Hidden Discrimination in Frictional Labor Markets

Elisa Macchi* and Claude Raisaro†

Version: November 22, 2026.

Abstract

The production environment determines which attributes firms value in workers and, when employers infer these attributes from identity, shapes hiring discrimination even when beliefs and preferences over identity stay the same. We test this prediction with 915 employers in male-dominated sectors in Uganda. Under business-as-usual conditions, female trainee profiles are 10.6 percentage points less likely to be selected than otherwise identical male ones. Gender-neutral monitoring visits that mitigate one friction at a time, holding candidate information fixed, move this gap in opposite directions: theft monitoring widens it by 58%, while safety monitoring reduces it by 80%. Neither changes beliefs about candidates. Broad performance signals raise selection without narrowing the gap, and trainee data indicate that employers misperceive gender differences on both trust and technical skills. The hiring gender gap under business as usual combines inaccurate statistical discrimination against men with a larger residual penalty against women. This offset hides discrimination: estimated bias depends on which attributes researchers observe and treat as productive. It also creates a second-best problem: mitigating one friction moves hiring further from equal treatment.

*Brown University (email: elisa_macchi@brown.edu), BREAD, and CEPR

†Geneva Graduate Institute (email: claude.raisaro@graduateinstitute.ch)

‡We thank Siwan Anderson, Justus Bamert, Abhijit Banerjee, Arielle Bernhardt, Aditi Bhowmick, Aislinn Bohren, Laura Boudreau, Nina Buchmann, Lorenzo Casaburi, Pascaline Dupas, Ada González-Torres, Jonas Hjort, Peter Hull, Alex Imas, Supreet Kaur, Teddy Mekonnen, Rocco Macchiavello, Kristina Manysheva, Muriel Niederle, Rohini Pande, Frank Schilbach, Anna Vitali, and Roberto Weber for helpful discussions. Josh Bwiira and Carlotta Riva provided outstanding field management and research assistance. We also thank audience participants at BREAD, Brown, CEPR, EBRD, Geneva Graduate Institute, KU Leuven, MIT, NBER Development, NEUDC, NYU, Princeton, USI, and Zurich Behavioral Lab Meeting. We used Claude Code (Anthropic) for data analysis. The experiments were approved by Mildmay Uganda (0408-2023) and the Uganda National Council for Science and Technology (SS2574ES), and preregistered at the AEA RCT Registry (AEARCTR-0013698).

1 Introduction

Employers value social skills and personality traits alongside technical skills (Deming, 2017; Deming and Silliman, 2025) and commonly associate these attributes with group identity. U.S. employers describe immigrants as hardworking and Black workers as lacking work ethic (Moss and Tilly, 2001), while economics reference letters place greater emphasis on diligence for female candidates and on brilliance for male candidates (Eberhardt et al., 2023). Each association may generate either statistical discrimination, when beliefs are accurate and concern unobserved productive attributes, or bias, when beliefs are inaccurate or attributes are valued for reasons unrelated to productivity. When multiple associations operate simultaneously, however, perceived strengths on some attributes may offset perceived weaknesses on others. Which attributes drive hiring disparities, and what do outcome gaps imply about discrimination?

We posit that the production environment shapes the set of worker attributes employers value and, in turn, hiring disparities when employers infer these attributes from identity. An identity-neutral change in the production environment can shift hiring disparities by altering which attributes matter, even if beliefs and preferences over identity remain unchanged. In a given environment, the observed hiring gap reflects employers' beliefs about various worker attributes and any preferences over identity. Beliefs can favor different groups and may or may not be accurate. This has two implications for measurement. First, estimating bias requires researchers to distinguish each attribute's contribution to the hiring gap. Because these attributes extend beyond technical skills, estimates also depend on which attributes researchers treat as productive. Second, favorable treatment on one dimension can offset a penalty on another, including one driven by bias. The observed hiring gap can therefore hide the extent of discrimination.

In this paper, we study whether mitigating workplace frictions affects gender discrimination in hiring and what observed hiring gaps reveal about underlying discrimination. We conduct an experiment with 915 employers in motor mechanics, welding, and carpentry in Kampala, Uganda. Employers evaluate vocational trainee profiles with randomly assigned gender and technical performance signals, while we experimentally vary monitoring support to prevent theft or workplace harassment. We also test for misperceptions by comparing employers' gendered beliefs with trainees' skills and behavior, and examine how residual estimates of bias depend on which attributes researchers observe and treat as productive.

Uganda provides a useful setting to study these questions. Despite high female labor force participation—76.5%, above the sub-Saharan African average—employment remains highly segregated by gender.[1] The sectors we study are heavily male-dominated and among the country's highest-return vocational occupations. Efforts to increase women's entry into these occupations have largely focused on the supply side. The effectiveness of these interventions, however, depends on employers' willingness to hire women, which remains understudied in sub-Saharan Africa.[2]

To motivate our perspective, we present survey evidence on the attributes employers value in workers and their beliefs about how these attributes vary by gender. Employers rate trustworthiness as the most important attribute, followed by cooperation, which they rate as important as technical skills (Figure 1). Men are perceived as stronger on technical performance, while women are perceived as more trustworthy and less likely to steal. Employers worry about perceived gaps in women's technical skills, but even more about harassment risk. Employers may accurately perceive gender differences in attributes that matter for productivity when monitoring is difficult and teamwork is common.[3] However, the same assessments may reflect stereotypes about women's abilities, or they may serve to justify a preference for hiring men.

We turn to an experiment to test whether these gendered beliefs shape hiring decisions by independently varying the workplace frictions employers face and the information they receive about candidates' technical skills. Through partnerships with local training institutes, we offer 915 employers access to vocationally trained candidates, including qualified women.[4] We measure hiring preferences through an incentivized ratings design. Employers evaluate hypothetical profiles with randomly assigned gender and technical perfor-

[1]Segregation across occupations and firms contributes to gender earnings gaps (Card et al., 2016; Blau and Kahn, 2017; Goldstein et al., 2019; Morchio and Moser, 2026).

[2]Research on women's employment in low-income countries emphasizes barriers to participation (Heath et al., 2024), including mobility constraints (Garlick et al., 2025; Abou Daher et al., 2025), social norms (Bernhardt et al., 2018; Bursztyn et al., 2020), and limited pay transparency (Jalal, 2026). Training programs show mixed effects on women's employment (Groh et al., 2016; Bandiera et al., 2020). At the aggregate level, Goldberg et al. (2025) estimate that demand distortions account for most of the productivity losses from gender misallocation. Micro evidence shows pay and promotion disparities in sectors where women already work (Brown, 2023; Ashraf et al., 2025; Macchiavello et al., 2026).

[3]Monitoring and other agency frictions shape how firms organize work in low-income countries (see Breza and Kaur 2025 for a review). Firm-level studies show how agency frictions limit delegation (Bloom et al., 2013; Kotia, 2025) and cooperation frictions reduce team productivity (Hjort, 2014). Gendered trust appears related to these frictions: in Zambia, Ashraf et al. (2026) show that female entrepreneurs distrust and avoid dealing with men under weak contract enforcement.

[4]Firms in our sample otherwise recruit primarily through personal networks, where women are underrepresented.

mance signals, knowing that their choices guide referrals to similar candidates. Across employers, we randomize offers of monitoring support to prevent theft or workplace harassment, with business as usual as the control.[5] This varies the production environment while holding candidate information fixed across arms. To assess the accuracy of employers' gendered beliefs, we compare their incentivized predictions with the skills and behavior of 173 trainees from the partner institutes.

Employers report a stated demand for female workers that is larger than observed employment: 85% say their ideal workforce includes women, but only 13.6% currently employ any. Offering access to qualified female candidates therefore creates scope for hiring women. In the business-as-usual condition, 89.7% of employers select at least one female profile, though female profiles are 10.6 percentage points less likely to be selected than otherwise identical male profiles. These choices track stated preferences: the gap narrows as employers' preferred female share rises and disappears among those whose ideal workforce includes between 40% and 60% women.[6]

We next test whether the production environment affects disparate treatment by gender by reducing workplace frictions that make worker attributes valuable. Theft monitoring visits should lower the value of honesty, an attribute on which women hold a perceived advantage, and therefore widen the hiring gender gap. Safety monitoring visits should ease harassment concerns associated with hiring women and narrow the gap. Both treatments provide the same amount of monitoring and support. The opposite predictions distinguish our hypothesis from the main alternatives, such as scrutiny effects or experimenter demand, which predict effects in the same direction across the two types of visits. We hold candidate information fixed across treatment arms and use gender-neutral language in both scripts. Workers learn of the visits only after accepting a job, preventing selection into jobs based on treatment assignment. Employers and workers know visits will occur, but their timing is unannounced. Both scripts mention theft; the safety script explicitly states that visits cannot discourage stealing or dishonesty. Comparisons between the two types of visits therefore help distinguish theft salience from its mitigation. To test whether each treatment reduces its targeted concern without changing beliefs about work-

[5]For ethical reasons and as preregistered, after a placement we surprise all employers, including those in the business-as-usual arm, with visits covering both theft and safety.

[6]We follow employers after the experiment: two years later, the share of firms employing at least one woman has risen from 13.6% to 23.6% ($p < 0.001$), and the average preferred female share from 21.7% to 27.0% ($p < 0.001$). Employers who hired women do not lower their demand for them: their preferred female share rises by 6.8 percentage points, against 3.2 among other firms ($p = 0.011$). This is consistent with employers acting on the demand they expressed in the experiment once connected to trained female candidates.

ers, we ask employers to rate each candidate's skills and trustworthiness and to report their worries about theft, harassment, and two placebo risks (low effort and turnover).

The two treatments move the hiring gender gap in opposite directions, as predicted. Overall selection rises by 5.0% under theft monitoring visits and by 6.6% under safety monitoring visits, but through different margins. Theft monitoring visits raise male selection by 5.4 percentage points ($p < 0.001$), with no detectable effect on female selection ($p = 0.683$), widening the gap from 10.6 to 16.7 percentage points. Safety monitoring visits raise female selection by 7.3 percentage points ($p < 0.001$), with no detectable effect on male selection ($p = 0.467$), reducing the gap to 2.1 percentage points, statistically indistinguishable from zero. Two further results support the mechanism. First, each type of visit reduces worries only about the friction it targets. Theft monitoring visits reduce theft worries (Holm-adjusted $p = 0.004$), and safety monitoring visits reduce safety worries (Holm-adjusted $p = 0.068$), with no significant effect on the placebo worries. Second, neither treatment detectably changes employers' beliefs that women are less technically skilled and more trustworthy, implying that the hiring effects do not operate through new information about workers.

The hiring gap under business as usual understates the penalty women face because their perceived trustworthiness offsets other disadvantages. Theft monitoring reveals more of the underlying penalty against women by reducing the value of their perceived trustworthiness advantage, widening the hiring gap by 58%. The effect is strongest among employers who state gender-neutral hiring preferences: their gap widens by 17.0 percentage points ($p < 0.001$), nearly three times the average effect, reaching a level similar to that of other employers. This suggests that much of their apparent willingness to hire women reflects the value of women's perceived trustworthiness when monitoring is difficult. By a similar logic, the near-zero gap under safety monitoring does not imply that women face no penalty. Safety concerns explain about half of the penalty against women; once they are mitigated, the remaining penalty is offset by women's trust premium.

Performance signals do not explain the underlying penalty against women. A top rather than median ranking in the VTI cohort raises selection by 9.2 percentage points (22%, $p < 0.001$) and a certificate by 15.1 percentage points, yet neither detectably narrows the gender gap against women in any arm. The gap is largest among top-ranked candidates with a certificate, at 16.2 percentage points compared with 9.1 among median-ranked candidates without one ($p = 0.047$ for the difference). While perceived differences in other skills such as strength could explain the patterns, the selected trainee pool and the

practical skills assessed by the ranking suggest that the gap likely reflects a preference for hiring men or a tendency to give women less credit for the same qualifications.

We next ask how estimated bias against women changes when we account for trustworthiness and harassment risk alongside technical performance. Using a discrete-choice model, we decompose the hiring gap into a trust premium for women, a harassment penalty, and a residual estimate of bias.[7] Relative to the full model, omitting the trust premium understates bias against women, while omitting the harassment penalty overstates it. Which estimate best captures bias depends on which attributes researchers treat as productive. Employers' safety concerns, for example, may reflect expected output losses from harassment, reputational concerns, or a desire to protect women. Estimates of bias range from 2.1 to 16.8 percentage points across models.

We also test whether employers accurately perceive gender differences in technical skills and trustworthiness. We compare employers' incentivized predictions for male and female trainees with outcomes for 173 vocational trainees. We measure technical skills with a sector-specific theory exam and self-reported practical ability, following Alfonsi et al. (2020). We measure trustworthiness with a novel cheating task in which over-reporting output earns a bonus but may carry reputational costs. Employers expect women to perform worse on both skill measures, yet men and women report similar practical ability, and women score slightly higher on theory. Employers also expect nearly half of men to cheat when unsupervised (six times the observed rate) and fewer women to do so. In fact, cheating is rare and, if anything, more common among women.[8] These findings echo prior work on misperceptions (Bordalo et al., 2019; Bursztyn and Yang, 2022). Combined with the decomposition, they reveal a second-best problem. The trust premium rests on inaccurate beliefs, yet it partly offsets harassment concerns and any underlying bias against women. Making theft easier to monitor reduces the premium and moves hiring further from equal treatment.

This paper makes two contributions to the literature on hiring discrimination. Standard models emphasize employers' preferences over identity (Becker, 1957) and beliefs about worker attributes (Phelps, 1972; Arrow, 1973; Aigner and Cain, 1977; Bohren et al.,

[7]We interpret the residual as bias given that stronger performance signals do not close the hiring gap and under the assumption that no unmeasured differences in productivity between male and female trainees account for it.

[8]Previous work also finds similar technical skills among male and female vocational trainees in these sectors (Alfonsi and Ferreira, 2026). Evidence on gender differences in prosocial behavior is mixed (Niederle, 2016), though people often overestimate how much more prosocial women are than men (Exley et al., 2025).

2025a). Taking a multidimensional view of workers' skills (Woessmann, 2025), we show that, given the same beliefs and preferences, discrimination depends on the production environment, which shapes the returns to attributes employers infer from identity. This connects to Hurst et al. (2024), who use a model of occupational sorting to quantify how task returns interact with racial barriers to shape the Black–White wage gap in the U.S.. We randomly offer monitoring support to identify how changing returns affect direct discrimination in hiring, measuring employers' hiring preferences separately from supply responses. The design isolates each attribute's contribution to the gap without changing candidate information, since information about one skill can shift beliefs about others.

Our second contribution concerns the measurement of discrimination when identity signals multiple attributes. Experiments typically condition on qualifications or past performance and assess bias by varying what decision-makers learn about that same dimension (Bertrand and Mullainathan, 2004; Sarsons, 2017; Coffman et al., 2021; Kline et al., 2022). Such comparisons can miss earlier discrimination that shaped applicants' qualifications (Bohren et al., 2019, 2025b). We show experimentally that even within a single hiring decision, a perceived advantage on an unmeasured attribute can partly or fully offset a belief-based or taste-based penalty against minority applicants. Just as omitting a productive disadvantage can overstate bias (Lang and Spitzer, 2020), omitting a productive advantage can understate it. Even when all relevant attributes are observed, estimating bias requires assumptions about how much of each return reflects productivity and how much reflects employers' preferences.

The paper also connects to a growing micro literature on how frictions shape firms in poor countries (see Bergquist et al., 2026, for a review), including incentive contracts, technology adoption, and hiring (Atkin et al., 2017; Kelley et al., 2024; Houeix, 2026). When workers are hard to monitor, firms may rely on relatives they trust or use referrals to mitigate moral hazard (Beaman and Magruder, 2012; Heath, 2018; Chandrasekhar et al., 2020). We show experimentally that this logic extends beyond personal ties. When theft is hard to detect, employers place a premium on applicants they expect to act honestly without supervision, and because they use gender to infer whom they can trust, this provides an advantage for women. Previous work finds that employers' responses to frictions disadvantage women through referral hiring that excludes them (Beaman et al., 2018) or safety concerns that keep them out of jobs they would take (Buchmann et al., 2023). Our two ex ante gender-neutral interventions have opposite distributional effects, showing that which group benefits depends on which friction binds and what employers believe about each group.

2 Setting, Sample, and Motivating Evidence

2.1 Setting

We study motor mechanics, welding, and carpentry firms in Kampala, Uganda. These sectors are an important source of manufacturing and youth employment in the country.[9] Despite their importance as employment opportunities, women remain largely absent from these sectors.

Firms in this setting recruit both vocationally trained workers and workers whom they train on the job, largely through personal networks. Vocational training institutes (VTIs) generally provide training rather than job placement, so trained workers also depend on recruitment connections to access employment. Firms can develop comparable sector-specific skills through on-the-job training, but this requires the owner's time.

Training is one of several responsibilities that owners and managers combine. Firms exhibit limited labor specialization (Bassi et al., 2023): employers perform technical and administrative work while training and supervising workers. Production also requires workers to cooperate, especially in mechanics, where teamwork is common. Worker identity and social preferences can affect these interactions and their productivity (Hjort, 2014; Breza et al., 2018; Ashraf et al., 2026).

These competing responsibilities constrain supervision. Employers work an average of 10.4 hours per day, including 2.1 hours monitoring workers, their second most time-consuming activity after technical work. Yet over 90% report wanting to increase monitoring, and 83% have experienced theft by an employee. Workshop layouts add to this difficulty: firms typically combine indoor workspaces with large outdoor areas, making it hard to observe all workers at once (Appendix Figure A4, panel A). Four out of five employers use piece-rate pay, which helps address moral hazard related to effort but does not directly address theft or concerns that workers may take clients from the firm.

[9] Carpentry and welding, together with grain milling, account for 30% of manufacturing employment (Bassi et al., 2023). Motor mechanics and welding are among the highest paying of eight vocational sectors that together employ 25% of employed Ugandans aged 18–25 (Alfonsi et al., 2020).

2.2 Sample

Firms and Employers. Our employer sample consists of 915 employers, each representing a different firm: 313 at motor mechanics firms, 299 at carpentry firms, and 303 at welding firms (Table 1). We collected the data in two waves. The first wave, in 2024, covered all three sectors. The second wave, in 2025, added welding and carpentry firms to balance the sample across sectors (Appendix A.5).[10]

Our sample is equivalent in size to 30.0% of the firms in these sectors recorded in Kampala's 2020 business census and 27.8% in the 2010 census (Uganda Bureau of Statistics, 2011, 2020).[11] Study firms are broadly comparable to those in the 2020 census on observed characteristics: mean firm age is 10.1 years versus 11.5, and 88.0% report annual revenues above UGX 10 million versus 80.2%. Firms in our sample average 8.3 employees, excluding owners, compared with 7.4 in the 2020 census.[12]

At each firm, we interviewed the person with hiring authority (employer). Of these respondents, 85.7% are owners and 12% are managers. Almost all are men (96.4%). They average 37 years of age and 15 years of experience in the sector. Sixty-three percent have at least some secondary education, and 31% attended a vocational training institute.

Across the three sectors, employers report similar revenues and similarly high levels of concern about their capacity to monitor workers. Nearly half of all study firms recruit mainly through family referrals (45.6%), consistent with evidence that networks help address moral hazard when supervision is limited. One difference is that mechanics firms operate on a larger scale, employing 15.2 workers on average, roughly three times as many as in carpentry (4.5) or welding (5.0). In these larger workshops, owners more often share supervision with managers. Managers help monitor workers in 58% of mechanics firms, compared with 20% in carpentry and 27% in welding. The sectors also differ in how they find workers. In mechanics, 38% of firms recruit from VTIs, compared with just 9% in carpentry and 7% in welding.

[10] We selected neighborhoods in greater Kampala where these sectors are clustered (Appendix Figure A3). Within these neighborhoods, we listed all firms in the three sectors. Firms were eligible if they hired workers and had fewer than 100 employees. Respondents were at least 18 years old, spoke English or Luganda, and expressed interest in recruiting a new worker on probation. We interviewed employers at consenting firms.

[11] We benchmark study firms against fixed-location firms in Kampala's motor mechanics, welding, and carpentry sectors in the 2010 and 2020 Censuses of Business Establishments (Uganda Bureau of Statistics, 2011, 2020); see Table 1.

[12] Unlike our sample, the census includes firms with no employees besides the owner, which do not meet our eligibility criteria.

Women in the workforce are rare. They account for 2.3% of employees (172 of 7,589), and fewer than 14% of firms employ any women. Employers estimate that women make up only 3.6% of their applicants, and 80.3% report no female applicants. This pattern reflects the sharp occupational segregation documented in previous studies. In the firm census used in Alfonsi et al. (2020), women account for 1%–7% of employees in five male-dominated sectors (motor mechanics, welding, plumbing, construction, and electrical work), compared with 69%–83% in three female-dominated sectors (catering, tailoring, and hairdressing).

Trainees. We collaborate with seven VTIs in Kampala that train students across a range of sectors. Our partnerships allow us to collect information on trainees and connect employers with candidates seeking work. Our primary partner is the Smart Girls Foundation, a nonprofit offering nationally recognized vocational training through its "Girls With Tools" program. The program trains women in male-dominated sectors, including automotive mechanics, electrical installation, welding, and carpentry, and also trains male students (Appendix Figure A4, panel B).

Through these partnerships, we survey 182 trainees across seven institutes. Our sample consists of the 173 trainees whose training or desired occupation falls within the study sectors' hiring pools: 84 in mechanics, 36 in carpentry, and 53 in welding (Table 1, panel C).[13] Trainees are 20.8 years old on average, and their programs typically last 24 months. They receive formal training in sector-specific skills that firms also develop through on-the-job training.

VTIs offer a pool of skilled candidates with greater female representation than firms' usual applicant pools, potentially easing constraints on hiring women. Yet only 18.4% of employers report contacting or searching at VTIs; firms primarily recruit through referrals or candidates approaching them directly. Women account for 36.4% of our trainee sample, or 19.5% excluding Girls With Tools, compared with 3.6% of firms' usual applicants. Female trainees are unevenly distributed across partner institutes, so employers' access to them may depend on their links to particular centers.

[13]We assign trainees by training sector or, for related programs, stated desired occupation. The mechanics pool includes 21 electrician trainees, whom firms hire as mechanics. We exclude nine trainees in plumbing, hairdressing, tailoring, and information and communications technology who do not name a study sector as a desired occupation.

2.3 Demand for Worker Attributes

We ask employers to rate, on a 0–10 scale, the importance of several worker attributes when hiring.[14] Good behavior, described as trustworthiness, honesty, and not stealing, stands out: 68.2% of employers rate it 10 (Figure 1, panel A). This emphasis is consistent with the monitoring constraints and theft risk described in Section 2.1. The share rating each attribute 10 is 29.0% for technical skills, 28.3% for cooperation, 26.0% for effort, and 15.4% for interest in learning. The relatively lower priority given to technical skills when hiring is consistent with employers developing these skills through on-the-job training.

The attributes employers value most are also among the hardest to find or monitor, consistent with workplace frictions shaping demand for worker attributes (Figure 1, panel B). Good behavior ranks first, cited by 90.3% of employers, followed by cooperation (45.8%), interest in learning (41.6%), and technical skills (41.0%). Effort ranks last, at 33.9%, consistent with widespread piece-rate pay, which ties compensation to output and reduces the need to monitor effort (Foster and Rosenzweig, 1994). Both features are common in low-income labor markets, where firms have limited information about workers' skills (Carranza et al., 2022) and non-cognitive skills are hard to verify (Bassi and Nansamba, 2022).

2.4 Gender Signals Multiple Attributes

Employers associate gender with several worker attributes. Following Macchiavello et al. (2026), we ask whether men are better, women are better, or there is no difference for each of the five traits above.[15] Most employers perceive men as more technically skilled (79.3%), harder working (71.1%), and more interested in learning (53.8%). By contrast, 79.7% perceive women as better behaved or less likely to steal (Figure 1, panels C and

[14] We developed the list with local enumerators from pilot answers to the question "What is the single most important trait you look for when hiring a worker?" We validate the list by asking the same question in the main survey; only 0.97% of human-coded responses fall outside the pilot categories. In the 2024 wave, enumerators recorded answers verbatim. We classify these responses using GPT-5 mini, whose primary category matches at least one independently human-coded category for 96.6% of responses. In the 2025 wave, enumerators recorded the closest category without reading the options aloud. We pool the two waves. Appendix A.1 describes the procedure.

[15] We ask these questions after the hiring experiment to avoid priming gender before employers evaluate candidate profiles. The same qualitative patterns appear in pilot data collected without the experiment and when we restrict the employer sample to the business-as-usual arm.

D). Employers also more often favor women on cooperation: 45.8% say women are better, compared with 29.8% who favor men.

These five traits capture employers' priorities in a predominantly male workforce. To identify concerns specific to hiring women, we ask employers an open-ended question about what worries them when doing so (Figure 1, panel E).[16] Harassment of women by male workers is the most common concern, cited by 39.4% of employers. Viewed through the five attributes above, this concern reflects a perceived disadvantage in cooperation for women arising from the firm's predominantly male workforce. The answers also clarify the technical skill gaps that employers perceive: 26.0% raise concerns about women's technical skills, mostly mentioning physical strength. Not a single employer mentions physical strength when asked which worker attribute matters most, suggesting that employers apply different criteria when assessing female workers. Effort is cited by 21.6%, while 15.1% mention learning or interest, mostly because they expect women to leave. Only 4.4% mention good behavior, and none specifically mentions stealing. Gender thus signals multiple attributes and workplace risks, with perceived advantages to hiring women on some dimensions and disadvantages on others.

3 Hiring Experiment

This section describes the experiment testing whether the production environment shapes the hiring gender gap. Within an incentivized ratings exercise, we randomize candidate gender and technical performance signals. Across employers, we randomize monitoring support to prevent theft or harassment, holding candidate information fixed. We then introduce a simple framework to explain how each type of monitoring should affect the gap and present the empirical strategy.

3.1 Incentivized Ratings Design

We offer employers easier access to vocationally trained candidates whose qualifications and performance are documented. These referrals supplement firms' usual recruitment channels and help them find candidates matching their hiring needs, including qualified women who rarely appear in their applicant pools. We do not emphasize women's inclusion

[16] Appendix A.1 describes how we classify the open-ended responses.

when presenting the referral opportunity. Gender-related questions are embedded among general survey questions, and beliefs about gendered worker attributes are elicited after the rating task.[17] Employers spend close to two hours on the interview on average, suggesting that they value the opportunity to receive referrals.

We measure demand for these candidates using an incentivized ratings design (Kessler et al., 2019). Employers know that the profiles are hypothetical, but their choices guide referrals to candidates from partner VTIs who match their preferences. Before rating profiles, employers are told that accurate ratings lead to better matches, giving them an incentive to report their preferences accurately, but we do not disclose the exact rating-to-referral rule.[18]

Profiles. We build the hypothetical profiles from administrative records on 472 trainees enrolled with partner VTIs in Kampala since 2016. Using these data, we create 24 base profiles that include age, marital status, possession of a driver's license, language spoken, motivation for career choice, education, and VTI training (institute and duration). All profiles include the VTI director's contact details. Profiles vary in whether they report formal certification from Uganda's Directorate of Industrial Training (DIT).

We generate four versions of each base profile by cross-randomizing gender and trainee performance. Gender is signaled by a male or female avatar. Performance is shown as a star rating anchored to the actual performance distribution of the trainee's VTI cohort: three stars indicate median performance, and five stars indicate performance above the 95th percentile. Training is strongly practical, with trainees performing tasks used in these occupations (Appendix Figure A4, panel B). The ranking therefore provides a broad signal of performance, summarizing instructors' assessments of theoretical knowledge and practical ability, as well as effort and interest in learning. Crossing the two genders with the two performance levels yields 96 unique profiles (4 versions $\times$ 24 base profiles; see Figure 2 for an example).

Each employer evaluates all 24 base profiles during regular work hours, seeing one randomly assigned version of each. The pool contains equal numbers of male and female candidates, with performance varying within employer. Because the gender mix of available

[17] The stated ideal gender mix is the only gender-related question asked before the experiment, approximately 30 questions before the rating task.

[18] The wording is: "When you rate an applicant positively, it is more likely that we refer to you one or more workers with the characteristics you like.[...] It is in your best interest to select those you actually would like to hire."

VTI candidates depends on employers' links to particular institutes, we use a balanced pool to measure preferences given equal access to qualified men and women. This also increases statistical power to detect gender differences in selection.

Outcomes. Our primary outcome is whether the employer wants to hire a worker matching the profile for a probationary period:

- *Hire*: "Do you want us to refer you a similar worker to start the probation period at your firm?" (Yes/No).

A "yes" means the employer wants a similar worker to start a probationary job, not attend another interview. All hiring in this setting begins with probation.[19] The question is administered in Luganda ("Wandyagadde tukuweeyo omukozi eyefananyirizako ono okutandiika okugezesebwa mu business yo?"). The terms *tukuweeyo* ("send/place with you") and *okugezesebwa* ("to be placed on probation") make this meaning explicit. Wages are not a primary outcome because compensation is difficult to compare across firms: apprentices may be unpaid or receive piece rates, stipends, or in-kind payments that are difficult to value.

Secondary outcomes measure employers' assessments of each candidate along four dimensions. All questions are administered in Luganda; the English translations are the following:

- *Technical performance*: "How would you rate the worker profile's technical skills and work quality?" Responses range from 0 (very low quality) to 10 (very high quality).
- *Trustworthiness*: "How would you rate the worker profile's trustworthiness and honesty (good behavior)?" Responses range from 0 (not at all trustworthy/honest) to 10 (very trustworthy/honest).
- *Effort*: "How likely is this worker profile to put in effort (hardworking, not lazy, concentrated, punctual)?" Responses range from 0 (very unlikely) to 10 (very likely).

[19] We preregistered a second primary outcome measuring expected retention after probation. A programming error omitted it from the 2024 wave, and therefore we did not collect it in 2025. Referrals were unaffected because they depend on the probationary hiring choice. See Appendix A.5.

- *Expected earnings*, a proxy for perceived labor market outside options: "What is your best guess of the monthly earnings of this worker profile a year from now?"

We also track employers' choices beyond the rating exercise, as detailed in Section 6. We recorded referral, matching, and monitoring-visit outcomes through December 2025. In August 2026, we conducted a follow-up phone survey of all sample firms to measure subsequent hiring.

3.2 Experimental Variation in Workplace Monitoring

We randomly assign employers to a business-as-usual condition or offers of monitoring support to prevent theft or workplace harassment. Employers evaluate the same profiles across the three arms. Both treatments offer weekly visits by trained VTI staff during the one-month probation period, at no cost to the firm. Theft monitoring visits (MT) focus on new workers' conduct, discouraging theft and dishonesty. Safety monitoring visits (MS) focus on how new workers are treated, preventing harassment and disrespect. Both supplement employers' own supervision and increase their capacity to monitor the targeted behavior. Employers are told that support is allocated by lottery and does not reflect characteristics of the firm or workers.[20]

Monitoring support may be particularly valuable when firms hire outside familiar networks. Profiles provide information about candidates' training and technical performance, but their behavior and workplace interactions remain uncertain. The visits mirror the supervision VTIs normally provide during internships. We fund them, but partner VTIs implement them independently. Staff know the trainees, whose standing with their institute matters for future placements. Employers and workers also know that visits will occur, but their timing is unannounced, giving them an incentive to avoid misconduct between visits.

Before rating profiles, employers hear a script describing their assigned condition. After hearing the script and before rating profiles, employers in both treatment arms answer the same comprehension check, asking which concerns the visits address: theft and dishonesty,

[20] Assignment is stratified by enumerator, survey wave, and sector. Workers learn about the visits only after accepting a job, preventing selection into jobs based on treatment assignment. For ethical reasons and as preregistered, after a placement all employers receive the same visits, covering both theft and safety, regardless of experimental arm. Employers learn this only after the experiment.

safety and harassment, or effort (a placebo option).[21] The treatment scripts use the same wording where possible, varying the stated purpose of the visits, and are as follows:

- *Business as usual*: "We are committed to ensuring that managers and workers will have a positive experience if they end up being matched."
- *Theft monitoring*: "We are committed to ensuring that managers and workers will have a positive experience if they end up being matched. Our team members will conduct unannounced weekly visits to some of the firms where new workers are placed during the probation period. These workers will also be informed about the visits. Your firm may be randomly selected to receive such support visits via a lottery. Receiving visits does not reflect any characteristics of the firm or the workers. These visits discourage new workers from practices such as stealing, dishonesty, and disrespect by way of monitoring."
- *Safety monitoring*: "We are committed to ensuring that managers and workers will have a positive experience if they end up being matched. Our team members will conduct unannounced weekly visits to some of the firms where new workers are placed during the probation period. These workers will also be informed about the visits. Your firm may be randomly selected to receive such support visits via a lottery. Receiving visits does not reflect any characteristics of the firm or the workers. Visits are to ensure that the new workers are safe (not harassed) and treated with respect. These visits cannot discourage new workers from practices such as stealing, dishonesty, and disrespect by monitoring."

The two treatment arms hold the format, frequency, and cost of visits fixed. Thus, they introduce the same amount of information. Both scripts use gender-neutral language: the safety script refers to safety, respect, and harassment without mentioning women or specifying sexual harassment. We also highlight that both treatment scripts mention theft: the safety script explicitly states that visits cannot discourage stealing or dishonesty. This is intended to keep theft salience similar across the two treatment arms, helping separate the effect of mentioning theft from that of mitigating it. The two scripts therefore differ only in the friction they target and are otherwise as similar as possible so that the opposite predictions identify the mechanism rather than features of the scripts, following the design-by-elimination logic in Niederle (2025). The risk of this choice is that the statement itself

[21] After the treatment script, we ask, "Just to recap, what do you think our visit will help with?" Respondents can select safety and harassment; bad behavior, including stealing and dishonesty; or laziness and lack of concentration. Business-as-usual employers are not asked this question.

moves employers' theft concerns in either direction. We test for this in the mechanism analysis using employers' reported worries.

Because the visits mirror familiar VTI supervision, we expect employers to believe they can meaningfully mitigate workplace frictions. The identification concern is whether employers expect them to mitigate only the targeted friction, without also changing beliefs about candidates. In particular, three concerns arise. The first concern is spillovers: employers may expect visits to reduce both theft and harassment risks. Spillovers would make opposite effects harder to detect but would also make it harder to isolate the effect of each friction. The second concern is priming: explicitly stating that safety visits do not address theft may increase the value of women's perceived trustworthiness. This could explain increased demand for women under safety monitoring without any reduction in harassment concerns. The third concern is belief updating: either treatment may change employers' assessments of candidate quality or trustworthiness.

To detect spillovers and priming, we use worries elicited after the rating task. Employers rate their worries about theft and dishonesty and about worker safety. They also rate two placebo worries, about workers' effort and about workers leaving the firm (turnover). To avoid linking these questions too closely to the visits, we include the placebo worries and ask about workers in general, not specifically about new workers. We therefore treat the worries as a test for spillovers and not as a first stage. In particular, we test whether safety monitoring, whose script also mentions theft, increases theft worries. To test for belief updating, we examine whether the offers change employers' ratings of candidate quality and trustworthiness, our secondary outcomes.

Several features of the design limit experimenter demand: assignment is between employers, the scripts differ in one statement, choices are incentivized, gender-related questions follow the rating task, and partner institutes rather than the research team implement the visits. Although experimenter demand common to both treatment arms would not explain their predicted opposite effects, we also measure it directly using an approach inspired by de Quidt et al. (2018). At the end of the survey, we ask whether employers thought the research team preferred them to hire men or women or had no preference. We use responses indicating a preference for women to test whether perceived researcher preferences differ across arms, a version of the check suggested by de Quidt et al. (2026), and, as a further check, whether excluding these employers changes the results.

3.3 Conceptual Framework

We develop a simple model of employers' hiring decisions to show how monitoring capacity affects the hiring gender gap and what observed gaps reveal about discrimination. Motivated by the survey evidence, we model employers as evaluating candidates along three dimensions: technical performance, trustworthiness, and the ability to cooperate.[22] Vulnerability to harassment reduces a worker's ability to cooperate and depends on coworkers' behavior and the firm's gender composition. Appendix A.2 presents the full model.

An employer observes a candidate's gender g and performance signal q and forms a posterior belief $\hat{p}_g^S(q)$ that the candidate is highly technically skilled. The employer's expected value of hiring the candidate is

$$V_g(q, \mu, \eta) = y_L + \Delta y\, \hat{p}_g^S(q) - \theta \max\{\tilde{s}_g(\mu) - \bar{s}, 0\} - \tau \max\{\tilde{h}_g(\eta) - \bar{h}, 0\} - b\, \mathbf{1}\{g = f\}. \quad (1)$$

Here y_L is the value of low technical performance, and $\Delta y > 0$ is the additional value of high performance. Monitoring capacity determines the probabilities that theft and harassment are detected, μ and η. The terms $\tilde{s}_g(\mu)$ and $\tilde{h}_g(\eta)$ are the perceived probabilities that the new hire steals and that a coworker harasses the new hire. Harassment risk comes from male coworkers and scales with the male share of the workforce, m, which we take as given because our experiment focuses on the marginal hire. A favorable performance signal raises $\hat{p}_g^S(q)$ and therefore the candidate's value, yet employers may still infer other productive attributes from gender.

The costs of theft and harassment to the employer, θ and τ, may reflect production losses or nonproductive concerns, such as social preferences. Each risk is costly to the employer only above a threshold, $\bar{s}$ for theft and $\bar{h}$ for harassment, the same for men and women. Employers believe that women have a higher moral cost of stealing and face greater harassment risk from male coworkers. Theft risk therefore exceeds its threshold only for male candidates, and harassment risk only for female candidates. An increase in μ raises the value of male candidates and leaves the value of female candidates unchanged, and an increase in η does the reverse. The term b captures a residual preference against or for hiring women.

[22] Technical performance includes attributes relevant to completing production tasks, including technical skills, learning interest, effort, and physical strength.

The framework makes two additional implicit assumptions. First, the performance signal leaves gender differences in theft and harassment risk unchanged. Second, greater monitoring capacity reduces theft and harassment risk without changing beliefs about underlying worker characteristics or preferences over gender. We discuss the evidence on both in Section 4.

The difference in employers' valuation of otherwise identical male and female candidates, defined as male minus female, is

$$\Delta(q,\mu,\eta) = \underbrace{\Delta y\, \Delta^S(q)}_{\text{technical performance}} - \underbrace{T(\mu)}_{\text{trust premium}} + \underbrace{H(\eta)}_{\text{harassment penalty}} + \underbrace{b}_{\text{residual preferences}}, \quad (2)$$

where $\Delta^S(q) = \hat{p}^S_m(q) - \hat{p}^S_f(q)$ is the gender gap in perceived technical performance conditional on the signal. The trust premium $T(\mu) = \theta\big[\max\{\tilde{s}_m(\mu) - \bar{s}, 0\} - \max\{\tilde{s}_f(\mu) - \bar{s}, 0\}\big]$ favors women and falls as theft monitoring capacity μ increases. Because theft risk binds only for men, the trust premium is the theft cost that employers attach to male candidates and not to female ones. The harassment penalty $H(\eta) = \tau\big[\max\{\tilde{h}_f(\eta) - \bar{h}, 0\} - \max\{\tilde{h}_m(\eta) - \bar{h}, 0\}\big]$ works against women. It falls as safety monitoring capacity η increases.

The framework shows how mitigating monitoring frictions changes the demand for workers by gender through the demand for worker attributes. It also shows that women's trust premium offsets penalties arising from perceived technical performance differences, harassment risk, or residual preferences against women, so a small hiring gender gap can coexist with substantial penalties against women.

The main prediction is that mitigating theft and harassment frictions moves the hiring gender gap in opposite directions. Greater capacity to monitor theft reduces women's trust premium and increases the relative demand for men over women. Greater capacity to monitor harassment reduces the harassment penalty and increases the relative demand for women over men.[23]

[23] These predictions do not depend on the thresholds, which simplify the model. The same predictions hold as long as monitoring reduces the gender difference in perceived theft or harassment risk, and they are consistent with models of hiring risky workers (Lazear, 1998).

3.4 Empirical Strategy

Analysis Sample. Our main sample includes profile evaluations from 912 of the 915 employers in the employer sample. We apply one preregistered exclusion, an attention check that applies to all employers in the same way. The check operates at the level of eight-profile blocks: we exclude any block in which the employer selects either all eight profiles or none. These blocks, in which all eight answers are identical, account for 4.8% of blocks, a share that does not detectably differ across arms ($p = 0.319$), and three employers with no variation in any block leave the sample. To avoid differential selection across arms, we control for the comprehension check (Section 3.2), which only employers in the treatment arms answer, and do not exclude on it, as described below. The results are robust to including the blocks excluded by the attention check and to excluding employers who fail the comprehension check, as preregistered (Appendix Table A3). Table 2 shows covariate balance across the three experimental arms, including employer characteristics and firm-level attributes. We only find mild imbalances in three baseline variables: whether the employer has formal vocational training, firm capital, and the employer's daily hours worked.

Estimating Hiring Gender Gaps. We estimate the hiring gender gap and how monitoring support changes it using a pooled regression of profile evaluations across all three experimental arms. Each observation is an employer–profile evaluation. Our primary outcome, *Hire*, equals one if employer j wants to hire a worker matching base profile i for a probationary period. Random assignment of gender across otherwise identical profiles identifies the hiring gender gap. The index q denotes the displayed performance signal, while s, w, and e denote the sector, survey wave, and enumerator. We estimate a linear probability model:

$$\begin{aligned} \text{Hire}_{ij} = \beta_0 &+ \beta_1 \text{Female}_{ij} + \beta_2 \text{MT}_j + \beta_3 \text{MS}_j \\ &+ \beta_4 \text{Female}_{ij} \times \text{MT}_j + \beta_5 \text{Female}_{ij} \times \text{MS}_j \\ &+ \mathbf{C}_j' \kappa + \delta_{iq} + \gamma_s + \omega_w + \nu_e + u_{ij}. \end{aligned} \tag{3}$$

Female_{ij} indicates a female profile, while MT_j and MS_j indicate assignment to theft monitoring and safety monitoring. Business as usual is the omitted category. The coefficient β_1 measures the female-minus-male difference in hiring under business as usual, so the baseline hiring gender gap, defined as male minus female, is $-\beta_1$. Our main coefficients of

interest, β_4 and β_5, measure how theft and safety monitoring change the hiring of women relative to men. A positive interaction narrows the hiring gender gap, whereas a negative interaction widens it.

The vector $\mathbf{C}_j$ contains two indicators related to the comprehension test. Only employers in the treatment arms take this test, so excluding those who fail would select the treated samples on comprehension while leaving the business-as-usual sample unselected. We therefore retain all treated employers, which preserves the comparison created by random assignment. The indicators flag the 20 employers who select an incorrect or additional purpose for their assigned visits and the six interviewed before the question was introduced. We preregistered that we would exclude employers who fail the comprehension test; the results are robust to applying this exclusion (Appendix A.5; Appendix Table A3).

We include base profile × performance signal fixed effects (δ_{iq}) so that comparisons hold the candidate's background and performance signal fixed. We also include sector (γ_s), survey wave (ω_w), and enumerator (ν_e) fixed effects, corresponding to the dimensions used to stratify assignment. The error term is u_{ij}. We check robustness using a probit model. Following Abadie et al. (2023), we cluster standard errors at the employer level to allow evaluations by the same employer to be correlated. The main qualitative conclusions are unchanged under two-way clustering by employer and profile.

Heterogeneity by Stated Preferences for Workforce Gender Mix. We examine whether treatment effects vary with employers' stated preferences for workforce gender composition. Before the experiment, we ask employers, "What is the best gender composition of the workers in your firm?" and measure their preferred female share on a 10-worker scale. Using a standardized workforce of 10 lets employers express gender-mix preferences that would be difficult to capture with only a few workers.[24]

We first examine whether the effect of theft monitoring is muted among employers who prefer an all-male workforce. These employers may already be unlikely to hire women, leaving little scope for theft monitoring support to reduce their demand further. We therefore re-estimate equation (3) excluding employers who state a preference for an all-male workforce before the experiment.

[24]Employers answer on a scale ranging from 10 men out of 10 workers to 0 men out of 10 workers. We recode responses as the preferred female share, and we frame the scale in terms of men to make its focus on women less salient.

We then examine whether stated preferences predict experimental choices and whether workplace frictions affect employers similarly across the preference distribution. We estimate the specification separately by preference quintile, corresponding to preferred female shares of 0%–10%, 10%–20%, 20%–30%, 30%–40%, and 40%–60%. In the appendix, we report complementary specifications interacting treatment and candidate gender with the preferred female share, measured continuously or split at the median. We also report preregistered heterogeneity by employer gender and recruitment from VTIs. In an exploratory analysis that was not preregistered, we examine whether the effects of the visits vary with the severity of firms' monitoring constraints, using the presence of a manager who monitors workers as a proxy.

Technical Performance Signals. We examine whether stronger or additional performance signals narrow the hiring gender gap. The first test varies the quality of performance reported by showing a top or median VTI class rank. The second varies the amount of information employers receive by showing or omitting DIT certification. We extend equation (3) to interact each signal with candidate gender and monitoring assignment, including their three-way interaction. For VTI class rank, we replace base profile $\times$ performance signal fixed effects with base profile fixed effects to estimate the randomized signal's effect. Certification varies across base profiles, so its specification omits base profile fixed effects and estimates associations rather than randomized effects.

4 Experimental Hiring Gender Gap

4.1 Demand for Female Workers Under Business as Usual

Employers express demand for women when asked about their ideal workforce composition: 85% include at least one woman, and 70.4% include at least 2 out of 10 workers (Appendix Figure A2). The average preferred female share is 21.7%, compared with an actual firm-level average of 3.2%. Most nevertheless prefer a male-majority workforce: only 13.4% report an ideal female share between 40 and 60%. Employers most often attribute the gap between their preferred and actual gender mix to a scarcity of female applicants.

Employers also show substantial demand for trained female workers in the experiment. In the business-as-usual arm, 89.7% of employers ($N = 311$) select at least one female profile. Women account for 41.9% of the average employer's selected profiles (median 43.8%). A back-of-the-envelope calculation compares this revealed demand with stated preferences. We estimate each firm's workforce composition if it filled its stated trainee vacancies, hiring women at the average share among selected profiles. Women would account for 20.6% of the resulting workforce on average, close to the 23.0% ideal female share reported by the same employers.[25] These results show that employers are willing to hire trained women when given access to qualified candidates.

We next examine whether they are equally willing to hire otherwise identical men and women. The randomized comparison between male and female versions of the same profiles identifies this hiring gender gap. Under business-as-usual conditions, female profiles are 10.6 percentage points less likely to be selected than otherwise identical male profiles ($p < 0.001$; Table 3, column 1). This represents a meaningful hiring penalty against women: a 20% gap relative to the male selection rate, equivalent to 1.32 standard deviations of the trainee exam score distribution.[26]

The average hiring gender gap masks substantial heterogeneity that tracks employers' stated preferences. The gender gap under the business-as-usual arm narrows with employers' stated preference for women (Appendix Figure A5, panel A). Excluding employers who prefer an all-male workforce reduces the gap to 8.3 percentage points ($p < 0.001$; Table 3, column 3). More generally, a 10-percentage-point increase in the preferred female share is associated with a 4.82-percentage-point reduction in the gender gap ($p < 0.001$; Appendix Table A4, panel A, slope row). In the top quintile, where employers state an ideal workforce between 40% and 60% women, the hiring gender gap is not statistically significant.[27]

We also confirm that employers' assessments of otherwise identical profiles reflect the gendered beliefs documented in the motivating survey. Under the business-as-usual arm, female profiles receive lower ratings for technical performance and effort (both Holm-

[25]The calculation includes only business-as-usual employers who report their vacancies; their average preferred female share differs from the employer sample average of 21.7%.

[26]The baseline hiring gender gap is robust to using a probit model rather than a linear probability model. The estimated gap remains similar in magnitude and is statistically significant (Table 3, column 2).

[27]The heterogeneity by stated gender-mix preference is preregistered; the quintile cutoffs are not pre-specified. Results are robust to alternative cutoffs and to the continuous specification (Appendix Table A4).

adjusted $p < 0.001$) but higher trustworthiness ratings (Holm-adjusted $p = 0.020$; Table 5). Employers also expect women to earn less per month one year later, approximately 11% below expected male earnings (Holm-adjusted $p < 0.001$). However, employers more willing to hire women do not expect a larger female earnings disadvantage, suggesting that their greater demand is not driven by beliefs that women have weaker outside options.

4.2 Effects of Monitoring Visits on the Hiring Gender Gap

We next test whether greater capacity to monitor theft or harassment changes the hiring gender gap. The framework predicts opposite effects: theft monitoring reduces women's trust premium and should widen the gap, while safety monitoring reduces harassment concerns and should narrow it.

Average Treatment Effects. We find that the two types of visits move the hiring gender gap in opposite directions (Figure 3). Both increase overall selection relative to the business-as-usual selection rate, by 5.0% under theft monitoring visits ($p = 0.048$) and 6.6% under safety monitoring visits ($p = 0.010$), but through different gender margins.

Theft monitoring widens the hiring gender gap from 10.6 to 16.7 percentage points, a 58% increase relative to business as usual ($p = 0.010$). This increase comes from increased hiring of men: the share of male profiles selected rises by 5.4 percentage points ($p < 0.001$), while the share of female profiles selected does not detectably change ($p = 0.683$). Safety monitoring reduces the hiring gender gap by 80%, from 10.6 to 2.1 percentage points ($p < 0.001$), leaving it statistically indistinguishable from zero. This reduction comes from increased hiring of women: the share of female profiles selected rises by 7.3 percentage points ($p < 0.001$), while the share of male profiles selected does not detectably change ($p = 0.467$). The opposite effects are consistent with each monitoring support treatment changing the value employers place on a gendered worker attribute: theft monitoring reduces women's trust premium, while safety monitoring eases concerns about harassment and disrespect associated with hiring women.

Mechanism Tests. We examine this interpretation through the three concerns discussed in the design section: visits may affect both frictions (spillovers), mentioning theft

in the safety script may increase demand for women (priming), and either offer may change beliefs about candidates (belief updating).

Employers' worries about workplace risks let us test for spillovers and priming. First, we find that each type of visit reduces worries only about the friction it targets, without detectable spillovers to other frictions. Theft monitoring reduces theft worries by 0.22 standard deviations of the business-as-usual distribution (Holm-adjusted $p = 0.004$), while safety monitoring reduces worries about worker safety by 0.16 standard deviations (Holm-adjusted $p = 0.068$; Table 4). Worries about the other targeted friction and the placebo concerns of effort and turnover do not detectably change. Second, theft worries under safety monitoring are statistically indistinguishable from business as usual (Holm-adjusted $p = 0.157$), despite the script mentioning that the visits cannot discourage theft. The statement therefore does not lead employers to worry more about theft, which weighs against theft risk priming as an explanation for the increase in female hiring under safety monitoring. Taken together, these results suggest that the safety arm primarily affects perceived harassment risk rather than theft concerns, which is what the design-by-elimination logic rests on.

Employers' ratings of candidate attributes allow us to test for belief updating. Neither treatment detectably changes the gender gaps in technical performance, effort, trustworthiness, or expected earnings (Table 5). Female profiles remain rated as more trustworthy but lower in technical performance and effort than otherwise identical male profiles. Hiring disparities therefore change without detectable changes in these beliefs, supporting a mechanism operating through the value employers place on worker attributes.

Competing explanations such as scrutiny and experimenter demand predict effects in the same direction across the two treatments rather than the opposite effects we observe. The safety monitoring results provide further evidence against scrutiny. Employers may expect visits to monitor their own conduct toward female workers and therefore become less willing to hire women. Instead, female hiring increases, with no clear evidence that safety monitoring widens the gap among a subset of employers (Appendix Figure A6). We also find little evidence of experimenter demand in general. Only 7.0% of employers believe the research team prefers them to hire women, and this share does not detectably differ across treatment arms. Excluding these employers leaves the estimates similar (Appendix Table A2), suggesting that perceived experimenter demand does not explain the results.

If the visits affect the hiring gender gap by easing monitoring constraints, their effect should be larger where these constraints are more severe. Stated constraints offer little variation to test this prediction: nearly all respondents would like to monitor workers more than they do. Firms differ, however, in who monitors workers (Table 1). The owner monitors alone in 19% of firms, with other workers in 46%, and with at least one manager in 35%. We expect theft monitoring to matter less where managers share the task. Indeed, theft monitoring widens the hiring gender gap by 12.4 percentage points where the owner monitors alone ($p = 0.018$) and by 8.8 percentage points where the owner monitors with other workers ($p = 0.014$). It does not detectably change the gap where managers are involved ($p = 0.738$), and this estimate differs from the other two ($p = 0.036$ and $p = 0.057$; Appendix Table A4, panel B). The prediction for safety monitoring is less clear, because managers and coworkers can both deter harassment and commit it. The effect of safety monitoring does not detectably differ across the three groups ($p = 0.307$). While this analysis supports our hypothesis, it is exploratory as it was not preregistered. Manager involvement is also correlated with firm size and with recruitment from VTIs, both more common in the mechanics sector.

Finally, the treatment effects are robust across designs, samples, and specifications. Within-employer comparisons show that theft monitoring widens the gap by 11.9 percentage points, while safety monitoring narrows it by 15.4 percentage points (both $p < 0.001$; Appendix Table A5). The results are similar using probit (Table 3, column 2) or excluding employers who prefer an all-male workforce (column 3). They also persist when including the blocks excluded by the attention check, excluding employers who fail the comprehension check, or including soft-launch firms (Appendix Table A3), and under every combination of these choices (Appendix Figure A9). Controlling for the three baseline variables that are imbalanced across arms leaves the estimates nearly unchanged (Appendix Table A3, column 5).

Hidden Hiring Penalties. Women's perceived trustworthiness offsets other hiring penalties when theft and other worker misconduct are difficult to monitor. Theft monitoring reduces this advantage, revealing penalties partly hidden under business as usual.

The effects are largest among employers who appear gender-neutral under business as usual. In the top quintile of stated preferences for hiring women, theft monitoring increases the gap by 17.0 percentage points ($p < 0.001$), from a level statistically indistinguishable

from zero to 14.8 percentage points.[28] Once constraints on monitoring theft and other worker misconduct are mitigated, these employers exhibit a gender gap similar to that of other employers (Table 3, columns 1 and 8; Figure 3, panel B).

The relationship between stated preferences and the hiring gender gap is less than half as strong under theft monitoring as under business as usual, though the difference is imprecisely estimated.[29] This suggests that employers' greater willingness to hire women partly reflects the value of women's perceived trustworthiness when monitoring is difficult.

Safety monitoring effects do not detectably vary with stated preferences, consistent with harassment concerns constraining demand for women across the preference distribution (Appendix Table A4, panel A, column 3). Under a simple additive decomposition, harassment concerns explain roughly half of the total penalty revealed by theft monitoring. Safety monitoring removes that component but leaves women's trust premium intact. The lack of a detectable gender gap under safety monitoring thus also masks a remaining penalty against women, offset by their perceived trustworthiness advantage given existing theft risk.

Because we randomly assign gender to otherwise identical profiles, the hiring gender gap measures direct discrimination (Bohren et al., 2025b), net of any offsetting advantage. Theft monitoring increases net direct discrimination against women by reducing the trust premium that offsets other penalties. The penalties it reveals may reflect employers' beliefs about other worker attributes, preferences over gender, or both.

Other Preregistered Dimensions of Heterogeneity. The remaining preregistered tests examine whether the effects of monitoring vary by recruitment practice or employer gender (Appendix Table A4, panel A). We detect no corresponding difference by employer gender for theft monitoring. Differences by recruitment practice are imprecisely estimated for both treatments, but the point estimates suggest that theft monitoring widens the hiring gap more among firms that do not usually recruit from VTIs, for which experimentation costs may be higher. The estimated reduction in the gap under safety monitoring is more precise in this group but similar in magnitude across groups. Safety monitoring narrows the gap more among male than female employers ($p = 0.037$ for the difference), though only 33 of the 912 employers are women. This pattern could reflect

[28] This increase is significantly larger than among all other employers ($p = 0.022$).

[29] The coefficient falls from 0.482 to 0.217, a change of −0.265 (standard error 0.170; Appendix Table A4).

stronger paternalistic concerns among male employers, a perception that women face greater harassment risk in male-run firms, or both.

4.3 Effects of Technical Performance Signals

We test whether beliefs about technical skills explain the remaining hiring penalty against women. If employers use gender to infer technical ability, more precise information about individual skills can narrow the hiring gap, provided it is equally informative for men and women (Phelps, 1972; Aigner and Cain, 1977). We vary the level of performance reported by showing a top or median VTI class rank. We also vary the amount of information employers receive by showing or omitting DIT certification across profiles.

Employers respond to both signals. Under business as usual, a top rather than median rank raises selection by 9.2 percentage points (21.8%, $p < 0.001$), and reporting DIT certification raises it by 15.1 percentage points. However, neither signal detectably narrows the hiring gender gap in any arm, and their effects on the gap do not detectably vary across monitoring arms (Table 6, columns 2 and 4).

Women appear to benefit less from performance signals. Their hiring response is 4.2 percentage points smaller ($p = 0.066$; Table 6, column 2; Figure 4). Consistent with the selection results, top-ranked candidates receive higher work-quality ratings, but the gender gap in these ratings also does not detectably change. The hiring gender gap is largest among the candidates with the strongest performance signals. Under business as usual, women with a top-ranked profile and a certificate are 16.2 percentage points less likely to be selected than similar male profiles, compared with 9.1 percentage points among median-ranked profiles without one ($p = 0.047$ for the difference).[30]

A preference for hiring men could explain why stronger performance signals do not narrow the hiring gap. This form of bias may stem from gender norms or taste. Employers may also regard the same credentials as less informative for women or use gender to infer attributes the credentials reveal only imperfectly, such as physical strength.[31] Either

[30] Estimates are from a regression of *Hire* on candidate gender, rank, certification, and monitoring arm, fully interacted, with strata fixed effects and standard errors clustered at the employer level. The certificate is fixed within profile, and 7 of the 24 profiles carry one (4 male and 3 female), so comparisons by certification rest on differences across profiles.

[31] Employers often cite physical strength as a concern about hiring women. When asked which attributes matter for hiring in general, none mentions it. This suggests that employers may apply different standards to men and women.

belief-based explanation amounts to accurate statistical discrimination only if employers base hiring on expected productivity and correctly assess how informative the credentials are for each gender, and how those differences affect productivity.

5 Hidden Gender Discrimination

We next test whether employers accurately perceive gender differences in technical skills and trustworthiness by comparing their predictions with the skills and behavior of the trainees they could hire. We find significant misperceptions on both dimensions. Employers underestimate women's technical skills relative to men's and overestimate their relative trustworthiness. These errors can generate biases that partly offset each other in the observed hiring gap, suggesting that at least part of the hidden gap we uncover with theft monitoring reflects bias.

We then use the monitoring experiment to estimate how much of the business-as-usual hiring gap remains after accounting for the premium on perceived trustworthiness and the penalty associated with harassment risk. We compare this residual with those from simpler models that account for technical performance alone or include only one of these additional attributes. We show that omitting attributes employers value can lead researchers to overstate or understate bias. Estimates also depend on the extent to which researchers attribute each return to productivity.

5.1 Accuracy of Employers' Beliefs

We ask whether employers misperceive gender differences in technical skills and trustworthiness in ways that could bias hiring. We compare their incentivized beliefs with outcomes from a complementary experiment with 173 vocational trainees seeking work in the study sectors. The trainees come from seven partner VTIs (Table 1, panel C). To measure technical skills, we administer a sector-specific theory exam. We also ask trainees whether they can perform a standard practical task in their sector. To measure honesty in a work-like setting, we ask trainees to check VTI records for errors and offer a bonus for reporting more errors than the records contain. Both exercises carry reputation stakes because trainees know their results may be used for job referrals.

We compare the gender gaps employers predict in their own sector with those we observe among trainees on the same measures. We describe the trainee pool and each measure without naming a specific VTI, then ask each employer to predict outcomes separately for male and female trainees seeking work in that sector. Employers receive a bonus for accuracy.[32] Because male and female trainees are unevenly distributed across institutes, we include VTI fixed effects when estimating the gender gaps in trainee outcomes.

We find systematic misperceptions about both women and men, consistent with broader evidence on gendered beliefs (Bursztyn and Yang, 2022). Employers, most of whom are men, underestimate women's technical ability relative to men's. They also substantially overestimate how often men cheat and expect women to cheat less, a prediction the trainee data do not support. Together with the hiring results, these findings suggest that the business-as-usual hiring gap masks inaccurate statistical discrimination in both directions. Perceived differences in technical ability penalize women, while perceived differences in trustworthiness penalize men.

Technical Skills. The trainee outcomes show that employers underestimate women's technical ability relative to men's (Table 8 and Figure 5). Across the full theory exam, women score 6.2 percentage points higher than men on average, though the difference is not statistically significant ($p = 0.198$). On the randomly selected theory question shown to employers, women are 21.4 percentage points more likely than men to answer correctly ($p = 0.022$). We find no statistically significant gender difference in the share of trainees who report that they can perform the practical task in their sector (the estimated share is 7.0 percentage points higher among women, $p = 0.124$). Appendix Figure A8, panel A, shows the distribution of the combined skill score by gender. These outcomes are consistent with previous evidence of similar technical skills among vocationally trained men and women in these sectors (Alfonsi and Ferreira, 2026). Employers' predictions are consistent with experimental evidence that people underestimate women's ability relative to men's in male-typed domains (Bordalo et al., 2019).

Trustworthiness. We test whether employers correctly predict how often male and female trainees over-report in a record-checking task, which we use to measure honesty in a work setting. We adapt the production task of Caria and Falco (2024), which measures

[32]Employers were told, "We are going to ask you a set of questions, and we are going to compare your answers with the behavior of vocationally trained workers looking for a job in your sector. When your answers match the true behavior, you will receive a bonus" (up to UGX 10,000).

shirking, to study over-reporting for pay. Trainees complete two five-minute rounds, one supervised and one unsupervised.[33] In each round, they check 12 VTI registration forms for mismatched identifiers and record which forms they check and how many errors they find. Only five forms contain an error. Following the logic of the "claim your output" design of Mazar et al. (2008), we measure cheating by whether trainees claim a bonus of UGX 1,000 for finding at least six errors.[34] For each round, we ask employers how many out of ten male trainees and how many out of ten female vocational trainees in their sector would claim the bonus.

Two features of the task help us interpret over-reporting as a measure of workplace honesty. First, trainees stand to gain financially from over-reporting, but doing so may put their reputation at risk. They check VTI records and know that information from the study may be used for job referrals.[35] Second, we measure effort because it could confound our measure of cheating. Trainees who check more forms have more opportunities to mistake a correct entry for an error, which could lead them to claim the bonus in good faith. We therefore compare over-reporting by male and female trainees controlling for effort, measured by the number of forms checked.

Employers expect 46.3% of male trainees and 35.0% of female trainees to claim the bonus without supervision. Under supervision, they expect those shares to fall to 25.5% for men and 19.8% for women. Employers thus expect supervision to reduce over-reporting by 5.6 percentage points more among men than among women ($p < 0.001$), roughly halving the predicted gender gap (Table 8 and Figure 5). This belief is consistent with their hiring choices in the experiment, where theft monitoring increases selection of male trainees without a detectable change in selection of female trainees.

Employers' beliefs substantially overestimate cheating, especially among men. Only 23 of 173 trainees (13.3%) claim the bonus in at least one round. Without supervision, 7.3% of men claim it, compared with the 46.3% employers predict. If anything, claims are more common among women (15.9%), although the gender difference in cheating is imprecisely estimated (Fisher's exact test, $p = 0.12$). As employers anticipate, bonus

[33] In the unsupervised round, anonymous tracking sheets record effort and gender but cannot be linked to individual trainees. We therefore control for gender-group average effort in the unsupervised round.

[34] We randomize the order of the rounds and select one at random for payment. In the unsupervised round, trainees receive UGX 2,000 regardless of performance. In the supervised round, they earn up to UGX 2,000 depending on how many forms they check correctly. The comparison across rounds therefore captures the joint effect of supervision and pay for accurate work.

[35] Tracking sheets identify trainees by name and ID in the supervised round and are anonymous in the unsupervised round. The link to referrals mirrors the role of reputation in a labor market where firms often hire through networks.

rates are significantly lower in the supervised round (10.4% to 4.0%; $p = 0.016$). Yet, we also detect no gender difference in the response to supervision ($p = 0.68$), against employers' expectations.

Thus, employers substantially overestimate how often trainees cheat, particularly among men, and how much cheating declines under supervision. They also overstate gender differences in cheating in a way that appears to penalize men in hiring. The trainee results are consistent with the mixed evidence on gender differences in prosocial behavior (Niederle, 2016). The gap between employers' predictions and trainees' behavior echoes Exley et al. (2025), who find that people systematically expect women to be more prosocial than men, even though these beliefs are largely inaccurate in the settings they study.

5.2 Decomposing the Hiring Gender Gap

Following the framework in Section 3.3, we use the two monitoring treatments to decompose the business-as-usual hiring gap into a trust premium for women, a harassment penalty, and a residual. Given the evidence that technical performance signals do not detectably narrow the hiring gap and that employers underestimate women's technical skills relative to men's, we interpret the residual as bias against women, assuming that no other differences in productivity account for the remaining gap.

We estimate these components with a discrete-choice model; the estimation is described in Appendix A.3. Identification rests on the assumption that each treatment affects only its targeted risk. Consistent with this assumption, each treatment reduces employers' worries about the risk it targets, with no detectable change in worries about the other risk (Table 4). The decomposition also maintains the framework's assumption that monitoring does not change employers' beliefs about underlying worker attributes. We find no detectable treatment effects on the gender gaps in employers' ratings of candidates' technical skills or trustworthiness (Table 5).

Our benchmark assumes that each intervention fully resolves its targeted friction, so the estimated reductions equal the full premium and penalty under business as usual. Under this benchmark, we compare the residual from the model that includes technical performance, trustworthiness, and harassment risk with the residuals from models that include fewer attributes. Because the extent of the friction mitigated by the treatment

is not known, we also trace the implied residual for each model across the full range of assumed intervention effectiveness (Appendix Figure A10).

The model that accounts for all three attributes yields a trust premium for women of 6.1 percentage points and a harassment penalty of 8.5 percentage points, leaving a residual of 8.3 percentage points (Table 7). We interpret each model's residual as bias against women under the assumption that no omitted productivity differences explain it. The model with technical performance alone thus attributes the entire 10.6-percentage-point hiring gap to bias. Relative to the full model, omitting trustworthiness understates bias (the estimate falls to 2.1 percentage points), while omitting harassment risk overstates bias, raising the estimate to 16.8 percentage points.[36]

Omitting attributes employers value can lead researchers to overstate or understate bias. In a one-dimensional framework, observed and omitted measures are often positively correlated proxies for the same underlying attribute, such as technical performance. A familiar concern is that the observed measure captures only part of a productive disadvantage, leading researchers to overstate bias (Lang and Spitzer, 2020). When employers value multiple attributes, a group facing a hiring penalty can also have a productive advantage on an omitted dimension. That advantage can offset part of the penalty in the observed hiring gap, so omitting it can lead researchers to understate bias.

Even when researchers observe all the attributes employers value, measuring bias requires assumptions about which returns reflect productivity. Violence and harassment can impose economic costs on workers and firms (Adams and Nix, 2026), and employers' hiring decisions may reflect accurate expectations of output losses. Evidence from Bangladesh, however, shows that employers may also choose not to hire women out of a desire to protect them from harm (Buchmann et al., 2023). Employers could also worry that a harassment incident would damage the firm's reputation. Which estimate best captures bias therefore depends on how much of the harassment penalty researchers treat as a production cost.

The decomposition quantifies how a technological constraint can mitigate the hiring consequences of other barriers women face. When theft is difficult to monitor, the trust premium can be large enough that the business-as-usual hiring gap substantially understates the underlying bias against women (Appendix Figure A10, panel B). Easing the constraint

[36]For a given hiring response, a less effective treatment implies a larger underlying premium or penalty. Omitting the corresponding attribute therefore moves the estimate further from the full model's estimate (Appendix Figure A10).

reduces that premium and widens the hiring gap while other barriers remain, providing direct evidence of a second-best problem under an equal-treatment benchmark (Lipsey and Lancaster, 1956; Bergquist et al., 2026).

6 Referrals and Female Hiring After the Experiment

We follow employers beyond the rating exercise to examine whether their expressed demand translates into hiring women. We first describe the referrals and resulting placements and then report changes in female employment in the follow-up survey.

Referrals and Implementation. At the end of each wave, we connected employers to partner VTIs with candidates matching their hiring preferences. We matched firms to institutes using geographic proximity and institutes' female trainee shares; employers seeking only men were connected to all-male institutes. We shared employers' preferences with VTI focal points and sent employers an SMS in English and Luganda with the institute's contact details, explicitly mentioning female trainees where relevant.

Institutes and employers arranged placements directly, subject to trainee availability. VTI focal points identified and contacted suitable graduates, and institutes received a monthly fee to manage referrals and visits.[37] The referral SMS reached 99.2% of employers contacted, and 81% confirmed receipt. VTI focal points reported 62 placements. This is roughly one-third of our 173 trainee survey sample, providing a sense of the scale relative to the trainee pool, though institutes could also refer graduates outside this sample.

Following a placement, partner VTIs implemented the visits independently of the research team, using the supervision format they normally provide during internships. Staff visited weekly during the one-month probation period, at unannounced times and at no cost to the firm. For ethical reasons and as preregistered, visits focused on both preventing theft and ensuring safety in all experimental arms, including the business-as-usual arm. We verified 57 placements and their visits through phone interviews with employers.

[37] Graduates can be difficult to locate: many come from outside Kampala and, according to VTI staff, return home after training.

Female Hiring Two Years Later. To obtain systematic information on subsequent hiring and employment by gender, we conducted a follow-up phone survey. About two years after the start of the experiment, our team re-interviewed 776 firms, covering 84.8% of the employer sample. These firms are similar to the employer sample as a whole (Table 1).

Female employment is substantially higher at follow-up. Among the 771 firms with employment data at both dates, the number of women employed more than doubled, from 140 to 297. The share of firms employing at least one woman rose from 13.6% to 23.6%, and the average firm-level female workforce share rose from 2.7% to 5.6% (Appendix Table A6, panel A). Firm size did not grow, so the rise reflects a shift in workforce composition rather than expansion.

These figures capture employment at the time of the follow-up and may therefore miss women who were hired but had already left the firm. Indeed, significantly more firms, 59.4%, report having hired at least one woman since the experiment. Reported tenure averages 7.1 months, indicating substantial employment spells despite worker turnover (Appendix Table A6, panel B).

Stated demand for women also increased, with the average preferred female share rising from 21.7% to 27.0%. It rose most among firms that employed women: by 6.8 percentage points among firms that hired at least one woman since the experiment, compared with 3.2 percentage points among the other firms ($p = 0.011$). Employers who hired women did not update negatively about female workers.

7 Conclusion

This paper shows that the production environment shapes discrimination through employers' demand for worker attributes, including technical skills, social skills, and personality traits. In Uganda's male-dominated manufacturing sectors, theft monitoring increases relative demand for men, consistent with employers expecting men to be more likely to steal. Safety monitoring increases relative demand for women, consistent with concerns about their exposure to harassment. These shifts occur without detectable changes in employers' assessments of candidates, which points to monitoring changing the value of attributes rather than providing information about workers. The production environment is therefore a determinant of discrimination, beyond beliefs and preferences over identity.

The results also show why observed gender gaps can understate discrimination when employers care about multiple worker attributes. Women's perceived trustworthiness offsets other hiring penalties when monitoring is difficult. Trainee data do not support employers' beliefs that women are more trustworthy or less technically skilled. The hiring gender gap therefore conceals inaccurate statistical discrimination against men alongside bias against women. The near-zero gap under safety monitoring does not imply an absence of penalties: women's trust premium remains in place and offsets the underlying disadvantage.

Failing to account for all the attributes employers value can lead researchers to overstate or understate bias. Among candidates with the same observed technical skills, a perceived advantage on an omitted attribute can mask part of a hiring penalty, while a perceived disadvantage on an omitted attribute can be mistaken for bias. Even when researchers observe all relevant attributes, estimates depend on how much of each return they attribute to productivity. Harassment concerns, for example, may reflect expected output losses, paternalism, or reputational motives.

Disparate treatment can itself be a response to frictions in the production environment and can favor minority applicants when employers believe they have an advantage in traits that help firms manage those frictions. When theft is difficult to monitor, women's perceived trustworthiness raises the expected cost to employers of acting on bias against them, consistent with Becker's insight that discrimination responds to its economic cost (Becker, 1957). Easing the monitoring constraint lowers that cost and can therefore widen hiring disparities when other sources of disadvantage remain.

Our findings have two main policy implications. First, apparently gender-neutral interventions that change workplace frictions can amplify or mitigate the hiring consequences of beliefs and bias. Institutional design thus shapes minority outcomes through employers' demand for attributes as well as through its recognized effects on workers' choices (Niederle, 2016). Second, the effectiveness of supply-side policies depends on this demand. Predicting whether training and improved access to jobs translate into employment, and whether these effects generalize across settings, requires understanding which attributes employers value and which workplace frictions bind. An important next step is to evaluate the distributional effects of interventions that ease workplace frictions by examining when they generate sustained employment gains for underrepresented groups and how they interact with supply-side policies.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge, "When Should You Adjust Standard Errors for Clustering?," *Quarterly Journal of Economics*, 2023, *138* (1), 1–35.

Adams, Abi and Emily Nix, "The Economic Impacts of Gender-Based Violence and Harassment," *Annual Review of Economics*, 2026, *18*, 343–367.

Aigner, Dennis J. and Glen G. Cain, "Statistical Theories of Discrimination in Labor Markets," *Industrial and Labor Relations Review*, 1977, *30* (2), 175–187.

Alfonsi, Livia and Pedro Ferreira, "Who You Know, Who You Recommend: Referrals and Occupational Gender Segregation," CEPR Discussion Paper 21964, Centre for Economic Policy Research September 2026.

_ , Oriana Bandiera, Vittorio Bassi, Robin Burgess, Imran Rasul, Munshi Sulaiman, and Anna Vitali, "Tackling youth unemployment: Evidence from a labor market experiment in Uganda," *Econometrica*, 2020, *88* (6), 2369–2414.

Arrow, Kenneth J., "The Theory of Discrimination," in Orley Ashenfelter and Albert Rees, eds., *Discrimination in Labor Markets*, Princeton: Princeton University Press, 1973, pp. 3–33.

Ashraf, Nava, Alexia Delfino, and Edward Glaeser, "Female Entrepreneurship and Trust in the Market," 2026. Conditionally accepted, *Journal of Political Economy*.

_ , Oriana Bandiera, Virginia Minni, and Victor Quintas-Martínez, "Gender Gaps across the Spectrum of Development: Local Talent and Firm Productivity," April 2025. Revise and Resubmit at *Review of Economics and Statistics*.

Atkin, David, Azam Chaudhry, Shamyla Chaudry, Amit K. Khandelwal, and Eric Verhoogen, "Organizational Barriers to Technology Adoption: Evidence from Soccer-Ball Producers in Pakistan," *Quarterly Journal of Economics*, 2017, *132* (3), 1101–1164.

Bandiera, Oriana, Niklas Buehren, Robin Burgess, Markus Goldstein, Selim Gulesci, Imran Rasul, and Munshi Sulaiman, "Women's Empowerment in Action: Evidence from a Randomized Control Trial in Africa," *American Economic Journal: Applied Economics*, 2020, *12* (1), 210–259.

Bassi, Vittorio and Aisha Nansamba, "Screening and Signalling Non-Cognitive Skills: Experimental Evidence from Uganda," *The Economic Journal*, 2022, *132* (642), 471–511.

_ , **Jung Hyuk Lee, Alessandra Peter, Tommaso Porzio, Ritwika Sen, and Esau Tugume**, "Self-Employment Within the Firm," Working Paper 31740, National Bureau of Economic Research September 2023.

Beaman, Lori and Jeremy Magruder, "Who gets the job referral? Evidence from a social networks experiment," *American Economic Review*, 2012, *102* (7), 3574–3593.

_ , **Niall Keleher, and Jeremy Magruder**, "Do job networks disadvantage women? Evidence from a recruitment experiment in Malawi," *Journal of labor economics*, 2018, *36* (1), 121–157.

Becker, Gary S, *The economics of discrimination*, University of Chicago Press, 1957.

Bergquist, Lauren Falcao, Danial Lashkari, and Eric Verhoogen, "Wedges: A Microeconomic Perspective on Misallocation," in "Handbook of Development Economics," Vol. 6, Elsevier, 2026. Forthcoming.

Bernhardt, Arielle, Erica Field, Rohini Pande, Natalia Rigol, Simone Schaner, and Charity Troyer-Moore, "Male social status and women's work," in "AEA Papers and Proceedings," Vol. 108 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2018, pp. 363–367.

Bertrand, Marianne and Sendhil Mullainathan, "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," *American economic review*, 2004, *94* (4), 991–1013.

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

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts, "Does Management Matter? Evidence from India," *Quarterly Journal of Economics*, 2013, *128* (1), 1–51.

Bohren, J Aislinn, Alex Imas, and Michael Rosenberg, "The dynamics of discrimination: Theory and evidence," *American economic review*, 2019, *109* (10), 3395–3436.

_ , **Kareem Haggag, Alex Imas, and Devin G Pope**, "Inaccurate Statistical Discrimination: An Identification Problem," *Review of Economics and Statistics*, 2025, *107* (3), 605–618.

_ , **Peter Hull, and Alex Imas**, "Systemic discrimination: Theory and measurement," *The Quarterly Journal of Economics*, 2025, *140* (3), 1743–1799.

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

Breza, Emily and Supreet Kaur, "Labor Markets in Developing Countries," *Annual Review of Economics*, 2025, *17*, 747–776.

_ , _ , and Yogita Shamdasani, "The Morale Effects of Pay Inequality," *The Quarterly Journal of Economics*, 2018, *133* (2), 611–663.

Brown, Christina, "Understanding Gender Discrimination by Managers," 2023. Unpublished manuscript.

Buchmann, Nina, Carl Meyer, and Colin D Sullivan, "Paternalistic discrimination," Technical Report, Working Paper 2023.

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.

_ and David Y. Yang, "Misperceptions About Others," *Annual Review of Economics*, 2022, *14*, 425–452.

Card, David, Ana Rute Cardoso, and Patrick Kline, "Bargaining, Sorting, and the Gender Wage Gap: Quantifying the Impact of Firms on the Relative Pay of Women," *The Quarterly Journal of Economics*, 2016, *131* (2), 633–686.

Caria, Stefano A and Paolo Falco, "Skeptical employers: Experimental evidence on biased beliefs constraining firm growth," *Review of Economics and Statistics*, 2024, *106* (5), 1352–1368.

Carranza, Eliana, Robert Garlick, Kate Orkin, and Neil Rankin, "Job Search and Hiring with Limited Information about Workseekers' Skills," *American Economic Review*, 2022, *112* (11), 3547–3583.

Chandrasekhar, Arun G, Melanie Morten, and Alessandra Peter, "Network-based hiring: Local benefits; global costs," Working Paper 26806, National Bureau of Economic Research 2020.

Coffman, Katherine, Christine Exley, and Muriel Niederle, "The role of beliefs in driving gender discrimination," *Management Science*, 2021, *67* (6), 3551–3569.

Daher, Chaza Abou, Erica Field, Kendal Swanson, and Kate Vyborny, "Drivers of Change: Employment Responses to the Lifting of the Saudi Female Driving Ban," *American Economic Review*, 2025, *115* (9), 3248–3271.

de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth, "Measuring and bounding experimenter demand," *American Economic Review*, 2018, *108* (11), 3266–3302.

_ , **Lise Vesterlund, and Alistair Wilson**, "How Serious Are Experimenter Demand Effects? Evidence and Mitigation Strategies," Working Paper 35350, National Bureau of Economic Research June 2026.

Deming, David and Mikko Silliman, "Skills and human capital in the labor market," in "Handbook of Labor Economics," Vol. 6, Elsevier, 2025, pp. 115–157.

Deming, David J, "The growing importance of social skills in the labor market," *The quarterly journal of economics*, 2017, *132* (4), 1593–1640.

Eberhardt, Markus, Giovanni Facchini, and Valeria Rueda, "Gender differences in reference letters: Evidence from the economics job market," *The Economic Journal*, 2023, *133* (655), 2676–2708.

Exley, Christine L, Oliver P Hauser, Molly Moore, and John-Henry Pezzuto, "Believed gender differences in social preferences," *The Quarterly Journal of Economics*, 2025, *140* (1), 403–458.

Foster, Andrew D and Mark R Rosenzweig, "A test for moral hazard in the labor market: Contractual arrangements, effort, and health," *The Review of Economics and Statistics*, 1994, *76* (2), 213–227.

Garlick, Robert, Erica Field, and Kate Vyborny, "Women's Mobility and Labor Supply: Experimental Evidence from Pakistan," IZA Discussion Paper 17883, IZA Institute of Labor Economics 2025.

Goldberg, Pinelopi, Charles Gottlieb, Somik Lall, Meet Mehta, Michael Peters, and Aishwarya Lakshmi Ratan, "The Global Gender Distortions Index (GGDI)," Working Paper 34142, National Bureau of Economic Research 2025.

Goldstein, Markus, Paula Gonzalez Martinez, and Sreelakshmi Papineni, "Tackling the global profitarchy: Gender and the choice of business sector," Policy Research Working Paper 8865, World Bank 2019.

Groh, Matthew, Nandini Krishnan, David McKenzie, and Tara Vishwanath, "The Impact of Soft Skills Training on Female Youth Employment: Evidence from a Randomized Experiment in Jordan," *IZA Journal of Labor & Development*, 2016, *5*, 9.

Heath, Rachel, "Why do firms hire using referrals? Evidence from Bangladeshi garment factories," *Journal of Political Economy*, 2018, *126* (4), 1691–1746.

_ , **Arielle Bernhardt, Girija Borker, Anne Fitzpatrick, Anthony Keats, Madeline McKelway, Andreas Menzel, Teresa Molina, and Garima Sharma**, "Female labour force participation," *VoxDevLit*, 2024, *11* (1), 1–43.

Hjort, Jonas, "Ethnic Divisions and Production in Firms," *The Quarterly Journal of Economics*, 2014, *129* (4), 1899–1946.

Houeix, Deivy, "Asymmetric Information and Digital Technology Adoption: Evidence from Senegal," 2026. Working Paper, Harvard University.

Hurst, Erik, Yona Rubinstein, and Kazuatsu Shimizu, "Task-Based Discrimination," *American Economic Review*, 2024, *114* (6), 1723–1768.

Jalal, Amen, "Screening Women Out? Pay Transparency in Job Postings," Working Paper 104, G2LM|LIC, IZA 2026.

Kelley, Erin M., Gregory Lane, and David Schönholzer, "Monitoring in Small Firms: Experimental Evidence from Kenyan Public Transit," *American Economic Review*, 2024, *114* (10), 3119–3160.

Kessler, Judd B, Corinne Low, and Colin D Sullivan, "Incentivized resume rating: Eliciting employer preferences without deception," *American Economic Review*, 2019, *109* (11), 3713–3744.

Kline, Patrick, Evan K Rose, and Christopher R Walters, "Systemic discrimination among large US employers," *The Quarterly Journal of Economics*, 2022, *137* (4), 1963–2036.

Kotia, Ananya, "When Competition Compels Change: Trade, Management, and Productivity," 2025. Working Paper, London School of Economics.

Lang, Kevin and Ariella Kahn-Lang Spitzer, "Race discrimination: An economic perspective," *Journal of Economic Perspectives*, 2020, *34* (2), 68–89.

Lazear, Edward P., "Hiring Risky Workers," in Isao Ohashi and Toshiaki Tachibanaki, eds., *Internal Labour Markets, Incentives and Employment*, London: Palgrave Macmillan, 1998, pp. 143–158.

Lipsey, Richard G and Kelvin Lancaster, "The general theory of second best," *The Review of Economic Studies*, 1956, *24* (1), 11–32.

Macchiavello, Rocco, Andreas Menzel, Atonu Rabbani, and Christopher Woodruff, "Promoting Women to Managerial Roles in the Bangladeshi Garment Sector," *Econometrica*, 2026, *94* (5), 1685–1716.

Mazar, Nina, On Amir, and Dan Ariely, "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance," *Journal of Marketing Research*, 2008, *45* (6), 633–644.

Morchio, Iacopo and Christian Moser, "The Gender Pay Gap: Micro Sources and Macro Consequences," *American Economic Review*, 2026, *116* (5), 1765–1810.

Moss, Philip and Chris Tilly, *Stories Employers Tell: Race, Skill, and Hiring in America*, New York: Russell Sage Foundation, 2001.

Niederle, Muriel, "Gender," in John H. Kagel and Alvin E. Roth, eds., *The Handbook of Experimental Economics*, Vol. 2, Princeton, NJ: Princeton University Press, 2016, pp. 481–553.

_ , "Experiments: Why, How, and A Users Guide for Producers as well as Consumers," Working Paper 33630, National Bureau of Economic Research 2025. Handbook chapter.

Phelps, Edmund S., "The Statistical Theory of Racism and Sexism," *American Economic Review*, 1972, *62* (4), 659–661.

Sarsons, Heather, "Interpreting Signals in the Labor Market: Evidence from Medical Referrals," 2017. Working paper.

Uganda Bureau of Statistics, "Census of Business Establishments, 2010/11," Census Report, Uganda Bureau of Statistics, Kampala, Uganda 2011.

_ , "Census of Business Establishments, 2020," Census Report, Uganda Bureau of Statistics, Kampala, Uganda 2020.

Woessmann, Ludger, "Skills and earnings: A multidimensional perspective on human capital," *Annual Review of Economics*, 2025, *17*, 397–425.

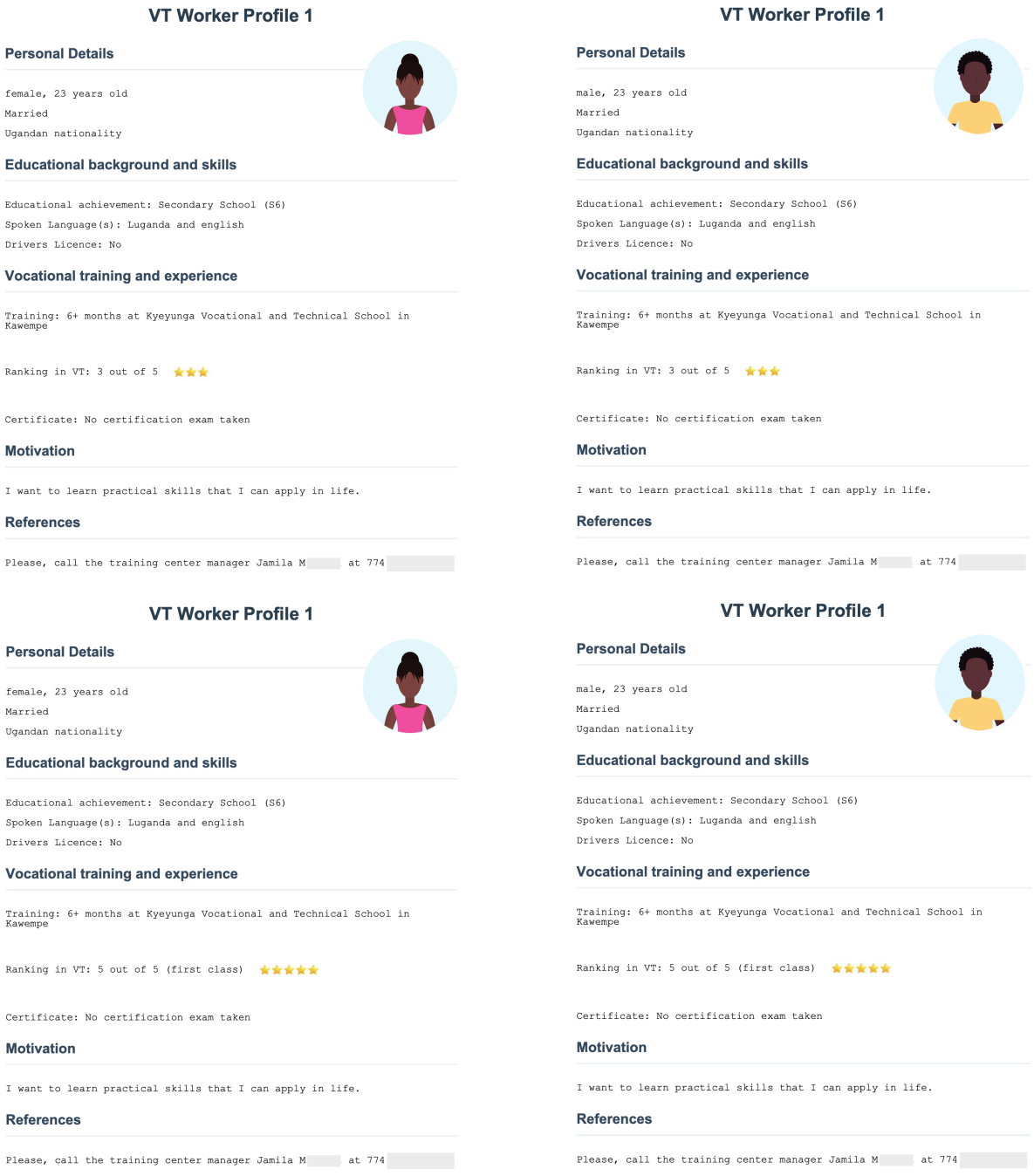
Figures

Figure 1: Demand for Worker Attributes and Employers' Gendered Perceptions



Notes: Employer sample ($N = 915$). Panel A: share of employers rating each trait 10 on a 0–10 scale, excluding 73 employers who gave all six traits the same rating ($N = 842$); "Behavior / No Stealing" combines the honesty and items about handling money. Panels B–D: closed-choice questions with the same categories. Panel E: worries about hiring women from the open-ended question, for employers with coded responses ($N = 589$), classified into the same categories; darker portions are women-specific sub-worries (strength, leaving, harassment); "Nothing" (10.70%) and "Other" (1.53%) are omitted. Open-ended answers are classified with GPT-5 mini (Appendix A.1).

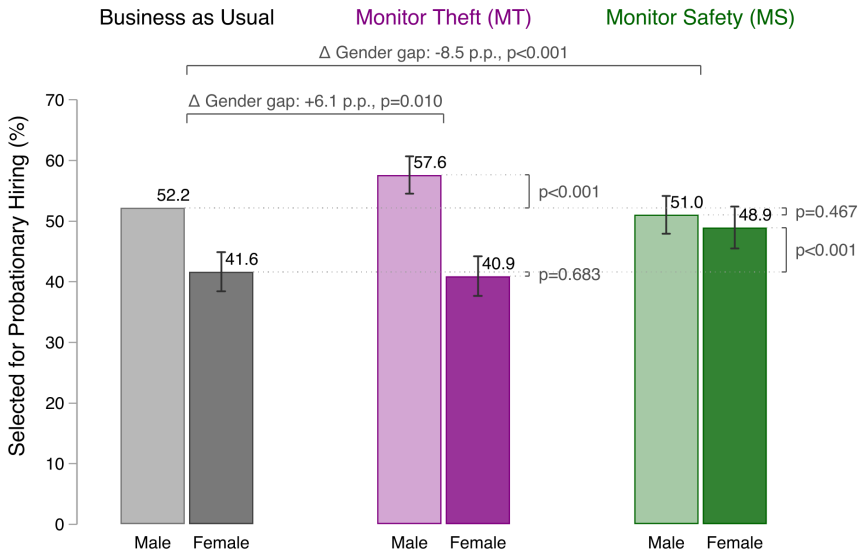
Figure 2: CV Examples



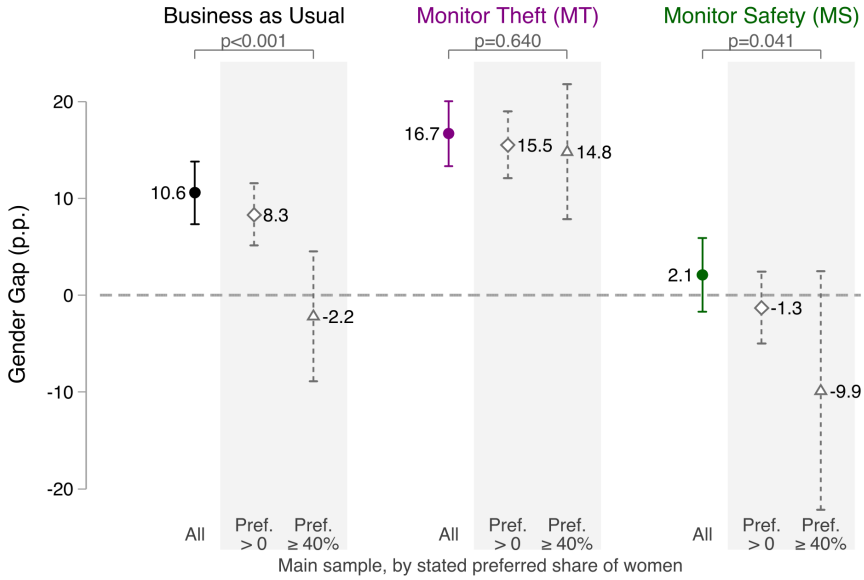
Notes: One base profile in its four versions, varying gender and performance. There are 24 base profiles and 96 profiles in total. Each employer evaluates all 24 base profiles, one randomly assigned version each.

Figure 3: Effect of Monitoring Visits on Experimental Hiring Choices

(A) Share of Profiles Selected for Probationary Hiring



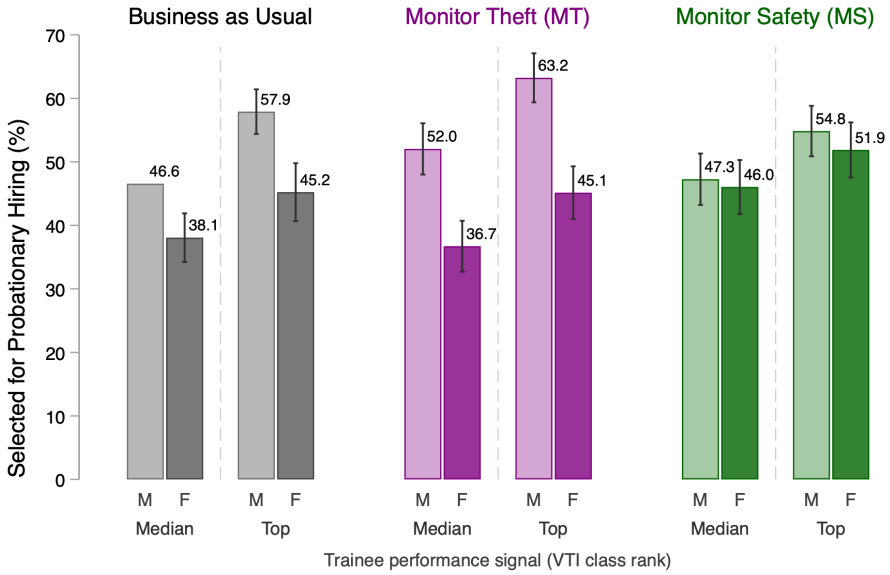
(B) Hiring Gender Gap, by Stated Preference for Hiring Women



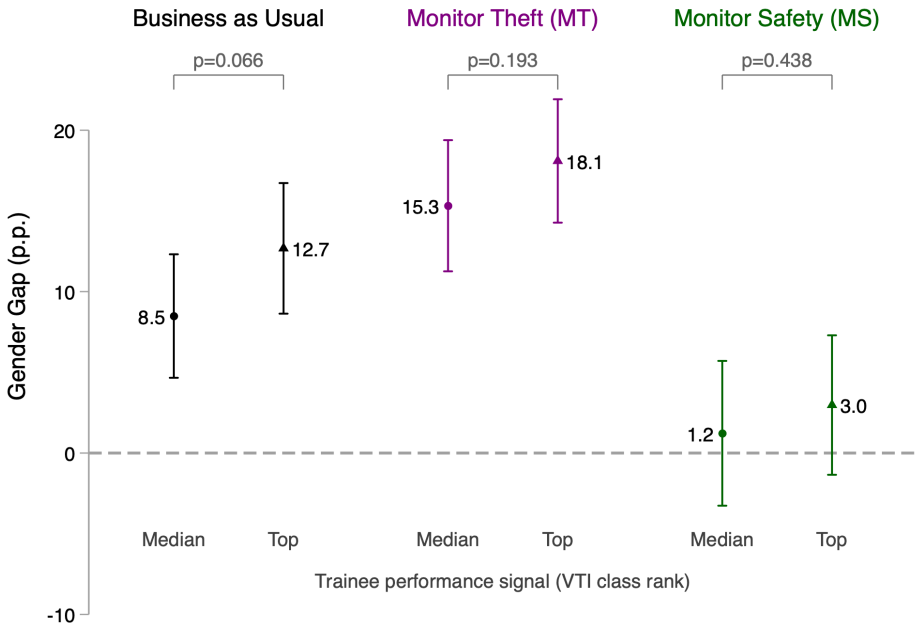
Notes: Estimates from equation (3). Panel A: share of profiles selected for probationary hiring, in percent, by candidate gender and monitoring arm: business as usual, theft monitoring (MT), and safety monitoring (MS). Panel B: hiring gender gap, male minus female, in percentage points. Circles: main sample (N = 20,759 evaluations, 912 employers). Diamonds: employers with a stated preferred share of women above zero (N = 17,790). Triangles: top quintile of stated preferences, 40% or more women (N = 2,981). Each set of estimates is a separate regression on the corresponding sample; a fully saturated model yields equivalent estimates. Brackets report p-values for the change relative to business as usual, by candidate gender (panel A), and for the difference between the top quintile and all other employers (panel B). In panel A, the top brackets report the change in the hiring gender gap relative to business as usual, in percentage points, with its p-value. Bars are 95% confidence intervals. Standard errors are clustered at the employer level.

Figure 4: Effect of Monitoring Visits, by Performance Signal

(A) Share of Profiles Selected for Probationary Hiring

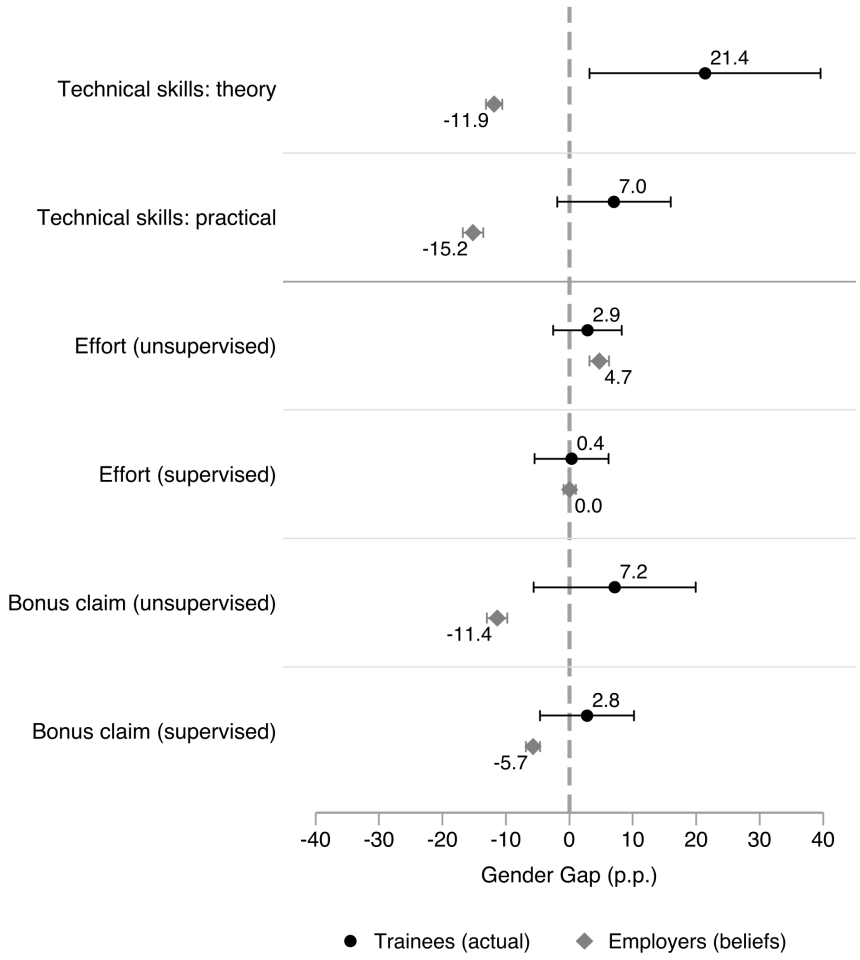


(B) Hiring Gender Gap



Notes: Estimates from a single fully interacted model of hiring on candidate gender, performance signal, and monitoring arm, with profile and strata fixed effects. Panel A: regression-adjusted share of profiles selected for probationary hiring, in percent, by candidate gender (M, F) and performance signal (Median, Top); the reference cell (male, median, business as usual) has no confidence interval. Panel B: implied hiring gender gap in percentage points, by performance signal; positive values are gaps against women. Brackets report p-values for the difference between top- and median-performance profiles within each arm. Bars are 95% confidence intervals. Standard errors are clustered at the employer level.

Figure 5: Trainee Outcomes and Employers' Beliefs by Gender



Notes: Female-minus-male differences, in percentage points, in trainee outcomes (circles) and employers' beliefs about them (diamonds). Theory: share answering correctly the exam question on which beliefs were elicited. Practical: share reporting they can perform a standard sector task. Effort: share of 12 VTI registration forms checked in a five-minute round. Bonus claim: share reporting six or more errors when five are planted, which earns a bonus. Employers predict each outcome separately for men and women. Specifications follow Table 8: unsupervised is the *Female* coefficient, supervised adds *Female* × *Supervised*. Trainee regressions include VTI fixed effects; the bonus claim also controls for effort. Employer estimates are within-employer differences, main sample (912 employers). Bars are 95% confidence intervals. Standard errors are robust, clustered at the trainee level for the trainee bonus claim.

Tables

Table 1: Summary Statistics

	Experimental sample					Census
	All	Mechanics	Carpentry	Welding	Follow-up	UBOS 2020
Panel A: Firms						
Employees (excl. owners)	8.32	15.21	4.54	5.00	8.29	7.40
Female workers (N)	0.19	0.19	0.26	0.12	0.18	—
Trainees (N)	2.48	5.00	0.97	1.41	2.55	—
Family workers (share)	0.14	0.07	0.16	0.20	0.14	—
Age (years)	10.12	11.28	9.40	9.63	10.22	11.50
Firm capital (UGX m)	5.29	7.35	4.49	4.12	5.34	—
Firm revenue (UGX m)	3.23	3.19	2.73	3.82	3.25	—
Annual Rev. > UGX 10m	0.88	0.90	0.86	0.88	0.89	0.80
Firm profits (UGX m)	1.26	1.28	1.07	1.45	1.26	—
Customers (N/day)	3.55	6.38	2.03	2.10	3.51	—
Any stealing	0.83	0.91	0.82	0.77	0.84	—
Monitor constrained	0.91	0.89	0.93	0.91	0.91	—
Monitors: owner alone	0.19	0.10	0.28	0.19	0.18	—
Monitors: owner with workers	0.46	0.32	0.53	0.54	0.47	—
Monitors: manager(s) involved	0.35	0.58	0.20	0.27	0.35	—
Recruits through family referrals	0.46	0.40	0.47	0.50	0.46	—
Recruits from VTIs	0.18	0.38	0.09	0.07	0.18	—
Panel B: Employers						
Male	0.96	0.96	0.96	0.97	0.96	0.96
Age (years)	37.15	40.29	35.80	35.24	37.39	—
Experience (years)	15.25	18.81	13.21	13.60	15.44	—
Secondary education	0.63	0.64	0.62	0.65	0.63	—
VTI trained	0.31	0.36	0.27	0.29	0.31	—
Hours worked (per day)	10.42	10.46	10.43	10.38	10.42	—
Hours monitor (per day)	2.06	2.02	2.10	2.07	2.09	—
Observations	915	313	299	303	776	3,045
Panel C: Trainees						
Female (share)	0.36	0.45	0.11	0.40	—	—
Age (years)	20.80	20.30	21.28	21.26	—	—
Completed S4+ (share)	0.71	0.58	0.64	0.96	—	—
Training duration (months)	25.05	27.04	25.50	21.62	—	—
Theory exam (share correct)	0.53	0.46	0.59	0.61	—	—
Practical skill (share)	0.94	0.98	0.89	0.91	—	—
Exp. income, 1yr (UGX m)	0.98	1.05	0.80	1.00	—	—
Observations	173	84	36	53	—	—

Notes: Sample means, overall and by sector; monetary values in millions of UGX. Employees include managers and trainees and exclude owners. *Any stealing*, *Monitor constrained*, *Secondary education*, and *VTI trained*: indicators for experienced worker theft, wanting more monitoring, secondary education, and VTI training. *Monitors*: who is involved in monitoring workers: the owner alone; the owner or other workers, with no manager; or one or more managers (shares sum to one). *Recruits through family referrals* and *Recruits from VTIs*: the firm searches for workers through referrals from family members or by contacting VTIs. *UBOS 2020*: 2020 Census of Business Establishments (Uganda Bureau of Statistics, 2020), Kampala, three study sectors; census employees are total employment minus one working owner, and *Annual revenue > UGX 10m* is the census's top turnover band. Panel C: the 173 surveyed trainees in the study sectors' labor pools, assigned by training sector (electricians to mechanics) or by stated desired occupation. *Theory exam*: share correct on a sector-specific exam. *Practical skill*: self-assessed. *Expected income*: expected monthly earnings one year after training. *Follow-up*: baseline means for firms reached in the 2026 follow-up phone survey.

Table 2: Balance Across Treatment Arms

	C (1)		MT (2)		MS (3)		*p*-value	*p*-value	N
	Mean	SD	Mean	SD	Mean	SD	(1)-(2)	(1)-(3)	
Panel A: Firms									
Workers (N)	8.04	(8.49)	8.64	(8.65)	8.27	(9.44)	0.125	0.539	909
Female workers (N)	0.19	(0.55)	0.21	(0.56)	0.17	(0.49)	0.450	0.607	906
Trainees (N)	2.49	(3.47)	2.56	(3.25)	2.40	(3.38)	0.678	0.894	907
Family workers (share)	0.15	(0.25)	0.15	(0.23)	0.13	(0.22)	0.663	0.201	907
Age (years)	10.06	(7.86)	10.61	(8.07)	9.66	(7.11)	0.395	0.513	912
Capital (UGX m)	6.19	(14.29)	5.80	(11.23)	3.93	(4.30)	0.959	0.054	598
Revenue (UGX m)	3.10	(2.99)	3.17	(2.86)	3.40	(2.91)	0.828	0.463	545
Profits (UGX m)	1.16	(1.18)	1.25	(1.17)	1.34	(1.35)	0.493	0.224	542
Customers (N/day)	3.51	(3.51)	3.74	(3.40)	3.39	(3.02)	0.208	0.479	907
Any stealing	0.84	(0.37)	0.84	(0.36)	0.81	(0.39)	1.000	0.225	912
Monitor constrained	0.90	(0.30)	0.92	(0.27)	0.90	(0.29)	0.456	0.767	905
Recruits from VTIs	0.19	(0.40)	0.17	(0.38)	0.18	(0.38)	0.665	0.680	912
Panel B: Employers									
Male	0.97	(0.18)	0.95	(0.23)	0.98	(0.15)	0.220	0.461	912
Age (years)	37.31	(9.85)	36.52	(9.25)	37.68	(10.57)	0.248	0.601	912
Experience (years)	15.24	(9.00)	15.40	(9.10)	15.15	(9.27)	0.811	0.874	910
Secondary education	0.62	(0.49)	0.62	(0.49)	0.66	(0.47)	0.933	0.395	912
VTI trained	0.26	(0.44)	0.33	(0.47)	0.33	(0.47)	0.028	0.046	912
Hours worked (/day)	10.53	(1.53)	10.44	(1.70)	10.30	(1.62)	0.682	0.086	911
Hours monitor (/day)	2.08	(1.59)	2.00	(1.29)	2.10	(1.51)	0.486	0.825	908

Notes: Baseline balance, main sample (912 employers). Columns 1–3: means and standard deviations by arm: business as usual (C), theft monitoring (MT), and safety monitoring (MS). Columns (1)–(2) and (1)–(3): randomization-inference *p*-values from 10,000 permutations of assignment within strata, from regressions of each variable on a treatment indicator with strata fixed effects and robust standard errors. The last column is the sample size of each test; observations are missing due to nonresponse.

Table 3: Effects of Monitoring Visits on the Hiring Gender Gap

				Stated Preferences for Workforce Gender Mix				
Hire (0-1)	(1) Main Sample LPM	(2) Main Sample Probit	(3) Gender-Mix Pref. > 0	(4) Bottom QT	(5) 2nd QT	(6) 3rd QT	(7) 4th QT	(8) Top QT
Female	-0.106***	-0.106***	-0.083***	-0.226***	-0.154***	-0.107***	-0.090***	0.022
	[0.016]	[0.016]	[0.016]	[0.054]	[0.044]	[0.028]	[0.027]	[0.034]
Monitor Theft	0.054***	0.054***	0.063***	-0.014	0.046	0.051*	0.057*	0.106***
	[0.016]	[0.016]	[0.016]	[0.045]	[0.044]	[0.027]	[0.030]	[0.040]
Monitor Safety	-0.012	-0.012	-0.018	0.014	-0.044	-0.017	-0.033	-0.002
	[0.016]	[0.016]	[0.017]	[0.046]	[0.041]	[0.029]	[0.028]	[0.047]
Female x Monitor Theft	-0.061***	-0.061***	-0.072***	-0.000	-0.053	-0.054	-0.035	-0.170***
	[0.024]	[0.024]	[0.024]	[0.075]	[0.066]	[0.038]	[0.046]	[0.049]
Female x Monitor Safety	0.085***	0.085***	0.096***	-0.002	0.134**	0.112***	0.099**	0.077
	[0.025]	[0.025]	[0.025]	[0.090]	[0.058]	[0.042]	[0.043]	[0.071]
Monitor Theft, pooled	0.023** [0.012]							
Monitor Safety, pooled	0.031** [0.012]							
Observations	20,759	20,759	17,790	2,969	3,106	6,517	5,186	2,981
R-squared	0.127		0.128	0.195	0.174	0.146	0.126	0.124
Control Mean	0.522	0.522	0.517	0.548	0.493	0.525	0.540	0.487

Notes: Estimates of equation (3); the dependent variable, *Hire*, equals 1 if the employer selects the profile for probationary hiring. Column 1: main sample. Column 2: probit, average marginal effects. Column 3: excludes employers who prefer an all-male workforce. Columns 4–8: quintiles of the stated preferred share of women. *Monitor Theft* and *Monitor Safety* indicate the assigned visit type; business as usual is omitted. The bottom panel reports the monitoring effects pooled across candidate genders, from the column 1 regression (arm coefficient plus one half the interaction, profiles being half female); the pooled business-as-usual selection rate is 0.468. Profile and strata fixed effects are included, as are indicators for treated employers who failed or were not asked the comprehension check. Standard errors are clustered at the employer level in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Effects of Monitoring Visits on Employers' Theft, Safety, and Placebo Worries

			Placebo	
	(1) Stealing worry	(2) Safety worry	(3) Effort worry	(4) Turnover worry
Monitor Theft	-0.225	-0.050	-0.063	-0.161
	[0.065]	[0.062]	[0.063]	[0.082]
Holm p-value	0.004	0.949	0.949	0.244
WY p-value	0.004	0.657	0.657	0.212
Monitor Safety	-0.144	-0.164	-0.079	0.007
	[0.065]	[0.063]	[0.067]	[0.083]
Holm p-value	0.157	0.068	0.949	0.949
WY p-value	0.145	0.071	0.628	0.931
Observations	896	908	911	609
R-squared	0.416	0.458	0.397	0.370
Control mean	5.511	4.656	4.965	4.052
Control SD	2.722	2.872	2.517	2.581

Notes: Post-experiment worries, elicited on a 0–10 scale after the profile evaluations (Section 3.2) and expressed in standard deviations of the business-as-usual distribution. Each column regresses the worry on indicators for theft monitoring and safety monitoring, business as usual omitted, with sector, survey wave, and enumerator fixed effects and indicators for treated employers who failed or were not asked the comprehension check. Robust standard errors are in brackets. Control mean and Control SD: mean and standard deviation of the worry in the business-as-usual arm, in points on the 0–10 scale. Below each coefficient: p-values adjusted across the family of eight tests (two arms $\times$ four outcomes), Holm step-down and Westfall–Young (WY, 2,000 bootstrap replications); no stars. Effort and turnover worries are placebo outcomes. The turnover question was asked in the 2024 wave only (column 4). Results are robust to adding the baseline variables imbalanced across arms (vocational training, firm capital, daily hours worked).

Table 5: Effects of Monitoring Visits on Beliefs (Secondary Outcomes)

	(1) Quality	(2) Trustworthiness	(3) Effort	(4) Expected monthly earnings (UGX m)
Female	-0.171	0.058	-0.203	-0.046
	[0.022]	[0.025]	[0.026]	[0.004]
Holm p-value	0.000	0.020	0.000	0.000
WY p-value	0.000	0.021	0.000	0.000
Monitor Theft	-0.014	-0.020	0.001	0.008
	[0.044]	[0.051]	[0.045]	[0.011]
Monitor Safety	-0.031	-0.029	-0.005	-0.010
	[0.046]	[0.051]	[0.044]	[0.011]
Female × Monitor Theft	-0.039	-0.009	-0.057	-0.003
	[0.034]	[0.038]	[0.038]	[0.006]
Holm p-value	1.000	1.000	1.000	1.000
WY p-value	0.779	0.962	0.632	0.953
Female × Monitor Safety	0.024	0.049	-0.001	-0.003
	[0.033]	[0.036]	[0.039]	[0.006]
Holm p-value	1.000	1.000	1.000	1.000
WY p-value	0.931	0.665	0.979	0.953
Observations	20,785	20,712	20,761	20,786
R-squared	0.227	0.085	0.174	0.209
BAU male mean	0.087	-0.027	0.103	0.427

Notes: The table reports how employers rate the profiles they evaluate, by candidate gender and treatment arm, for the main sample. Columns 1–3 are standardized on the business-as-usual distribution. Column 1 is perceived quality and skills: "Based on your first impression, how would you rate the worker's skills and work quality? Please rate on a scale from 0 to 10, where 0 is very low quality and 10 is very high quality." Column 2 is perceived trustworthiness: "Based on your first impression, how would you rate the workers' behavior (trustworthiness and honesty)? Please rate on a scale from 0 to 10, where 0 is not at all trustworthy/honest and 10 is very trustworthy/honest." Column 3 is perceived effort: "Based on your first impression, how likely is this worker to put in effort (hardworking, not lazy, concentrated, punctual)? Please rate on a scale from 0 to 10, where 0 is very unlikely and 10 is very likely." Column 4 is expected monthly earnings in millions of UGX: "What is your best guess of the monthly earnings of this worker a year from now?" Each column reports one regression of the indicated rating on an indicator for female candidates, indicators for the two monitoring arms, and their interactions, with the business-as-usual arm omitted. Profile × performance and strata fixed effects are absorbed, and indicators for treated employers who failed or were not asked the comprehension check are included. Standard errors, in brackets, are clustered at the employer level. Stars are omitted; inference should be read from the adjusted p-values, which we report for two families of hypotheses. The first family asks whether ratings are gendered and consists of the *Female* coefficient in each of the four columns. The second asks whether the gender gap in ratings changes with the monitoring regime and consists of the two *Female* × arm interactions in each of the four columns. We treat them as separate families because they answer separate questions; pooling them would penalize the first for tests that belong to the second. Within each family, we report Holm step-down and Westfall–Young (WY) adjusted p-values, the latter from 2,000 bootstrap replications resampled by respondent. The monitoring main effects are shown for completeness and belong to neither family. The final row reports the mean of the outcome for male candidates under business as usual.

Table 6: Hiring and the Technical Performance Signal

Signal:	Top Performance		Certificate	
Hire (0-1)	(1)	(2)	(3)	(4)
Female		-0.085***		-0.101***
		[0.019]		[0.018]
Monitor Theft	0.020	0.055***	0.034**	0.066***
	[0.015]	[0.021]	[0.014]	[0.019]
Monitor Safety	0.043***	0.007	0.038***	-0.000
	[0.015]	[0.021]	[0.015]	[0.020]
Female × Monitor Theft		-0.068**		-0.064**
		[0.028]		[0.026]
Female × Monitor Safety		0.073**		0.076***
		[0.030]		[0.029]
Signal	0.092***	0.113***	0.151***	0.159***
	[0.013]	[0.018]	[0.016]	[0.022]
Female × Signal		-0.042*		-0.018
		[0.023]		[0.027]
Monitor Theft × Signal	0.006	-0.001	-0.040*	-0.048
	[0.019]	[0.025]	[0.023]	[0.031]
Monitor Safety × Signal	-0.025	-0.038	-0.026	-0.040
	[0.019]	[0.025]	[0.022]	[0.030]
Female × Monitor Theft × Signal		0.014		0.018
		[0.031]		[0.039]
Female × Monitor Safety × Signal		0.024		0.028
		[0.032]		[0.037]
Observations	20,759	20,759	20,759	20,759
R-squared	0.113	0.126	0.022	0.035
Control Mean	0.422	0.466	0.475	0.475

Notes: Main sample; the dependent variable is *Hire*. *Signal*, in columns 1 and 2, is an indicator for a top rank in the training cohort; in columns 3 and 4, an indicator for a displayed certificate. Odd columns pool male and female profiles; even columns are fully interacted in gender, signal, and monitoring arm. Columns 1 and 2 include profile and strata fixed effects; columns 3 and 4 include strata fixed effects. The control mean is for business-as-usual candidates without the signal: median-performance candidates (1), median-performance male candidates (2), and non-certified male candidates (3)–(4). All columns include indicators for treated employers who failed or were not asked the comprehension check. Standard errors are clustered at the employer level in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Decomposition of the Hiring Gender Gap

	(1) Latent index	(2) Probability (pp)
Panel A: Hiring gender gap by arm		
Business as usual	0.30	10.64
	[0.05]	[1.65]
Monitoring Safety	0.06	2.14
	[0.05]	[1.93]
Monitoring Theft	0.47	16.77
	[0.05]	[1.70]
Panel B: Decomposition components		
Trust premium (T)	0.17	6.13
	[0.07]	[2.37]
Harassment penalty (H)	0.24	8.50
	[0.07]	[2.54]
Residual (B)	0.23	8.27
	[0.09]	[3.05]
Panel C: Residual by attributes treated as productive		
Technical performance only	0.30	10.64
	[0.05]	[1.65]
+ trustworthiness	0.47	16.77
	[0.05]	[1.70]
+ safety	0.06	2.14
	[0.05]	[1.93]
+ trustworthiness and safety	0.23	8.27
	[0.09]	[3.05]
Employers	912	912
Observations	20,759	20,759

Notes: Probit decomposition of the business-as-usual hiring gender gap into a trust premium (T), a harassment penalty (H), and a residual (B). Column 1: latent index, in probit units. Column 2: average male-minus-female differences in predicted selection probability, in percentage points. Panel A: the gap in each arm. Panel B: the components, $T = \mathrm{MT} - \mathrm{C}$, $H = \mathrm{C} - \mathrm{MS}$, and $B = \mathrm{MT} + \mathrm{MS} - \mathrm{C}$, assuming the treatments do not interact. Panel C: the residual attributed to bias as further attributes are treated as productive (δ_{C}, δ_{MT}, δ_{MS}, B). Main sample, with profile × performance signal and strata fixed effects and indicators for treated employers who failed or were not asked the comprehension check. Delta-method standard errors clustered at the employer level in brackets. A linear probability model on the same sample yields the same decomposition.

Table 8: Trainee Attributes and Employers' Beliefs

	Trainee		Employer	
	Actual		Beliefs	
Panel A: Technical skills				
	Theory	Practical	Theory	Practical
Female	0.214**	0.070	-0.119***	-0.152***
	[0.092]	[0.045]	[0.007]	[0.008]
Mean, men	0.773	0.918	0.578	0.606
Observations	173	173	595	596
Panel B: Cheating task				
	Effort	Claimed (≥ 6)	Effort	Claimed (≥ 6)
Pooled across supervision conditions				
Female	0.017	0.050	0.024***	-0.085***
	[0.023]	[0.043]	[0.005]	[0.006]
By supervision condition				
Female	0.029	0.072	0.047***	-0.114***
	[0.027]	[0.065]	[0.008]	[0.008]
Supervised	-0.003	-0.055*	0.233***	-0.209***
	[0.007]	[0.030]	[0.006]	[0.009]
Female × Supervised	-0.025	-0.044	-0.047***	0.056***
	[0.035]	[0.062]	[0.008]	[0.007]
Mean, unsupervised men	0.989	0.073	0.649	0.463
Observations	355	346	907	904

Notes: *Trainee* columns report trainee outcomes; *Employer* columns report employers' beliefs about trainees, elicited within employer by trainee gender and, in panel B, by supervision condition, rescaled to shares. Panel A: theory is the exam question on which beliefs were elicited; practical is self-reported ability to perform a standard sector task. Across all theory questions, the trainee *Female* coefficient is 0.062 [0.048]. Panel B: effort is the share of 12 registrations checked; the bonus claim is reporting six or more mistakes when five are planted, with effort as a control. *Female* and *Supervised* indicate the trainee's gender and condition or the group the employer was asked about. Trainee regressions include VTI fixed effects: panel A at the trainee level with robust standard errors; in panel B, effort at the sheet level (unsupervised sheets are anonymous) with robust standard errors, and the bonus claim at the trainee level with two observations per trainee and standard errors clustered by trainee. Employer estimates are within-employer contrasts for the 912 employers in the main sample, one observation per employer, with robust standard errors. Mean rows report the omitted-category mean: male trainees (panel A) and unsupervised male trainees (panel B). Standard errors are in brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A Appendix

A.1 Additional Details on Measurement and Coding

We use open-ended survey responses to measure the attributes employers value and their concerns about hiring women. The main text reports answers to "What is the single most important trait you look for when hiring a worker?" Figure 1, panel E, uses answers to "If you think of hiring a woman, what do you worry about?" We classify these answers into worker-attribute categories to make responses comparable across employers and across panels.

The categories build on pilot data. In the pilot, we classified open-ended answers manually with local enumerators. We then used the most common categories to structure the closed-choice questions used in panels B–D, following the approach in Macchiavello et al. (2026). Panel B asks which traits are hardest to find or monitor. Panels C–D ask whether men or women are better at each trait. These panels therefore require no text classification.

For the most-important-trait question, enumerators recorded open-ended answers for 618 employers, 592 of them in the employer sample. A research assistant manually classified these answers using the same categories. Only 0.97% of responses were classified as other, suggesting that the categories capture the main worker traits employers mention in both the pilot and the main survey. For the other 303 employers, enumerators assigned each answer directly to the closest category during the survey. Enumerators did not read the categories aloud. Employers answered in their own words, and enumerators coded the response using the categories derived from earlier open-ended answers: behavior/respect/trust/no stealing, effort/hardworking/discipline, skills/education, learning/flexibility/motivation/interest, background/history, age, cooperation, and other.

The validation in the main text uses the manual coding; for replicability, we also classify the answers with a language model and benchmark it against the manual coding. We classify the 618 open-ended answers using GPT-5 mini and the pre-specified pilot codebook. The codebook allows one answer to enter more than one category if the employer mentions multiple attributes, and assigns the answer to "other" if no listed category applies. We benchmark the classifier against the human coding. The classifier's primary category belongs to the human-coded category set for 96.6% of responses. Category shares are also similar across human and model-assisted coding: behavior/no stealing is 41.91% under human coding and 40.61% under model-assisted coding; effort is 24.43% and 25.40%; skills is 17.31% and 17.80%; learning is 14.24% in both; and cooperation is 5.66% and 7.12%. Combining the model-coded open-ended answers in the employer sample with the 303 mechanically mapped responses yields coded answers for 895 of the 915 employers.

For panel E, we classify the 615 nonmissing open-ended answers, 589 of them in the employer sample and shown in the figure, using GPT-5 mini and the same broad worker-attribute categories used in panel A. This harmonizes worries about hiring women with the attributes employers mention as important in workers. The classifier allows one answer to enter more than one category, and assigns answers to nothing or other when no substantive category applies. We

keep the most common women-specific worries visible inside the broad categories. In the figure, darker bar portions identify strength within skills/education, leaving within learning/interest, and harassment within cooperation/respect.

We benchmark the classifier against the human-coded source used in the previous version of the figure. The classifier's primary category belongs to the human-coded category set for 92.4% of responses. Category shares on the classified responses remain similar: effort is 20.16% under human coding and 20.81% under model-assisted coding; harassment is 38.86% and 38.37%; leaving is 12.20% and 13.33%; strength is 24.55% and 22.28%; and nothing and other are unchanged at 10.73% and 1.79% (of the 615 classified answers).

A.2 Model Details

This appendix formalizes how monitoring changes the hiring consequences of gendered beliefs. Employers assess candidates' technical performance and the theft and harassment risks associated with hiring them, given the production environment. Monitoring capacity changes these risks without changing beliefs about underlying worker characteristics or preferences over gender.

Set-up. An employer decides whether to hire a candidate of gender $g \in \{m, f\}$. The employer forms beliefs about the candidate's technical performance, the likelihood that the candidate steals, and the likelihood that the candidate will be harassed or treated disrespectfully by at least one coworker.

The production environment determines how these beliefs affect the candidate's value. Theft risk depends on the employer's capacity to monitor workers, while harassment risk also depends on the firm's gender composition. We denote the probabilities of detecting theft and harassment by $\mu \in [0, 1]$ and $\eta \in [0, 1]$, respectively, and the male share of the existing workforce by $m \in (0, 1]$. We consider one hire at a time and take m as given, so the model leaves out how hiring changes the composition of the workforce.

Theft risk. A worker decides whether to steal by comparing the expected gain with a personal (moral) cost. Stealing yields $\pi > 0$ if undetected, while the personal cost is $c_g \geq 0$. With detection probability μ, the worker steals when $(1 - \mu)\pi > c_g$. Monitoring increases the chance that theft is detected, reducing the expected gain from stealing. A detected worker forfeits the gain; a further sanction, such as dismissal, would strengthen this deterrence effect. Detection does not undo the employer's loss, for example when stolen tools have been sold or a client has been taken; recovering part of it would only lower the cost of theft further.

The employer does not observe this cost and forms beliefs about its distribution by gender. We assume that the employer perceives c_g as uniformly distributed on $[0, \kappa_g]$, with $\kappa_g \geq \pi$. The perceived probability of theft is therefore

$$\tilde{s}_g(\mu) = \frac{(1 - \mu)\pi}{\kappa_g}, \qquad g \in \{m, f\}, \tag{4}$$

where the tilde denotes the employer's belief. The condition $\kappa_g \geq \pi$ ensures that this probability does not exceed one. This expression separates beliefs about workers from the production environment: κ_g captures perceived differences in workers' willingness to steal, while μ captures the employer's monitoring capacity. Greater monitoring reduces theft risk without changing these beliefs.

Harassment risk. Employers' beliefs about a candidate's exposure to harassment depend on the behavior they expect from coworkers and the firm's gender composition. We model this interaction by pairing the new hire with a coworker, who is male with probability m. We assume that employers believe male coworkers may harass new hires, particularly women, while female coworkers do not.

A male coworker gains $\beta_g > 0$ from harassing a worker of gender g if undetected and bears a personal cost $d \geq 0$. With detection probability η, he harasses when $(1-\eta)\beta_g > d$. The employer perceives the gain as $\tilde{\beta}_g$ and believes that d is uniformly distributed on $[0, \bar{d}]$, with $\bar{d} \geq \tilde{\beta}_g$. The perceived harassment risk is therefore

$$\tilde{h}_g(\eta) = m \, \frac{(1-\eta)\tilde{\beta}_g}{\bar{d}}, \qquad g \in \{m, f\}. \tag{5}$$

In the employer's assessment, greater monitoring reduces this risk, while a larger male share of the workforce increases it. Differences in $\tilde{\beta}_g$ capture employers' beliefs about whether male or female hires are more likely to be targeted.[38]

Payoffs. A hire produces y_L, plus $\Delta y > 0$ if the worker is skilled. The employer believes that a candidate of gender g with performance signal q is skilled with probability $\hat{p}^S_g(q)$, his posterior given the signal, which rises with q for both genders. A theft costs him $\theta > 0$ and a harassment incident costs him $\tau > 0$. The employer bears these risks only above a threshold, $\bar{s}$ for theft and $\bar{h}$ for harassment, the same for male and female hires. Below the threshold, he manages the risk with the supervision he already provides. Let b denote his residual preference against women: any difference in value between a man and an otherwise identical woman that the three attributes do not explain. The values of a male and a female hire are

$$V_m = y_L + \Delta y \, \hat{p}^S_m(q) - \theta \max\{\tilde{s}_m(\mu) - \bar{s}, 0\} - \tau \max\{\tilde{h}_m(\eta) - \bar{h}, 0\}, \tag{6}$$

$$V_f = y_L + \Delta y \, \hat{p}^S_f(q) - \theta \max\{\tilde{s}_f(\mu) - \bar{s}, 0\} - \tau \max\{\tilde{h}_f(\eta) - \bar{h}, 0\} - b. \tag{7}$$

[38] In 86% of firms the workforce is all male, so $m = 1$. Hiring a man may also raise the harassment risk for the women already at a firm, which would add a cost to a male hire. We omit this term because the sectors are highly segregated. The survey likewise asks employers what worries them about hiring a woman, not about the risk a male hire poses to female staff. In an exploratory check, safety monitoring visits raise the share of male profiles selected by 5.2 percentage points in firms that already employ women and lower it by 2.3 percentage points in all-male firms (difference $p = 0.077$). This pattern is consistent with the omitted term, but imprecise.

We posit that, without monitoring visits, employers believe that men's theft risk exceeds the threshold while women's does not, $\tilde{s}_f(\mu) \leq \bar{s} < \tilde{s}_m(\mu)$, and that women's harassment risk exceeds the threshold while men's does not, $\tilde{h}_m(\eta) \leq \bar{h} < \tilde{h}_f(\eta)$. Thus women have a higher personal cost of stealing than men ($\kappa_f > \kappa_m$) and face greater harassment risk from male coworkers ($\tilde{\beta}_f > \tilde{\beta}_m$). These beliefs match the survey evidence: employers perceive women as less likely to steal and cite harassment as their main concern about hiring women (Figure 1, panels D and E).

Hiring. The employer hires when the candidate's expected value, V_g, plus an idiosyncratic match component, ε, is nonnegative. We assume $\varepsilon \sim \mathcal{N}(0,1)$, independently of candidate characteristics and monitoring. The hiring probability is therefore $H_g = \Phi(V_g)$, where Φ is the standard normal cumulative distribution function. We derive predictions for the difference in employers' valuations of male and female candidates:

$$\Delta \equiv V_m - V_f, \tag{8}$$

which the probit coefficients measure. The gender gap in hiring, $H_m - H_f = \Phi(V_m) - \Phi(V_f)$, has the same sign. A positive gap means that the employer values the male candidate more than an otherwise identical female candidate.

Assumptions. The model embeds two assumptions.

Assumption 1 (performance signal). The signal q affects only $\hat{p}_g^S(q)$, not perceived theft or harassment risk.

Assumption 2 (monitoring and beliefs). Monitoring capacity does not change employers' beliefs about underlying worker characteristics, the costs they attach to theft and harassment, or their preferences over gender.

Prediction: effects of monitoring capacity. Under Assumptions 1 and 2 and these beliefs:

(i) greater capacity to monitor theft widens the gender gap:

$$\frac{\partial \Delta}{\partial \mu} = \frac{\theta\, \pi}{\kappa_m} > 0; \tag{9}$$

(ii) greater capacity to monitor harassment narrows the gender gap:

$$\frac{\partial \Delta}{\partial \eta} = -\frac{\tau\, m\, \tilde{\beta}_f}{\bar{d}} < 0. \tag{10}$$

Identifying assumption. To interpret the effects of the monitoring visits as a test of the predicted effects of monitoring capacity, we add one identifying assumption.

Assumption 3 (targeted visits). Theft monitoring visits raise the capacity to monitor theft, μ, and leave the capacity to monitor harassment, η, unchanged. Safety monitoring visits raise η and leave μ unchanged.

Section 4 discusses the evidence on all three assumptions.

A.3 A Discrete-Choice Decomposition of the Hiring Gender Gap

This appendix derives the decomposition used in Section 5.2 from the model, describes its estimation, and states its assumptions.

Components. Substituting equations (6) and (7), the gender gap is

$$\Delta = B - T + H, \qquad T = (1-\mu)\,\theta\,\pi\left(\frac{1}{\kappa_m} - \frac{1}{\kappa_f}\right), \qquad H = (1-\eta)\,\frac{\tau\, m}{\bar{d}}\left(\tilde{\beta}_f - \tilde{\beta}_m\right). \tag{11}$$

Here T is women's trust premium, H is their harassment penalty, and B combines the technical-performance belief gap and residual preferences over gender. Averaging over employers before the visits, and when each type of visit fully removes its targeted friction, the three arms give

$$\delta_{\mathrm{C}} = B - T + H, \qquad \delta_{\mathrm{MT}} = B + H, \qquad \delta_{\mathrm{MS}} = B - T, \tag{12}$$

where δ_a denotes the gender gap in valuations in arm a, and B, T, and H now denote averages over employers. Solving the three equations gives

$$T = \delta_{\mathrm{MT}} - \delta_{\mathrm{C}}, \qquad H = \delta_{\mathrm{C}} - \delta_{\mathrm{MS}}, \qquad B = \delta_{\mathrm{MT}} + \delta_{\mathrm{MS}} - \delta_{\mathrm{C}}. \tag{13}$$

Estimating equation. We estimate the latent value of hiring candidate i for employer j in arm $a \in \{\mathrm{C}, \mathrm{MT}, \mathrm{MS}\}$ as

$$U^*_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} - \mathbf{1}[g_i = f]\,\delta_a + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim \mathcal{N}(0,1), \tag{14}$$

and the employer selects the candidate when $U^*_{ij} > 0$. The controls $\mathbf{x}_{ij}$ replace the value of a male candidate. They are base-profile-by-performance-signal fixed effects, arm indicators, and sector, survey-wave, and enumerator fixed effects. The arm indicators capture that each type of visit raises the value of every candidate. The coefficient δ_a is the average gender gap in valuations in arm a, on the probit scale. The hiring gender gap does not detectably vary with the performance signal (Section 4.3). We therefore estimate equation (14) without an interaction between gender and the signal and report components pooled across performance levels.[39]

[39]The value of a male candidate, and every other part of the value that is the same for men and women, varies across employers beyond what the fixed effects capture. These differences enter the error term. They affect male and female candidates of the same employer equally, so they change the scale of

Assumptions. Beyond Assumptions 1–3 from the model, the decomposition uses the additive form of the valuation: the trust premium and harassment penalty enter employers' valuations additively. Our benchmark is that each type of visit fully removes its targeted friction; Appendix Figure A10 relaxes this.

Estimates. We estimate equation (14) by probit using the main sample, with the same fixed effects and employer-level clustering as Table 3. Table 7 reports the decomposition on the latent valuation scale in column 1 and in percentage points in column 2. For column 2, we calculate the average male–female difference in predicted hiring probabilities in each arm and apply equation (13) to these gaps. The estimated hiring gender gaps in all three arms differ by less than 0.1 percentage points between the probit and linear probability models.

Partial removal of the targeted frictions. In the benchmark, each type of visit removes the entire friction it targets. Suppose instead that it removes a share $\lambda \in (0, 1]$. The differences across arms then equal λT and λH, and

$$T = \frac{\delta_{\mathrm{MT}} - \delta_{\mathrm{C}}}{\lambda}, \qquad H = \frac{\delta_{\mathrm{C}} - \delta_{\mathrm{MS}}}{\lambda}, \qquad B = \delta_{\mathrm{C}} + T - H. \tag{15}$$

Appendix Figure A10 plots T, H, and B for λ between 1 and 0.2, using the estimated hiring gender gaps by arm in panel A, column 2 of Table 7, and the same λ for both types of visit.

the latent index, not the gap between men and women. The components in percentage points are robust to this scale.

A.4 Within-Subject Design Results

Following the pre-registration amendment, employers in carpentry and welding evaluate profiles under two of the three experimental conditions. The two conditions and their order are randomized across all six possible pairs. The design serves two purposes. First, it increases statistical power in carpentry and welding, where we added it (Section A.5). Second, it serves as a robustness check, estimating the visit effects within the same employer. After 24 evaluations under his first condition, each of these employers rates 12 more profiles, randomly assigned from the same base profiles in another version, under the second condition. We estimate

$$\begin{aligned}\text{Hire}_{ijb} = \beta_1 \text{Female}_{ij} + \beta_2 \text{MT}_{jb} + \beta_3 \text{MS}_{jb} \\ + \beta_4 \text{Female}_{ij} \times \text{MT}_{jb} + \beta_5 \text{Female}_{ij} \times \text{MS}_{jb} + \delta_{iq} + \alpha_j + u_{ijb},\end{aligned} \tag{16}$$

where b indexes the two sets of profiles an employer rates, and MT_{jb} and MS_{jb} indicate the condition under which employer j rates set b. The employer fixed effect α_j replaces the sector, survey-wave, and enumerator fixed effects of equation (3). The visit coefficients therefore come only from comparing the two sets rated by the same employer. As in the main specification, β_4 and β_5 measure how each type of visit changes female selection relative to male selection, and standard errors are clustered at the employer level.

Within the same employer, the two types of visits move the gender gap in opposite directions, as in the main design. Both effects are larger than in the main design. Theft monitoring visits increase the hiring gender gap by 11.9 percentage points, and safety monitoring visits reduce it by 15.4 percentage points (both $p < 0.001$; Table A5, column 1). The opposite effects hold when we exclude employers who prefer an all-male workforce (column 2) and in each of the top four quintiles of stated preferences (columns 4–7).[40]

[40] In the bottom quintile, neither effect is statistically significant (column 3).

A.5 Deviations from Preregistration

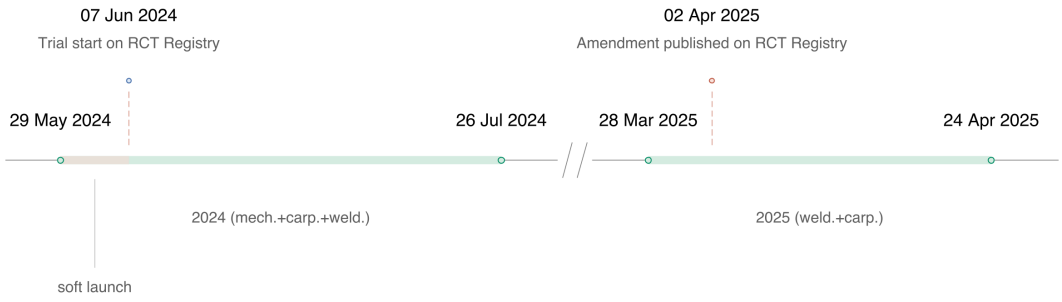
We note three main deviations from the first preregistration. First, the second primary outcome was not recorded as intended due to a coding error in the survey. The variable *Retain* — elicited using the question, "Okumuwa: Kisooboka kyenkanaki gwe okuwa omukozi ono omulimu nga [job] ku bizinensi yo? Geregeranya ku lupimo oluva ku 0 okutusa ku 10, nga 0 etegeza nti tekisoboka ddala ate 10 kisoboka ddala", meaning "How likely would you be to give this worker a (post-probation/stable) position as a mechanic at your firm? Please rate on a scale from 0 to 10, where 0 is very unlikely and 10 is very likely" — was intended to be a primary outcome to capture retention post-probation but was not recorded in the first wave; we therefore chose not to collect it in the second wave.

Second, we expanded the study to include two additional sectors, carpentry and welding, in addition to mechanics. The first preregistration anticipated expanding to additional sectors if we could not reach the target sample in mechanics; we reached the target but expanded for power and external validity, as documented in the amendment. We observed substantial heterogeneity in stated preferences for hiring women, and the additional sectors provide sufficient power to study treatment effect heterogeneity by stated preferences.

Third, we added a *within-subject* component to improve statistical power in carpentry and welding. This component was implemented in both the 2024 and 2025 waves. After completing the initial 24 evaluations under the main treatment condition, these employers rated 12 additional profiles under a second condition.

The preregistration amendment documents these changes. Estimates are similar when we include soft-launch interviews; both visit effects retain their signs and are larger in magnitude in the post-amendment sample (Table A3; Figure A1). Table A1 lists each outcome and analysis with its preregistration status.

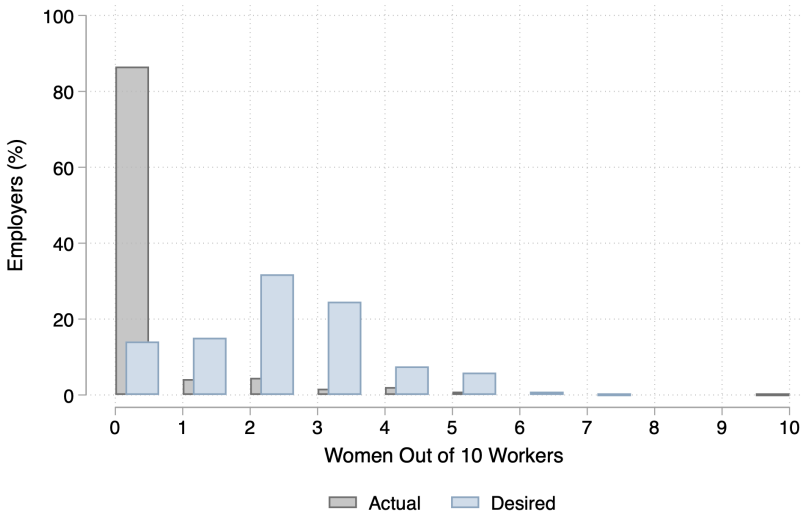
Figure A1: Timeline of Data Collection and Registry Milestones



Notes: Data collection by survey date; the horizontal axis is broken between waves. The 2024 wave (29 May–26 July 2024) covers motor mechanics, then carpentry and welding; the 2025 wave (28 March–24 April 2025) covers welding and carpentry. The soft launch (29 May–07 June 2024) precedes the registered trial start. Markers indicate the trial start (07 June 2024) and the preregistration amendment (02 April 2025).

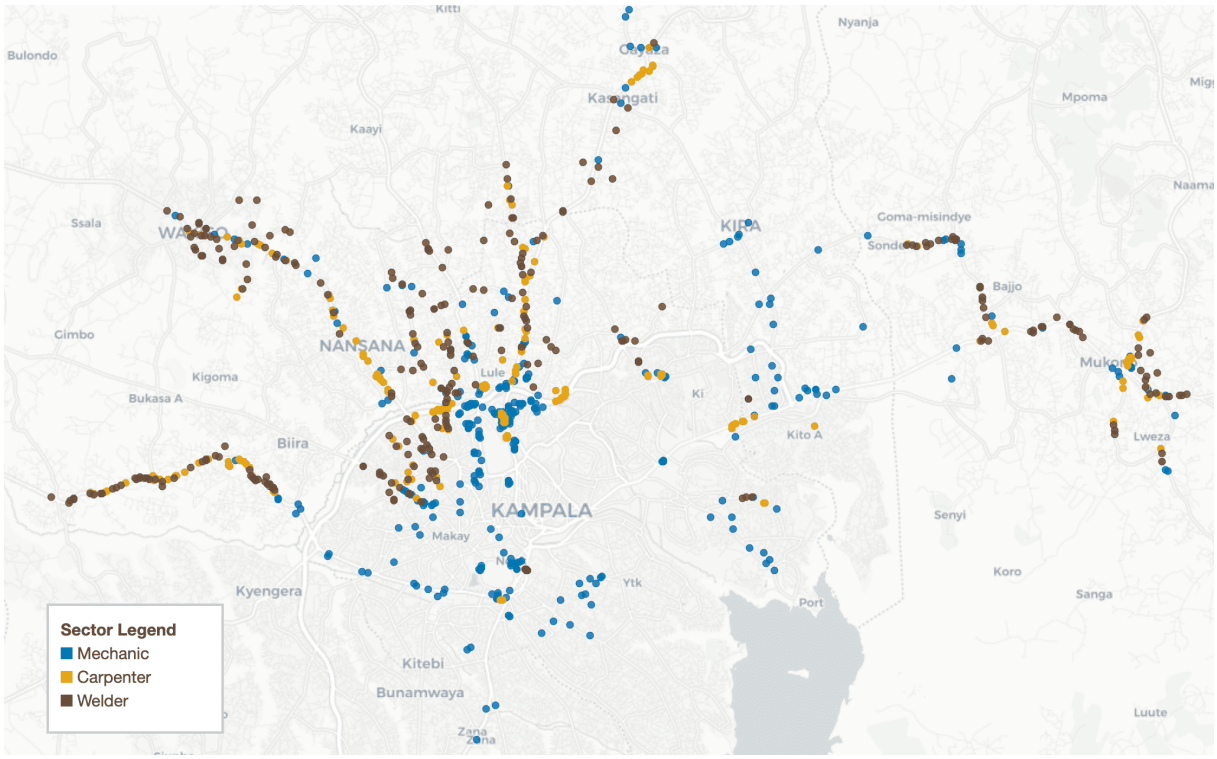
A.6 Appendix Figures

Figure A2: Employers' Desired and Actual Gender Mix in the Workforce



Notes: Employer sample (N=915; the actual share is available for 906). Distributions of the desired gender composition, from 0 to 10 women out of 10 workers, and of the actual share of women among current workers, rescaled to 10.

Figure A3: Firm Spatial Distribution



Notes: Locations of the 915 study firms in the Kampala metropolitan area, by sector, each randomly displaced by up to 100 meters.

Figure A4: Firm and Vocational Training Environment

(A) Firm Layout



(B) Practical VTI Class

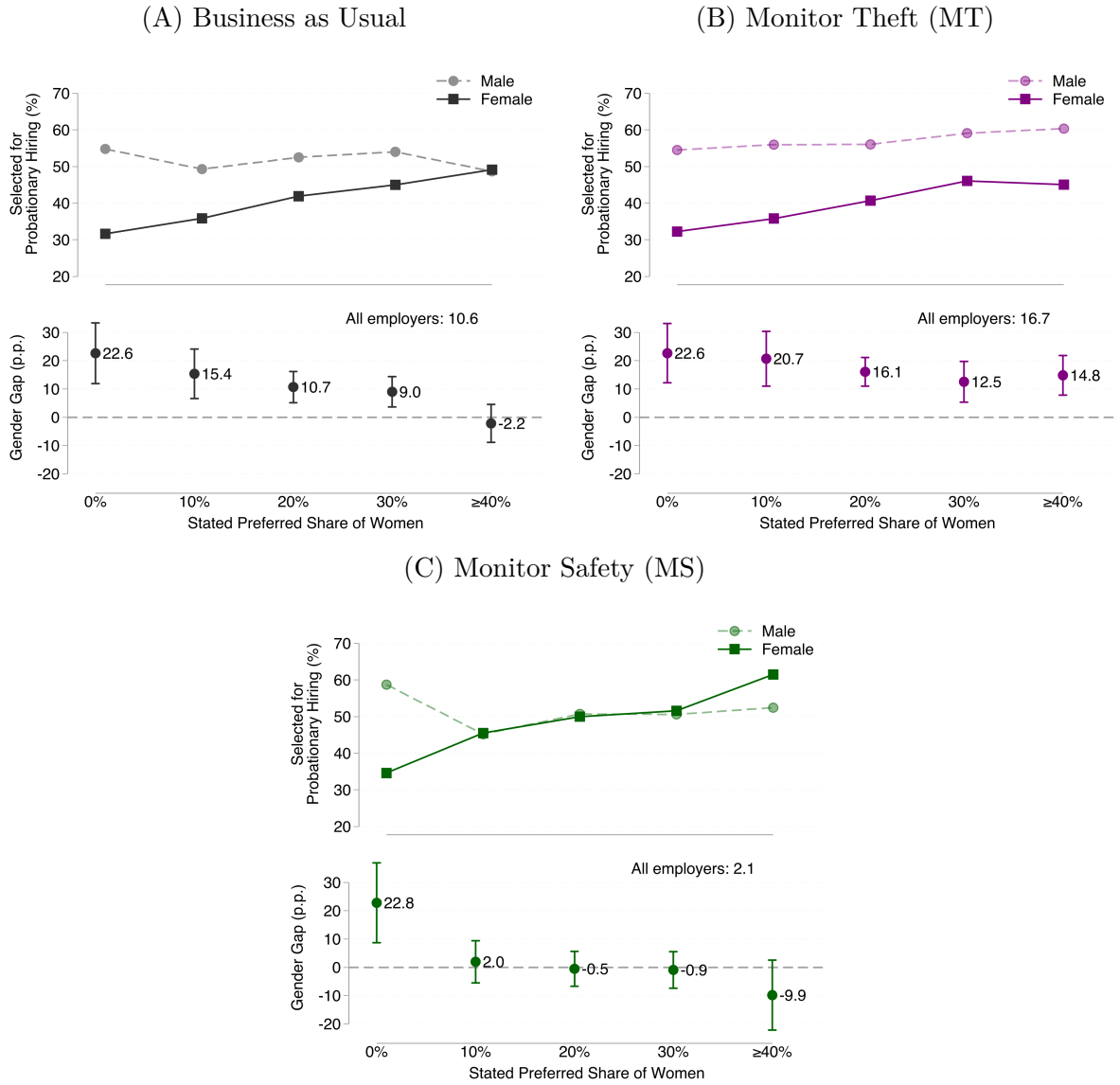


(C) Internship



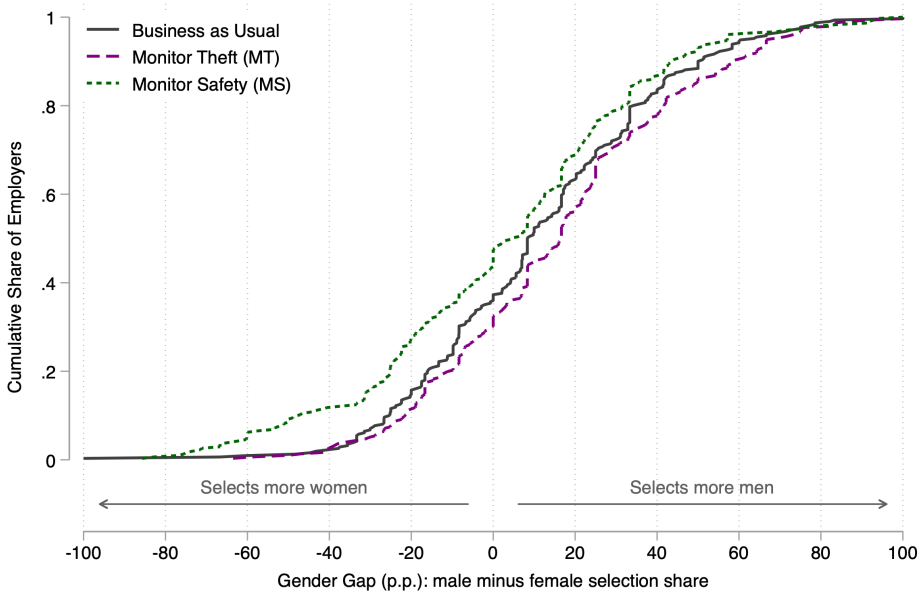
Photo credit: © Mariajose Silva-Vargas.

Figure A5: Hiring Gender Gap by Treatment Arm and by Stated Preferences for Hiring Women (Raw Data)



Notes: Hiring gender gap by stated preferred share of women, by arm: business as usual (A), theft monitoring (B), and safety monitoring (C). Top graphs: raw share of profiles selected for probationary hiring, by candidate gender and preferred female share. Bottom graphs: the female coefficient from equation (3), by preferred female share. Bars are 95% confidence intervals.

Figure A6: Distribution of Employer-Level Gender Gaps by Monitoring Arm (Raw Data)



Notes: Empirical CDFs of the within-employer hiring gender gap (male minus female selection rate, in percentage points) by arm: business as usual (solid grey), theft monitoring (dashed purple), and safety monitoring (dotted green).

Figure A7: Theory Technical Exam — An Example (Motor Mechanics)

FULL NAME:_____ ☐ MAN

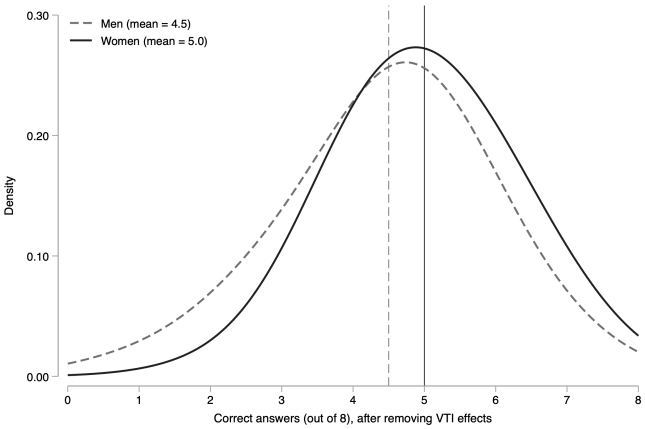
ID:_____ ☐ WOMAN

MOTOR-MECHANICS

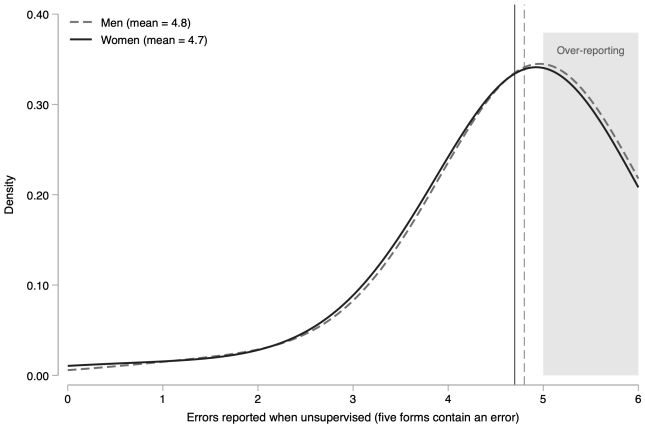
	Question	Answers
1	*multiple-choice* What are you advised to do when servicing the engine by changing oil?	A. Top up lubricating oil B. Replace oil filter C. Over hand engine D. Over hand cylinder head
2	*multiple-choice* What immediate remedy can you give to a vehicle with a problem of excessive tyre wear in the center more than other parts?	A. Increase tyre pressure B. Reduce tyre pressure C. Inflate pressure D. Remove the vehicle tire
3	*Multiple-choice* If a customer reports to you that his/her vehicle charging system works at lower rate, how can you help him?	A. Replacing the charging system B. Adjusting the alternator tension C. Replacing alternator housing D. Renewing wire insulator
4	*Multiple-choice* Which of the following set of systems or component call for mechanical adjustment during general vehicle service?	A. Tyres, cooling system, master cylinder B. Break shoes, alternator, and valve clearance C. Distributor, radiator, propeller shaft D. Tank, crank shaft, Turbo charger
5	*Multiple-choice* What solution would you give a customer with a vehicle engine producing blue smoke?	A. Top up lubricant B. Time the engine C. Replace piston rings D. Remove carbon deposits
6	*Matching* What should you do to stop the following vehicle troubles?	1. Battery over charging — A. Leaking fuel tank 2. Engine over heating — B. Renew regulator 3. Lubricant leakage — C. Reduce oil to the correct level 4. Smoke in exhaust — D. Renew piston rings 5. Engine fails to start — E. Charge the battery ____ ; ____ ; ____ ; ____ ; ____ ; ____ .
7	*Order* When changing engine oil, in which order should you perform the following steps?	A. Drain oil through drain plug B. Remove oil filter cup C. Run engine to check leaks D. Fill new oil through filler cup to level E. Remove oil filter F. Warm up the engine 1____ ; 2____ ; 3____ ; 4____ ; 5____ ; 6____ .

Figure A8: Distribution of Trainee Outcomes by Gender

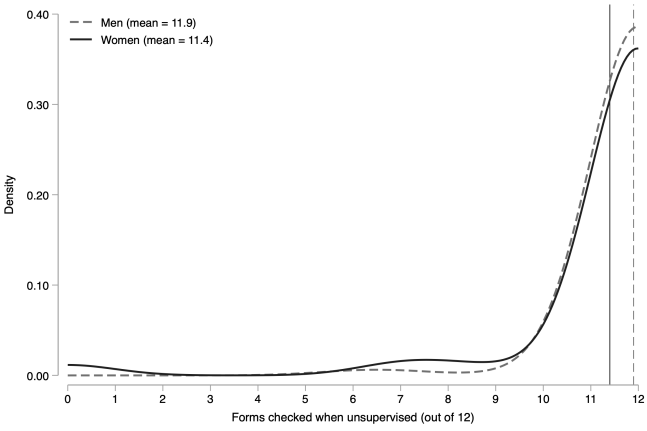
(A) Technical Skills



(B) Trustworthiness

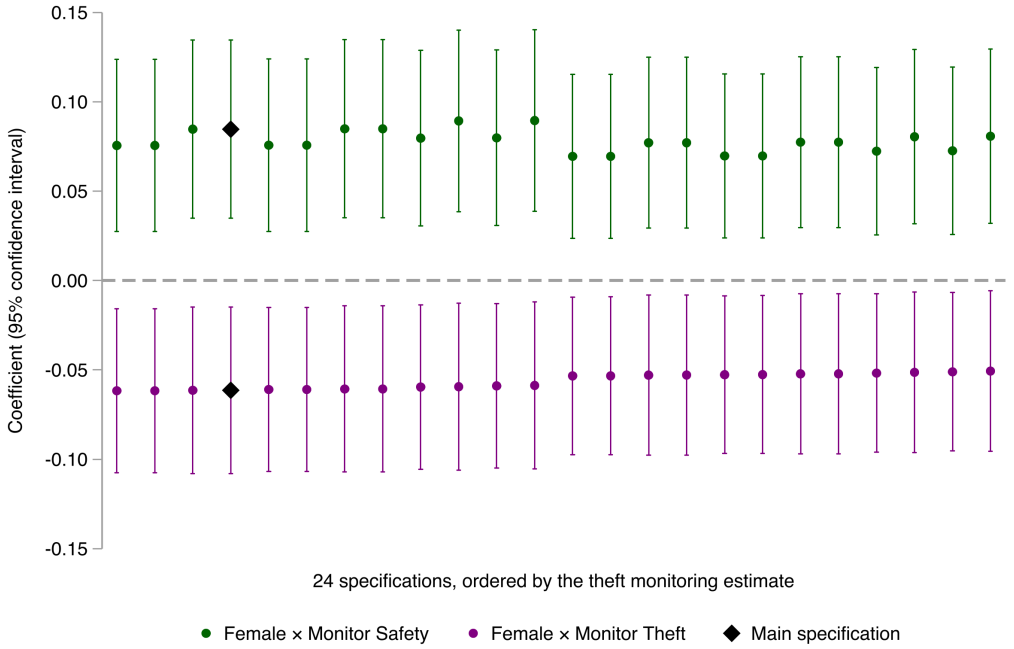


(C) Effort



Notes: Each panel compares male trainees (dashed line) with female trainees (solid line); vertical lines mark the means. Panel A: number of correct answers to the seven theory exam questions, plus one if the trainee answers yes to the question on the practical task (maximum eight), after removing VTI fixed effects. Panel B: number of errors trainees report finding when unsupervised. Only five forms contain an error, so a report above five is an over-report (shaded area). Panel C: number of forms trainees check when unsupervised, out of 12. Kolmogorov–Smirnov tests of equal distributions for men and women: $p = 0.09$ (panel A), $p = 1.00$ (panel B), and $p = 0.99$ (panel C).

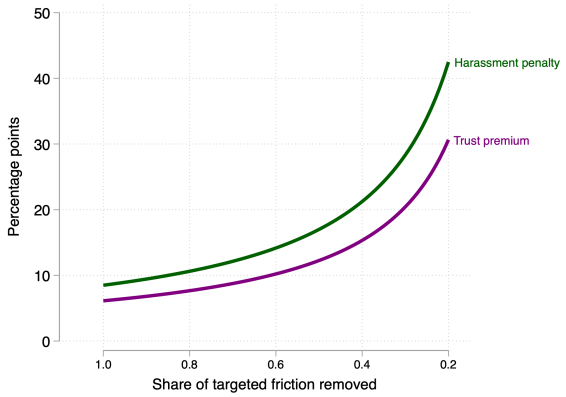
Figure A9: Robustness: All Combinations of Sample and Specification Choices



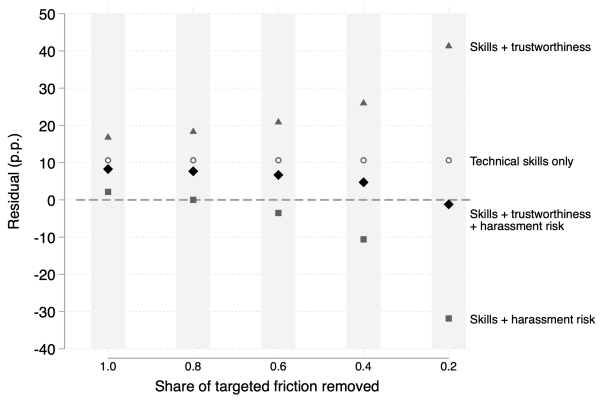
Notes: Estimates of the *Female* × *Monitor Safety* and *Female* × *Monitor Theft* coefficients from column 1 of Table 3 under all 24 combinations of four choices: excluding or including the blocks that fail the attention check; controlling for the comprehension check, excluding employers who fail it as preregistered, or neither; excluding or including soft-launch firms; and omitting or adding the three baseline controls of Appendix Table A3, column 5. Diamonds mark the main specification. Soft-launch firms were not asked the comprehension check and are never excluded on it. Bars are 95% confidence intervals. Standard errors are clustered at the employer level.

Figure A10: Decomposition under Partial Removal of the Targeted Frictions

(A) Trust Premium and Harassment Penalty



(B) Residuals



Notes: Components of the hiring gender gap under business as usual when each type of visit removes a share of the friction it targets (horizontal axis). Both types of visit remove the same share. Panel A: trust premium and harassment penalty. Panel B: at each share, residual compared across models that account for different attributes. The one-dimensional model accounts for technical skills only, so its residual is the gap under business as usual. The two-dimensional models add either trustworthiness or harassment risk. The three-dimensional model accounts for all three attributes.

A.7 Appendix Tables

Table A1: Deviations from the Preregistration

Item	Change	Registry date
Retention (second primary outcome)	Not collected: survey coding error in wave 1; dropped in wave 2	—
Sample expansion: carpentry and welding	Anticipated in the first preregistration; added by amendment for external validity and statistical power	02 Apr 2025
Within-subject second monitoring condition	Added from the second sector onward (carpentry and welding, both waves); documented in the amendment	02 Apr 2025
Heterogeneity by workforce gender-diversity preference	Measure and analysis added by amendment	02 Apr 2025

Notes: All other outcomes and analyses follow the preregistration (AEA RCT Registry AEARCTR-0013698, registered 19 June 2024; trial start 07 June 2024). One amendment, published 02 April 2025, covers the rows dated accordingly. Table A3 shows the results are robust to restricting the sample to observations collected after the trial start and after the amendment.

Table A2: Experimenter Demand Checks

Panel A: Experimenter-demand measure and correlation with treatment

	Share	Monitor Safety	Monitor Theft	Joint p
States that team prefers hiring of:				
Men	0.244	0.019 [0.036]	-0.032 [0.035]	0.34
No preference	0.686	0.007 [0.038]	0.023 [0.038]	0.82
Women	0.070	-0.026 [0.020]	0.009 [0.022]	0.20

Panel B: Results are robust to dropping employers who believe the team prefers they hire women

	(1) Baseline	(2) Excl. believers	(3) p-value
Female	-0.106***	-0.121***	<0.01
	[0.016]	[0.017]	
Monitor Safety	-0.012	-0.014	0.70
	[0.016]	[0.016]	
Monitor Theft	0.054***	0.060***	0.22
	[0.016]	[0.016]	
Female × Monitor Safety	0.085***	0.096***	0.37
	[0.025]	[0.026]	
Female × Monitor Theft	-0.061***	-0.067***	0.51
	[0.024]	[0.025]	
Observations	20,759	19,336	
Control mean	0.522	0.529	
Share female among selected	0.448	0.440	

Notes: Checks that experimenter demand does not drive the results, following de Quidt et al. (2026). At the end of the survey, employers state whether they thought the research team preferred them to hire men, women, or had no preference; 897 of the 912 main-sample employers answer. Panel A: for each answer, the share of employers and the coefficients from regressing its indicator on the monitoring arms (business-as-usual omitted) at the firm level, robust standard errors in brackets; *Joint p* tests that both arm coefficients are zero. Panel B: the main hiring regression at the profile level, with profile-by-performance and strata fixed effects and indicators for treated employers who failed or were not asked the comprehension check; column (2) drops employers who believe the team preferred women (non-respondents to the item retained); column (3) reports, for each coefficient, the p-value that it differs for the excluded employers, from the model fully interacted with the believer indicator. *Share female among selected* is the female share of selected profiles in each sample. Interacting the gender gap and the arms with the item leaves the triple interactions insignificant ($p = 0.72$ for Monitor Theft, $p = 0.81$ for Monitor Safety), as does adding the stated preference gradient ($p = 0.67$, $p = 0.67$). Standard errors in brackets are clustered at the employer level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Robustness: Alternative Samples

Hire (0-1)	(1) Main sample	(2) Attention check	(3) Comprehension screen	(4) Soft launch	(5) Baseline controls
Female	-0.106***	-0.103***	-0.106***	-0.102***	-0.106***
	[0.016]	[0.016]	[0.016]	[0.016]	[0.016]
Female × Monitor Safety	0.085***	0.077***	0.089***	0.076***	0.085***
	[0.025]	[0.024]	[0.026]	[0.025]	[0.025]
Female × Monitor Theft	-0.061***	-0.053**	-0.059**	-0.062***	-0.061**
	[0.024]	[0.023]	[0.024]	[0.023]	[0.024]
Employers	912	915	886	951	912
Observations	20,759	21,810	20,170	21,668	20,759
Control mean	0.522	0.513	0.522	0.522	0.522

Notes: Column (1) of Table 3 re-estimated on alternative samples. Column (1): main sample. Column (2): adds back the profile blocks excluded by the attention check. Column (3): applies the preregistered comprehension screen, excluding the treated firms that failed, or were not asked, the comprehension check. Column (4): adds the soft-launch firms interviewed before the registered trial start (07 June 2024). Column (5): main sample, controlling for the three baseline variables imbalanced in Table 2 (vocational training, firm capital, daily hours worked), with missing values set to the median and flagged. The dependent variable is *Hire*; monitoring main effects are estimated but not reported. Profile and strata fixed effects included, as are indicators for treated firms that failed or were not asked the comprehension check, except in column (3). Standard errors clustered at the employer level in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Heterogeneity of Treatment Effects

Hire (0-1)	(1) Female	(2) Female × Monitor Theft	(3) Female × Monitor Safety	Employers
Panel A: Preregistered dimensions				
Preferred female share:				
Below 30%	-0.146***	-0.042	0.090***	556
	[0.023]	[0.032]	[0.034]	
30% or more	-0.044**	-0.089***	0.077**	356
	[0.021]	[0.034]	[0.037]	
p-value, equal across groups	0.001	0.320	0.789	
Slope in the share (0–1)	0.482***	-0.265	0.136	912
	[0.126]	[0.170]	[0.211]	
Employer gender:				
Male employer	-0.107***	-0.058**	0.092***	879
	[0.017]	[0.024]	[0.026]	
Female employer	-0.056	-0.144	-0.234	33
	[0.104]	[0.123]	[0.154]	
p-value, equal across groups	0.631	0.493	0.037	
Recruitment from VTIs:				
Does not recruit from VTIs	-0.090***	-0.080***	0.083***	747
	[0.018]	[0.026]	[0.028]	
Recruits from VTIs	-0.172***	0.021	0.091	165
	[0.042]	[0.059]	[0.059]	
p-value, equal across groups	0.070	0.119	0.902	
Panel B: Exploratory dimension (not preregistered)				
Who monitors workers:				
Owner alone	-0.043	-0.124**	0.021	170
	[0.034]	[0.052]	[0.046]	
Owner with workers	-0.128***	-0.088**	0.111***	421
	[0.024]	[0.036]	[0.040]	
Manager(s) involved	-0.122***	0.013	0.096**	321
	[0.030]	[0.039]	[0.040]	
p-value, equal across groups	0.106	0.061	0.307	
p-value, managers = owner alone	0.084	0.036	0.224	
p-value, managers = owner with workers	0.860	0.057	0.792	

Notes: Main sample. Each heterogeneity dimension is analyzed in a separate regression: column (1) of Table 3 fully interacted with that dimension. Each cell is the estimate for that group. Column (1): the *Female* coefficient under business as usual. Columns (2) and (3): the *Female* × *Monitor Theft* and *Female* × *Monitor Safety* coefficients. The *p*-value rows test that the estimates are equal across groups. Panel A: preregistered dimensions. *Preferred female share*: the share of women in the employer's ideal workforce, split at 30% (the preregistered median split; the mass point at 20%, the median, falls in the low group) and, in the *Slope* row, entered continuously (0 to 1); that row reports the interaction with the share. *Recruitment from VTIs*: the firm recruits from vocational training institutes. Panel B: exploratory dimension, not preregistered. *Who monitors workers*, as reported by the respondent: the owner alone; the owner or other workers, with no manager; or one or more managers. The dependent variable is *Hire*. Profile and strata fixed effects included, as are indicators for treated employers who failed or were not asked the comprehension check. Standard errors clustered at the employer level in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Within-Subject Results

			Stated Preferences for Workforce Gender Mix				
Hire (0-1)	(1) Main Sample	(2) Gender-Mix Pref. > 0	(3) Bottom QT	(4) 2nd QT	(5) 3rd QT	(6) 4th QT	(7) Top QT
Female	-0.117***	-0.098***	-0.243***	-0.128**	-0.100***	-0.125***	-0.050*
	[0.015]	[0.015]	[0.060]	[0.053]	[0.026]	[0.026]	[0.028]
Monitor Theft	0.092***	0.102***	0.034	0.071*	0.120***	0.084***	0.107***
	[0.013]	[0.013]	[0.043]	[0.036]	[0.022]	[0.026]	[0.029]
Monitor Safety	-0.047***	-0.047***	-0.016	-0.098***	-0.041*	-0.052**	-0.027
	[0.013]	[0.013]	[0.049]	[0.036]	[0.024]	[0.024]	[0.028]
Female x Monitor Theft	-0.119***	-0.122***	-0.111	-0.144**	-0.126***	-0.092**	-0.123***
	[0.021]	[0.021]	[0.074]	[0.066]	[0.034]	[0.039]	[0.039]
Female x Monitor Safety	0.154***	0.168***	-0.017	0.179***	0.164***	0.187***	0.170***
	[0.023]	[0.023]	[0.091]	[0.062]	[0.040]	[0.040]	[0.050]
Observations	20,698	18,379	2,319	2,447	6,573	5,654	3,705
R-squared	0.219	0.213	0.313	0.252	0.232	0.207	0.191
Control Mean	0.531	0.530	0.540	0.533	0.513	0.551	0.523
p-value MT vs. MS	0.000	0.000	0.302	0.000	0.000	0.000	0.000

Notes: Employers in the within-subject design, who rated a second set of profiles under a different condition; the two conditions and their order are randomized across all six pairs. Column (1): all such employers. Column (2): excludes employers who prefer an all-male workforce. Columns (3)–(7): quintiles of the stated preferred share of women. The dependent variable is *Hire*; *Monitor Theft* and *Monitor Safety* indicate the condition under which the set was rated. The last row reports the p-value of the test that the two *Female* $\times$ arm interactions are equal. Employer and profile-by-performance fixed effects included. Standard errors clustered at the employer level in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: Female Hiring at Follow-Up

	Baseline (1)	Follow-up (2)	Difference (3)	p-value (4)	N (5)
Panel A: Workforce at the time of the survey					
Firms with at least one woman (share)	0.136	0.236	0.100	0.000	771
Women employed (N)	0.182	0.385	0.204	0.000	771
Female workers (share)	0.027	0.056	0.029	0.000	737
Workers (N, excl. owners)	8.29	7.14	-1.15	0.000	774
Ideal female share (0–1)	0.217	0.270	0.053	0.000	776
Panel B: Hiring since the end of the experimental wave					
Firms that hired at least one woman (share)		0.594			776
Women employed at a given time (N)		0.51			776
Female workers at a given time (share)		0.100			744
Women hired (N per firm-year)		0.95			776
Female worker tenure (months)		7.1			461

Notes: Employer phone follow-up about two years after the experiment: 776 firms reached, consent rate 97.5% among those called. Columns: baseline value for the same firms, follow-up value, within-firm difference, Wilcoxon signed-rank p-value of no within-firm change, and number of firms; Panel B has no baseline counterpart. Panel A: the workforce at each survey. *Workers* excludes owners and includes trainees; *Female workers (share)* requires at least one worker at both surveys; *Ideal female share* is the stated ideal share of women. Panel B: hiring between the firm's data-collection round and the follow-up call (24 months for 2024 firms, 15 for 2025, 21.0 on average). *Women hired* counts women who worked at the firm at any point in the window, harmonized by window length; *Female worker tenure* is defined for firms that hired a woman; *Women employed at a given time* is women hired times average tenure over the window length, with running spells right-censored; *Female workers at a given time (share)* divides it by the workforce. Selection into the follow-up: *Follow-up* column of Table 1. Unweighted firm-level means.