Discrimination from Below: Experimental Evidence on Female Leadership in Ethiopia*

Shibiru Ayalew Shanthi Manian Ketki Sheth

We propose and test an understudied explanation for the under-representation of women in management: gender discrimination by subordinates may reduce the effectiveness of female leadership. In a lab-in-the-field experiment in Ethiopia, subjects discriminate against female leaders, causing female-led subjects to perform worse. Subjects also give lower evaluations to hypothetical female managerial candidates. However, when the leader is presented as highly competent, subjects are *more* likely to follow women. We show that this reversal implies statistical discrimination, and can be explained by a model where the same signal is interpreted differently for each gender.

JEL Codes: O10, O15, J71

*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: Arsi University (shibekoo84@gmail.com), Manian: Washington State University (shanthi.manian@wsu.edu), Sheth: University of California Merced (ksheth@ucmerced.edu).

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. Representation falls even further in low-income countries (African Development Bank, 2015).[1] In addition to equity considerations, these gaps suggest that the productivity potential of the labor force is not fully utilized. Existing explanations for these gaps have often focused on supply-side differences between male and female candidates.[2] 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.

Successful performance in management and leadership depends in large part on how well others adhere to one's advice and direction. Thus, even if women are equally skilled and have similar leadership styles, female-led teams may perform worse if team members are less likely to heed their advice. 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 also implies that even if a woman alters her leadership style or increases her human capital, she may still fall short of her male counterparts. However, little well-identified evidence exists on whether individuals in the workforce respond differently to women due to gender discrimination, and evidence is particularly scarce for developing countries.

Using a novel lab-in-the-field experiment in Ethiopia, we study whether individuals respond differently when they are randomly assigned to a male versus female leader. We use a unique sample of high-skilled employees who are unfamiliar with research experiments.

[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] In developing countries, supply-side explanations typically focus on lower educational attainment or skill accumulation among women (African Development Bank, 2015). In high income countries, the literature has discussed differences in preferences or leadership styles, and the notion that women are less likely to "lean in" and go after management positions (Niederle and Vesterlund, 2011; Sandberg and Scovell, 2013). In addition, a large literature documents discrimination from "above" in the hiring and promotion processes (Bertrand and Duflo, 2016; Neumark, 2018).

Importantly, our design allows us to hold leader ability and communication style 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. Strikingly, although the female and male leaders are otherwise identical, we find that subjects are 10 percent less likely to follow the same guidance when provided by a woman rather than a man. As a result, female-led subjects earn fewer total points, a reduction of 0.34 standard deviations.

Interestingly, the gender gap in the experiment is not only mitigated, but is actually reversed, in a cross-randomized information treatment where subjects are told that their leader is highly trained and competent. The observed pattern allows us to characterize the discrimination as statistical, where beliefs about a group are used to solve a signal extraction problem, and rule out "taste-based" discrimination, in which individuals simply dislike female leadership (Becker, 1957; Aigner and Cain, 1977; Guryan and Charles, 2013).[3] Moreover, we show that this reversal can be explained by a model in which the same information about leader ability is interpreted differently for men versus women.

We then provide additional evidence for discrimination from below using a hypothetical resume evaluation experiment. Subjects provided lower evaluations of female candidates for a senior management position, despite the resumes being identical with candidate gender randomly assigned.

We close with a discussion of the implications of discrimination from below for the representation of women in senior management. We adapt the canonical model of Coate and Loury (1993) to show that because discrimination from below reduces the performance of female-led teams, women with the same ex-ante qualifications as men are less likely to be promoted. In addition, women who nevertheless succeed in attaining management positions will be positively selected—that is, the underlying ability of an accomplished woman will

[3]We do not claim that these beliefs are necessarily accurate reflections of differences between the two groups; 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.

generally be higher than the underlying ability of an accomplished man.

We contribute to the literature in four ways. First, we provide clean evidence for discrimination from below, an understudied form of discrimination. Second, we provide evidence on the existence and patterns of gender discrimination in leadership and labor markets in a low-income country, where the literature is particularly scarce. Third, we show that these patterns are driven by statistical discrimination, whereas most of the literature in low-income country contexts focuses on violations of social norms. And fourth, we document a reversal of gender discrimination conditional on an ability signal. The primary existing explanation for reversals of discrimination relies on the dynamics of discrimination; we demonstrate that discrimination reversals can occur outside a dynamic setting.

Several recent experiments have documented differential responses by individuals to randomly assigned female versus male leaders, advisers, and experts, particularly in low-income countries.[4,5] Similarly, observational studies in high-income countries show that female experts are more likely to be punished for negative shocks, even when they are random.[6] The consistent differential response to women documented in this recent literature raises the question of *why* individuals are responding differently to women. Are they prejudiced against women? Are they relying on their beliefs about women on average, because they do not have enough information about the ability of individual women? Or are they responding to

[4]For example, Yishay et al. (2018) show that subjects assigned to female trainers in Malawi are less likely to adopt a new agricultural technology, despite the fact that female trainers are more knowledgeable. Macchiavello, Menzel and Woodruff (2014) find that female manager trainees in Bangladeshi garment factories are seen as less effective, along with suggestive evidence that their production lines under perform. Hardy (2018) shows that female businesses receive fewer customers because they are demand-constrained in Ghana. In an artefactual field experiment in India, Gangadharan et al. (2016) find that males are less likely to adhere to the suggested contribution of a female leader in a public goods game.

[5]There is also an earlier literature on gender discrimination in low-income countries primarily focused on early childhood investments and son preference (Bharadwaj and Lakdawala, 2013; Jayachandran and Kuziemko, 2011; Jayachandran, 2015; Jayachandran and Pande, 2017), documenting gender gaps in the labor market (Jensen, 2012; Heath, 2014; Heath and Mobarak, 2014; ILO, 2016), and exploring how gendered networks and peers create and perpetuate gender gaps in the labor market (Beaman, Keleher and Magruder, 2017; Field et al., 2016; Hardy, 2018).

[6]See Egan, Matvos and Seru (2017), Landsman (2017), and 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. In high income countries, there is also an extensive literature on gender discrimination in other contexts, such as hiring and promotion decisions, evaluations, and credit or rental offers (Bertrand and Duflo, 2016; Mengel, Sauermann and Zölitz, 2017; Boring, 2017).

other characteristics that are correlated with being female? The answer leads us to different policy solutions: should policies focus on improving gender attitudes and relaxing gender norms, on improving signals of ability, or on making women more similar to men, such as by increasing female educational attainment or confidence when asserting their authority?

Because existing experiments document differential responses to gender in natural settings, the men and women in their samples often differ on a number of characteristics in addition to gender. This is especially true for the studies in low-income countries, where gender differences tend to be larger. For example, Macchiavello, Menzel and Woodruff (2014) find that randomly assigned female manager trainees are seen as less effective, but are also younger, less experienced, less educated, less interested in being promoted, and have more children than their male counterparts. And in a lab experiment in the United States that also finds differential responses to advice by gender, Grossman et al. (2017) provide leaders with "talking points", but encourage them to provide the advice "in their own words". In addition to the observed differences between genders in many of these studies, a significant literature documents average differences by gender in communication style, confidence, and risk preferences (see Niederle (2016) for a review), all of which are likely to influence how others respond to female authority.

Our tightly identified, lab-based evidence of discrimination from below is a strong complement to these experiments documenting differential responses to female leadership. We advance the literature by showing that individuals are responding to gender itself, as opposed to correlates of gender. Our results yield support to interpreting the gaps documented in field experiments as statistical gender discrimination; likewise, the field experiments highlight the external validity and real-world consequences of our lab-based findings.[7]

[7]In addition, because discrimination from below is discrimination to those in more senior positions, it is difficult to test using correspondence or audit studies, and in most cases, field experiments cannot hold constant the myriad of differences across genders. Thus, a lab-in-the-field experiment is a particularly useful method to provide clean identification of such discrimination. There is a psychology literature that has used lab experiments to study discrimination toward female leaders, primarily in high-income countries, but generally does not involve real stakes. See Eagly (2013) for a review, and Beaman et al. (2009) for an example in India.

We also advance the literature by finding support for statistical discrimination, and not taste-based discrimination, in a context with rigid gender norms and high gender inequality, like many low-income countries. To date, the growing literature on gender discrimination in low-income countries has largely characterized discrimination as a consequence of strong gender norms or of violations of those norms. An exception is Beaman, Keleher and Magruder (2017), who also find that gender differentials in job referrals in Malawi are more consistent with statistical discrimination.

Moreover, although our results are broadly consistent with statistical discrimination, the reversal of discrimination that we document is *not* consistent with the simplest and most standard model of statistical discrimination in which beliefs are normally distributed, ability signals are uncorrelated with gender, and subjects update their beliefs using Bayes' rule. While our design does not allow us to pin down which of these assumptions is violated, we show that a model in which signals are interpreted differently by gender can explain our results. A recent paper by Bohren, Imas and Rosenberg (2017) provides one explanation for why signals might be interpreted differently by gender: dynamic discrimination. In an elegant online experiment, they find a reversal in gender discrimination conditional on an ability signal, and show that this can be explained by subjects accounting for discrimination faced by women in obtaining the ability signal.[8] We document a similar reversal in which the ability signal was interpreted more favorably for women, but 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. Thus, our results, combined with those of Bohren, Imas and Rosenberg (2017) may suggest a broader phenomenon in which subjects respond particularly favorably to women of high ability, especially in contexts like Ethiopia where women generally face barriers to attaining skills or accolades. Importantly, both our

[8]Although this online experiment was conducted in an advice-giving context (an online math forum), they do not find evidence for discrimination from below. The study finds no gender discrimination in the ratings of answers, only in questions posed, and there is no inherent hierarchy among users and raters. Nevertheless, it provides a convincing theory and empirical results for why gender discrimination may reverse conditional on the same signal of ability.

results and those of Bohren, Imas and Rosenberg (2017) suggest that positive discrimination in favor of high-ability women does not preclude the existence of discrimination against women in the labor market more generally.

In summary, this paper provides a well-identified estimate of gender discrimination from below that *cannot* be attributed to unobservable differences between men and women, in an environment with real stakes. It additionally shows that this discrimination is statistical in nature and documents a reversal in discrimination conditional on a signal of ability.

The concept of discrimination from below and our results are important in several ways. First, this mechanism highlights that the performance metric itself may be a function of discrimination from below, and that women face differential barriers to effectiveness in leadership. While others have also described how objective evaluation metrics can be biased through discrimination (Glover, Pallais and Pariente, 2017), we provide a different, less discussed mechanism of how this may occur. We also show theoretically how this can explain the empirical fact that women are under-represented in senior management. Discrimination, and resulting gender gaps, may go undetected if this mechanism is not considered in anti-discriminatory policies. In particular, the mechanism of discrimination from below implies that providing anti-discrimination interventions only to those involved in hiring decisions may not be sufficient to remedy gender disparities. Ability information must be conveyed and believed by subordinates as well. Finally, discrimination from below highlights discriminatory concerns in advice-giving contexts more generally. If female advice is less likely to be followed when offered, then simply giving women the opportunity to "sit at the table" may not be sufficient to overcome gender disparities. In general, though we focus on the context of management in this paper, discrimination from below can generate gender disparities in any position in which successful performance requires individuals to follow one's advice or direction.

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 leadership game

and the supporting resume evaluation. In Section 4, we present our findings and in Section 5 we present a model of the dynamic implications of discrimination from below. Section 6 concludes and discusses policy implications of the results.

2 Theory

In this section, we develop a model incorporating both taste-based and statistical discrimination. We then generate testable predictions that will allow us to distinguish between these two sources of discrimination using our experimental results. 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 $\theta \sim N(\bar{\theta}_g, \sigma_g^2)$.[9] We focus on female and male managers 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 $E(\theta|g)$ 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 $f(E(\theta|g))$. We also allow the employee's utility from following the advice to depend directly on the manager's gender, as in a model of "taste-based" discrimination (Becker, 1957). Thus, the employee has utility function $u(g, f(E(\theta|g)))$. To focus on the core predictions of our model, we assume rational expectations, that utility is linear in payoffs, and that taste-based utility and utility from payoffs are additively separable. This yields $u(g, f(E(\theta|g))) = f(\bar{\theta}_g) - c_g$, where c is the "taste-based" cost associated with following each gender. We standardize the utility of not following the 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's directions:

[9] 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.

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

We allow employees to be heterogeneous in these taste-based costs, where c_g has the cumulative distribution function $D_g(x)$. We assume that the taste-based cost of following a female manager first order stochastically dominates the taste-based cost of following a male manager: $D_f(x) \leq D_m(x) \; \forall \; x$.

Discrimination occurs when, for a male and female manager of equal ability θ and an employee with the information set $\mathbf{S}$, we have:

$$D_f(f(E(\theta|f, \mathbf{S})) < D_m(f(E(\theta|m, \mathbf{S}))$$

That is, discrimination occurs when employees are strictly less likely to follow the advice of a female manager than a male manager of equal ability.

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 ($\mathbf{S} = \emptyset$), both taste-based discrimination and statistical discrimination toward women result in employees being less likely to follow the female manager relative to the male manager.[10] 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.

[10]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 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 for a male and female manager 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 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 signal of high ability, 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 $E(\theta|s, g) \geq E(\theta|g)$ and the expected payoff from following the manager increases.

We now consider the role of a high ability signal 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 $f(E(\theta|s)) > c_m$ if the manager is male and $f(E(\theta|s)) > c_f$ if the manager is female.

Proposition 1 *Under only taste-based discrimination, $c_f > c_m$, signals of high ability cannot reverse the gender gap in following the manager.*

A high 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 gen-

der, any given expected payoff is weakly more likely to exceed the distaste for following a male manager than a female manager by assumption. Thus, under taste-based discrimination, the share following the female manager can never exceed the share following the male manager.

Proposition 1 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. Therefore, we focus on beliefs, the basis for statistical discrimination, for the remainder of this section. We now return to our initial assumption that the priors on ability may vary by gender. In this case, after observing a signal of high ability, the gender gap in beliefs is:

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

Holding taste preferences constant ($D_m(x) = D_f(x)$ for all x), any reduction in the gender gaps in beliefs will translate into a corresponding reduction in discrimination from below. If the prior is that male managers 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 reverse 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 managers:

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

However, in the special case of $s = \bar{\theta}_m$, that is, the signal indicates that the manager is of average male ability, even differences in prior variances in ability cannot reverse the gender gaps in beliefs. In such a case, the signal will have no effect of employees' response to a male manager, but will increase beliefs about the ability of a female manager.[11]

Proposition 2 *A signal indicating that a female manager is equal to the average male man-*

[11] 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., $s = \bar{\theta}_m$.

ager, $s = \bar{\theta}_m$, can reduce, but cannot reverse, the gender gap in following the manager.

The gender gap in following the manager can reverse only if there is a reversal in the gender gap in beliefs. When the signal indicates that the female manager is equal to the average male manager, $s = \bar{\theta}_m$, the gender gap in beliefs is $\lambda_f(\bar{\theta}_m - \bar{\theta}_f)$, which is weakly positive by assumption.

Discussion: understanding a belief reversal

Propositions 1 and 2 show that the standard models of taste-based and statistical discrimination we have considered so far cannot explain a reversal in beliefs when $s = \bar{\theta}_m$. Here, we provide one example of how a reversal can be obtained within our framework. We consider a model in which employees interpret the same signal differently based on the gender of the manager. As a simple example, 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.

There may be several reasons that employees would interpret the same signal differently for male and female manager. One is gender stereotypes (Bordalo et al., 2016): for example, employees may expect female managers to perform worse on math or logic problems. Another is the dynamic model of discrimination described by Bohren, Imas and Rosenberg (2017), where signals are interpreted differently because of barriers to entry in obtaining those signals. For example, in Ethiopia, as in many places around the world, barriers to entry for women in education are well documented. The World Economic Forum's 2016 Global Gender Gap Report ranked Ethiopia 132, out of 144 countries evaluated, for educational attainment. Thus, it may be rational in the Ethiopian context to infer different levels of ability for the same signal, such as an advanced degree.

If employees believe that the signal mean differs by gender, we then have:

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

For $s = \bar{\theta}_m$, the gender gap in beliefs is now $E(\theta|s,m) - E(\theta|s,f) = \lambda_f(s - \bar{\theta}_f) - (1-\lambda_f)\gamma_f$. 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. Thus, 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.

Summary of testable predictions

The model developed in this sections makes the following testable 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.

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

3. If there is taste-based discrimination only, under reasonable assumptions on preferences, a signal of high ability cannot reverse the gender gap in following the leader. Thus, a reversal indicates that discrimination is driven by beliefs.

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) 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 sample itself is quite novel: the subjects are high-skilled, 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.

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.[12]

Adama Science and Technology University (ASTU) is an elite public university located about 100 km from the capital, Addis Ababa. Table I shows summary statistics for all administrative employees at ASTU, based on institutional data from the human resources department. 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 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 in

[12]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 I: 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)	
Salary|BA or higher	3613.11	3681.16	3525.79	155.37
	(1624.55)	(1769.13)	(4161.84)	
Observations	1685	746	939	1685

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

Ethiopia were almost equally likely to be in the labor force, although the female labor force participation rate is about 15 percentage points lower overall (World Bank, 2017). We observe significant differences in job tenure and salary by gender: women have been with the institution longer but are paid less on average.

Importantly for the interpretation of our model, women in the sample 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 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.[13] Furthermore, the salary gap we observe on average disappears when limiting attention to those with advanced degrees.

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

3.2.1 Sample

Using a list of employees provided by the human resource department, we contacted all administrative employees with a BA or higher ($n = 500$)[14], and implemented the experiment until we reached 150 female subjects and 150 male subjects. Thus, relative to all university employees shown in Table I, those in the experiment were more educated, had higher salaries, and were balanced on gender. Within this sample, there is no salary or tenure differences across subject gender, though we continue to see that females have fewer years of education than males even conditional upon obtaining a bachelors degree.

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.[15,16]

[13]For example, in primary and secondary school, the gender parity index of gross school enrollment is 1. But for tertiary school, the gross enrollment gender parity index is .5 (World Bank, 2017).

[14]We restricted the experimental game to highly educated employees because we were concerned that the game may be too complicated for subjects with lower levels of education and literacy.

[15]Most eligible subjects who did not participate (about 40 percent) could not be located during the week of the study. Only one subject refused to participate.

[16]Unlike in the United States, recruitment of subjects in this lab-in-the-field experiment was not routine, making it difficult to increase our sample size to more than 300. There was no systematic recruitment pool or reliable method to recruit subjects in advance of the experiment. Instead, enumerators would go to the

3.2.2 Overview of design

The basic setup of the experiment is that subjects are randomly assigned to either a male or female "leader", subjects 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 identical across all leaders. In this way, we are able to hold the leader's behavior constant across male and female leaders.[17] 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 being able.[18]

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

In the logic game, subjects are asked to solve the Tower of Hanoi logic game, (see Appendix Figure A.1 for details of the puzzle and Appendix Figure B for compensation schedule).[19]

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.

[17]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 selected one male and one female leader who played in the same way and had the same outcomes to be matched to subjects. The purpose of using real individuals as leaders was to not deceive our subjects.

[18]We recognize that a manager's or leader's role is more than just providing advice, but we maintain the "leader" descriptor, instead of "advisor" for example, because the experiment was framed as a leader relationship with the "leader" explicitly described as such to subjects in Amharic. It may be the case that leadership performance on net does not exhibit gender differences even if there is a differential response to accepting advice, due to such effects being offset by a differential response to other leadership activities, such as monitoring.

[19]How well a person solves the puzzle is measured by the number of moves required, in which fewer moves are better. Prior to actually playing, we asked subjects how many moves they think *they* will require to solve the puzzle, how many moves they think *their leader* will require to solve the puzzle, and finally how

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 I: Signaling Game Payoffs (colors and expected payoffs not shown to subjects)

The second component was a signaling game adapted from Cooper and Kagel (2005). We selected this game because it has a clear correct answer, but it is quite complex and the correct answer is difficult to guess. This is particularly true for subjects with no previous exposure to game theory. Thus, there is a clear and important role for leader advice in this setting. 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 I.[20]

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

many moves they think their leader guessed *they* would require to solve the puzzle. These responses 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 did not appear to be an effective question for precisely eliciting beliefs. We therefore do not include these questions in our final analysis. In general, 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.

[20]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.

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 I). A similar logic could be applied to playing 4.

The leader provides advice to play strategically 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 computer 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, prior to being given any advice from their leader and without seeing 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 the 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. However, we note that 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.

The subject then played 10 rounds on the game. Prior to each round, the subject observes how their assigned leader played for that given round.[21] In addition, subjects are told that

[21]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

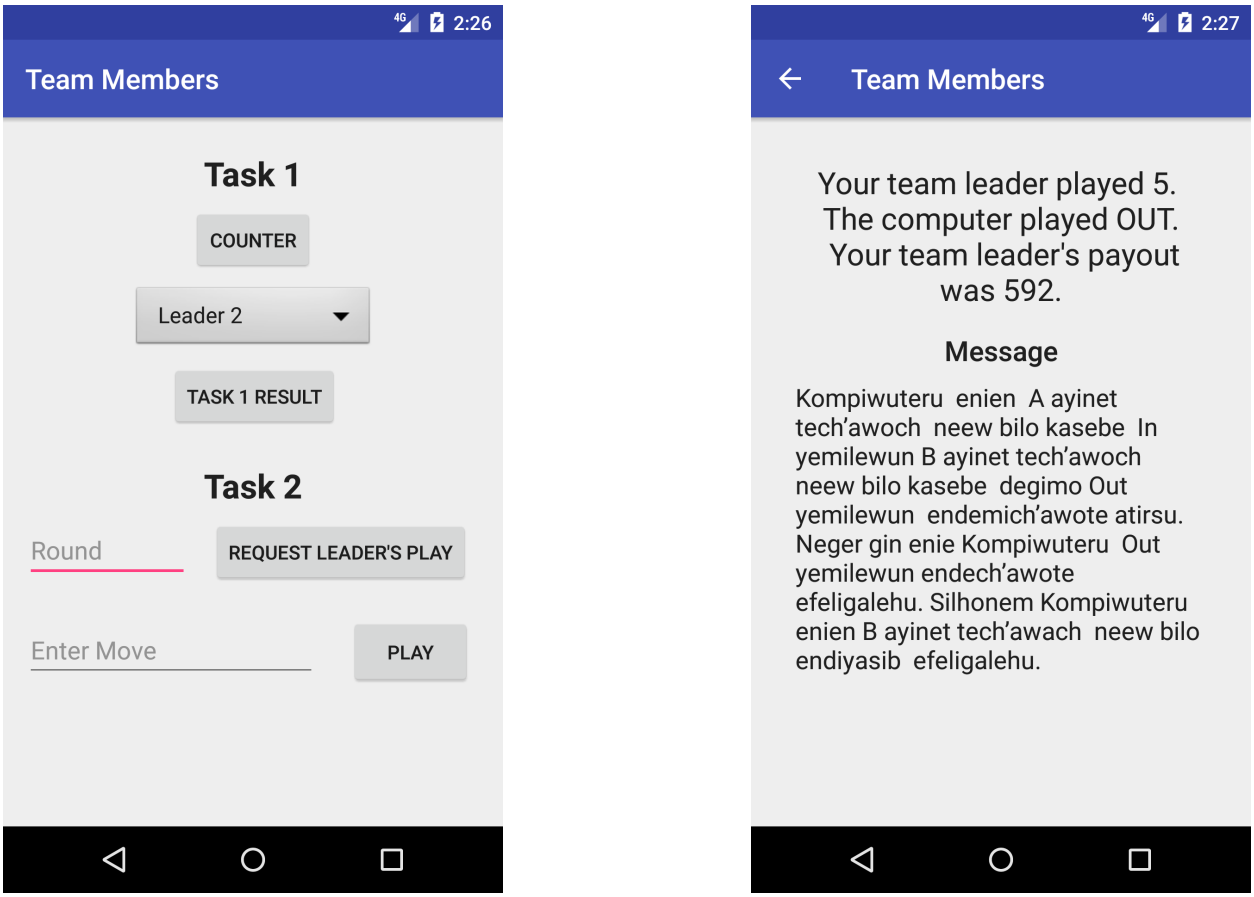

Figure II: Leader result and messages as shown to subjects

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.[22] The messages were displayed on an Android app by the enumerator (Figure II), and became increasingly informative over the rounds of the game. The enumerator additionally recorded the leader's play and outcome for each round on a piece of paper in front of the subject.

Figure III provides an overview of the experiment. We completed the game in a span of 6 days. Due to subjects potentially discussing the game with colleagues, we relabeled the choices for Day 5 and Day 6.[23] Specifically, Player 1 selected from two different sets of letters

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. Analysis on the sample of recruited leaders is not possible as less than 20 persons were recruited to be potential leaders.

[22] All leaders chose to send the messages.

[23] The purpose of the relabeling was not due to concerns of subjects discussing the purposes of the research experiment. Rather, due to the relatively significant monetary incentives and somewhat fun nature of

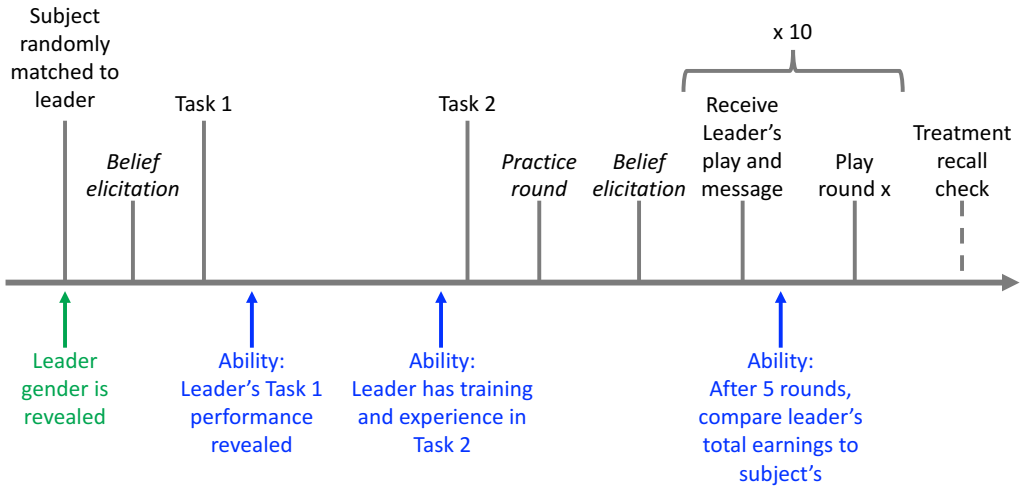

Figure III: Timeline of Leadership Game

for Days 5 and 6, and the computer 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 the leader being of high 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 high ability, and male leader with information on high ability.[24]

the logic games, there was a potential concern of subjects discussing strategies on maximizing payouts for the game, potentially attenuating estimated effects. We therefore relabeled choices for later days of the experiment. However, we observe no consistent differential pattern of choices for subjects playing later in the study. We are relatively confident that subjects were unaware of the purpose of the experiment as enumerators themselves were not made aware of the study's primary purpose, and upon introduction of the Treatment Recall Check we realized that the enumerators themselves had been unaware of the gender salience of the leader in their own scripts. In addition, treatment assignment was done in small batches such that there is balance across days, and all comparisons are made across subjects rather than within subjects.

[24]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 information on ability: n = 85.

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 pseudonyms[25] for the leader (mentioned 23 times in the enumerator's script) and 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 the end of the game 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 ability treatment consists of three components. First, after the "Tower of Hanoi" logic game, the enumerator informed the subject that the leader solved the puzzle with the minimum number of moves possible, and noted how many moves fewer this was than their own performance.[26] 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. And third, after 5 rounds of play, the enumerator totalled the points earned by the leader versus the subject to highlight the (expected) point advantage by their 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

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

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

Table II: 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

Standard errors in parentheses

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

assignments per male and female subjects 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, 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 using inverse probability weights.

Table II 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 (gender, salary, job level, education, and tenure). We also confirm pairwise balance in the bottom three rows of Table II.

In addition to balance across subject characteristics, we may be concerned that the pseudonyms we used to connote gender also contained information on other important char-

Table III: 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

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

acteristics (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).[27] We use 193 unique names to reduce the 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 III confirms that the characteristics associated with the pseudonym assigned to each subject in a given treatment are balanced across treatment arms.[28]

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

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

[28]The results in Table II and Table III are robust to the exclusion of day fixed effects.

Table IV: 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

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

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 IV 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 such gender gaps.

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 an indicator for playing strategically (i.e., selecting 4 or 5)[29] for subject i in round r (of 10 rounds). 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 high ability, and $FL * Ability$ is the interaction of the two indicators.[30] We additionally include an indicator of whether the individual chose to play strategically in their practice round selection, 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).

Based on our model, we note the following hypotheses:

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

- $\beta_2 > 0$: Informing subjects that the leader is of high ability increases the likelihood that subjects follow the leader's directions.

- $\beta_3 > 0$: The return to a signal of high ability is higher for female leaders that for male leaders. That is, the gender gap in following the leader narrows in the ability treatment.

- $\beta_1 + \beta_3 < 0$: The gender gap in following the leader conditional on receiving a signal of high ability reduces, but does not eliminate, the gender gap. Recall from Section

[29] 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). However, our results are also qualitatively similar when using an indicator for selecting 5 only as the dependent variable.

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

2 that a reversal in the gender gap, i.e., $\beta_1 + \beta_3 > 0$ and $\beta_1 < 0$, is not consistent with a model of taste-based discrimination. In addition, if $\beta_2 = 0$, this suggests that $s = \bar{\theta}_m$: the signal indicated that the leader was of average male ability. In such a case, models of statistical discrimination predict that an unbiased signal will mitigate, but not reverse, the gender gap. Thus, if we do observe a reversal of the gender gap, it is consistent with statistical discrimination in which the signal is being interpreted differently for men and women.

3.3 Resume Evaluation

Upon completion of the experimental game, we implemented a resume evaluation experiment that began the following week. 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. This resume evaluation exercise is an additional test of discrimination from below in that the large majority of our subjects are low-level administrative employees, and the job description represents one of the most senior management positions in the organization.[31]

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, indicating that they read the resumes carefully. Subjects were randomly assigned one of four possible resumes: two different "candidates" that were designed to be comparable in quality, each of which was presented as either representing a

[31] After completing the resume evaluation, subjects were immediately given a detailed survey questionnaire on gender and the workplace. There was thus a greater concern that discussion about the survey by subjects, again partly driven by the relatively large monetary incentive given for participation, would reveal that the underlying purpose of the research was related to gender. For this reason, we did not introduce the resume or survey until the lab experiment was completed among all subjects.

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

male candidate or a female candidate. 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.[32]

After reviewing the resume and completing the comprehension test, subjects evaluated

[32]In the experiment, 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. These results are shown in Appendix Table A.2. Thus, 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 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.

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.[33]

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.[34] We successfully followed up with 74 percent of the experimental subjects who complete the resume evaluation component in its entirety.[35,36] Table V confirms the validity of our randomization by documenting that subject characteristics were balanced across treatment arms.

The resume evaluation provides an additional test of gender discrimination by lower-level employees towards potential managers. We test for this using the following linear regression model:

$$Outcome_i = \alpha + \gamma_1 * FC_i + \gamma_2 * ResumeType_i + \epsilon_i \quad (2)$$

where *Outcome* is competence, likeability, hireability, or salary offer (in logs); FC is an indi-

[33]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?"

[34]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.

[35]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 marginal 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 based on resume experiment completion.

[36]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 V: Resume Experiment Balance

	(1) Fem. subject	(2) ln(Salary)	(3) Level	(4) Years Ed.	(5) MA or higher	(6) Job tenure
Female Version A	0.0160	-0.0897	-0.574	0.00447	-0.00639	449.8
	(0.0953)	(0.0685)	(0.503)	(0.110)	(0.0545)	(375.2)
Male Version B	-0.0364	-0.0596	-0.431	-0.0780	-0.0390	-467.2
	(0.0962)	(0.0682)	(0.493)	(0.100)	(0.0501)	(350.7)
Female Version B	-0.00975	-0.0219	-0.223	-0.0448	-0.0224	-92.62
	(0.0958)	(0.0724)	(0.512)	(0.105)	(0.0524)	(394.6)
Observations	225	225	225	225	225	225
p-val: Fem = Male	0.752	0.598	0.607	0.792	0.886	0.122

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

cator of whether the resume was randomly assigned to be a female candidate, *ResumeType* is a control for which of the two "candidate" resume was given; and i represents subject. The coefficient of interest is γ_1. Based on our model, we hypothesize $\gamma_1 < 0$.[37]

4 Results

4.1 Leadership Game

Table VI, Column 1, shows our primary results from estimating equation (1).[38] We find that in the absence of information on ability, subjects with female leaders were 6 percentage points less likely to play in accordance with their leader's directions (see β_1). Relative to subjects with male leaders and no information on ability, this reflects a 10 percent reduction in adherence to the leader's recommendation. This pattern of gender discrimination is further supported by the resume evaluation, discussed in further detail in the following section.

[37]The pre-analysis plan uses a different estimating equation based on within subject comparisons; however, as previously discussed, we use across subject comparisons due to evidence of social desirability bias in evaluations of the second resume.

[38]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 VI: Leadership Game Results

Dependent Variable:	Strategic Play		
	(1)	(2)	(3)
	All Rounds	Round 1	Rounds 1-5
(β_1) Fem. Leader	-0.0590*	-0.0573	-0.0813**
	(0.0352)	(0.0822)	(0.0406)
(β_2) Ability	-0.00301	-0.0353	-0.0461
	(0.0350)	(0.0781)	(0.0399)
(β_3) Fem. leader × Ability	0.115**	0.274**	0.147***
	(0.0479)	(0.113)	(0.0551)
Day FE	X	X	X
Round FE	X		X
Practice round	X	X	X
Observations	3020	302	1510
Control group mean	0.618	0.479	0.614
$\beta_1 + \beta_3$	0.0561	0.217	0.0657
P-val.: $\beta_1 + \beta_3$	0.0891	0.00583	0.0825

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

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 suggests that the signal indicated an ability level approximately equal to the expected group mean for 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.

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 after receiving information that leader was of high 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 contained in the signal is identical.

This pattern of discrimination against female leaders in the absence of ability information, and a reversal of discrimination with ability information, emerges from the first round of play. Columns 2 and 3 of Table VI present results for earlier rounds in the game (Round 1 and Rounds 1-5) to highlight that the discrimination begins early and that learning appears to reduce the effects in later rounds.[39] The coefficient estimate on discrimination from below (β_1) is remarkably stable across rounds; while it is not statistically significant in the first round due to lower power, it is statistically significant for rounds 1-5. The large return to ability signals for female leaders (β_3) diminishes over rounds.

The discrimination against female leaders in the absence of ability information is costly. In the absence of information on high ability, having a female leader reduced total points earned by .34 standard deviations, which is statistically significant at the 5 percent level. In contrast, when provided information on high ability and the discrimination from below is reversed, we no longer observe a statistically significant difference in performance by leader

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

Table VII: Beliefs about leaders

Dependent Variable:	Leader's performance
	(1)
(β_1) Fem. Leader	-5.812
	(9.056)
(β_2) Ability	6.362
	(9.527)
(β_3) Fem. leader × Ability	14.39
	(12.98)
Day FE	X
Observations	301

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

gender.[40]

We estimate our results separately for male and female subjects in Appendix Table A.1. Though less precise, the estimates suggest that the general pattern is quite robust across subject genders.[41]. If anything, the reversal of discrimination appears to be somewhat stronger among female subjects.[42]

Our estimates of belief expectation on how well the leader will perform in Task 2 can also act as a robustness check for our results, and for our conclusion that the results are more consistent with statistical discrimination. Unfortunately, the belief expectation exercises were difficult for subjects to understand and thus were likely very noisy estimates of belief. However, as Table VII shows, the pattern of the magnitudes of the beliefs elicited for Task 2 align with the pattern of following the leader's directions in Table VI. Female leaders (relative to male leaders) were expected to perform more poorly (i.e., lower expected value) when no

[40]However, the only reason for this difference between the subject's selection and their final points earned is chance, since there was randomness in how the computer responded to each play.

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

[42]Using the decision to play 5 as the dependent variable, we see much stronger results for female subjects - $\beta_3 = .213$, is statistically significant at the .01 level, and is statistically different from β_3 for male subjects, which falls to 0. This suggests that female subjects were more likely to mimic the leader, and were more sensitive to female leaders and if female leaders were presented as high-ability.

Table VIII: Resume Evaluation Results

	(1) Competence	(2) Likeability	(3) Likelihood of Hire	(4) Log Salary Offer
Female Resume	-0.0732	-0.0286	-0.152	-0.124**
	(0.118)	(0.108)	(0.142)	(0.0518)
Male Resume Mean	3.75	3.89	4.28	8,085 Birr
Observations	225	225	225	225

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Regression specifications include the resume version as a covariate.

information was provided on ability—their expected performance was 5.81 fewer points. However, when leaders were presented as high-ability, female leaders' expected performance was 8.58 more points than male leaders.[43] 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 a similar pattern to our primary results is suggestive of the robustness of our results in Table VI.

4.1.1 Resume Evaluation

The discrimination we observe in the absence of high ability information is supported by our results from the resume evaluation experiment. On all measures, female candidates were evaluated more poorly than male candidates. Female candidates were rated less competent, less likeable, less likely to be hired, and were offered a 12 percent lower salary. Only this last result is statistically significant, at the 5 percent level.We observe no statistically significant difference by candidate gender when using only within subject variation across two different resumes. We expect this is due to social desireability bias, based on how subjects rated the second resume relative to the first as shown in Appendix Table A.2. However, we should expect discrimination to be difficult to detect and results to be relatively imprecise given the crude evaluation measures. Nonetheless, the pattern of lower evaluations of female candidates

[43]These estimated effects on leader's expected performance use the same estimating model as in VI.

is quite stark, and consistent across all measures, providing additional evidence of employees discriminating against potential female managers relative to male counterparts.[44,45] The lack of a gender wage gap among those who hold advanced degrees at the university suggests that the difference in salary offered is less likely to reflect differences in expectations of the candidate's outside option.

This exercise differs from typical correspondence studies in that our sample is not involved with human resources or hiring decisions. Instead, we interpret our results as suggestive survey evidence on how the subjects may generally view managers.

5 Dynamic Implications of Discrimination from Below

In the preceding sections, we have documented the existence of discrimination from below and shown that it is driven primarily by statistical discrimination—that is, beliefs that women have lower ability than men on average. Now, we discuss the theoretical implications of discrimination from below for the representation of women in management positions. We show that discrimination from below can generate disparate promotion probabilities for male versus female managers even when the employer is unbiased. In addition, we show that female managers who are promoted are positively selected.

We adapt Coate and Loury (1993) to demonstrate the implications of discrimination from below on promotion probabilities and selection of managers. 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

[44] We do not observe statistically significant differences by subject gender.

[45] Among those who did not respond to salary (39 subjects), the same pattern is observed for competency and likelihood of hiring, though likeability goes in the opposite direction, and all results remain statistically insignificant. Our results are also robust to using enumerator reported treatment, as opposed to assigned treatment. And finally, the estimated effects from the experimental game display the same pattern when restricted to this subsample.

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 improve 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 advice of female managers due to discrimination, as shown above. As in our experiment, this reduces the performance of teams led by both qualified and unqualified female managers relative to teams led by male managers of equal ability. We assume $F_{q,m}(\phi) \leq F_{q,f}(\phi)$ and $F_{u,m}(\phi) \leq F_{u,f}(\phi)$ for all ϕ.

Now suppose employers are unbiased and know that the share π of both male and female 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 3 *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 $1 - \left[(1-\pi)F_{u,f}(\underline{\phi}) + \pi F_{q,f}(\underline{\phi})\right]$ and men are promoted with probability $1 - \left[(1-\pi)F_{u,m}(\underline{\phi}) + \pi F_{q,m}(\underline{\phi})\right]$. The difference between men and women in promotion probabilities is $(1-\pi)(F_{u,f}(\underline{\phi}) - F_{u,m}(\underline{\phi})) + \pi(F_{q,f}(\underline{\phi}) - F_{q,m}(\underline{\phi}))$, which is strictly positive by assumption.

Proposition 4 *Promoted female managers are more likely to be qualified than promoted male managers.*

A promoted female manager is more likely to be qualified than a promoted male manger if:

$$\frac{(1-F_{q,f}(\underline{\phi}))\pi}{(1-F_{q,f}(\underline{\phi}))\pi+(1-F_{u,f}(\underline{\phi}))(1-\pi)} > \frac{(1-F_{q,m}(\underline{\phi}))\pi}{(1-F_{q,m}(\underline{\phi}))\pi+(1-F_{u,m}(\underline{\phi}))(1-\pi)}$$

Simplifying, this condition holds when the gender gap in team performance is smaller for qualified than unqualified managers:

$$\frac{1-F_{q,f}(\underline{\phi})}{1-F_{q,m}(\underline{\phi})} > \frac{1-F_{u,f}(\underline{\phi})}{1-F_{u,m}(\underline{\phi})}$$

Thus, this section shows that discrimination from below can generate both under-representation of women in senior management, and positive selection of female leaders in high level management positions. Given our finding that discrimination from below is statistical in nature, an interesting implication of this last result is that conditional on obtaining a high enough management position, female leaders may see a reduction or even a reversal in discrimination from below.

6 Conclusion

This paper uses a novel experimental design to study how leader gender influences the way individuals respond to leadership. We find striking evidence for discrimination against female leaders: subjects are less likely to follow the same advice from a female leader than an otherwise identical male leader. Because this discrimination from below lowers the performance of the female-led team, our results raise concerns about how best to evaluate female leaders and highlight a tension between gender equity and successful performance. Discrimination from below implies that qualified women may not be promoted even when those making

the promotion decision do not engage in discrimination. Thus, the paper highlights that in general, performance metrics that are based on subordinate or client responsiveness may be problematic in reaching equity goals.

We also show the gender gap in following the leader reverses when the leader is presented as highly trained and competent. Conditional on signaling high ability, female leaders are *more* likely to be followed. We show that this pattern of empirical results implies statistical discrimination. Despite strong gender norms and severe gender inequality in Ethiopia, we find that a general distaste for taking advice from females cannot explain our results. Instead, our results imply that subjects are using gender as a proxy for quality of the advice.

Global development goals have focused on improving gender parity in low-income countries, making it particularly important to understand the role and sources of gender discrimination in the labor market in these countries. Our results suggest that in order to achieve improvements in gender equity in developing countries, it is not sufficient to change norms about the appropriate roles for women in society; beliefs about women's ability must also change.

Finally, we show that discrimination from below implies positive selection of female leaders who succeed in rising to higher levels of a hierarchy. 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.[46] 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.

The discrimination we observe against female leaders is a potential explanation for why female representation in top management remains low globally despite large country-to-country variation in gender norms, female educational attainment and female labor force participation. Our results suggest that discrimination from below will be most prominent at

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

lower stages in the management pipeline, and reduce for those women who are able to move up the pipeline.

Given the statistical nature of this discrimination, our findings imply that providing women with credible signals of their ability and skill that can be communicated widely can improve their performance by reducing such discrimination from below. It follows that sensitivity training should not be limited to only those who hire and evaluate employees, but changing gendered beliefs of *all* employees is important for reducing gender equities. A better understanding of how ability can be communicated to a broad audience is an important area for future research.

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.

Beaman, Lori, Niall Keleher, and Jeremy Magruder. 2017. "Do Job Networks Disadvantage Women? Evidence from a Recruitment Experiment in Malawi." *Journal of Labor Economics*, forthcomin: 1–49.

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.

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.

Bharadwaj, Prashant, and Leah K. Lakdawala. 2013. "Discrimination Begins in the Womb: Evidence of Sex-Selective Prenatal Investments." *Journal of Human Resources*, 48(1): 71–113.

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.

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.

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."

Field, Erica, Seema Jayachandran, Rohini Pande, and Natalia Rigol. 2016. "Friendship at Work: Can Peer Effects Catalyze Female Entrepreneurship?" *American Economic Journal: Economic Policy*, 8(2): 125–153.

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.

Glover, Dylan, Amanda Pallais, and William Pariente. 2017. "Discrimination as a Self-Fulfilling Prophecy_ Evidence from French Grocery Stores _ Amanda Pallais." *The Quarterly Journal of Economics*, 132((3)): 1219–1260.

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.

Hardy, Morgan. 2018. "If she builds it, they won't come: The gender profit gap." *VoxDev.org*.

Heath, Rachel. 2014. "Women's Access to Labor Market Opportunities, Control of Household Resources, and Domestic Violence: Evidence from Bangladesh." *World Development*, 57: 32–46.

Heath, Rachel, and A. Mushfiq Mobarak. 2014. "Manufacturing Growth and the Lives of Bangladeshi Women."

ILO. 2016. *Women at Work: Trends 2016 - Executive Summary*. Vol. 42.

Jayachandran, S., and I. Kuziemko. 2011. "Why Do Mothers Breastfeed Girls Less than Boys? Evidence and Implications for Child Health in India." *The Quarterly Journal of Economics*, 126(3): 1485–1538.

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

Jayachandran, Seema, and Rohini Pande. 2017. "Why Are Indian Children So Short? The Role of Birth Order and Son Preference." *American Economic Review*, 107(9): 2600–2629.

Jensen, R. 2012. "Do Labor Market Opportunities Affect Young Women's Work and Family Decisions? Experimental Evidence from India." *The Quarterly Journal of Economics*, 127(2): 753–792.

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

Macchiavello, Rocco, Andreas Menzel, and Christopher Woodruff. 2014. "Managerial Capital and Productivity: Evidence from a Training Program in the Bangladeshi Garment Sector."

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. 2017. "Gender Bias in Teaching Evaluations."

Neumark, David. 2018. "Experimental Research on Labor Market Discrimination." *Journal of Economic Literature*, 56(3): 799–866.

Niederle, Muriel. 2016. "Gender."

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. 2017. "Interpreting Signals in the Labor Market: Evidence from Medical Referrals." 1–71.

World Bank. 2017. "World Development Indicators."

Yishay, Ariel Ben, Maria Jones, Florence Kondylis, and Ahmed Mushfiq Mobarak. 2018. "Are Gender Differences in Performance Innate or Socially Mediated ?"

For Online Publication

A Tower of Hanoi

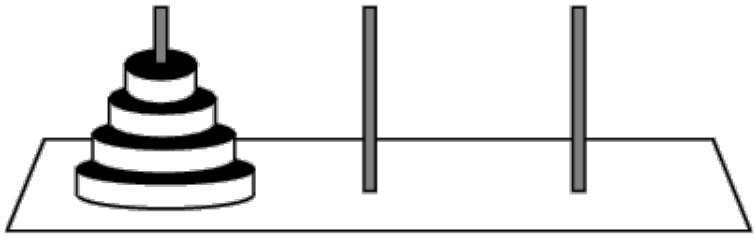

Figure A.1: Tower of Hanoi

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 Leadership Game Heterogeneity

Table A.1: Leadership Game: Results by subject gender

Dependent Variable:	Strategic Play		
	(1)	(2)	(3)
	All subjects	Male Subjects	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

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

E Resume Experiment Robustness Checks

Table A.2: Resume Evaluation Results: Social Desireability Bias

	(1) Competence	(2) Likeability	(3) Likelihood of Hire	(4) Log Salary Offer
Male Resume	0.0763	0.0300	0.151	0.124**
	(0.120)	(0.108)	(0.142)	(0.0518)
Reviewed Second	0.222*	0.103	0.255*	0.114**
	(0.124)	(0.112)	(0.139)	(0.0543)
Male * Reviewed Second	-0.237	-0.142	-0.402*	-0.227**
	(0.210)	(0.193)	(0.242)	(0.0992)
Observations	450	450	445	441

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered at the subject level and are in parentheses. Male is an indicator for the resume belong to a randomly assigned male candidate. Reviewed Second is an indicator for whether the candidate was reviewed second.