Distorted Signals and Evaluator Bias:
A Two-Stage Experiment on Gendered Hiring in Malawi*

Gabriella Fleischman† and Mansa Saxena‡

January 9, 2026

Click here for the latest version

Abstract

Female labor force participation is constrained by supply-side frictions (lack of qualified female applicants) and demand-side frictions (biases in hiring decisions). How do supply- and demand-side distortions interact to affect gender gaps in employment? We partner with a firm to conduct two experiments: The first experiment addresses supply-side behavioral gender gaps during a recruitment drive through female-directed advertising, while the second experiment is a resume audit study among evaluators to isolate biases in hiring. Despite no treatment effects on the objective skills of male or female applicants, the treatment has the perverse effect of leading to a *reduction* in female hiring. In the resume audit study, evaluators place greater weight on noisy signals of soft skills for women than men when shortlisting candidates, while, in the recruitment experiment, the treatment leads low-ability women to inflate their signals of soft-skills. Evaluators subsequently shortlist lower-ability women in treated areas, who crowd out higher-ability female candidates from the shortlisted pool. Our estimates suggest that mitigating *either* supply-side distortions in application signals or evaluator bias in signal weighting would close the gender gap in the ability of shortlisted candidates by 51-55%, with only modest additional gains to mitigating both.

*We are thankful for useful comments and suggestions from seminar participants at the Northwestern University Development Economics Workshop, the Harvard Kennedy School Economic and Social Policy Workshop, Advancements in Field Experiments 2025 and NEUDC 2025. This study received IRB approval from Harvard University protocol number IRB23-1605. This experiment was pre-registered under RCT ID AEARCTR-0015284.

†Fleischman: Harvard Kennedy School, gfleischman@g.harvard.edu

‡Saxena: Northwestern University

1 Introduction

Female labor force participation (FLFP) has been a central issue to international development and economic policy in low- and middle-income countries throughout the twenty-first century, resulting in notable decreases in gender gaps in labor force participation (Heath and Jayachandran, 2017). However, even in places where norms permit women to take up jobs and gender gaps in education are closing, gender gaps persist, particularly in *formal* labor force participation (Arbache et al., 2010). Women are underrepresented in the labor force due to a lack of qualified female applicants (supply-side) (Angrist and Evans, 1998; Bertrand et al., 2015), and biases in hiring decisions (demand-side) (Beaman et al., 2018; Buchmann et al., 2024; Goldin and Rouse, 2000). Our study asks: how do supply-side and demand-side constraints *interact* to affect the gender gap in employment? Does alleviating constraints on the supply side amplify or mitigate the importance of demand side biases in hiring decisions?

We conduct two sequential experiments to study constraints to female employment in the formal sector in Malawi. The first experiment focuses on reducing behavioral gender gaps that screen women out or affect how they fill out application forms, such as confidence gaps and self-stereotyping, through female-directed advertising. Our second experiment is a resume audit study among evaluators in the firm where we use real applications from stage one and manipulate application features, while holding qualifications constant, to isolate biases in hiring evaluations. Together, these experiments allow us to observe how demand-side biases interact with supply-side constraints to affect female hiring.

First, we partner with a firm in Malawi to investigate supply-side barriers to formal female labor force participation by implementing a cluster-randomized controlled trial. We experimentally vary the job advertisements that the firm posts during a recruitment drive across geographic areas, where each geographic area has one open position. The treatment area advertisements are designed to encourage more female applicants. The treatment advertisement implicitly or explicitly conveys three pieces of information to potential applicants: (1) the position is currently held by both men and women, (2) men and women are equally likely to report that the job is less challenging than they expected, and (3) that the firm is interested in specifically recruiting women.

The treatment advertisement leads to economically meaningful, but noisy and non-detectable, increases in the number of female applicants in treatment areas, and no changes in qualifications along objective margins. However, application evaluators are more likely to screen in *less* objectively qualified female applicants in treated areas for a limited number of interviews, while screening in *more* qualified men. Women in treated areas ultimately perform worse on interviews and written tests, and are less likely to be hired. We hypothesize that the treatment advertisement encourages less qualified female applicants to change the way they signal skills, particularly by putting more effort into signaling their soft-skills, which are not objective nor verifiable. As a result, the correlation between soft-skill signals and objective qualifications declines for female applicants in treated areas, which we refer to as a "signal distortion" effect.

In the presence of evaluator bias in screening, that is, when evaluators apply different weights to soft-skill signals for male and female applicants and are more likely to rely on soft-skills signals when screening women, signal distortion leads evaluators to screen in less-qualified female applicants, while screening remains more skill-based for other groups.

To test this hypothesis, we implement a hypothetical resume audit study among the application evaluators within the firm. While the randomized resume experiment is hypothetical, evaluators are informed that their responses will be used to determine policies and procedures to improve evaluation and recruitment processes at the firm, and thereby have an incentive to respond honestly in order to ultimately make their own jobs easier. Evaluators each review six applications drawn from the recruitment drive almost one year after the recruitment concludes. We randomly select these applications from real applications from the recruitment drive, and then we randomly vary six key variables: whether or not the candidate (1) can speak English, (2) can use a smartphone, (3) lists two references (rather than one reference), (4) uses the term 'Mr.' when listing their references' names, (5) has a more expensive phone plan (as indicated by the first two digits of the phone number on the application), and (6) is male or female. We consider English-speaking and smartphone capability, which are both immediately relevant to the job, as objective, technical skills. We consider listing two references and using honorifics such as 'Mr.' in reference names as signals of soft skills. Finally, we vary applicants' phone plan to analyze evaluators' potential for implicit bias with respect to the applicant's income, which might otherwise confound how they interpret smartphone capability.

While we find that evaluators consistently reward English proficiency, evaluators appear to interpret these signals differently by gender: among women, soft-skills signals, particularly listing two references, increases composite evaluation scores when they *can not speak English*, while among men, evaluators only reward such signals if they *can* speak English. Furthermore, there is suggestive evidence that evaluators use smartphone capability as a screener for male applicants, whereas they never use smartphone capability to screen female applicants.

Taken together, these results suggest that areas with the treatment advertisement are less likely to hire female candidates because of a combination of latent evaluator bias and treatment-induced signal distortion, or the erosion of the correlation between the most- and least-informative signals. Because evaluators overweight less informative signals (soft-skills signals) when screening women, even when they are no longer predictive of underlying ability, this leads to the selection of a shortlisted female candidate pool that is, on average, less objectively qualified.[1] Ultimately, the firm conducts hiring on the basis of interview and test scores, for which objective skills are highly predictive.

Because we have causal estimates of the effects of supply-side behavioral distortions *and* demand-side biases on who is selected for shortlisting, our treatment design enables us to ask: how would women have fared in treatment areas if their application signals were weighted the same way as men's? First, we compute a total application score for each applicant, weighting

[1] We define a signal as informative if it is predictive of hiring decisions. We do not observe job performance.

each application signals with the weights applied to the randomized signals for male applicants in the resume audit study. We construct a counterfactual group of shortlisted candidates: all candidates who rank in the top-five on the male-weighted application score within their geographic area. We then compare predicted aptitude scores of men and women who are actually shortlisted with men and women who are shortlisted using the unbiased counterfactual decision rule. Using an unbiased decision rule improves the selection of female shortlisted candidates and closes the gender gap in the ability of shortlisted applicants by 55% without changing the average ability or gender composition of the applicant pool. This effect is indistinguishable from the effect of mitigating supply-side behavioral distortions: the control condition closes the gender gap in predicted aptitude scores by 51%. In other words, the negative impact of the treatment on the shortlisting of high-ability women is mitigated when their applications are evaluated in the same way that men's applications are evaluated, suggesting that supply-side distortions only bite when evaluators are biased.[2]

We highlight three central contributions of our paper. First, we contribute to the literature on gender gaps in employment. While many papers study supply-side and demand-side constraints in isolation, our paper is unique in evaluating the importance of the interaction between the two, and is one of few papers in the literature that observes behavior under varying conditions on both sides of the market (Boring et al., 2025). We find that, in our setting, the gender gap in hiring induced by the treatment is *entirely* explained by the interaction between demand-side bias and supply-side distortions: a simple counterfactual exercise suggests that supply-side distortions only lead to worse shortlisting decisions when evaluators are biased. Furthermore, mitigating behavioral distortions on *either* the supply or demand side would significantly improve the selection of shortlisted women relative to men, and mitigating both has only modest additional benefits. This finding implies that, for some sets of frictions, policy-makers or firms need only resolve distortions on one side of the market to make significant gains in equitable hiring.

Our experiments also make important contributions to our understanding of supply-side and demand-side constraints to FLFP in isolation. We find powerful evidence that evaluator bias can act as an important constraint on formal FLFP, contributing to a small literature identifying the underlying causes of firm bias against hiring women in lower- and middle-income countries (Beaman et al., 2018; Buchmann et al., 2024; Macchi and Raisaro, 2025).[3] We also expand on a large literature identifying supply-side constraints to FLFP.[4] While numerous studies identify

[2]Similarly, applying male weights to female application signals in control areas has only a modest additional benefit for women. This may explain why the firm had not created strict evaluation guidelines. In status quo, men's and women's more- and less-informative signals are highly correlated, so evaluators' biases only minimally alter their ultimate shortlisting decisions.

[3]Our results are most consistent with Macchi and Raisaro (2025), which finds that evaluators value difficult-to-observe soft-skills, such as trust. In the absence technologies to improve observing these signals, firms evaluate male and female trustworthiness differentially.

[4]Studies show that FLFP is constrained by forces such as gender roles and gender norms (Bertrand et al., 2015; Bursztyn et al., 2020, 2017; Kleven et al., 2019; Borker et al., 2021; Boudet, 2013; Dean and Jayachandran, 2019; Giuliano, 2020; McKelway and Lowe, 2024), human capital accumulation (Autor et al., 2016; Fiala et al., 2022; Gallus and Heikensten, 2020; Goldin et al., 2006; Archibong and Annan, 2017; Behrman and Knowles,

the role of confidence gaps and self-stereotyping in differential employment outcomes in high-income countries (Bordalo et al., 2019; Coffman et al., 2024; Exley and Kessler, 2022; Lundeberg et al., 1994; Möbius et al., 2022; Niederle and Vesterlund, 2007; Samek, 2019), we are one of few papers that evaluate behavioral mechanisms such as confidence and self-stereotyping in low- and middle-income countries (McKelway, 2025).

Second, we contribute to a growing literature on information asymmetries about applicant quality as an important hiring friction in lower-income countries. Our results are consistent with a growing literature that finds providing firms with credible signals of applicant quality can shift interviewing and hiring decisions (Carranza et al., 2022; Fernando et al., 2023; Macchi and Raisaro, 2025). Fernando et al. (2023) finds that expanding the pool of applicants through advertising in an online hiring portal in India has no effect on firm hiring *unless* it is paired with a screening service. Merely expanding the applicant pool – even when this means increasing the number of qualified applicants – does not change hiring outcomes because large applicant pools are a significant burden on firm capacity. We replicate this pattern in a very different context, and show that this phenomenon has implications for diversity in hiring. Applicants adjust how they present themselves in response to recruitment interventions, exacerbating information frictions.

Lastly, we contribute to the literature on statistical discrimination and affirmative action policy. We show causal empirical evidence of the implications of the theoretical model proposed by Fershtman and Pavan (2021), whereby soft affirmative action policies – or affirmative action policies that commit to more diverse *candidates*, rather than more diverse *hires* – can backfire, and ultimately *reduce* the probability of hiring a diverse candidate. Consistent with the theoretical predictions of Fershtman and Pavan (2021), soft affirmative action backfires as a result of evaluator's difficulty assessing minority candidates coupled with a lack of clear evaluation procedures. We find that the principal reason why evaluators struggle to assess applications from minority candidates is because they weigh signals differently across candidates, putting heavier weights among minority candidates on signals that are inherently less informative—or, signals where there is more variation between the signals and the true characteristics that they approximate. This is an important and policy-relevant insight, because this is a choice that evaluators make that can be corrected, rather than a fundamental difference in the noisiness of signals across majority and minority candidates. We also document that, in our case, candidates respond to soft affirmative action policies in a way that distorts signals, making the realizations of those inherently uninformative signals even *less* informative, thereby exacerbating the role of evaluator biases (the condition that gives rise to statistical discrimination).

The paper proceeds as follows: Section 2 describes design of the recruitment experiment and the randomized resume experiment, Section 3 presents results from both experiments, Section 4 discusses the results, and Section 5 concludes.

1999; Björkman-Nyqvist, 2013; Rakshit and Sahoo, 2023), and differential job search (Eriksson and Lagerström, 2012; Cortés et al., 2021; Jalal, 2025; Jensen, 2012).

2 Experimental Design

We partner with a non-profit firm in Malawi that provides trainings and loans to small-holder farmers. First, we ran an experiment in March 2024. In this experiment, we randomize advertisements that are posted across geographic areas during a recruitment drive for the Field Officer position. Then, in February 2025, we ran a resume audit study with the firm's employees who had ever evaluated Field Officer applications.

Field Officer Position

The Field Officer job is an entry-level position in the firm. One Field Officer oversees all of the beneficiary farmers and on-the-ground activities within a "section", a geographic area used for agricultural services in Malawi that is comprised of a small collection of villages.[5] Field Officers are responsible for recruiting new beneficiary farmers, enrolling farmers in benefits and loan products each year, and ensuring farmer compliance with farming techniques. These elements of the job require that a Field Officer possesses social skills to engage face-to-face with farmers on a day-to-day basis, and the leadership skills to be taken seriously by farmers and respected members within their communities. Field Officers are also responsible for ensuring farmer's timely repayment of loans. Farmers repay their loans digitally directly to the firm, so Field Officers never handle cash directly, but this element of the job does mean that trustworthiness and ability to handle uncomfortable situations are also key job criteria.

While "soft skills" are perhaps more important job qualifications than technical skills, there are four objective job qualifications required and listed on the recruitment advertisement: secondary school passing grades (an MSCE certificate), ability to read and write in English, ability and willingness to travel by bicycle, and being above eighteen years old. In practice, field officers also need to be able to use a tablet or smartphone, but this is an easier skill to train and therefore is considered a preferred skill rather than a required skill. Field Officers are sometimes hired without an MSCE certificate in exceptional circumstances.

We conducted a survey of 206 existing Field Officers in January 2024, two months before the recruitment drive experiment. Prior to our recruitment experiment, 38% of Field Officers are women (Table C.1). In the survey, we ask participants questions about the top stressors they face in their professional and personal lives, the most important perks of the job, and how the job compares to their prior expectations. Men and women are both likely to state that top job stressors include meeting repayment targets (ranked 3.04 out of 5 on a stress scale), traveling by bicycle between farmers (ranked 3.06 out of 5 on a stress scale), and non-compliant farmers (ranked 2.99 out of 5 on a stress scale) (Table C.7). 46% of men and 41% of women report that, taken as a whole, the job is harder than they had anticipated when applying for the job. Much of this challenge comes from the unexpected difficulty of riding a bike between farmers each day, which 53% of men and 51% of women report is more challenging than they expected.[6]

[5] In the recruitment drive, each section receives applications from residents of 10 villages on average (25th percentile = 7, median = 9, 75th percentile = 14).

[6] Field Officers have to bike far distances between farmers, sometimes in challenging weather conditions such

However, Field Officers report that the interpersonal aspects of the job are usually easier than they expected. 74% of men and women say that recruiting and training farmers is easier than they expected, and 82% percent of men and women say that earning the respect of farmers and other community members is easier than they expected. More details on the Field Officer survey results are in Appendix Section C.

2.1 Recruiting Experiment

In 2024, the firm expanded their services into 59 new areas of Malawi where they did not previously operate. They ran a recruitment drive in March 2024 to fill 59 open Field Officer positions (one per section). The firm recruits for the Field Officer position by enlisting Village Chiefs to advertise the job, holding town halls, and posting paper advertisements within schools and on trees around villages including markets.

Randomized Advertisements

We conduct a cluster-randomized controlled trial (RCT) across agricultural sections to test whether gender-coded job advertisements highlighting the on-the-job experiences of current employees influence applicant behavior. The first part of the advertisement is held constant across all sections and includes information about the firm, the job role, requirements, and benefits. The experimental variation is introduced in the final section of the advertisement, which features an infographic differing by treatment arm:

Control (30 sections) – Infographic with gender-neutral content and imagery

Treat W (14 sections) – Gendered information included in the advertisement, with graphics maximizing the female-coding

Treat M (15 sections) – Gendered information included in the advertisement, with graphics minimizing the female-coding

We reproduce the advertisements used in each experimental arm in Appendix A. The Control advertisement (Figure A.1) includes gender-neutral imagery and phrasing. The treatment advertisements—Figure A.2 for Treat W and Figure A.3 for Treat M—disaggregate the same statistic by gender and differ only in visual emphasis.

The control advertisement still includes status-quo efforts to encourage female applicants. It features the statement "Women are welcome and encouraged to apply" at the top and includes a dedicated "Benefits for Women" section outlining maternal and child benefits. The only element that varies across treatment arms is the infographic panel at the bottom of the advertisement, which presents a statistic from the January 2024 Field Officer survey. In the Control

as intense direct sunlight or heavy rain, often on muddy roads.

advertisement, the panel states:

"4 out of 5 Field Officers say that ***earning the respect*** *of the community and* ***engaging with farmers*** *as a Field Officer was* ***easier*** *than they expected"*

Below this statistic is a graphic image depicting five gender-neutral human-like figures, with four of them shaded to reflect the statistic.

In the treatment arms, the infographic is split into two columns, showing the same statistic disaggregated by gender. In the Treat W condition, the female statistic appears on the left; and in Treat M, the male statistic appears on the left. The statistic itself remains the same across genders, differing only in phrasing: "4 out of 5 women" versus "4 out of 5 men". Each column features five gender-coded figures (female or male), with four shaded to reflect the statistic.

The treatment advertisement conveys two key messages: first, that women are already employed as Field Officers at the firm; and second, that male and female Field Officers report similar experiences on the job. Additionally, the use of gender-coded graphic figures may visually reinforce the firm's interest in hiring women—particularly in the Treat W condition, where female-coded figures appear on the left-hand side (in this setting, people read left-to-right, so this increases the salience of female representation). The small change between the Treat W and Treat M conditions allows us to use left-hand-side bias—or, the tendency for people who read left-to-right to process left-hand-side information with higher salience than right-hand-side information (Román et al. (2015)) to disentangle the roles of pure information from norms.

Recruitment Format

After evaluators collect applications in each section, they screen applications for shortlisting. Although evaluators know that there are two different recruitment advertisements, they are not aware of the treatment status of each section at the time of screening. Evaluators shortlist five candidates in each section, regardless of the number of applications that the section receives. The firm practices soft affirmative action at this stage, giving preferential treatment to female applicants in shortlisting. Then, the firm meets the applicants in person for a written aptitude test and an interview. The test and interview are conducted in English. Interviews include standard interview questions, as well as role-play scenarios where the applicants are asked to respond to hypothetical situations as if they are a Field Officer and the interviewer is a farmer. One applicant is hired from each section, almost always strictly on the basis of the test and interview scores.

Data

We evaluate the efficacy of the treatment advertisements using two primary datasets: all of the applications that the firm receives across Treatment and Control areas (1,255 applicants), and the test and interview scores achieved by the candidates who are shortlisted and come to their in-person interview (289 candidates). We also ask shortlisted candidates to fill out a short questionnaire to provide insight into how the advertisements change the selection or

perceptions of candidates. We cannot use these responses to make any definitive statements about how the Treatment advertisement works, because the treatment also could change the way that evaluators select candidates for shortlisting. However, we use candidate's answers to this questionnaire to provide suggestive evidence about the mechanisms behind the Treatment advertisement effects, which we discuss in Section 4.

Outcomes

We evaluate a range of outcomes corresponding to key stages in the hiring pipeline, as well as measures of applicant qualifications and signal quality. First, we analyze application volume outcomes by measuring the total number of applications received in each section, disaggregated by gender. Second, we construct two composite indices that each collect related variables from the application. One index is comprised of variables that we consider to be highly "informative"—those where, in expectation, any given realization of that signal closely approximates the real-world characteristic that it maps onto. The other index is comprised of variables that we consider to be "uninformative"— those where, in expectation, any given realization of that signal less reliably approximates the real-world characteristic that it maps onto. We call these indices "objective technical skills" and "soft-skills signals".

Our index of *objective technical skills* is comprised of verifiable application features that reflect job-relevant qualifications (Table 1). As an example of our concept of "informativeness", if somebody writes on the application that they speak English, this most likely maps on to English-speaking capability with a reasonable amount of consistency. To construct the index, we first standardize each component using the mean and standard deviation of male applicants in control areas. We then take the unweighted average across the seven standardized components to form a composite index of objective technical skills for each applicant. Next, we regress this index on district and strata fixed effects, clustering standard errors at the section level. We extract the residual from this regression and re-standardize it using the control group distribution (mean and standard deviation among male applicants in control areas). This residualization ensures that comparisons across treatment arms are not confounded by geographic variation in applicant quality and is used in subsequent analyses and figures.

Our index of *signals of soft skills* is comprised of variables that may positively indicate an applicant's social capital, leadership, or work ethic through the application, but that are not verifiable or strictly job relevant on their own (Table 1). To illustrate why signals of soft skills are less "informative" according to our definition, consider an applicant who indicates that she heard about the job from the village chief. In doing so, she may be signaling that she is a trustworthy and well respected leader in the community. However, an equally respected and trustworthy leader who heard about the job through a family member would be no less qualified. This index captures variation in social connectedness, application effort, and leadership signaling, but small adjustments in effort can distort the relationship between these signals and true underlying leadership and connectedness (without resorting to lying, as would have to be the case to distort signals of objective skills). Similar to the objective skills index, we stan-

Table 1: Index Components

Index	*Variables*
Objective Technical Skills	Ability to use a smartphone
	Ability to cycle long distances
	Ability to speak English
	Ability to speak another local language
	Completion of secondary school (MSCE)
	Completion of secondary school before age 20
	Any education beyond MSCE
Signals of Soft Skills	The number of questions answered on the application form
	An indicator for if the applicant listed the maximum number of references allowed
	An indicator for if the applicant listed any reference with a formal title (e.g., "Mr." or "Mrs.")
	An indicator for if the applicant heard about the job from, or listed as a reference, the village chief
	An indicator for if the applicant uses a more-expensive phone plan (identifiable through the phone number)
	An indicator for if the applicant listen reference with a TNM number

dardize each component using the control group distribution, compute the unweighted average across components, and then residualize the index on district and strata fixed effects. The resulting residual is re-standardized using the control group mean and standard deviation, and used in all subsequent analyses.

We then evaluate a series of hiring outcomes. At the screening stage, we consider a binary indicator for whether the applicant is *shortlisted*. Once applicants are shortlisted, they are invited to take an in-person aptitude test and appear for an interview. The aptitude test assesses applicants' arithmetic skills, including basic calculations related to loan amounts and repayment balances. The interview is conducted by a panel of the firm's staff and combines personal interaction with on-the-spot role-play exercises that simulate real scenarios encountered on the job. We report results on standardized *aptitude test score* and an *interview score*. Finally, we analyze a binary *hiring* outcome indicating whether the applicant was ultimately selected for the job.

2.2 Randomized Resume Experiment

To examine how evaluator make shortlisting decisions in this setting, we conduct an unincentivized audit experiment using mock applications. Evaluators are informed that the survey aims to help the firm improve its recruitment processes, that their responses will remain anonymous, and that their responses will not be linked to their individual performance.

Experimental Variation

The mock applications are based on real submissions from the firm's March 2024 Field Officer

recruitment. We randomly select nine real applications and randomly assign each six of them to each evaluator for review. Within each application, we independently randomize six applicant attributes: sex (male or female); the first two digits of the phone number, which signals the phone carrier (a potential proxy for income); number of references (one or two, as a signal of soft skills or social capital); whether the applicant lists their references' names with formal honorifics (e.g.,"Mr." or "Miss", another soft-skill or social capital signal); ability to use a smartphone (a technical skill and key job component); and ability to speak English (a technical skill and job requirement). After reviewing each application, evaluators are asked a series of questions to elicit their interpretation of the applicant's qualifications and potential to be a successful Field Officer. This design allows us to isolate the causal effect of each randomized signal on evaluator perceptions, and to test whether there is evaluator bias wherein different weights are placed on these signals by applicant gender.

Sample

Our sample consists of 57 evaluators, one third of whom are female. For two-thirds of the sample, their highest level of education secondary school completion, which is the same level of education required for the Field Officer job. The remaining third of evaluators hold a tertiary degree. At least half of all Field Officers who participated in the 2024 recruitment drive are included in this sample. Approximately one-quarter of the evaluators had never previously participated in recruitment, but were slated to be an application evaluator for the firm's upcoming 2025 recruitment drive.

Outcomes

For each mock application, evaluators answer a standardized set of questions: (1) if they would shortlist the candidate; (2) to predict their interview and aptitude test performance; and (3) to rate the applicant's expected social skills, technical skills, and work ethic on a Likert scale. First, they indicate if they would hypothetically shortlist the candidate ("Based on the application above, would you shortlist this candidate?"). While this question is most directly comparable with an evaluators actions when they shortlist a candidate, we do not limit the number of applicants that they can shortlist from the six applications that they review. This is distinct from the actual shortlisting procedure, where evaluators always choose five applicants in each section.

Thus, in addition to hypothetical shortlisting, we ask evaluators to predict the candidates' interview score and aptitude test score (on a scale of 1-10). This allows us to rank-order all of the applicants that each evaluator reviews, and thus determine which candidates are crowded out from the top of the pool. Finally, evaluators rate their expectation of the applicant's technical skills, social skills, and work ethic. These assessments are collected on a five-point Likert scale, ranging from "completely lacking any skill (ethic)" to "very high levels of skill (ethic)". We convert these responses to numeric values ranging from 1 (lowest) to 5 (highest) to create standardized outcome variables. These six evaluator-reported outcomes constitute

our primary outcome measures.

2.3 Empirical Specifications

2.3.1 Recruiting Experiment

We estimate treatment effects using two primary specifications. First, we estimate outcomes at the section level using OLS with strata and district fixed effects, and robust standard errors. Second, we estimate applicant-level outcomes using OLS with the same fixed effects and standard errors clustered at the section level to account for correlated outcomes within treatment clusters.

$$Y_s = \beta_0 + \beta_1 \text{TreatW}_s + \beta_2 \text{TreatM}_s + \gamma_s + \delta_d + \varepsilon_s \tag{1}$$

Here, Y_s denotes the outcome of interest at the section level (e.g., total number of female applicants). TreatW_s and TreatM_s are binary indicators for whether section s is randomly assigned to the Treat W or Treat M recruitment advertisement, respectively. We include strata fixed effects γ_s and district fixed effects δ_d.

$$\begin{aligned} Y_{is} = \beta_0 &+ \beta_1 \text{TreatW}_s + \beta_2 \text{TreatM}_s + \beta_3 \text{Female}_i \\ &+ \beta_4 (\text{TreatW}_s \times \text{Female}_i) + \beta_5 (\text{TreatM}_s \times \text{Female}_i) \\ &+ \gamma_s + \delta_d + \varepsilon_{is} \end{aligned} \tag{2}$$

Here, Y_{is} denotes the outcome of interest for applicant i in section s (e.g., smartphone ownership, education level, or a composite skill index). TreatW_s and TreatM_s are binary indicators for whether the applicant's section s was assigned to the Treat W or Treat M advertisement. Female_i is a binary indicator equal to 1 if the applicant is female. The interaction terms $\text{TreatW}_s \times \text{Female}_i$ and $\text{TreatM}_s \times \text{Female}_i$ capture whether the impact of the treatment varies by gender. We include strata fixed effects γ_s and district fixed effects δ_d to absorb residual variation across geographic areas. Standard errors are clustered at the section level, the level of randomization.

2.3.2 Randomized Resume Experiment

We analyze our randomized resume experiment using the following empirical specification:

$$\begin{aligned} Y_{ia} = \alpha_0 &+ \alpha_1 \text{Female} + \alpha_2 \text{TwoRef} + \alpha_3 \text{ProfRef} + \alpha_4 \text{Smartphone} \\ &+ \alpha_5 \text{English} + \alpha_6 \text{TNM} + \delta_a + \delta_o + \theta_i + \varepsilon_{ia} \end{aligned} \tag{3}$$

In equation (3), Y_{ia} denotes the evaluation outcome for application a scored by evaluator i (e.g. a shortlist indicator, predicted interview score, or a 1–5 rating of technical ability,

social skills, or work ethic). The coefficients α_1–α_6 capture the causal contribution of each randomly assigned applicant attributes: *Female* (= 1 if the applicant's name is female), *TwoRef* (= 1 if two referees are listed), *ProfRef* (= 1 if referees are presented with formal titles), *Smartphone* (= 1 if the applicant reports being able to use a smartphone), *English* (= 1 if the applicant reports speaking English), and *TNM* (= 1 if the application includes a TNM mobile number, a locally salient income signal). The constant α_0 is the baseline evaluation for a male applicant that omits all six signals. Three sets of fixed effects absorb systematic variation that is orthogonal to the experimental treatment: δ_a are application fixed effects (holding constant the underlying objective qualifications of each base application), θ_i are evaluator fixed effects (controlling for differences in average stringency across screeners), and δ_o are review-order fixed effects (capturing learning or fatigue as evaluators progress through their six applications). The disturbance term ε_{ia} collects idiosyncratic factors affecting evaluator i's assessment of application a. Standard errors are clustered at the evaluator level to allow arbitrary correlation across the multiple ratings submitted by the same screener.

Secondly, we consider: how do evaluators interpret each signal for male versus female applicants? More precisely, we ask, for men and women, do evaluators incorporate signals of soft skills (less-informative) *instead* of or *conditional* on objective skills (more-informative)? We consider English-speaking as our "objective skills" measure because it is a non-negotiable skill for the job. Furthermore, the aptitude test and interview are both conducted in English, so shortlisting or highly-ranking any non-English-speaking candidate indicates significant bias towards other favorable attributes. To answer this question we conduct the following analysis for the male and female sub-samples separately:

$$\begin{aligned} Y_{ia} = {} & \omega_0 + \omega_1 \text{English} + \omega_2(\text{English} \times \text{Attribute}) \\ & + \omega_3 \text{TwoRef} + \omega_4 \text{ProfRef} + \omega_5 \text{Smartphone} + \omega_6 \text{TNM} \\ & + \delta_a + \theta_i + \varepsilon_{ia} \end{aligned} \tag{4}$$

In this specification, Y_{ia} denotes the outcome of the evaluation of the application a reviewed by the evaluator i. The variable Attribute_{ia} is one of the randomized skill indicators, specifically, either *TwoRef*, *ProfRef*, *Smartphone* or *TnM*, interacted with the English-speaking indicator *English*. All randomized attributes are included as additive controls, and the regression is estimated separately for male and female applicants. Application fixed effects δ_a control for the base resume assigned to each evaluator and evaluator fixed effects θ_i account for variation in evaluator stringency. Standard errors are clustered at the evaluator level.

3 Results

This section presents findings from the recruiting experiment and resume audit study. We first show that gender-targeted recruitment messages modestly shift applicant pools, encouraging more women to apply without reducing average female applicant qualifications. However, in

sections with female-directed recruitment, evaluators ultimately select less-qualified women, leading to worse hiring outcomes. We then explore mechanisms, using both the recruiting experiment and resume audit study to show that this backfiring is driven by two phenomenon that compound on one another: "evaluator bias", whereby evaluators consider less-informative soft-skill signals from female applicants in lieu of the more-informative objective skills, but only consider these soft-skills signals for male applicants *conditional* on objective skills; and "signal distortion", whereby women change the way that they fill out the applications in response to the treatment, reducing the cross-signal correlation between over-weighted, less-informative soft-skills signals and under-weighted, more-informative objective skills.

3.1 Recruiting Experiment

Candidate Gender Composition

First we evaluate the efficacy of the treatment advertisement in encouraging more women to apply, the objective it was designed for. In Control sections, 7.1 women and 17.4 men apply on average. Both Treat W and Treat M sections receive one additional application from a woman on average, which is a 16% increase (Table 2). While we are not powered to be able to statistically distinguish this change in the number of applications, this is still an encouraging result for such a light-touch experiment. Interestingly, Treat W sections received 0.8 *fewer* applications from men (a 4.6% decrease), and Treat M sections received 1.6 *more* applications from men (a 9.4% increase). Again, these figures are not statistically distinguishable from Control or from one another, but it suggests that men might be sensitive to the gender-coding of the job. Taken together, there is no change in the number of applications in Treat W areas, while Treat M areas have 2.8 more applicants per section on average (an 11.3% increase, though not statistically distinguishable from Control).

Table 2: Treatment Effects on Number of Applications by Section

	(1) Total Number of Applicants		(2) Number of Female Applicants		(3) Number of Male Applicants	
Treat W	0.293	(2.866)	1.133	(1.445)	-0.840	(2.092)
Treat M	2.781	(2.832)	1.143	(1.427)	1.639	(2.067)
Observations	59		59		59	
Control Mean	24.512		7.105		17.407	

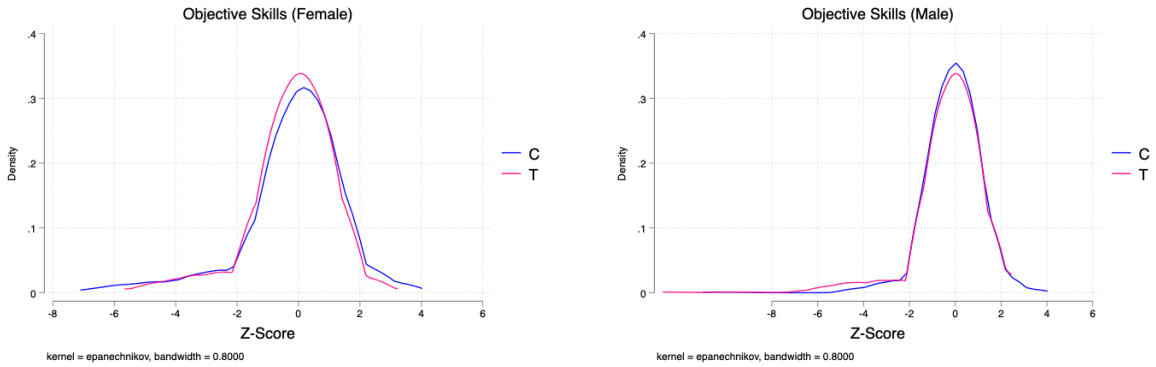
This table reports treatment effects on the number of job applications received per section. The dependent variables are: (1) total number of applicants, (2) number of female applicants, and (3) number of male applicants, measured at the section level using specification in equation (1). The key independent variables are indicators for whether the section was assigned to the Treat W or Treat M advertisement, with Control sections as the omitted category. Each column reports results from a separate regression. All regressions include strata and district fixed effects. Robust standard errors are reported in parentheses. Statistical significance is denoted as: $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

Objective Technical Skills

Next, we evaluate changes in the candidate composition across Treatment and Control areas using the objective technical skills index described earlier. In Figure 1, we present kernel density plots of the objective skills index, disaggregated by gender and treatment status. These curves allow us to visualize shifts in the distribution of applicant quality induced by the treatment.

Overall, we observe very few changes in candidate composition between treatment and control areas, for both male and female applicants. The most meaningful changes we observe are reductions in MSCE completion and English-speaking rates for male applicants in Treat W areas. Specifically, there is a 2.3 percentage point decline in MSCE completion on a control mean of 99% and a 1.8 percentage point decline in English-speaking on a control mean of 100%, among male applicants in Treat W sections—both statistically significant at the 10% level (Table B.1). A back of the envelope calculation indicates that this treatment effect represents 0.82 fewer qualified male applicants in each Treat W section.

Figure 1: Objective Skills Distribution



Although we cannot say with any certainty that there was an increase in the number of female applications, the fact that there are *no* differences in female qualifications is still noteworthy. It implies that there were 33 qualified women available who were potentially encouraged to apply through a simple, costless nudge. Even in a setting where women face systemic inequities in accumulating human capital, and face norms-based barriers to seeking a job that requires tremendous leadership skills, this suggests there can be room for simple interventions to encourage qualified women to apply for good jobs.

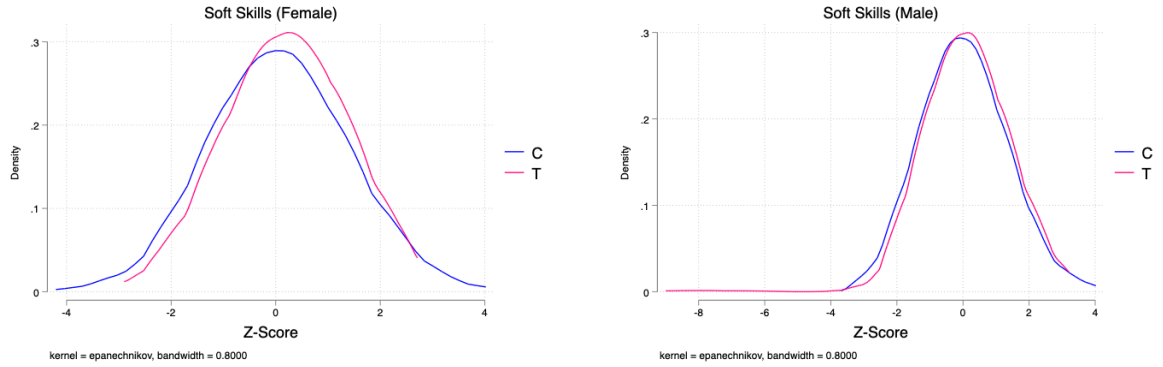
Signals of Soft Skills

Next, we evaluate the index of signals of soft skills. This index combines signals of income, application effort, and social connectedness. If we observe changes in some of these signals in response to the treatment, such as income, this would suggest that the treatment drives a different selection of candidates to apply. Conversely, if we observe changes in effort or social connectedness, it may imply that the treatment leads candidates to change how they fill out the application form.

We find a significant increase on the index of soft skills signals among female applicants, driven by women in the Treat W sections (Figure 2 and Table B.2). This shift is driven by changes in how women filled out information about their references.

It is possible that the treatment application brings in a different selection of female candidates who are more connected, enthusiastic, or detail-oriented. Alternatively, the treatment application might lead similar candidates to change the way that they fill out the application. While

Figure 2: Signal of Soft Skills Distribution



we cannot say with certainty, since we do not have data on the full pool of *potential* candidates, the evidence is more consistent with similar candidates changing the way in which they fill out the application. First, we see few changes in signals of soft skills that are hardest to manipulate, such as income proxies. Changes in listing two references, the signal that moves the most in response to the treatment, might represent fixed social connectedness (implying differential selection), or might represent variable effort on the application or in seeking approval from existing social ties to act as a reference (implying treatment effects without differential selection). We do not see any treatment effects on connectedness to higher-income references, and increases in listing two references among women in Treat W areas changes invariably with other signals of connectedness (Table B.2).

Furthermore, reference-listing as a measure of effort is supported by Abel et al. (2020), which finds that information about the value of providing a reference alone induces job-seekers to attach a reference letter to an application in South Africa, implying that access to a reference is not the primary barrier to listing a reference. In our setting, applicants only need to list a name and phone number for their references, and do not need to obtain a letter. The application form explicitly suggests that applicants can list their secondary school Head Teacher as a reference, and 97% of applicants have completed secondary school. Consequently, margins of effort—reaching out the Head Teacher to ask their consent to be listed as a reference, for example—are feasibly impacted.

Shortlisting and Hiring

At the shortlisting stage, women in control areas are 13.3 percentage points more likely to be shortlisted than men (column 1), a difference that is statistically significant. However, there are no significant differences in the shortlisting rates between the treated and control areas for women. This pattern is consistent with the firm's implementation of a shortlisting quota that requires that approximately 50% of the shortlisted candidates in each section be female. Since this quota is applied uniformly in treatment and control areas, treatment does not lead to differential shortlisting of women.

We next examine outcomes at the subsequent stage of the hiring process. Female applicants

in Treat W sections perform worse on average in the standardized test score distribution. Specifically, there is a statistically significant decline of 0.61 standard deviations in test scores for women relative to men in the same sections, compared to the gender gap in control areas (column 2). There is no significant difference in test scores for female applicants in Treat M sections. The overall p-value for the joint test of the total treatment effect on female applicants' test scores in Treat W sections is 0.14, indicating the effect is not statistically significant at conventional levels.

In Treat W sections, male applicants receive significantly higher interview scores on average, with an increase of 0.41 standard deviations relative to control areas. However, female applicants in these sections score 0.74 standard deviations lower than men in the same sections compared to the gender gap in control areas (column 3), a difference that is statistically significant. As a result, the net effect of the treatment on women's interview performance is negative. There is no significant difference in interview scores for women in Treat M sections. The p-value for the joint test of the total treatment effect on women in Treat W sections is 0.08, indicating a marginally significant decline.

These performance differences translate into disparities at the final stage of the hiring process. In control areas, women are slightly more likely to be hired than men, but this difference is not statistically significant. In Treat W sections, male applicants are 10.8 percentage points more likely to be hired compared to men in control areas, a difference significant at the 10% level. In contrast, women in Treat W sections are 31.2 percentage points less likely to be hired than men in the same sections, a difference that is statistically significant at the 1% level. The p-value for the joint test of the total treatment effect on women's hiring in Treat W sections is less than 0.01, indicating a strongly significant negative effect on women's hiring outcomes and that the treatment backfired for female applicants.

3.1.1 Mechanisms

Why does the treatment backfire? Importantly, evaluators did not know which sections received which advertisements, making it unlikely that observed screening differences reflect conscious reactions to the recruitment message. It is also unlikely due to changes in the size or composition of the candidate pool. We don't find any reductions in the number of qualified female applicants in Treat W areas, suggesting that the worse performance that women exhibit on tests and interviews in these areas is not because there simply are no qualified women available. Furthermore, Treat W areas are precisely where *men* are less likely to meet the job qualifications, but it is men in Treat W areas who do better on the test and interview than any other group. Even if the suggestive changes to the size of the candidate pool that we observe are substantively important for evaluators, this should have the largest effect on screening in Treat M areas, where the candidate pool increases the most because *both* men and women are more likely to apply. Instead, we see screening changes in Treat W areas.

Then, the treatment must backfire because evaluators shortlist a less-qualified subset of women

Table 3: Shortlisting and Hiring

	Shortlisted (1)	Test Score (2)	Interview Score (3)	Hired (4)
Treat W	-0.002	0.231	0.412**	0.101**
	[0.030]	[0.186]	[0.190]	[0.047]
Treat M	0.000	0.201	-0.071	-0.045
	[0.026]	[0.159]	[0.170]	[0.057]
Female	0.133***	-0.402**	-0.200	0.012
	[0.034]	[0.170]	[0.138]	[0.072]
Female X Treat W	0.026	-0.612**	-0.744**	-0.294***
	[0.057]	[0.295]	[0.292]	[0.100]
Female X Treat M	-0.037	-0.317	0.083	0.050
	[0.054]	[0.270]	[0.273]	[0.119]
Mean	0.191	0.000	0.000	0.225
p-value: W + F x W = 0	0.64	0.14	0.08*	0.00***
p-value: M + F x M = 0	0.50	0.59	0.94	0.94
Observations	1255	283	283	298

This table reports treatment effects on four stages of the hiring process: (1) whether the applicant was shortlisted, (2) standardized score on aptitude test, (3) standardized score in the interview, and (4) whether the applicant was hired, using specification in equation (2). The key independent variables are treatment assignment (Treat W or Treat M), applicant gender, and interactions between treatment and gender. The omitted category is male applicants in the control sections. Each column presents results from a separate regression. All regressions include fixed effects for strata and district. Robust standard errors clustered at the section level are shown in brackets. The bottom rows report means of the dependent variable in the control group, and p-values for joint tests of significance of the treatment and interaction coefficients for each arm. Statistical significance is denoted as: $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

in Treat W areas, crowding out the more-qualified women in the applicant pool. Treatment-driven signal distortion, especially in Treat W areas, could only have led to worse screening if evaluators use these signals to shortlist women. This requires two conditions: first, that evaluators rely on soft-skill signals for screening women; second, that these signals are distorted to become less predictive of objective skill in treated areas.

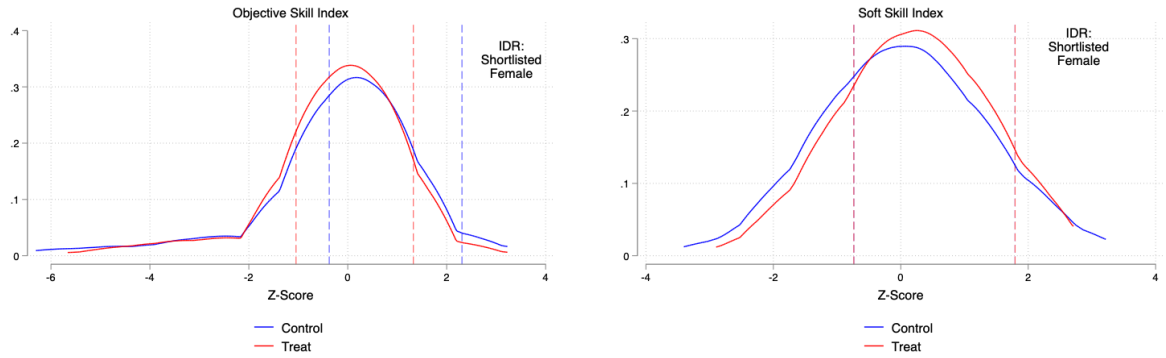
With these two conditions in mind, we consider three possible hypotheses to explain why the treatment backfires: (1) evaluators apply the same screening rule to all applicants, i.e. there is no evaluator bias, but the treatment weakens the cross-signal correlation for women only; (2) the treatment leads to signal distortion for all, but latent evaluator bias leads this distortion to only matter for how evaluators screen women; or (3) there is latent evaluator bias and the treatment leads to signal distortion for women only, creating a compounding effect.

Screening

We take a closer look at evaluator screening practices by examining the distributions of objective and soft-skill signal indices for male and female applicants separately. In Figures 3 and 4, we plot the density of these indices for all applicants, with dotted lines indicating the interdecile range of shortlisted candidates. This allows us to see where in the overall applicant distribution evaluators are selecting from.

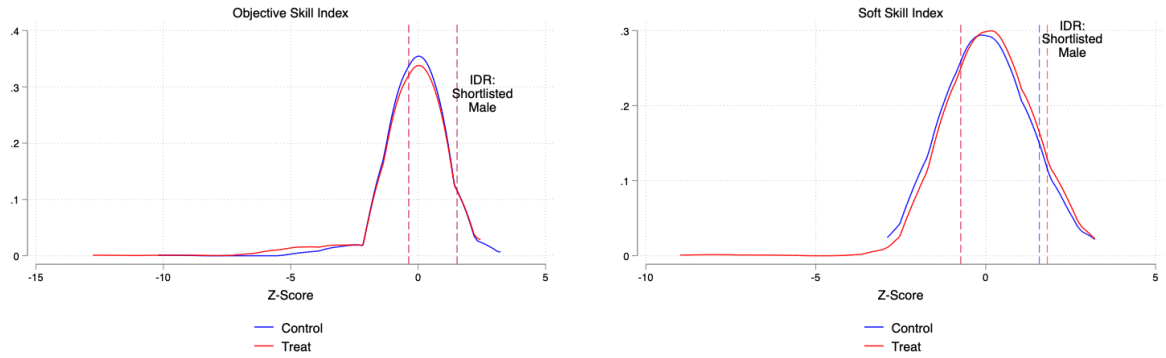
For female applicants (Figure 3), the distribution of objective skill in treated areas shifts slightly leftward, and not significantly so. However, the interdecile range of shortlisted female candidates shifts significantly leftward, suggesting that evaluators are selecting women with lower objective skills in treated areas compared to control, even conditional on the distribution of

Figure 3: Interdecile range of indices for female applicants



qualified women available. In contrast, the interdecile range on the soft-skill index is indistinguishable between treated and control areas, despite a rightward shift in the overall distribution of soft-skill signals. For male applicants (Figure 4), the interdecile ranges from which male ap-

Figure 4: Interdecile range of indeices for male applicants



plicants are selected in treated and control areas almost completely overlap. This indicates that screening practices for men remain stable. Together, these patterns suggest that the treatment weakened the screening process for women, leading evaluators to select lower-skilled female candidates, while the screening process for male candidates remained unaffected. These results then raise the question of *why* female screening changes so dramatically in treated areas. Do evaluators change the way that they conduct screening, or do they always screen on the basis of signals of soft skills, which become less correlated with objective skills in treated areas?

Correlation Between Signals of Soft Skills and Objective Skills

To assess whether the observed gender differences in screening outcomes are driven by differences in the quality of applicant signals, we examine whether the treatment induced gender-specific changes in the correlation between soft-skill signals and objective skills. If the predictive relationship between the signals and skills declines for women but not for men, evaluators applying the same screening criteria across genders may still end up selecting lower-skilled female applicants. Identifying whether the cross-signal correlation differs by gender and treatment status allows us to assess the extent to which the observed screening patterns can be explained by changes in signal informativeness rather than by evaluator bias.

Figure 5: Correlation between Objective Skills and Soft-Skills Signals Index

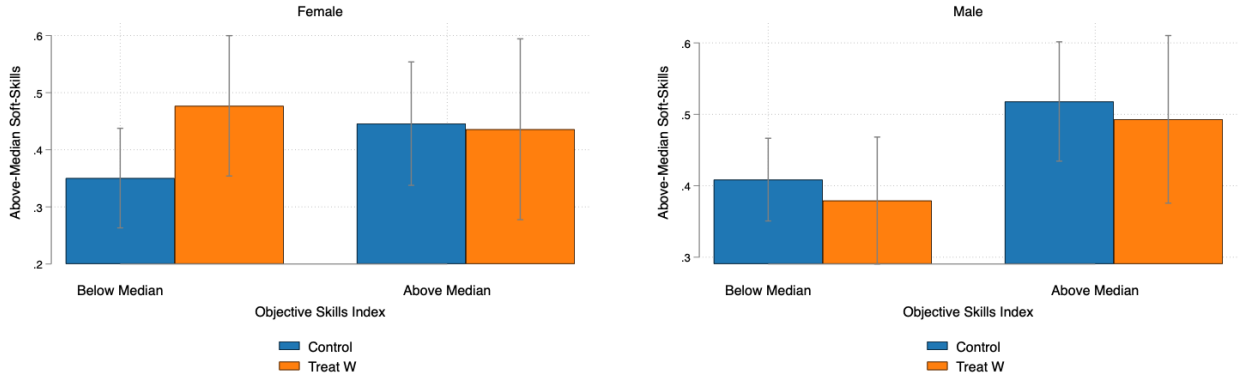


Figure 5 reports, for each sex, the proportion of applicants whose soft-skills index lies above the median (vertical axis) across two objective-skill categories—below-median and above-median (horizontal axis)—separately for control and Treat W sections.

Among female applicants (left-hand-side panel), the control group exhibits a clear positive gradient: the share of applicants with above-median soft-skill signals is substantially higher in the above-median objective-skill group than in the below-median group, indicating a positive association between the two indices. In Treat W sections, however, the proportion of below-median objective-skill women exhibiting high soft-skill signals increases sharply and nearly matches that of the above-median group. As a result, the relationship between objective skill and soft-skill signaling weakens substantially for women in Treat W areas. To quantify the visual pattern shown in Figure 5, we examine the correlation between objective skills and soft-skill signaling among female applicants. In the control group, the correlation is modestly positive (ρ=0.22), indicating that applicants with higher objective skills are more likely to signal soft skills. Under treatment, this relationship weakens (ρ=0.14), consistent with the visual flattening observed in the left-hand panel.

For male applicants, the positive correlation between the soft-skills index and the objective-skill index remains intact. Men with above-median objective skills are consistently more likely to exhibit above-median soft-skill signals in both control and Treat W sections. The correlation between objective skills and soft-skill signaling is ρ=0.19 in control areas and ρ=0.18 in Treat W sections, indicating that the relationship remains stable for male applicants across treatment conditions Thus, the treatment does not alter the relationship between soft-skill signaling and objective ability for male applicants.

These findings point to gender-biased signal distortion. The treatment decreases the cross-signal correlation among female applicants, thereby reducing the informativeness of signals of soft skills for evaluators. What remains to be established is whether evaluators apply screening criteria uniformly across genders or whether the screening process itself is gender-biased, thereby exacerbating screening mis-optimization induced by the signal distortion.

3.2 Randomized Resume Experiment

Table 4 presents results from the randomized audit study, where we estimate the effect of experimentally assigned application attributes on evaluator assessments. Each regression includes evaluator, application, and review-order fixed effects, and standard errors are clustered at the evaluator level.

We find consistent evidence that evaluators respond to objective skill signals. Applicants who report speaking English are significantly more likely to be shortlisted and receive higher predicted interview and test scores, as well as higher ratings on technical skills, social skills, and work ethic. All effects are significant at the 5 percent confidence level. Applicants who report being able to use a smartphone receive higher ratings on technical skills (0.441, $p < 0.05$) and work ethic (0.220, $p < 0.10$), with a smaller effect on predicted interview scores (0.215, $p < 0.10$). These results suggest that evaluators value objective skill-related information on average, especially English-speaking.

Table 4: Resume Audit Results

	(1) Shortlisted	(2) Interview Score	(3) Test Score	(4) Technical Skills	(5) Social Skills	(6) Work Ethic
Female	-0.008	0.053	0.177	0.185	0.052	-0.030
	(0.046)	(0.109)	(0.120)	(0.131)	(0.162)	(0.112)
Mock Applicant Listed Two References	-0.072	0.050	-0.052	0.249*	0.072	0.050
	(0.046)	(0.122)	(0.116)	(0.145)	(0.169)	(0.139)
Mock Applicant Used Formal Terms (Refs)	-0.045	-0.048	-0.096	0.089	0.195	0.042
	(0.054)	(0.113)	(0.106)	(0.167)	(0.177)	(0.101)
Mock Applicant can Use a Smartphone	0.013	0.215*	0.176	0.441**	0.071	0.220*
	(0.055)	(0.125)	(0.125)	(0.182)	(0.158)	(0.127)
Speaks English	0.398***	1.111***	1.120***	0.586***	0.434**	0.732***
	(0.074)	(0.156)	(0.155)	(0.197)	(0.185)	(0.187)
Mock Applicant has a TNM Number	-0.050	-0.055	-0.008	0.041	0.089	0.181
	(0.051)	(0.099)	(0.096)	(0.150)	(0.147)	(0.114)
Observations	292	291	291	291	291	291
Overall Mean	0.597	2.967	3.027	3.110	3.284	3.241
Evaluator Fixed Effects	X	X	X	X	X	X
Application Fixed Effects	X	X	X	X	X	X
Order Fixed Effects	X	X	X	X	X	X

This table reports results from a resume audit experiment using the specification described in equation (3). The dependent variables include: (1) an indicator for whether the applicant was shortlisted, (2) predicted interview score, (3) predicted test score, and evaluator ratings of (4) technical skills, (5) social skills, and (6) work ethic, all measured on a scale from 1 to 5. Independent variables include applicant attributes that were randomly assigned: gender, English-speaking ability, listing two references, listing references professionally, indicating smartphone capability, and including a TNM phone number. Each column presents results from a separate regression. All regressions include fixed effects for evaluator, application, and application review order. Standard errors are clustered at the evaluator level. Statistical significance is denoted as: $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

By contrast, we find little evidence that soft skill signals such as listing two references, using formal honorifics for referees, or having a TNM phone number, systematically affect assessments. One exception is that applicants who list two references receive slightly higher technical skill ratings (0.249, $p < 0.10$), but this effect does not extend to other outcomes. On the whole, soft-skill signals appear weakly informative and inconsistently used by evaluators. We also find no evidence of direct gender-based discrimination in this setting, which is consistent with the recruitment experiment, where women are more likely to be shortlisted. The coefficient on the female applicant indicator is statistically insignificant across all outcomes.

Next, we examine whether evaluators' interpretation of application signals varies by gender. Specifically, we assess whether evaluators use soft-skill signals differently for male and female

applicants, conditional on both reporting English proficiency. This exercise is motivated by the results from the recruitment experiment. If evaluators consider signals of soft skills *conditional* on objective skills, then they will still shortlist a qualified sample. However, the results in the female treatment group suggests that this cannot be the case. Rather, signals of soft skills must be considered *instead* of objective skills among women, and signals of soft skills and objective skills happen to be correlated among control women. Thus, the following exercise represents an analogous analysis in the resume audit study to further validate that this friction contributes to the treatment backfiring. These results suggest that evaluator bias, manifested in the differential weighting of signals by applicant gender, interacts with treatment induced signal distortion for women. This compounds the effect of reduced cross signal correlation and contributes to the observed backfiring.

Heterogeneity by Applicant Gender and English-Speaking Ability

Speaking English is the most consistently rewarded signal in our setting, with large and significant effects on shortlisting, predicted performance, and subjective ratings. Conditioning on English allows us to focus on applicants who meet a baseline threshold of objective competence. For ease of interpretation and to mitigate concerns about multiple hypothesis testing, we collapse all outcomes into a composite score, referred to as the Total Score in the results, which sums the evaluator's predicted interview score, predicted test score, and ratings on technical skills, work ethic, and social skills. This composite is then standardized using the relevant control group mean and standard deviation.

Table 5: Heterogeneity in Total Score by English-Speaking Status: Female Applicants

	(1) Two References		(2) Formal References		(3) Smartphone		(4) TNM Phone	
Speaks English	1.270***	(0.244)	0.858***	(0.292)	0.940***	(0.237)	1.133***	(0.242)
Attribute	0.315*	(0.186)	-0.046	(0.240)	0.120	(0.254)	0.066	(0.184)
Speaks English X Attr.	-0.354	(0.328)	0.374	(0.394)	0.005	(0.293)	-0.210	(0.314)
Observations	140		140		140		140	
Control Mean	0.000		0.000		0.000		0.000	
Controls for Other Chars.	X		X		X		X	
All Fixed Effects	X		X		X		X	
Estimates (P-values):								
Attr. + Attr. × English	-0.039 [p=0.882]		0.328 [p=0.267]		0.125 [p=0.559]		-0.145 [p=0.596]	

This table reports heterogeneity in evaluator screening scores for female applicants based on English-speaking and other individual application attributes, using the specification described in equation (4). The primary dependent variable is the Total Score, a standardized composite score constructed by summing the evaluator's predicted interview score, predicted aptitude test score, and ratings of the applicant's technical skills, work ethic, and social skills. Each column presents results from a separate regression in which the listed attribute (e.g., listing two references, owning a smartphone) is interacted with an indicator for English proficiency. All regressions include controls for other application characteristics, evaluator and application fixed effects. Standard errors are clustered at the evaluator level. Statistical significance is denoted as: $^{*}p < .10$,$^{**}p < .05$,$^{***}p < .01$.

Table 5 presents results for female applicants, examining how evaluators respond to other signals conditional on English proficiency. Because the total score is standardized, all coefficients are expressed in standard deviation (SD) units. Across all columns, the coefficient on *Speaks English* is large, positive, and statistically significant at the 1% level, indicating that evaluators consistently reward English-speaking with higher total scores, regardless of the accompanying signal. The coefficient on *Attribute* alone captures how evaluators weigh each signal among non-English-speaking women. The interaction term (*Speaks English* × *Attribute*) captures how evaluators interpret the signal when the applicant speaks English.

Table 6: Heterogeneity in Total Score by English-Speaking Status: Male Applicants

	(1) Two References		(2) Formal References		(3) Smartphone		(4) TNM Phone	
Speaks English	0.672**	(0.296)	0.987***	(0.354)	0.987***	(0.331)	0.597**	(0.256)
Attribute	-0.062	(0.274)	-0.107	(0.248)	0.529*	(0.313)	-0.200	(0.224)
Speaks English X Attr.	0.396	(0.322)	-0.126	(0.427)	-0.174	(0.386)	0.546	(0.374)
Observations	145		145		145		145	
Control Mean	0.000		-0.000		0.000		0.000	
Controls for Other Chars.	X		X		X		X	
All Fixed Effects	X		X		X		X	
Estimates (P-values):								
Attr. + Attr. × English	0.333 [p=0.200]		-0.233 [p=0.431]		0.355 [p=0.133]		0.347 [p=0.119]	

This table reports heterogeneity in evaluator screening scores for male applicants based on English-speaking and other individual application attributes, using the specification described in equation (4). The primary dependent variable is the Total Score, a standardized composite score constructed by summing the evaluator's predicted interview score, predicted aptitude test score, and ratings of the applicant's technical skills, work ethic, and social skills. Each column presents results from a separate regression in which the listed attribute (e.g., listing two references, owning a smartphone) is interacted with an indicator for English proficiency. All regressions include controls for other application characteristics, evaluator and application fixed effects. Standard errors are clustered at the evaluator level. Statistical significance is denoted as: $^{*}p < .10$,$^{**}p < .05$,$^{***}p < .01$.

Two striking results emerge when analyzing these coefficient. In Column (1), listing two references is associated with a 0.32 SD increase in total score, significant at the 10% level, suggesting that evaluators reward listing two references *even among non-English speaking women.* Recall that non-English speakers stand at a significant disadvantage to ultimately be hired, since the aptitude test and interview are conducted in English. Secondly, the interaction term coefficients show no consistent pattern. The signs across columns are inconsistent, and none of the interaction terms are statistically significant. This includes smartphone-capability, an objectively valuable skill that will help women in their job performance. The linear combination test of the main effect and interaction term confirms that the total effect of these soft-skill signals among English-speaking women is not distinguishable from zero in any case.

Taken together, these results suggest that while evaluators place weight on English proficiency, they place value on less-informative signals even in the absence of English-speaking among female applicants, and that their interpretation of other signals *conditional* on English-speaking is inconsistent. The lack of a clear pattern across signal types, and the absence of meaningful effects even when applicants speak English, point to unstable use of other signals in screening decisions. Given the small sample size and limited precision, these findings are suggestive rather than definitive, but they are consistent with our hypothesis: among female candidates, evaluators use signals of soft skills *instead* of objective skills, rather than conditional on objective skills.

Table 6 presents results for male applicants, analyzing how evaluators respond to different application signals, conditional on English proficiency. Across all columns, the coefficient on *Speaks English* is positive, large, and statistically significant, confirming that English is a strong predictor of total evaluation scores. These coefficients range from 0.67 to 0.99 SD, significant at the 5% or 1% level, and suggest that evaluators place substantial weight on this objective signal for male applicants.

The second row (*Attribute*) captures the effect of each individual signals among non-English-speaking men. The coefficients on listing two references, listing references formally, and includ-

ing a TNM phone number are small and statistically insignificant. In contrast, the coefficient on smartphone use in Column (3) is positive and significant at the 10% level, indicating that evaluators reward this objective technical skill among men, even when the applicant does not speak English.

In contrast to how evaluators assess female applications, several attributes appear to be weighted among men conditional on English-speaking, though the results are imprecise. The linear combination test (*Attr.* + *Attr.* × *English*) shows suggestive evidence that the total effect of smartphone use, two references, and having a TNM phone number is positive for English-speaking men (p-values of 0.133, 0.200, and 0.119, respectively).

Taken together, this exercise offers suggestive evidence of evaluator bias i.e. evaluators may interpret application signals differently for male and female applicants, even when objective qualifications are held constant. For both genders, English proficiency emerges as a consistently rewarded signal. However, the use of other signals appears to diverge: among male applicants, evaluators respond more systematically to both objective signals (e.g., smartphone use) and soft-skill signals (e.g., TNM phone number) when English is also present. In contrast, for female applicants, these additional signals do not meaningfully affect evaluation scores once English proficiency is known, and in some cases, unreliable signals are used *instead* of English proficiency. The estimated effects among English-speaking women are inconsistent in sign and statistically insignificant, suggesting that evaluator bias may result in less stable signal use for women.

While the estimates are imprecise and drawn from a relatively small sample, the findings are consistent with a form of gender-biased screening that does not stem from overt discrimination, but rather from an inconsistent and sometimes misdirected interpretation of signals for women relative to men. These patterns have important implications for the design of application materials and screening processes, particularly in contexts where evaluators must make high-stakes decisions based on limited or ambiguous information.

3.3 Shortlisting Decisions with Re-Weighted Application Signals

We can estimate the extent to which distorted signaling from the supply-side increases the gender gap in ability of the shortlisted candidate pool by comparing treatment and control areas. What would happen if the demand side did not evaluate candidates with a gender bias, and male and female application signals were weighted equally? How would the effect of resolving the demand side compare with the effect of preventing supply-side behavioral distortions? Although we do not observe variation in demand-side bias in the recruitment experiment, we are able to use the resume audit study to obtain the weights that evaluators place on uncorrelated signals for candidates randomly assigned to be male or female. We can then compare actual shortlisting decisions with counterfactual shortlisting decisions across a number of alternative decision rules that utilize the weights that evaluators apply to uncorrelated variables for male and female candidates in the resume audit study.

Table 7: Correlation between Actual Shortlisting Decision and Counterfactual Decision Rules

Application Variable Weights	ρ
Gender-Bias-Weighted Score	0.2065
Uniform Female-Weighted Score	0.1495
Uniform Male-Weighted Score	0.0792
Reverse Gender-Bias-Weighted Score	0.0265

Weights are constructed for five variables utilizing the regression in Equation 5 and the resume audit experiment. Female-weights are constructed using the sample of applications that are randomized to have a female name, and male-weights are constructed using the sample of applications that are randomized to have a male name. Using the real applications in the recruitment experiment, each applicant is assigned a reweighted score by applying the weights from the resume audit experiment to the variables from the real applications. We construct four scores: a uniform female-weighted score, where all variables are weighted with female weights for all applications; a uniform male-weighted score, where all variables are weighted with male weights for all applications; a gender-bias-weighted score, where all variables are weighted with female weights for female applicants, and all variables are weighted with male weights for male applicants; and a reverse gender-bias-weighted score, where all variables are weighted with female weights for male applicants, and all variables are weighted with male weights for female applicants. For each score, we select the top-five highest scorers within each section, selecting female candidates in the case of ties, and randomly breaking ties afterwards. Column (2) presents the Pearsons' correlation coefficient between the an indicator variable for being selected in the top-five using the score in variable in column (1) and the true shortlisting outcome for each applicant.

Counterfactual Shortlisting Decision Rules Using Application Scores

We construct application scores for each actual application in the recruitment experiment, using the variables that we randomize in the resume audit experiment: English-speaking ability, smartphone capability, having a TNM phone number, listing two references, and listing references' names with formal honorifics. We assign weights to those variables by using the coefficients from the following regression in the resume audit study:

$$\begin{aligned} Y_{ia} = \alpha_0 &+ \alpha_2 \text{TwoRef} + \alpha_3 \text{ProfRef} + \alpha_4 \text{Smartphone} \\ &+ \alpha_5 \text{English} + \alpha_6 \text{TNM} + X_{ia} + \delta_a + \delta_o + \theta_i + \varepsilon_{ia} \end{aligned} \tag{5}$$

where Y_{ia} is an indicator that takes the value of 1 if the evaluator reports that they would shortlist the candidate, and takes the value of 0 otherwise; X_{ia} is the set of two-by-two interaction terms between the five indicator variables in the regression; and all other variables are as in Equation 3. We run the regression for male and female mock applicants separately, so that we can construct male-weights and female-weights separately.

We assign all candidates who score in the top-five on the male-weighted applicant score within their section as shortlisted (recall that, in practice, five candidates are shortlisted within each section). We break ties by selecting female candidates over male candidates, consistent with the firm's "soft affirmative action" policy at the shortlisting stage, and randomly break any remaining ties. To test the validity of these weights for actual evaluator behavior, we create several versions of the weighted score, and compare shortlisting decisions based on our procedure with actual shortlisting decision. First, we create a "gender-bias-weighted" score, where we

Table 8: Shortlisted Candidate Comparisons for Policy Counterfactuals

Supply-Side Behavioral Variation ⇒ **Demand-Side Behavioral Variation ⇓**	*Treatment* *(Distorted)*	*Control* *(Not Distorted)*
Candidates Selected: *Actual Shortlisting Decisions* *(Biased)*	Distorted Supply-Side, Biased Demand-Side	**Policy:** Mitigate Supply-Side Distortions
Candidates Selected: *Top-5 with Uniform Male-Weighted Application Scores* *(Not Biased)*	**Policy:** Mitigate Demand-Side Bias	**Policy:** Mitigate Supply-Side Distortions and Demand-Side Bias

assign female-weights to female candidates, and male-weights to male candidates. Ranking in the top-five on this score is strongly correlated with actual shortlisting decisions, suggesting that evaluators use weights in the recruitment experiment and in the real-life recruitment similarly (Table 7). Next, we test uniform female-weighted scores and uniform male-weighted scores, which are each weakly correlated actual shortlisting decisions, consistent with evaluators using each of these weights for some candidates but not all (Table 7). Finally, we create a reverse gender-bias-weighted score, where female-weights are assigned to male candidates, and male-weights are assigned to female candidates, which is uncorrelated with shortlisting decisions.

Counterfactual Ability of the Shortlisted Applicant Pool

We can then compare candidate ability among those who are actually shortlisted with those who *would* have been shortlisted, if shortlisting were determined for all candidates using our male-weighted application score. Unfortunately, we only have actual interview and test scores for candidates who were shortlisted in the actual experiment. To create a measure of ability for the whole sample, we construct a lasso-predicted aptitude score, using variables from the application to predict a combined interview and test score among shortlisted Control participants.

Policy Counterfactuals

What would be the effect of mitigating supply-side behavioral distortions, mitigating demand-side behavioral distortions, or both, on the gender gap in ability and gender composition of the shortlisted candidate pool? Actual shortlisting decisions in the treatment condition, where supply-side application behaviors are distorted and evaluators select candidates with bias, represents an environment where supply-side and demand-side bias both influence shortlisting decisions. The control condition in the recruitment experiment offers a no-supply-side-distortions policy counterfactual, while the shortlisted candidate pool that would be selected under an unbiased decision rule offers a no-demand-side-bias counterfactual. Table 8 visually presents the comparisons we make to evaluate policy counterfactuals.

Our main analysis utilizes the following empirical specifications in the recruitment experiment data:

$$\begin{aligned}\hat{Y}_{is} = \beta_0 &+ \beta_1 \text{Treat} + \beta_2 \text{Female} + \beta_3 \text{Shortlisted} \\ &+ \beta_4(\text{Treat} \times \text{Female}) + \beta_5(\text{Treat} \times \text{Shortlisted}) + \beta_6(\text{Female} \times \text{Shortlisted}) \\ &+ \beta_7(\text{Treat} \times \text{Female} \times \text{Shortlisted}) + \gamma_s + \delta_d + \epsilon_{is} \end{aligned} \tag{6}$$

$$\begin{aligned}\hat{Y}_{is} = \beta_0^{MW} &+ \beta_1^{MW} \text{Treat} + \beta_2^{MW} \text{Female} + \beta_3^{MW} \text{Shortlisted}^{MW} \\ &+ \beta_4^{MW}(\text{Treat} \times \text{Female}) + \beta_5^{MW}(\text{Treat} \times \text{Shortlisted}^{MW}) + \beta_6^{MW}(\text{Female} \times \text{Shortlisted}^{MW}) \\ &+ \beta_7^{MW}(\text{Treat} \times \text{Female} \times \text{Shortlisted}^{MW}) + \gamma_s + \delta_d + \epsilon_{is} \end{aligned} \tag{7}$$

where $\hat{Y}_{is}$ is individual i's predicted aptitude score; Shortlisted takes the value of 1 if applicant i is actually shortlisted in the recruitment experiment, and zero otherwise; Shortlisted^{MW} takes the value of 1 if applicant i's male-weighted total application score is in the top five within their section, and zero otherwise; and all other variables are as in Equation 2. Then we can conduct the following tests to evaluate and compare the impacts of supply-side and demand-side distortions for the gender gap in the ability of shortlisted candidates:

1. $-(\beta_4 + \beta_7)$: What is the effect of mitigating supply-side behavioral distortions, without mitigating evaluator bias, on shortlisted women's aptitude relative to men's?
2. $(\beta_2^{MW} + \beta_4^{MW} + \beta_6^{MW} + \beta_7^{MW}) - (\beta_2 + \beta_4 + \beta_6 + \beta_7)$: What is the effect of mitigating evaluator bias, without mitigating supply-side behavioral distortions, on shortlisted women's aptitude relative to men's?
3. $(\beta_2^{MW} + \beta_6^{MW}) - (\beta_2 + \beta_4 + \beta_6 + \beta_7)$ What is the effect of mitigating both evaluator bias and supply-side behavioral distortions on shortlisted women's aptitude relative to men's?

These tests amount to difference-in-differences estimates: female versus male predicted aptitude scores, under a policy regime where both sides of the market are distorted compared with a policy regime that mitigates distortions on one or both sides of the market.

Mitigating evaluator bias on the demand side reduces the gender gap in the ability of shortlisted candidates at least as much as mitigating supply-side distortions (Table 9). This reduction in the gender gap comes without any reduction in the percent of the pool of shortlisted candidates who are female, and without losses in the average quality of the applicant pool, suggesting that reweighting would not screen out the highest-ability men (Table 10). Omitting soft affirmative action from the shortlisting decision rule, while maintaining uniform male-weighting in the shortlisting decision rule, would lead to a modest improvement in the average predicted aptitude score of the applicant pool (a 0.069 standard deviation increase), but with a smaller impact on the gender gap in the quality of applicants, and a 30% reduction in the share of candidates who are female relative to a uniform male-weighting decision rule *with* soft affirmative action.

Table 9: Δ(Shortlisted Female – Shortlisted Male) Predicted Aptitude Scores

Empirical Test	Policy	Score DID (SDs)	% Female Pool
Shortlisting Decision Rule: Actual Shortlisting Decisions			
$-(\beta_4 + \beta_7)$	Mitigating Supply-Side Distortions	0.765^{***}	46%
Shortlisting Decision Rule: Uniform Reweighting Application Signals Using Male-Weights			
$(\beta_2^{MW} + \beta_4^{MW} + \beta_6^{MW} + \beta_7^{MW})$ $-(\beta_2 + \beta_4 + \beta_6 + \beta_7)$	Mitigating Demand-Side Bias	0.830^{***}	47%
$(\beta_2^{MW} + \beta_6^{MW}) - (\beta_2 + \beta_4 + \beta_6 + \beta_7)$	Mitigating Both Distortions	0.947^{***}	49%
Shortlisting Decision Rule: Uniform Reweighting Application Signals Using Female-Weights			
$(\beta_2^{FW} + \beta_4^{FW} + \beta_6^{FW} + \beta_7^{FW})$ $-(\beta_2 + \beta_4 + \beta_6 + \beta_7)$	Mitigating Demand-Side Bias	0.115	47%
$(\beta_2^{FW} + \beta_6^{FW}) - (\beta_2 + \beta_4 + \beta_6 + \beta_7)$	Mitigating Both Distortions	0.388^{*}	35%
Shortlisting Decision Rule: No Soft Affirmative Action (Random Tie-Breaking) Uniform Reweighting Application Signals Using Male-Weights			
$(\beta_2^{MW} + \beta_4^{MW} + \beta_6^{MW} + \beta_7^{MW})$ $-(\beta_2 + \beta_4 + \beta_6 + \beta_7)$	Mitigating Demand-Side Bias	0.669^{***}	33%
$(\beta_2^{MW} + \beta_6^{MW}) - (\beta_2 + \beta_4 + \beta_6 + \beta_7)$	Mitigating Both Distortions	0.782^{***}	35%

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Column (1) references the empirical tests from Equation 7 and Equation 6. Column (2) describes what policy impact this empirical test measures. Column (3) reports the coefficient from column (1), which represents the difference-in-differences impact of the policy described in column (2) on the gender gap in ability of the shortlisted candidate pool. The first difference is average shortlisted female predicted aptitude scores versus average shortlisted female predicted aptitude scores, and the second difference is under the policy regime described in column (2) versus in a policy regime that allows for evaluator bias and encourages distorted supply-side signals from female candidates. Predicted aptitude scores are reported in standard deviations from the population mean. Predicted aptitude scores are constructed by utilizing lasso to predict a combined interview and test score among shortlisted Control participants with variables from the application. Column (4) reports the percent of candidates who are female when selected for shortlisting under the policy described in Column (2) and the decision rule described in the decision rule sub-headings.

While mitigating both supply-side distortions and demand-side bias leads to the largest reductions in the gender gap of the shortlisted candidate pool, the combined gains are surprisingly limited relative to uniform male-weights in evaluations alone. In treated areas, women's predicted aptitude score is 1.5 standard deviations lower than men's, implying that mitigating supply-side distortions closes the gender gap in the skill of the shortlisted candidate pool by 51%; mitigating demand-side bias closes the gender gap by 55%; and mitigating both closes the gender gap by 63%. This is notable since it suggests the firm can accomplish a lot by simply implementing strict evaluation procedures, which is likely easier to control than supply-side behavior. Furthermore, the firm may also want to increase the size of the applicant pool, which is suggestively achieved through the treatment, and loses little in gender equity from running the treatment advertisements if they also institute uniform male-weighting in their evaluation

Table 10: Average Predicted Aptitude Score of Shortlisted Candidates, by Decision Rule

Decision Rule	**Score Mean** (SDs from Population Mean)	**Score SD**
Actual Shortlisting	-0.090	1.022
Gender-Biased Reweighting	-0.585	1.050
Uniform Male-Reweighting	-0.131	0.791
Uniform Female-Reweighting	-0.339	1.083
No Soft Affirmative Action: Uniform Male-Reweighting	-0.021	0.822

This table reports the mean and standard deviation of the predicted aptitude score of shortlisted candidates according to five different decision rules to select shortlisted candidates. Column (1) reports the decision rule, column (2) reports the mean of the predicted aptitude scores of shortlisted candidates, and column (3) reports the standard deviation of the predicted aptitude scores of shortlisted candidates. Row (1) reports statistics about the predicted aptitude scores of candidates who were actually shortlisted in the recruitment experiment. Rows (2)-(5) report statistics according to decision rules that shortlist the top five candidates in each section on a weighted application score. Weights are constructed for five variables utilizing the regression in Equation 5 and the resume audit experiment. Female-weights are constructed using the sample of applications that are randomized to have a female name, and male-weights are constructed using the sample of applications that are randomized to have a male name. Using the real applications in the recruitment experiment, each applicant is assigned a reweighted score by applying the weights from the resume audit experiment to the variables from the real applications. We construct four scores: a uniform female-weighted score, where all variables are weighted with female weights for all applications; a uniform male-weighted score, where all variables are weighted with male weights for all applications; a gender-bias-weighted score, where all variables are weighted with female weights for female applicants, and all variables are weighted with male weights for male applicants; and a reverse gender-bias-weighted score, where all variables are weighted with female weights for male applicants, and all variables are weighted with male weights for female applicants. For each score, we select the top-five highest scorers within each section, selecting female candidates in the case of ties (except in the row (5) decision rule), and randomly breaking ties afterwards. Predicted aptitude scores are reported in standard deviations from the population mean. Predicted aptitude scores are constructed by utilizing lasso to predict a combined interview and test score among shortlisted Control participants with variables from the application.

procedures.

Placebo Test

We argue that evaluators weight relevant signals of objective technical skills higher for men than for women in status quo, and noisy signals of soft skills higher for women than for men. An alternative hypothesis is that evaluators do not utilize these variables at all, and the reweighting decision rule reduces gender gaps because it prevents evaluators from differentially utilizing other variables that we do not randomize in our resume audit experiment (for example, salary expectations or current employment status). If this is the correct hypothesis, then using the variables that we randomize in the resume audit experiment should improve the gender gap, regardless of the relative weights across these variables.

We construct a placebo test where we use female-weighted scores in our shortlisting decision rule. This decision rule utilizes the same set of variables, but weights them differently, applying higher weights to variables such as listing two references, and lower weights to variables such as being able to use a smartphone. We find that the gender gap in the ability of the shortlisted applicant pool does not close significantly relative to actual shortlisting when evaluators make shortlisting decisions using female-weighted scores. Combining female-weighted scoring with mitigating supply-side distortions has modest positive effects on the gender gap, but with a reduction in the share of women in the applicant pool. Furthermore, this decision rule leads to a large decrease in the average score of the shortlisted pool. Taken together, this supports our hypothesis that, when evaluating male applicants, evaluators place more weight on variables that are predictive of ability for *the whole sample*. Conversely, when evaluating female applicants, evaluators place more weight on variables that are *less* predictive of ability.

4 Discussion

Using a randomized recruitment experiment, we show that providing more gendered information and gender-coding on a recruitment advertisement leads to changes in what candidates signal about themselves when filling out the job application. This change in signals could indicate that the treatment led a different type of woman to apply—for example, highly motivated, enthusiastic, detail-oriented, or well-connected women—or it could have led similar women to fill out the application in different ways. We believe that the change in signaling is most likely explained by a change in candidate effort.

The signal distortion is largest among *less*-qualified women, leading to a reduction in cross-signal correlation among female applicants in treated areas. We use a randomized resume experiment to show that evaluators screen women on less-informative signals that can be manipulated with effort and are not in themselves job-relevant, and in many cases use these signals *instead* of objective skills. Taken together, this implies that evaluators who screen in women that send positive signals of effort and connectedness in Treatment areas are shortlisting less-qualified women than when they screen in women who send similar signals in Control areas. This ultimately crowds out more-qualified female applicants from the candidate pool and reduces female hiring.

This collection of results raises two questions: Why does a simple change in a recruitment advertisement induce changes in candidate signaling? And why do evaluators use less-informative signals (such as soft-skills signals) to screen women, when precise, more informative signals of skill that are immediately job-relevant and verifiable are observable?

4.1 Marketing, Social Norms, and Confidence

Our experiment is not designed to say with any certainty why the treatment changed applicant's signaling. However, we have evidence suggesting that the application changes participant's view of social norms and female applicant's confidence. We ask participants who are shortlisted to fill out a brief questionnaire on their interview day that asks them about their perceptions about the job. Differences in respondent answers across Treatment and Control areas could be because the treatment changes participants' job perception, or because evaluators shortlist different types of candidates in Treatment and Control areas.

Respondents in treated areas report differences in their beliefs about social norms and their views about how challenging they think the job sounds. We think it is most likely that the differences in responses about social norms that we observe are due to changes in job perceptions that the treatment causes directly, rather than selection effects.[7] Conversely, treatment

[7]The treatment causes evaluators to shortlist *less* qualified women and *more* qualified men (on the basis of their interview and test scores) in Treatment areas relative to Control areas. If the differences in questionnaire responses are driven by selection, we should expect treatment effects on perceptions to go in opposite directions for male and female candidates, since the samples are selected differently. However, the treatment effects on social norms go in the same direction for male and female applicants.

effects on perceptions about the job difficulty are consistent with a story whereby the screening process led to a selection of candidates who are less-qualified but over-confidant.[8] Although this analysis is speculative, taken together, these results imply that the treatment advertisement led participants to view the job as more appropriate for women. This then led less-qualified but more confidant women to exert additional effort in the application.

We ask participants: Out of ten people in the village, how many would consider this job to be appropriate for a man? How many would consider this job to be appropriate for a woman? In the Treatment areas, *both* men and women report that more villagers would consider the job to be appropriate for women than in Control areas.[9] Among treated men, this is driven by an increase in viewing the job as equally appropriate for men and women, whereas for treated women this is driven by an increase in viewing the job as *more* appropriate for women than for men. This is true in both Treat M and Treat W areas, suggesting that the information in the infographic might be more important than the gender-coding of the graphic figures for changing job perceptions.[10]

We also ask participants to report how challenging the job sounds to them on a scale from 1 to 5. In Treat W areas, women were significantly less likely to rate the job with a 5 (18.6pp, $p = 0.043$), and significantly more likely to rate the job with a 1 (19.0pp, $p = 0.073$).[11] Conversely, men in Treat W areas are 21.1pp ($p = 0.045$) more likely to give the job an intermediate rating (2-4). They are 14.7pp less likely to give the job the easiest possible rating, and 6.5pp less likely to give the the hardest possible. Considering the selection of candidates into sections, these results are consistent with *less*-qualified applicants expressing overconfidence, and *more*-qualified applicants assigning intermediate-value ratings for the job.

Lastly, we ask: are the treatment effects on perceptions about the appropriateness of the job for women driven by implicit job perceptions from the gender-coding of the advertisement, or by information in the advertisement itself? One piece of explicit information that the treatment conveys is the existence of female Field Officers. In Treat W areas, women believe there are *fewer* existing Field Officers than any other group, indicating that providing information about female Field Officers' existence is likely not responsible for the changes we observe.

There are small changes in perceptions about the explicit job environment. Candidates in Treatment areas believe that Field Officers spend less time visiting farmers to encourage repayment

[8] Treatment effects on perceptions of job difficulty go in the same direction as the selection. Treatment effects go in opposite directions for men and women, and there are differences in Treat W and Treat M areas, where selection was also different.

[9] On average in Control areas, shortlisted candidates say that 4.7 villagers would consider the job to be appropriate for a woman, and 6.1 villagers would consider the job to be appropriate for a man. Shortlisted candidates in Treatment sections believe that 5.2 villagers would consider the job to be appropriate for a woman, and 5.8 villagers would consider the job to be appropriate for a man.

[10] Since the treatment only changes shortlisting selection in Treat W areas but reported perceptions about social norms change in both Treat W and Treat M areas, this further indicates that we are picking up a treatment effect on applicants' perceptions about the job rather than differential reporting due to differential selection.

[11] In Treat M areas, women were noisily less likely to rate the job with a 5 (13.3pp decrease, $p = 0.124$), and instead rated the job with an intermediate value of 2-4 (14.pp increase, $p = 0.249$).

compliance, which is arguably the most uncomfortable part of the job.[12] This change could be driven either by implicit belief updating on the basis of gender-coding, or by the information through the disaggregated statistic.[13] We note that the treatment effect on perceived time Field Officers spend encouraging repayment is driven by women in Treat W areas, who believed Field Officers spend 15% (18%) less time on repayment than Control men (women). Since Treat M and Treat W conveyed the same information, this suggests that implicit belief-updating on the basis of "female-coding" the job might play a larger role.

4.2 Evaluator Screening Bias

Although our study design and data do not allow us to precisely identify the evaluator beliefs that give rise to their biases, our qualitative and suggestive results propound that evaluator bias arises due to attempts to screen on characteristics that are hard to measure.

Social Skills and Work Ethic

First, evaluators are deeply concerned about applicant's social skills and trustworthiness, since these are crucial qualities for job success. However, these qualities are very difficult to assess in a simple application form. Through the text responses where evaluators write remarks on each application, we see that there is a significant degree of idiosyncrasy in how evaluators try to discern social skills and trustworthiness.

One strategy evaluators use is to look for signals of embeddedness in the local community, particularly through the presence and type of references. Applicants referred by village leaders or known to local authorities are often seen as more trustworthy and socially reliable. For instance, one evaluator described a candidate as "*a well-trusted member of a community, referred by local leaders*" while another noted that "*this one has been referred by a village elder, meaning he might be one of the entrusted individuals in the area*". These endorsements are interpreted as indicators that the applicant is likely to be socially effective and trustworthy. Conversely, when applicants list only one referee or omit reference details, evaluators often interpret this negatively. One comment reads, "*Having one referee... shows lack of association with community members,* " and another expresses concern that the candidate "*did not put all referees*". These assessments illustrate how, in the absence of formal tools to measure social skills, evaluators fall back on informal social markers, that are meaningful but also subjective and varies significantly across reviewers.

While this process is idiosyncratic, the resume audit study reveals two gendered patterns in

[12] It is worth recognizing that *all* groups over-estimate how much time Field Officers spend encouraging repayment on average.

[13] For example, if the image of a debt collector is incongruous with the stereotype of a "female" occupation, then gender-coding could be responsible for effects on beliefs about the time Field Officers spend encouraging repayment and on the appropriateness of the job for women. Conversely, applicants might believe that it would be impossible for women to have an easy time earning the respect of community members in the role of a debt collector. Then, they adjust their beliefs about the time that Field Officers spend encouraging repayment so that their beliefs are concordant with the information that the advertisement expresses.

these inferences: English-speaking women who use formal titles for their references are rated as having better social skills than English-speaking women who do not use formal titles for their references ($p = 0.035$), and English-speaking men who have a TNM phone number are rated as having better work ethic than English-speaking men who have an Airtel phone number ($p = 0.074$). Curiously, in Control areas in the recruiting experiment, it is actually *women* with TNM phone numbers who perform better on interviews and tests (men do not perform any differently according to their phone carrier). In Control areas, using formal titles for references is not associated with interview or test performance for men or women.

If evaluators are attempting to screen both men and women on their work ethic and social skills using criteria that are ultimately not important for the job – the applicant's phone carrier and their use of formal titles for references – why does this have different effects for the screening of men and women? Although screening men on TNM phone numbers amounts to income-based discrimination, it is true in practice that men with TNM phone numbers are 10 percentage points (13%) more likely to meet all of the firm's preferred applicant characteristics ($p = 0.037$).[14] Conversely, there is no correlation between using a formal title for references and meeting the firm's preferences among women. This implies that, although evaluators are using less informative, soft-skill signals in an attempt to screen both men and women on characteristics that they are concerned about but cannot observe, their strategy amounts to screening men on objective signals and a more-or-less random screening process for women.

It is important to note that the criteria that we use to determine that "TNM phone" is a better proxy for skill than "formal titles for references" are all observable to evaluators. Why don't they use these variables to screen applicants instead? It is possible that they give preference to what they interpret as signals of social skill and work ethic because these are more important job qualifications, and do not realize that these variables are correlated with other more reliable signals that they could use instead. Furthermore, because evaluators are faced with a high-dimensional problem, where they are trying to consider many objective characteristics and unmeasured characteristics at once, it is possible that they ultimately end up relying on heuristics in their selection process.

Equity-Driven Motives

A second mechanism that might give way to this evaluator bias arises from equity motives. The firm is conscientious of the systemic inequities that women face in obtaining the qualifying skills for the job, and has a desire to diversify their workforce. Thus, evaluators are conscious of screening women on their *potential*, rather than on their human capital accumulation to date. Since so much of Field Officer job ability relies on indistinct characteristics such as social skills, leadership, and trustworthiness, it is perfectly conceivable that a woman who has had fewer opportunities for technical skill development could possess these qualities and ultimately be a

[14]The firm's preferred applicant characteristics are variables that they explicitly consider to be useful screening criteria. Aside from "meeting the minimum requirements", these preferences are never communicated with potential candidates through the job advertisement. These variables include: meeting the minimum job requirements, able to use a smartphone, not seeking further education, and not currently employed.

remarkable candidate. The equity-based motive to give less-prepared women a chance could create differential screening processes for men and women. Unfortunately, this differential screening process does not have the intended effect for two reasons. First, it is very difficult to screen on potential, especially without a rigorous process to determine which variables are associated with and should be used instead of, or in addition to, signals of human capital accumulation. Ultimately, objective criteria of technical skills are more likely to be correlated with interview scores (the best measure of social skills and leadership that we have) than other characteristics that could signal personal character, such as effort on the application. Second, the firm is capacity-constrained in their ability to train Field Officers at great length, and ends up hiring candidates who meet an objective set of criteria.

5 Conclusion

Our results contribute to a broader literature on structured hiring and evaluation procedures, suggesting that clearer weighting of evaluation criteria can mitigate unintended consequences of diversity-focused recruitment. Our empirical results are consistent with standard models of statistical discrimination if we consider the evaluator's signal to be an aggregation of all the signals across the application, each weighted by the importance that the evaluator places on that signal. Then, minority signals are noisier because evaluators place higher weights on noisier signals for minority candidates, and higher weights on more precise signals for majority candidates. While this is still a model of statistical discrimination, it suggests that evaluator bias is the underlying driver of statistical discrimination, rather than fundamental differences in the noisiness of signals across majority and minority groups. Furthermore, this result points toward a simple policy solution, whereby standards are placed around the weight that evaluators should place on each signal.

Furthermore, our results have specific implications for hiring in the context of the rising importance of social skills in the formal labor market (Deming, 2017). Although economic development is often accompanied by increasing job formalization, the implications of the rising importance of social skills for employment has received little attention in low- and middle-income countries. We find suggestive evidence that qualified women self-select out of the formal labor force due to social norms or misperception, and that nudges can be utilized to encourage them to participate. In a setting where there are prohibitive gender norms and significant gender gaps in human capital accumulation, it is notable and surprising that there is slack in the supply of female applicants who are qualified for a high-human-capital job. Furthermore, we show that this labor supply slack can potentially be captured through simple policies, such as formalizing the application-evaluation process and even simple nudges through marketing. Our evidence suggests that job formalization and reliance on social skills may have important implications for gender-based discrimination. The role of social skills in low-income labor markets, firm biases in evaluating social skills, and implications for FLFP are all promising avenues of future research.

References

Abel, Martin, Rulof Burger, and Patrizio Piraino, "The Value of Reference Letters: Experimental Evidence from South Africa," *American Economic Journal: Applied Economics*, 2020, *12*, 40–71.

Angrist, Joshua D. and William N. Evans, "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, 1998, *88*, 450–477.

Arbache, Jorge Saba, Alexandre Kolev, and Ewa Filipiak, "Gender disparities in Africa's labor market," *World Bank Publications*, 2010.

Archibong, Belinda and Francis Annan, "Disease and Gender Gaps in Human Capital Investment: Evidence from Niger's 1986 Meningitis Epidemic," *American Economic Review*, 2017, *107* (5), 530–535.

Autor, David, David Figlio, Krzysztof Karbownik, Jeffrey Roth, and Melanie Wasserman, "School quality and the gender gap in educational achievement," *American Economic Review*, 2016, *106* (5), 289–295.

Beaman, Lori, Niall Keleher, and Jeremy Magruder, "Do Job Networks Disadvantage Women? Evidence from a Recruitment Experiment in Malawi," *Journal of Labor Economics*, 2018, *36*.

Behrman, Jere R and James C Knowles, "Household income and child schooling in Vietnam," *The World Bank Economic Review*, 1999, *13* (2), 211–256.

Bertrand, Marianne, Emir Kamenica, and Jessica Pan, "Gender Identity and Relative Income within Households," *Quarterly Journal of Economics*, 2015, *130*, 571–614.

Björkman-Nyqvist, Martina, "Income shocks and gender gaps in education: Evidence from Uganda," *Journal of Development Economics*, 2013, *105*, 237–253.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Beliefs about gender," *American Economic Review*, 2019, *109* (3), 739–773.

Boring, Anne, Katherine Coffman, Dylan Glover, and María José González-Fuentes, "Discrimination, Rejection, and Willingness to Apply: Effects of Blind Hiring Processes," *Working Paper*, 2025.

Borker, Girija et al., "Safety first: Perceived risk of street harassment and educational choices of women," 2021.

Boudet, Ana María Muñoz, *On norms and agency: Conversations about gender equality with women and men in 20 countries*, World Bank Publications, 2013.

Buchmann, Nina, Carl Meyer, and Colin D. Sullivan, "Paternalistic Discrimination," *Revise and resubmit, Econometrica*, 2024.

Bursztyn, Leonardo, Alessandra L González, and David Yanagizawa-Drott, "Misperceived social norms: Women working outside the home in Saudi Arabia," *American economic review*, 2020, *110* (10), 2997–3029.

_ , Thomas Fujiwara, and Amanda Pallais, "'Acting Wife': Marriage Market Incentives and Labor Market Investments," *American Economic Review*, 2017, *107*, 3288–3319.

Carranza, Eliana, Robert Garlick, Kate Orkin, and Niel Rankin, "Job Search and Hiring with Limited Information about Workseekers' Skills," *American Economic Review*, 2022, *112*, 3547–3583.

Coffman, Katherine, Manuela R Collis, and Leena Kulkarni, "Stereotypes and belief updating," *Journal of the European Economic Association*, 2024, *22* (3), 1011–1054.

Cortés, Patricia, Jessica Pan, Laura Pilossoph, and Basit Zafar, "Gender differences in job search and the earnings gap: Evidence from business majors," 2021.

Dean, Joshua T and Seema Jayachandran, "Changing family attitudes to promote female employment," in "AEA Papers and Proceedings," Vol. 109 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2019, pp. 138–142.

Deming, David J., "The Growing Importance of Social Skills in the Labor Market," *Quarterly Journal of Economics*, 2017, *132*, 1593–1640.

Eriksson, Stefan and Jonas Lagerström, "The labor market consequences of gender differences in job search," *Journal of Labor Research*, 2012, *33*, 303–327.

Exley, Christine L and Judd B Kessler, "The gender gap in self-promotion," *The Quarterly Journal of Economics*, 2022, *137* (3), 1345–1381.

Fernando, A. Nilesh, Niharika Singh, and Gabriel Tourek, "Hiring Frictions and the Promise of Online Job Portals: Evidence from India," *American Economic Review: Insights*, 2023, *5*, 546–562.

Fershtman, Daniel and Alessandro Pavan, ""Soft" Affirmative Action and Minority Recruitment," *American Economic Review: Insights*, 2021, *3*, 1–18.

Fiala, Lenka, John Eric Humphries, Juanna Schrøter Joensen, Uditi Karna, John A List, and Gregory F Veramendi, "How early adolescent skills and preferences shape economics education choices," in "AEA Papers and Proceedings," Vol. 112 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2022, pp. 609–613.

Gallus, Jana and Emma Heikensten, "Awards and the gender gap in knowledge contributions in STEM," in "AEA Papers and Proceedings," Vol. 110 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2020, pp. 241–244.

Giuliano, Paola, "Gender and culture," *Oxford Review of Economic Policy*, 2020, *36* (4), 944–961.

Goldin, Claudia and Cecilia Rouse, "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians," *American Economic Review*, 2000, *90* (4), 715–741.

_ , **Lawrence F Katz, and Ilyana Kuziemko**, "The homecoming of American college women: The reversal of the college gender gap," *Journal of Economic perspectives*, 2006, *20* (4), 133–156.

Heath, Rachel and Seema Jayachandran, "The Causes and Consequences of Increased Female Education and Labor Force Participation in Developing Countries," *The Oxford Handbook of Women and the Economy*, 2017.

Jalal, Amen, "Screening Women Out? Pay Transparency in Job Postings," *Working Paper*, 2025.

Jensen, Robert, "Do labor market opportunities affect young women's work and family decisions? Experimental evidence from India," *The Quarterly Journal of Economics*, 2012, *127* (2), 753–792.

Kleven, Henrik, Camille Landais, and Jakob Eghort Søgaard, "Children and Gender Inequality: Evidence from Denmark," *American Economic Journal: Applied Economics*, 2019, *11*, 181–209.

Lundeberg, Mary A, Paul W Fox, and Judith Punćcohaŕ, "Highly confident but wrong: Gender differences and similarities in confidence judgments.," *Journal of educational psychology*, 1994, *86* (1), 114.

Macchi, Elisa and Claudia Raisaro, "Hidden Discrimination in Frictional Labor Markets," *Working Paper*, 2025.

McKelway, Madeline, "Women's self-efficacy and women's employment: Experimental evidence from India," *Unpublished Working Paper*, 2025.

_ **and Matt Lowe**, "Coupling Labor Supply Decisions: An Experiment in India," *Revise and Resubmit, Journal of the European Economic Association*, 2024.

Möbius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat, "Managing self-confidence: Theory and experimental evidence," *Management Science*, 2022, *68* (11), 7793–7817.

Niederle, Muriel and Lise Vesterlund, "Do women shy away from competition? Do men compete too much?," *The quarterly journal of economics*, 2007, *122* (3), 1067–1101.

Rakshit, Sonali and Soham Sahoo, "Biased teachers and gender gap in learning outcomes: Evidence from India," *Journal of Development Economics*, 2023, *161*, 103041.

Román, Antonio, Andrea Flumini, Pilar Lizano, Marysol Escobar, and Julio Santiago, "Reading direction causes spatial biases in mental model construction in language understanding," *Scientific reports*, 2015, *5* (1), 18248.

Samek, Anya, "Gender differences in job entry decisions: A university-wide field experiment," *Management Science*, 2019, *65* (7), 3272–3281.

Online Supplementary Material

Appendix A Advertisements

Job Vacancy: Field Officer

[redacted] is a fast-growing non-profit organization that provides farmers with training, high-quality inputs on credit, and a flexible repayment schedule. Field Officers work directly with farmers and community leaders to enroll, train, and encourage repayment among farmers.

Women are welcome and encouraged to apply.

Qualifications:
- MSCE Certificate
- Able to read and write English
- Able and willing to travel by bicycle
- Above 18 years of age

Job Requirements:
- Engage directly with smallholder farmers in rural areas
- Forge strong ties within the community

Benefits for Women:
- 3 months time off with pay for new mothers
- Essential Items gift for the newborn babies
- Supportive items gift for expectant mothers
- A dedicated work assistant for returning mothers

Benefits:
- Pension fund
- Health Insurance
- Performance-based incentives
- Wedding stipend
- Bereavement support
- Counseling and mental health support
- Illness and emergency support

If you are a qualified and excited candidate, please collect our APPLICATION FORM from:

Deadline for Applications – 25 March, 2024

Please note that [redacted] will NEVER ask for money as part of the interview process.

4 out of 5 **Field Officers** say that **earning the respect** of the community and **engaging with farmers** as a Field Officer was **easier** than they expected:

Figure A.1: Advertisement: Control Sections

This advertisement was used in Control sections. The infographic features a gender-neutral image alongside the statement: "4 out of 5 Field Officers say that earning the respect of the community and engaging with farmers as a Field Officer was easier than they expected." Four of the five neutral figures are shaded to reflect the statistic. This design does not include gender-specific cues.

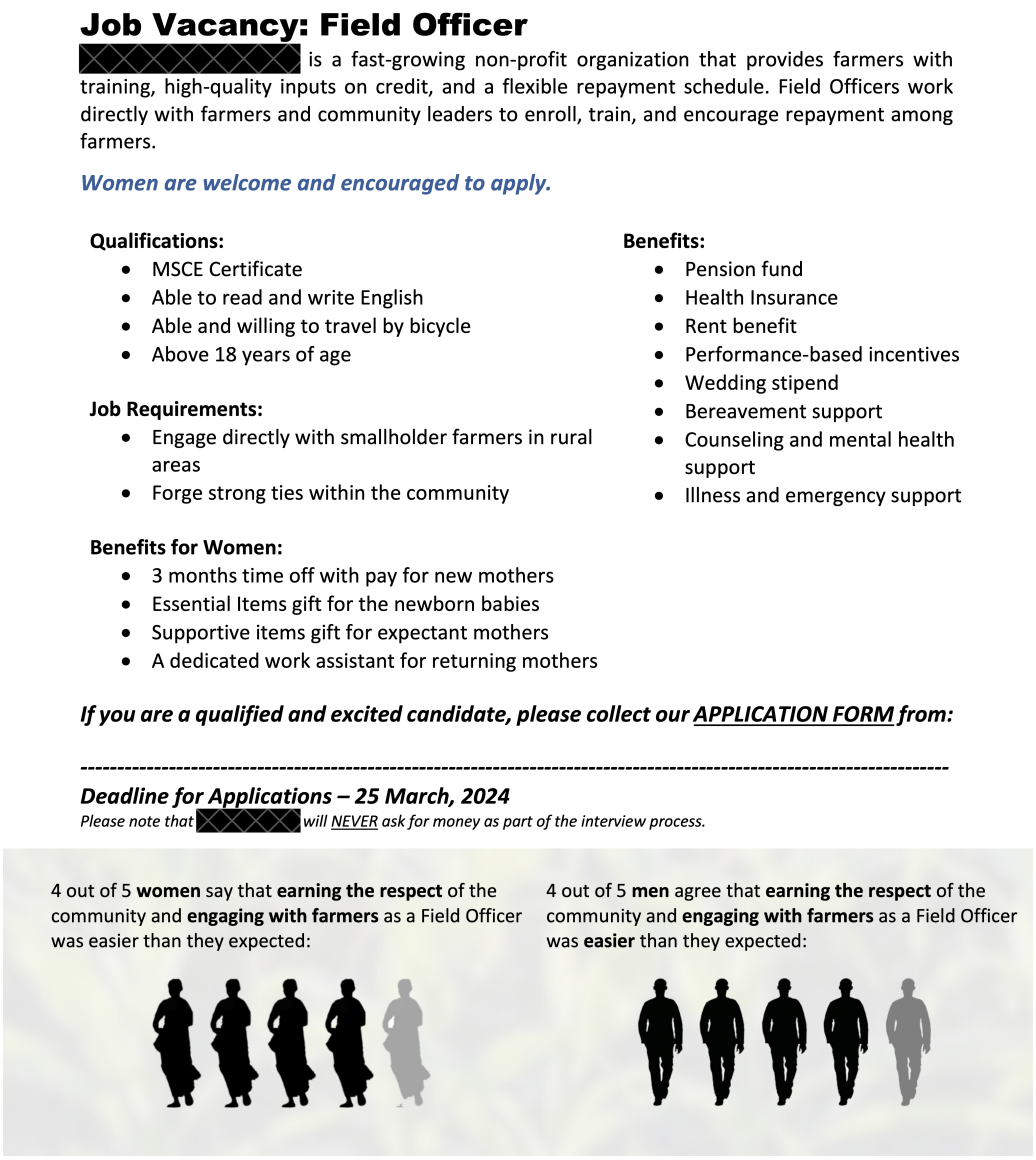
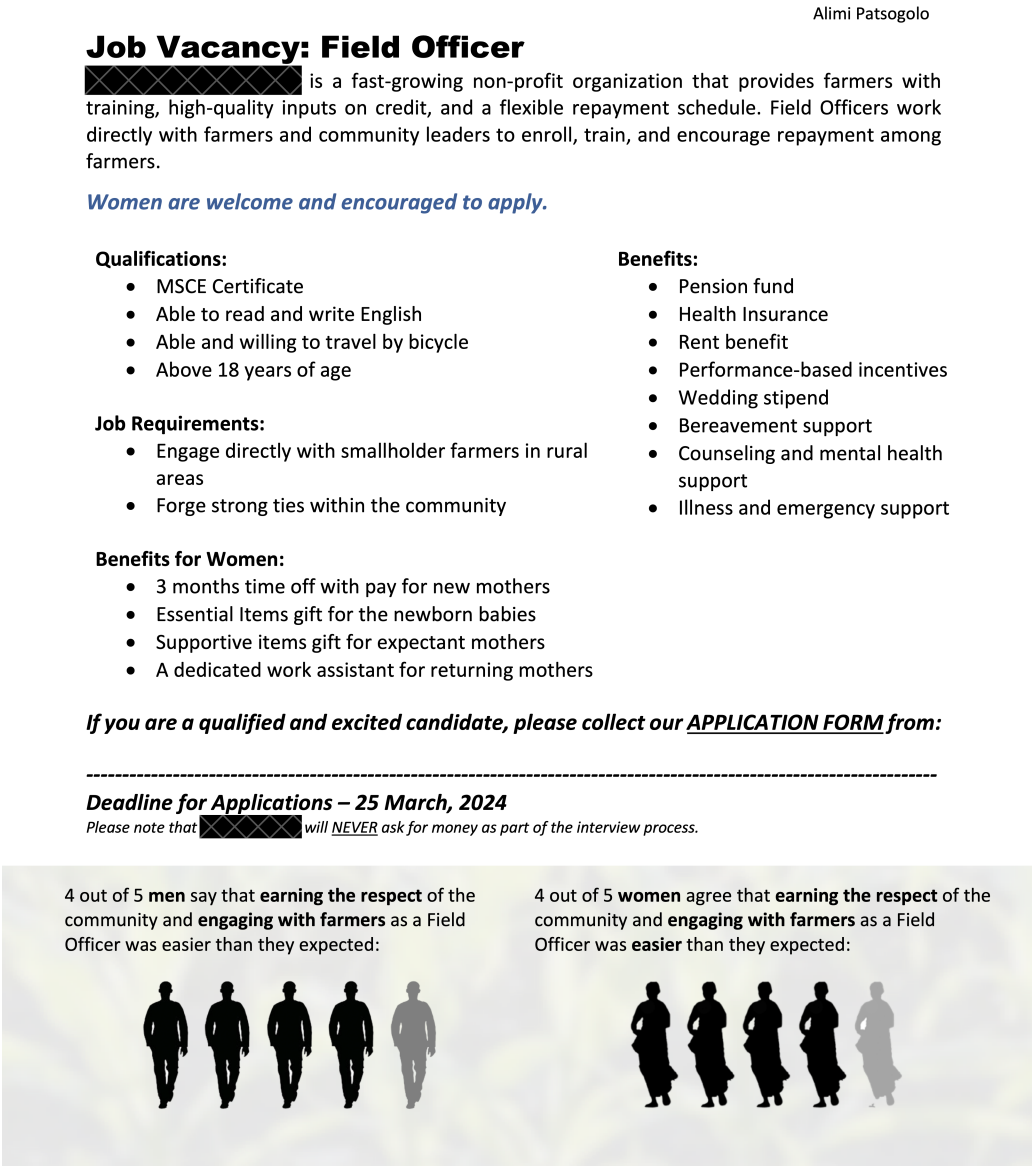


Figure A.2: Advertisement: Treat W sections

This advertisement was used in Treat W sections. The infographic is split into two columns and disaggregates the same statistic by gender: "4 out of 5 women say…" and "4 out of 5 men say…" The female-coded figures appear on the left, visually emphasizing female representation.


Alimi Patsogolo

Job Vacancy: Field Officer

[redacted] is a fast-growing non-profit organization that provides farmers with training, high-quality inputs on credit, and a flexible repayment schedule. Field Officers work directly with farmers and community leaders to enroll, train, and encourage repayment among farmers.

Women are welcome and encouraged to apply.

Qualifications:

- MSCE Certificate
- Able to read and write English
- Able and willing to travel by bicycle
- Above 18 years of age

Job Requirements:

- Engage directly with smallholder farmers in rural areas
- Forge strong ties within the community

Benefits:

- Pension fund
- Health Insurance
- Rent benefit
- Performance-based incentives
- Wedding stipend
- Bereavement support
- Counseling and mental health support
- Illness and emergency support

Benefits for Women:

- 3 months time off with pay for new mothers
- Essential Items gift for the newborn babies
- Supportive items gift for expectant mothers
- A dedicated work assistant for returning mothers

If you are a qualified and excited candidate, please collect our APPLICATION FORM from:

Deadline for Applications – 25 March, 2024

Please note that [redacted] will NEVER ask for money as part of the interview process.

4 out of 5 **men** say that **earning the respect** of the community and **engaging with farmers** as a Field Officer was easier than they expected:

4 out of 5 **women** agree that **earning the respect** of the community and **engaging with farmers** as a Field Officer was **easier** than they expected:

Figure A.3: Advertisement: Treat M sections

This advertisement was used in Treat M sections. It is identical in structure to the Treat W version, but with male-coded figures placed on the left-hand side of the infographic.

Appendix B Results

Table B.1: Objective Skills Index

	Index (1)	MSCE (2)	MSCE before 20yrs (3)	Smartphone (4)	English (5)	Bike (6)	Third Lang (7)	Other Quals (8)
Treat W	-0.077**	-0.023*	0.032	0.008	-0.018*	-0.008	0.005	-0.041
	[0.037]	[0.014]	[0.031]	[0.017]	[0.009]	[0.007]	[0.014]	[0.041]
Treat M	-0.055*	-0.003	0.056	-0.006	-0.010	-0.008	-0.009	-0.037
	[0.031]	[0.014]	[0.039]	[0.019]	[0.008]	[0.007]	[0.016]	[0.047]
Female	-0.026	-0.026*	0.272***	-0.023	-0.015	-0.008	0.005	-0.039
	[0.042]	[0.014]	[0.049]	[0.019]	[0.012]	[0.009]	[0.013]	[0.029]
Female X Treat W	0.041	-0.001	-0.082	0.007	0.028	0.014	-0.009	-0.034
	[0.072]	[0.033]	[0.069]	[0.030]	[0.018]	[0.010]	[0.020]	[0.063]
Female X Treat M	0.013	-0.004	-0.067	-0.012	0.018	0.014	-0.034	0.004
	[0.054]	[0.024]	[0.071]	[0.038]	[0.017]	[0.010]	[0.025]	[0.044]
Mean	-0.00	0.99	0.29	0.96	1.00	0.99	0.04	0.18
p-value: W + F x W = 0	0.55	0.38	0.42	0.59	0.53	0.55	0.82	0.14
p-value: M + F x M = 0	0.42	0.68	0.84	0.57	0.60	0.52	0.06*	0.48
Observations	1255	1255	1194	1255	1255	1255	1255	1250

Table B.2: Soft-Skills Signals Index

	Index (1)	No. answered (2)	Two Refs (3)	Prof Ref (4)	Knows Chief (5)	TnM (6)	TnM Ref (7)
Treat W	0.059	0.103	-0.009	0.099**	0.010	0.051	-0.049*
	[0.043]	[0.103]	[0.032]	[0.039]	[0.048]	[0.038]	[0.029]
Treat M	0.023	-0.048	0.019	0.019	0.042	0.023	-0.017
	[0.051]	[0.096]	[0.029]	[0.034]	[0.044]	[0.041]	[0.037]
Female	-0.047	-0.050	-0.047	-0.040	-0.002	-0.017	0.025
	[0.039]	[0.071]	[0.037]	[0.043]	[0.038]	[0.032]	[0.031]
Female X Treat W	0.072	-0.028	0.118**	0.085	0.104	-0.099*	-0.019
	[0.053]	[0.113]	[0.057]	[0.072]	[0.064]	[0.054]	[0.053]
Female X Treat M	0.042	0.054	0.102**	-0.027	-0.008	-0.015	-0.004
	[0.065]	[0.129]	[0.047]	[0.068]	[0.058]	[0.056]	[0.055]
Mean	-0.00	20.88	0.87	0.28	0.40	0.23	0.24
p-value: W + F x W = 0	0.00***	0.43	0.01**	0.00***	0.08*	0.18	0.07*
p-value: M + F x M = 0	0.26	0.94	0.00***	0.89	0.59	0.87	0.62
Observations	1255	1255	1255	1255	1255	1255	1255

Appendix C Field Officer Survey

Table C.1: Summary Statistics: By Gender

Variable	(1) Male N	(1) Male Mean/SE	(2) Female N	(2) Female Mean/SE	(3) Total N	(3) Total Mean/SE	T-test Difference (1)-(2)
What is your age?	128	32.641	78	31.500	206	32.209	1.141
		(0.603)		(0.704)		(0.460)	
Age when Hired	123	29.654	75	28.160	198	29.088	1.494
		(0.580)		(0.724)		(0.455)	
Highest Education: MSCE Level	128	0.164	78	0.141	206	0.155	0.023
		(0.033)		(0.040)		(0.025)	
Highest Education: MSCE Certificate	128	0.641	78	0.603	206	0.626	0.038
		(0.043)		(0.056)		(0.034)	
Highest Education: Tertiary Degree	128	0.180	78	0.244	206	0.204	-0.064
		(0.034)		(0.049)		(0.028)	
Married Before Hired as FO	128	0.695	78	0.487	206	0.617	0.208***
		(0.041)		(0.057)		(0.034)	
Married After Hired as FO	128	0.133	78	0.154	206	0.141	-0.021
		(0.030)		(0.041)		(0.024)	
Currently Married	128	0.836	78	0.641	206	0.762	0.195***
		(0.033)		(0.055)		(0.030)	
Do you have children?	128	0.852	78	0.821	206	0.840	0.031
		(0.032)		(0.044)		(0.026)	
Moved to a New Village when Hired as an FO	128	0.734	78	0.500	206	0.646	0.234***
		(0.039)		(0.057)		(0.033)	
Years as 1AF FO	123	3.054	75	3.120	198	3.079	-0.066
		(0.185)		(0.222)		(0.142)	
Worked Prior to Hired as an FO	125	0.984	78	0.897	203	0.951	0.087**
		(0.011)		(0.035)		(0.015)	
Prior Agricultural Work Experience	122	0.246	70	0.314	192	0.271	-0.068
		(0.039)		(0.056)		(0.032)	
Was a 1AF Farmer Prior to FO	128	0.258	78	0.192	206	0.233	0.066
		(0.039)		(0.045)		(0.030)	
Was a 1AF Group Leader Prior to FO	128	0.117	78	0.051	206	0.092	0.066*
		(0.029)		(0.025)		(0.020)	
F-test of joint significance (F-stat)							3.529***

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are the F-statistics. Standard errors are robust. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table C.2: Stressors (Scale of 1-5) – In Order of Average Score

Variable	(1) Male N	(1) Male Mean/SE	(2) Female N	(2) Female Mean/SE	(3) Total N	(3) Total Mean/SE	T-test Difference (1)-(2)
Stress Scale: Meeting Repayment Targets	123	3.024	75	3.067	198	3.040	-0.042
		(0.127)		(0.163)		(0.100)	
Stress Scale: Traveling to Farmers/Biking	125	3.024	75	3.120	200	3.060	-0.096
		(0.139)		(0.167)		(0.107)	
Stress Scale: Financial Support in the Community	125	3.080	75	2.880	200	3.005	0.200
		(0.127)		(0.165)		(0.101)	
Stress Scale: Non-compliant Farmers	123	3.041	73	2.904	196	2.990	0.137
		(0.122)		(0.165)		(0.098)	
Stress Scale: Elder Care	125	2.952	75	3.027	200	2.980	-0.075
		(0.125)		(0.165)		(0.100)	
Stress Scale: Distance to Home Village	123	2.943	74	2.932	197	2.939	0.011
		(0.126)		(0.156)		(0.098)	
Stress Scale: Compensation	124	2.952	75	2.773	199	2.884	0.178
		(0.132)		(0.171)		(0.105)	
Stress Scale: Pressure from Managers	125	2.632	75	2.667	200	2.645	-0.035
		(0.121)		(0.160)		(0.096)	
Stress Scale: Hours/Inflexibility	125	2.384	75	2.760	200	2.525	-0.376*
		(0.117)		(0.153)		(0.094)	
Stress Scale: Childcare	125	2.416	74	2.635	199	2.497	-0.219
		(0.128)		(0.173)		(0.103)	
Stress Scale: Emotional Support in the Community	124	2.379	74	2.459	198	2.409	-0.080
		(0.114)		(0.158)		(0.092)	
Stress Scale: Meeting Enrollment Targets	124	2.355	75	2.320	199	2.342	0.035
		(0.128)		(0.152)		(0.098)	
Stress Scale: Family Planning	124	2.145	75	2.427	199	2.251	-0.282
		(0.110)		(0.154)		(0.090)	
Stress Scale: Marital Challenges	125	2.184	75	2.307	200	2.230	-0.123
		(0.121)		(0.161)		(0.097)	
Stress Scale: Socializing in the Community	124	2.137	74	2.324	198	2.207	-0.187
		(0.111)		(0.150)		(0.089)	
Stress Scale: Training Farmers/Ensuring Adoption	123	1.992	75	2.147	198	2.051	-0.155
		(0.114)		(0.153)		(0.091)	
F-test of joint significance (F-stat)							0.726

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are the F-statistics. Standard errors are robust. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table C.3: Top 3 Stressors – In Order of Most Often Selected

Variable	(1) Male Mean/SE	(2) Female Mean/SE	(3) Total Mean/SE	T-test Difference (1)-(2)
Top 3 Job-Related Stressors: Repayment Targets	0.578	0.641	0.602	-0.063
	(0.044)	(0.055)	(0.034)	
Top 3 Life Challenge Stressors: Distance from Home Village	0.625	0.526	0.587	0.099
	(0.043)	(0.057)	(0.034)	
Top 3 Job-Related Stressors: Non-Compliant Farmers	0.547	0.500	0.529	0.047
	(0.044)	(0.057)	(0.035)	
Top 3 Life Challenge Stressors: Elder Care	0.555	0.474	0.524	0.080
	(0.044)	(0.057)	(0.035)	
Top 3 Life Challenge Stressors: Finding People for Financial Support	0.516	0.385	0.466	0.131*
	(0.044)	(0.055)	(0.035)	
Top 3 Life Challenge Stressors: Finding People for Advice/Emotional Support	0.328	0.423	0.364	-0.095
	(0.042)	(0.056)	(0.034)	
Top 3 Job-Related Stressors: Pressure from Managers to Meet Targets	0.406	0.295	0.364	0.111
	(0.044)	(0.052)	(0.034)	
Top 3 Job-Related Stressors: Enrollment Targets	0.312	0.397	0.345	-0.085
	(0.041)	(0.056)	(0.033)	
Top 3 Life Challenge Stressors: Childcare	0.258	0.474	0.340	-0.217***
	(0.039)	(0.057)	(0.033)	
Top 3 Job-Related Stressors: Compensation	0.312	0.282	0.301	0.030
	(0.041)	(0.051)	(0.032)	
Top 3 Life Challenge Stressors: Family Planning	0.312	0.244	0.286	0.069
	(0.041)	(0.049)	(0.032)	
Top 3 Life Challenge Stressors: Finding People to Socialize With	0.219	0.231	0.223	-0.012
	(0.037)	(0.048)	(0.029)	
Top 3 Life Challenge Stressors: Marital challenges	0.188	0.244	0.209	-0.056
	(0.035)	(0.049)	(0.028)	
Top 3 Job-Related Stressors: Long/Inflexible Hours	0.203	0.205	0.204	-0.002
	(0.036)	(0.046)	(0.028)	
Top 3 Job-Related Stressors: Biking	0.000	0.000	0.000	N/A
	(0.000)	(0.000)	(0.000)	
Top 3 Job-Related Stressors: Training Farmers	0.000	0.000	0.000	N/A
	(0.000)	(0.000)	(0.000)	
N	128	78	206	
F-test of joint significance (F-stat)				2.694***

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are the F-statistics. Standard errors are robust. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table C.4: Expectations versus Reality – In Order of Most Difficult Compared to Expectation

Variable	N	(1) Male Mean/SE	N	(2) Female Mean/SE	N	(3) Total Mean/SE	T-test Difference (1)-(2)
Biking	128	2.602	78	2.538	206	2.578	0.063
		(0.099)		(0.126)		(0.078)	
Safety of the Job	128	1.844	78	2.115	206	1.947	-0.272**
		(0.060)		(0.089)		(0.051)	
Recruiting/Training Farmers	128	1.805	77	1.909	205	1.844	-0.104
		(0.081)		(0.108)		(0.065)	
Earnings Farmers'/Community's Respect	128	1.797	78	1.769	206	1.786	0.028
		(0.069)		(0.098)		(0.056)	
Making Friends in the Village	93	2.054	39	1.923	132	2.015	0.131
		(0.107)		(0.139)		(0.086)	
F-test of joint significance (F-stat)							1.064

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are the F-statistics. Standard errors are robust. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table C.5: Top 3 Stressors Restrictive to a Potential Applicant – In Order of Most Often Selected

Variable	(1) Male Mean/SE	(2) Female Mean/SE	(3) Total Mean/SE	T-test Difference (1)-(2)
Top 3 Stressors Restricting Applications: Distance from Home Village	0.359	0.372	0.364	-0.012
	(0.043)	(0.055)	(0.034)	
Top 3 Stressors Restricting Applications: Pressure from Managers to Meet Targets	0.344	0.205	0.291	0.139**
	(0.042)	(0.046)	(0.032)	
Top 3 Stressors Restricting Applications: Non-Compliant Farmers	0.258	0.256	0.257	0.001
	(0.039)	(0.050)	(0.031)	
Top 3 Stressors Restricting Applications: Enrollment Targets	0.242	0.244	0.243	-0.001
	(0.038)	(0.049)	(0.030)	
Top 3 Stressors Restricting Applications: Long/Inflexible Hours	0.188	0.333	0.243	-0.146**
	(0.035)	(0.054)	(0.030)	
Top 3 Stressors Restricting Applications: Compensation	0.227	0.218	0.223	0.009
	(0.037)	(0.047)	(0.029)	
Top 3 Stressors Restricting Applications: Elder Care	0.188	0.154	0.175	0.034
	(0.035)	(0.041)	(0.027)	
Top 3 Stressors Restricting Applications: Marital challenges	0.141	0.192	0.160	-0.052
	(0.031)	(0.045)	(0.026)	
Top 3 Stressors Restricting Applications: Finding People for Advice/Emotional Su	0.141	0.115	0.131	0.025
	(0.031)	(0.036)	(0.024)	
Top 3 Stressors Restricting Applications: Finding People for Financial Support	0.164	0.064	0.126	0.100**
	(0.033)	(0.028)	(0.023)	
Top 3 Stressors Restricting Applications: Family Planning	0.109	0.103	0.107	0.007
	(0.028)	(0.035)	(0.022)	
Top 3 Stressors Restricting Applications: Childcare	0.102	0.090	0.097	0.012
	(0.027)	(0.033)	(0.021)	
Top 3 Stressors Restricting Applications: Finding People to Socialize With	0.055	0.090	0.068	-0.035
	(0.020)	(0.033)	(0.018)	
Top 3 Stressors Restricting Applications: Repayment Targets	0.000	0.000	0.000	N/A
	(0.000)	(0.000)	(0.000)	
Top 3 Stressors Restricting Applications: Biking	0.000	0.000	0.000	N/A
	(0.000)	(0.000)	(0.000)	
Top 3 Stressors Restricting Applications: Training Farmers	0.000	0.000	0.000	N/A
	(0.000)	(0.000)	(0.000)	
N	128	78	206	
F-test of joint significance (F-stat)				1.443

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are the F-statistics. Standard errors are robust. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table C.6: Top 3 Perks when Applied– In Order of Most Often Selected

Variable	(1) Male Mean/SE	(2) Female Mean/SE	(3) Total Mean/SE	T-test Difference (1)-(2)
Top 3 Job Appeal When Applying: Opportunity for Future Promotions	0.578	0.513	0.553	0.065
	(0.044)	(0.057)	(0.035)	
Top 3 Job Appeal When Applying: Pension Fund Benefit	0.336	0.333	0.335	0.003
	(0.042)	(0.054)	(0.033)	
Top 3 Job Appeal When Applying: Health Insurance Benefit	0.312	0.359	0.330	-0.046
	(0.041)	(0.055)	(0.033)	
Top 3 Job Appeal When Applying: Job Tasks	0.312	0.179	0.262	0.133**
	(0.041)	(0.044)	(0.031)	
Top 3 Job Appeal When Applying: Performance-Based Incentives	0.273	0.244	0.262	0.030
	(0.040)	(0.049)	(0.031)	
Top 3 Job Appeal When Applying: Income Certainty	0.242	0.282	0.257	-0.040
	(0.038)	(0.051)	(0.031)	
Top 3 Job Appeal When Applying: Opportunity to Move	0.219	0.205	0.214	0.014
	(0.037)	(0.046)	(0.029)	
Top 3 Job Appeal When Applying: Earning Respect in Community	0.148	0.244	0.184	-0.095
	(0.032)	(0.049)	(0.027)	
Top 3 Job Appeal When Applying: Child Benefit	0.141	0.141	0.141	-0.000
	(0.031)	(0.040)	(0.024)	
Top 3 Job Appeal When Applying: Regular Hours	0.109	0.179	0.136	-0.070
	(0.028)	(0.044)	(0.024)	
Top 3 Job Appeal When Applying: Rent Benefit	0.133	0.103	0.121	0.030
	(0.030)	(0.035)	(0.023)	
Top 3 Job Appeal When Applying: Competitive Compensation	0.094	0.154	0.117	-0.060
	(0.026)	(0.041)	(0.022)	
Top 3 Job Appeal When Applying: Opportunity to Build Resume	0.102	0.064	0.087	0.037
	(0.027)	(0.028)	(0.020)	
N	128	78	206	
F-test of joint significance (F-stat)				1.101

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are the F-statistics. Standard errors are robust. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table C.7: Top 3 Perks Now – In Order of Most Often Selected

Variable	(1) Male Mean/SE	(2) Female Mean/SE	(3) Total Mean/SE	T-test Difference (1)-(2)
Top 3 Job Appeal Now: Opportunity for Future Promotions	0.570	0.423	0.515	0.147**
	(0.044)	(0.056)	(0.035)	
Top 3 Job Appeal Now: Health Insurance Benefit	0.453	0.372	0.422	0.081
	(0.044)	(0.055)	(0.034)	
Top 3 Job Appeal Now: Performance-Based Incentives	0.367	0.295	0.340	0.072
	(0.043)	(0.052)	(0.033)	
Top 3 Job Appeal Now: Pension Fund Benefit	0.281	0.346	0.306	-0.065
	(0.040)	(0.054)	(0.032)	
Top 3 Job Appeal Now: Job Tasks	0.258	0.244	0.252	0.014
	(0.039)	(0.049)	(0.030)	
Top 3 Job Appeal Now: Earning Respect in Community	0.195	0.244	0.214	-0.048
	(0.035)	(0.049)	(0.029)	
Top 3 Job Appeal Now: Child Benefit	0.156	0.244	0.189	-0.087
	(0.032)	(0.049)	(0.027)	
Top 3 Job Appeal Now: Opportunity to Move	0.164	0.179	0.170	-0.015
	(0.033)	(0.044)	(0.026)	
Top 3 Job Appeal Now: Rent Benefit	0.125	0.179	0.146	-0.054
	(0.029)	(0.044)	(0.025)	
Top 3 Job Appeal Now: Income Certainty	0.148	0.141	0.146	0.007
	(0.032)	(0.040)	(0.025)	
Top 3 Job Appeal Now: Competitive Compensation	0.086	0.128	0.102	-0.042
	(0.025)	(0.038)	(0.021)	
Top 3 Job Appeal Now: Opportunity to Build Resume	0.125	0.064	0.102	0.061
	(0.029)	(0.028)	(0.021)	
Top 3 Job Appeal Now: Regular Hours	0.070	0.141	0.097	-0.071
	(0.023)	(0.040)	(0.021)	
N	128	78	206	
F-test of joint significance (F-stat)				1.273

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are the F-statistics. Standard errors are robust. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Appendix D Conceptual Framework

Model Setup

Let x denote a worker's true latent ability. This ability is unobserved by evaluators. Instead, evaluators observe two noisy signals:

$$\hat{x}_1 = x + e_1 + \varepsilon_1, \quad \varepsilon_1 \sim \mathcal{N}(0, \sigma_1^2)$$
$$\hat{x}_2 = x + e_2 + \varepsilon_2, \quad \varepsilon_2 \sim \mathcal{N}(0, \sigma_2^2), \quad \sigma_2^2 \gg \sigma_1^2$$

where e_1 and e_2 are the worker's effort to improve signals 1 and 2, respectively, and ϵ_1 and ϵ_2 are noise terms following normal distributions. Effort is costly and $c(e_1, e_2) = \frac{1}{2}\kappa_1 e_1^2 + \frac{1}{2}\kappa_2 e_2^2$ denotes the cost of effortful action to improve signals.

Signal Informativeness. The informativeness of a signal $\hat{x}_j$ is defined as the inverse of its noise variance:

$$I_{\hat{x}_j} := \frac{1}{\sigma_j^2}$$

This reflects how precisely the signal approximates the latent characteristic x in expectation.

Cross-Signal Correlation. Let $\hat{x}_1$ be the most informative signal, i.e., $\sigma_1^2 < \sigma_2^2$. The cross-signal correlation $\hat{x}_2$ is defined as:

$$R(\hat{x}_2; e_1, e_2) := Corr(\hat{x}_1, \hat{x}_2)$$

where the correlation is computed over the realized joint distribution of signal values in the data. Cross-signal correlations can then vary with effort. High cross-signal correlation indicates that realizations of the less-informative signal reliably track those of the more-informative signal. Signal distortion occurs when this correlation weakens, for example due to strategic effort.

Applicant's Problem

Applicants believe hiring is based on an index S and a benchmark τ, where they are hired if $S \geq \tau$ and:

$$S = w_1\hat{x}_1 + w_2\hat{x}_2$$

They choose (e_1, e_2) to maximize expected utility:

$$\max_{e_1, e_2 \geq 0} \quad \Phi\left(\frac{w_1(x + e_1) + w_2(x + e_2) - \tau}{\sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2}}\right) - \frac{1}{2}\kappa_1 e_1^2 - \frac{1}{2}\kappa_2 e_2^2$$

Let z denote the normalized signal index:

$$z(e_1, e_2) := \frac{w_1(x + e_1) + w_2(x + e_2) - \tau}{\sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2}}$$

The first-order conditions are:

$$\frac{\partial U}{\partial e_1} = \phi(z) \cdot \frac{w_1}{\sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2}} - \kappa_1 e_1 = 0$$
$$\frac{\partial U}{\partial e_2} = \phi(z) \cdot \frac{w_2}{\sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2}} - \kappa_2 e_2 = 0$$

Effort increases with x, and signal weights; effort decreases with cost and noise.

Comparative Statics: Uniform Shock to the Hiring Threshold

Let the perceived threshold τ decrease. That is, applicants' perceptions about the job change such that believe they need to cross a lower threshold of skill in order to be hired for the job. Note that:

$$\frac{\partial e_1^*}{\partial \tau} = -\frac{\phi'(z) \cdot \frac{\partial z}{\partial \tau} \cdot \frac{w_1}{\sigma_S}}{\kappa_1 + \phi''(z) \cdot \left(\frac{w_1}{\sigma_S}\right)^2} < 0$$
$$\frac{\partial e_2^*}{\partial \tau} = -\frac{\phi'(z) \cdot \frac{\partial z}{\partial \tau} \cdot \frac{w_2}{\sigma_S}}{\kappa_2 + \phi''(z) \cdot \left(\frac{w_2}{\sigma_S}\right)^2} < 0$$

where $\sigma_S^2 = w_1^2\sigma_1^2 + w_2^2\sigma_2^2$ and $\frac{\partial z}{\partial \tau} < 0$. Thus, a decrease in τ raises both e_1^* and e_2^*, with a larger proportional response in e_2^* if κ_2 is lower or $w_2/\kappa_2 > w_1/\kappa_1$.

Who Responds Most?

Recall that x denotes latent skill. Differentiating e_j^* with respect to x:

$$\frac{\partial e_j^*}{\partial x} = \frac{\phi'(z) \cdot \frac{\partial z}{\partial x} \cdot \frac{w_j}{\sigma_S}}{\kappa_j + \phi''(z) \cdot \left(\frac{w_j}{\sigma_S}\right)^2}$$

This expression is maximized when $z \approx 0$, i.e., when agents are near the hiring margin. Therefore, those with $x \approx \tau$ exert the most responsive effort to threshold changes. Thus, a negative shock to τ will result in a less-skilled set of applicants being closer to the threshold, and a less-skilled set of applicants being induced to exert more effort.

Evaluator Problem

Now we introduce differences by gender. Evaluators use different weighting rules for men and women. For men, evaluators only use signal $\mathbb{x}_2$ if the more informative signal $\mathbb{x}_1$ clears a threshold τ_1. Conversely, evaluators weight $\mathbb{x}_1$ and $\mathbb{x}_2$ simultaneously for women:

$$S^M = w_1\hat{\mathbb{x}}_1 + w_2\hat{\mathbb{x}}_2 \cdot \mathbb{1}\{\hat{\mathbb{x}}_1 \geq \tau_1\}$$
$$S^F = w_1\hat{\mathbb{x}}_1 + w_2\hat{\mathbb{x}}_2$$

Let x_g^{hire} be the expected value of x among hired applicants from group $g \in \{M, F\}$.

Assuming identical applicant responses to the hiring threshold, then:

- x_F^{hire} will be strictly lower than x_M^{hire} if *(i)* $\tilde{w}_2$ is large, and *(ii)* $\hat{\mathbb{x}}_2$ is inflated by low-x applicants in response to the hiring shock.
- Evaluators place weight on uninformative signals among women (regardless of $\hat{\mathbb{x}}_1$), so distorted signals reduce the skill level of selected applicants.

Comparative Statics and Final Result

- The threshold shock induces effort responses, especially among lower-skilled applicants.
- If evaluators interpret signals symmetrically, then most marginal hires are still reasonably skilled.
- If evaluators interpret signals differently (bias), and women's soft signals are overweighted, then many hires will be:
  - Low in true skill x
  - High in e_2 (inflated soft signal)
- Therefore, **the combination of signal distortion and evaluator bias leads to the worst hiring outcomes**: evaluators hire marginally more women, but from a lower-quality pool.