethiopia."

Aigner, Dennis J., and Glen G. Cain. 1977. "Statistical Theories of Discrimination in Labor Markets." *Industrial and Labor Relations Review*, 30(2): 175.

Alesina, Alberto, Paola Giuliano, and Nathan Nunn. 2013. "On the Origins of Gender Roles: Womenand the Plough." *The Quarterly Journal of Economics*, 128(May): 469–530.

Beaman, Lori, Esther Duflo, Rohini Pande, and Petia Topalova. 2012. "Female leadership raises aspirations and educational attainment for girls: a policy experiment in India." *Science (New York, N.Y.)*, 335(6068): 582–6.

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. 2009. "Powerful Women: Does Exposure Reduce Bias? *." *Quarterly Journal of Economics*, 124(4): 1497–1540.

Becker, Gary. 1957. "The Economics of Discrimination. Chicago: Univ."

Bertrand, Marianne, and Esther Duflo. 2016. "Field Experiments on Discrimination." *NBER Working Paper Series*, 22014: 110.

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

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg. 2017. "The Dynamics of Discrimination : Theory and Evidence."

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *Quartely Journal of Economics*, 1753–1794.

Chattopadhyay, Raghabendra, and Esther Duflo. 2004. “Women as Policy Makers: Evidence from a Randomized Policy Experiment in India.” *Econometrica*, 72(5): 1409–1443.

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.

Delavande, Adeline, and B Zafar. 2013. “Gender discrimination and social identity: experimental evidence from urban Pakistan.” *FRB of New York Staff Report.*

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.”

Egan, Mark L, Gregor Matvos, and Amit Seru. 2017. “When Harry Fired Sally: The Double Standard in Punishing Misconduct.”

Fryer, Roland G. 2007. “Belief flipping in a dynamic model of statistical discrimination.” *Journal of Public Economics*, 91(5-6): 1151–1166.

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. 2017. “It pays to be a man: Rewards for leaders in a coordination game.” *Working Paper*, , (00008).

Guryan, Jonathan, and Kerwin Kofi Charles. 2013. “Taste-based or statistical discrimination: The economics of discrimination returns to its roots.” *Economic Journal*, 123(572): 417–432.

Iyer, Lakshmi, Anandi Mani, Prachi Mishra, and Petia Topalova. 2012. “The Power of Political Voice: Women’s Political Representation and Crime in India.” *American Economic Journal: Applied Economics*, 4(4): 165–193.

Kalsi, Priti. 2017. "Seeing is believing- can increasing the number of female leaders reduce sex selection in rural India?" *Journal of Development Economics*, 126: 1–18.

Landsman, Rachel. 2017. "Gender Differences in Executive Departure."

Niederle, Muriel, and Lise Vesterlund. 2011. "Gender and Competition." *Annual Review of Economics*, 3(1): 601–630.

Sandberg, S, and N Scovell. 2013. *Lean in: Women, Work, and the Will to Lead. A Borzoi book*, Alfred A. Knopf.

Sarsons, Heather, Mitra Akhtari, Kai Barron, Anke Becker, Kirill Borusyak, Amitabh Chandra, Sylvain Chassang, John Coglianese, Oren Danieli, Krishna Dasartha, Ellora Derenoncourt, Raissa Fabregas, John Friedman, Roland Fryer, Peter Ganong, Edward Glaeser, Nathan Hendren, Simon Jäger, Anupam Jena, Jessica Laird, Jing Li, Jonathan Libgober, Matthew Rabin, Hannah Schafer, Frank Schilbach, Nihar Shah, Andrei Shleifer, Jann Spiess, Neil Thakral, Peter Tu, Shosh Vasserman, and Chenzi Xu. 2017. "Interpreting Signals in the Labor Market: Evidence from Medical Referrals." 1–71.

World Bank. 2017. "World Development Indicators."

A 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 7: Subject Compensation Schedule

B 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.

C Resume Experiment Robustness Checks

Table 12: Resume Evaluation Results: Lowest rating

	(1) Poor Competence	(2) Poor Likeability	(3) V. Unwilling to Hire
Female Resume	0.000333	0.00958	0.0116
	(0.0296)	(0.0199)	(0.0336)
Resume Version	-0.0309	0.0230	0.0227
	(0.0296)	(0.0199)	(0.0336)
Constant	0.0762***	0.0104	0.0628**
	(0.0259)	(0.0174)	(0.0294)
Observations	263	263	263

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Table 13: Resume Evaluation Results: Highest rating

	(1) Excellent Competence	(2) Excellent Likeability	(3) V. Willing to Hire
Female Resume	-0.0102	0.0233	0.0346
	(0.0445)	(0.0508)	(0.0620)
Resume Version	0.0441	-0.000647	0.0282
	(0.0445)	(0.0508)	(0.0620)
Constant	0.135***	0.202***	0.448***
	(0.0388)	(0.0444)	(0.0541)
Observations	263	263	263

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.