Hidden Discrimination in Frictional Labor Markets

Elisa Macchi* and Claude Raisaro†

Version: November 3, 2026.

Abstract

Organizational frictions shape which attributes firms value and affect hiring discrimination. Across three male-dominated sectors in Uganda, 915 employers evaluate vocational trainee profiles with cross-randomized gender and performance signals. We randomize gender-neutral monitoring support to prevent theft by workers or ensure safety and respectful treatment. Under business-as-usual, female profiles are 10.6 percentage points less likely to be selected. Theft monitoring widens this gap by 56%; safety monitoring reduces it by 84%. Neither detectably changes beliefs about attributes. The effects are consistent with theft monitoring reducing the value of trustworthiness, associated with women, and safety monitoring easing concerns about women's exposure to harassment. Skills signals raise selection but do not detectably narrow the gap. Trainee data do not support employers' beliefs that women are less skilled or men less honest, suggesting that the status-quo gap conceals inaccurate statistical discrimination against men and bias against women. Mitigating theft monitoring frictions amplifies the hiring consequences of bias, illustrating a second-best problem. When identity signals multiple attributes, offsetting advantages can hide discrimination, and measured bias depends on which attributes researchers observe and count as productive.

*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, 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 AER Registry (AEARCTR-0013698).

1 Introduction

Employers value social skills and personality traits alongside technical skills, and commonly associate these attributes with group identity. U.S. employers describe immigrants as hardworking and Black workers as lacking work ethic; economics reference letters describe female candidates as diligent, less often as brilliant.[1] In Ugandan male-dominated manufacturing sectors, our setting, employers see women as less technically skilled but more honest. When identity signals both strengths and weaknesses across technical and nontechnical attributes, which beliefs drive hiring disparities, and what do those disparities reveal about discrimination?

In this paper, we argue that hiring disparities depend not only on employers' beliefs and preferences, but also on which attributes they value given the production environment and its frictions. The production environment determines which identity beliefs affect hiring, so identity-neutral changes in production can alter disparities even when beliefs and preferences remain unchanged. These changes reflect statistical discrimination when beliefs are accurate and demand reflects productivity, and bias when beliefs are inaccurate or demand reflects personal or social preferences. This distinction has two implications for measurement. First, estimates of bias depend on which attributes researchers treat as productive. Second, favorable treatment on one attribute can offset unfavorable treatment on another, so outcome gaps can hide the extent of discrimination.

We study these questions by focusing on gender discrimination in hiring in motor mechanics, welding, and carpentry in Kampala, Uganda. Despite high female labor-force participation—76.5%, above the sub-Saharan African average—employment remains highly segregated by gender.[2] These sectors are heavily male-dominated and among the country's highest-return vocational occupations. Efforts to increase women's entry have largely focused on labor supply (McKenzie, 2017). Their effectiveness also depends on employers' demand for female workers, which remains understudied, particularly in sub-Saharan Africa (Heath et al., 2024).

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

[1] See Moss and Tilly (2001) and Eberhardt et al. (2023), respectively.

[2] 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; Bassier and Gautham, 2025). Whether this reflects women's preferences or barriers to better jobs remains an open question (Jalal, 2026).

trustworthiness as the most important worker attribute, followed by cooperation, which they rate as important as technical skills (Figure 1). Men are perceived as stronger on technical performance, women as more trustworthy and less likely to steal. Employers worry about perceived gaps in women's technical skills, particularly due to physical strength, but even more about harassment between male and female workers. These priorities may reflect productivity returns related to agency and teamwork frictions common in poor countries.[3] However, the same assessments may reflect stereotypes or rationalize underlying bias. We experimentally test whether changing workplace frictions changes the set of worker attributes employers value and, in turn, hiring disparities.

We measure hiring preferences by gender by offering 915 employers access to vocationally trained candidates, including qualified women, through partnerships with local training institutes.[4] In an incentivized ratings design, employers evaluate profiles with randomly assigned gender and technical performance signals, knowing their choices guide referrals to institutes with similar candidates. We randomize employers to a status-quo condition or to offers of monitoring support to prevent either theft or workplace harassment, holding candidate information fixed, so the treatments mitigate workplace frictions without changes in beliefs or preferences over identity. We assess the accuracy of employers' gendered beliefs in a lab-in-the-field experiment with 173 vocational trainees, benchmarking their incentivized predictions against trainees' actual skills and behavior.

Employers report demand for female workers that is largely absent from 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 status-quo condition, 89.7% of employers select at least one female profile, although 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.

We next test whether the production environment affects disparate treatment against women by reducing workplace frictions that make particular 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

[3]See Breza and Kaur (2025) for a review of frictions common in poor-country labor markets, Bloom et al. (2013) and Kotia (2025) on monitoring and delegation, and Hjort (2014) on teamwork frictions.

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

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. We test whether each treatment works through its targeted friction without changing beliefs about workers. We elicit employers' assessments of each candidate's skills and trustworthiness and, after the rating exercise, 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 4.9% under theft monitoring visits and 6.6% under safety monitoring visits, but through different margins. Theft monitoring visits raise male selection by 5.3 percentage points ($p = 0.001$), with no detectable effect on female selection ($p = 0.714$), widening the gap from 10.6 to 16.5 percentage points. Safety monitoring visits raise female selection by 7.6 percentage points ($p < 0.001$), with no detectable effect on male selection ($p = 0.404$), reducing the gap to 1.7 percentage points, statistically indistinguishable from zero. Two further results support the mechanism. First, theft monitoring visits reduce theft worries (-0.61 on a 0–10 scale, $p = 0.004$), while safety monitoring visits reduce safety worries (-0.48, $p = 0.056$). Other worries do not detectably change.[5] 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.

Our second result is that observed gender gaps understate the penalties against women because their perceived trustworthiness advantage offsets other disadvantages. Theft monitoring reduces the value of women's perceived trustworthiness advantage, widening the hiring gender gap by 56%. The effect is strongest among employers who state gender-neutral hiring preferences: their gap widens by 17.1 percentage points ($p < 0.001$),

[5]The safety script explicitly states that visits will not address theft, yet safety monitoring visits do not detectably increase theft worries. This supports an effect through reduced safety concerns rather than increased salience of women's perceived honesty.

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 mitigated, the remaining penalty is offset by women's trust premium. Without that underlying penalty, the trust premium would make women more likely to be hired than men. Technical performance signals do not close the remaining gender gap against women. A top rather than average VTI ranking raises selection by 9.3 percentage points (22%, $p < 0.001$), and a certificate by 15.0 percentage points, yet neither detectably narrows the gender gap against women in any arm. Beliefs about unmeasured technical attributes, such as physical strength, could still explain part of the penalty. However, candidates come from a highly selected pool of vocational trainees, and the ranking captures both theoretical and practical performance, including tasks that require strength. These findings suggest a role for bias in the criteria employers apply or how they weigh the same information. Hiring also responds less to women's performance signals ($p = 0.066$), consistent with stereotypes shaping judgments of productivity (Bertrand and Duflo, 2017).

Having established how employers' assessments affect hiring, we next show that they do not reflect actual skill differences between male and female trainees. We measure technical skills through sector-specific theory and practical tasks, and trustworthiness through a behavioral task with reputational stakes and financial incentives to over-report performance. We compare the outcomes of 173 vocational trainees with employers' incentivized predictions for men and women separately, testing whether their gendered beliefs are accurate (Bohren et al., 2025a). Employers expect women to perform worse on sector-specific theory and practical tasks, but women perform as well on practical tasks and, if anything, better on theory. Employers expect nearly half of men to cheat when unsupervised—six times the observed rate—and expect less cheating among women. In fact, cheating is rare and more common among women, although the gender difference is imprecisely estimated.[6] Thus, the status-quo hiring gap conceals inaccurate statistical discrimination against men that offsets the hiring consequences of bias against women.

[6]The evidence on technical skills is consistent with previous work documenting similar skills among vocationally trained men and women in these sectors (Alfonsi and Ferreira, 2026), while employers' inaccurate beliefs align with evidence that gender stereotypes distort assessments of ability (Bordalo et al., 2019). The misperception about trustworthiness is also consistent with the literature: evidence on actual gender differences in prosocial behavior is mixed (Niederle, 2016), yet women are consistently perceived as more prosocial, often inaccurately (Exley et al., 2025).

How much bias do women face? We show that residual estimates of bias depend on which attributes we observe and treat as productive. We estimate a discrete-choice hiring model with multiple worker attributes, preferences over identity, and equilibrium theft and harassment risks that vary by worker type. Using the experimental variation, we separate a trust premium, a harassment penalty, and a residual hiring penalty. We compare estimates from a model that includes only technical skills with those from models that also account for trustworthiness, harassment risk, or both. Even when all attributes are observed, expanding the analysis beyond technical skills raises the question of which attributes should count as productive. Safety concerns, for example, may reflect paternalism rather than effects on production, as in Buchmann et al. (2023). Estimates of bias therefore also depend on assumptions about these attributes' productive relevance. Across these assumptions, estimates range from 1.7 to 16.6 percentage points.

This paper connects the literature on hiring discrimination to recent work on how employers value multidimensional worker skills (Deming, 2017; Woessmann, 2025; Deming and Silliman, 2025). Experimental studies identify discrimination from outcome gaps between otherwise identical candidates, typically holding the production environment fixed and taking the set of valued skills as given (Goldin and Rouse, 2000; Bertrand and Mullainathan, 2004; Kline et al., 2022). Building on theories of taste-based and statistical discrimination (Becker, 1957; Phelps, 1972; Arrow, 1973), experiments often vary information about workers' skills, such as education or technical certifications, to distinguish beliefs about productivity from preferences over identity. We instead hold candidate information fixed and mitigate workplace frictions, showing that the production environment shapes disparate treatment through the demand for worker attributes. This complements the systemic view of discrimination, in which past evaluations shape the signals and attributes workers bring to subsequent evaluations (Bohren et al., 2019, 2025b). We show that measured disparate treatment is endogenous to the production environment, which determines which attributes are rewarded. Because employers value multiple attributes and infer them from identity, perceived advantages on some dimensions can offset penalties on others or underlying bias. The observed outcome gap can therefore understate penalties even among workers with identical histories.

This paper also contributes to a growing literature on how organizational frictions shape firm decisions, such as pay and technology adoption (Hjort, 2014; Atkin et al., 2017; Kelley et al., 2024; Houeix, 2026). We show that these frictions also shape disparate treatment in hiring: employers favor groups they believe have attributes that help overcome them. Existing work shows how frictions and firms' responses typically disadvantage women

through network-based recruitment into male-dominated jobs (Beaman et al., 2018; Alfonsi and Ferreira, 2026) and barriers to entrepreneurship in higher-return sectors (Ashraf et al., 2024; Chiplunkar and Goldberg, 2024). Our experiment shows that these responses can also favor women, depending on production needs and perceived skills. Easing one friction at a time reveals a second-best problem (Lipsey and Lancaster, 1956): easing monitoring constraints removes women's perceived advantage while leaving other disadvantages intact, amplifying the hiring effects of bias against women. Workplace frictions thus shape the effects of policies that expand women's access to male-dominated sectors, potentially explaining the mixed results of supply-side interventions (Alfonsi et al., 2024; Field and Vyborny, 2022).

2 Setting, Sample, and Motivating Evidence

2.1 Setting

We study motor mechanics, welding, and carpentry firms in Kampala, Uganda. These sectors are part of a substantial source of manufacturing and youth employment in the country.[7] 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 generally provide training rather than job placement (Alfonsi et al., 2020), 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 (Alfonsi et al., 2020).

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 iden-

[7] 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).

tity and social preferences can affect these interactions and their productivity (Hjort, 2014; Ashraf et al., 2024; Breza et al., 2018).

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 (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.

These constraints help explain firms' reliance on familiar workers. Networks reduce uncertainty about applicants' skills (Carranza et al., 2022; Bassi and Nansamba, 2022; Chandrasekhar et al., 2020) and help address moral hazard when supervision is limited (Heath, 2018). Hiring outside these networks requires costly experimentation (Groh et al., 2016), with implications for discrimination (Macchiavello et al., 2026). Network-based recruitment can therefore help firms manage workplace frictions while restricting women's access to male-dominated jobs (Beaman et al., 2018; Alfonsi and Ferreira, 2026).

2.2 Sample

Firms and Employers. Our main 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.6).[8]

Our sample of 915 firms is large relative to the number operating in these sectors in Kampala, equivalent to 30.0% of the firms recorded in the 2020 business census and 27.8%

[8]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, spoke English or Luganda, and expressed interest in recruiting a new worker on probation. We interviewed employers at consenting firms.

in 2010.[9] 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.[10] We interview one respondent per firm, the person with hiring authority: 85.7% are business owners and 12% are managers. Employers are almost all men (96.4%), 37 years old on average, with 15 years of experience in the sector. Most have at least some secondary education (63%), and 31% went through vocational training.

Women are rare both in the workforce and in the applicant pool. They account for only 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 vocational training institutes (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).[11] Trainees are 20.8 years old on average, and their programs typically last 24 months.

[9] 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.

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

[11] 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 ICT who do not name a study sector as a desired occupation.

They receive formal training in sector-specific skills that firms also develop through on-the-job training.

Vocational training institutes offer a pool of skilled candidates with greater female representation than firms' usual applicant pools, potentially easing constraints on hiring women. Yet only 18.3% 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.

2.3 Demand for Worker Attributes

We ask employers to rate, on a 0–10 scale, the importance of several worker attributes when hiring.[12] 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 above. 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).

[12] 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.

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.[13] 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 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).[14] 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 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. This suggests 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 introduce

[13] 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 main sample to the business-as-usual arm.

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

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 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.[15] 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.[16]

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 training background. 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 tied to class rank within the actual VTI cohort: three stars indicate median

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

[16] 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."

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 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.[17] 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 three dimensions. All questions are administered in Luganda; the English translations are:

[17] 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.6.

- *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).
- *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, holding candidate information fixed. 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.

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 know that visits will occur, but their timing is unannounced, giving them an incentive to avoid misconduct between visits.

Assignment is stratified by enumerator, survey wave, and sector. Employers are told that support is allocated by lottery and does not reflect characteristics of the firm or workers.

Workers learn about the visits only after accepting a job, preventing selection into jobs based on treatment assignment.

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, safety and harassment, or effort (a placebo option).[18] 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:

[18] 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.

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.

Mechanism Tests and Identification Concerns. 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. First, 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. Second, 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. Third, belief updating: either treatment may change employers' assessments of candidate quality or trustworthiness.

We use measures collected immediately after the rating task, and before requests for trainee contacts, to detect spillovers and priming. Employers rate their worries about theft and dishonesty, worker safety, and two placebo worries: effort and turnover. In particular, we examine whether safety monitoring increases theft worries. To test for belief updating, we examine whether the offers change employers' ratings of candidate quality and trustworthiness, our secondary outcomes.

Although experimenter demand common to both treatment arms would not explain their predicted opposite effects, we also measure it directly. At the end of the survey, we ask employers to report whether they thought the research team preferred them to hire men or women, or had no preference.

3.3 Conceptual Framework

The two treatments vary employers' monitoring capacity while holding candidate information fixed. We develop a simple framework to illustrate how monitoring support should affect the hiring gender gap and what observed gaps reveal about discrimination. Appendix A.2 presents the full model.

Motivated by the survey evidence, we model employers as evaluating candidates along three dimensions: technical performance, trustworthiness, and cooperation.[19] We model concerns about harassment risk, which depends on coworkers' behavior and the firm's gender composition, through cooperation. We allow these assessments to matter for productive or nonproductive reasons.

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 has high technical performance. The candidate's expected value is

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

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 probability that a theft is detected, μ, and the probability that harassment is detected, η. The terms $\tilde{s}_g(\mu)$ and $\tilde{h}_g(\eta)$ are the employer's expected theft and harassment risks. The harassment risk comes from male coworkers and scales with the male share of the workforce, m. The costs of theft and harassment to the employer, θ and τ, may reflect production losses or nonproductive concerns, such as paternalism. The term b captures a residual preference against or for hiring women.

A favorable performance signal raises $\hat{p}_g^S(q)$ and therefore the candidate's value, but need not fully reveal productivity. Employers may still infer productive attributes from gender, leaving a gap in expected performance among candidates with the same signal. Employers also believe that women have a higher moral cost of stealing and face greater harassment risk from male coworkers. Monitoring deters theft and harassment by increasing the chance of detection. Greater monitoring capacity reduces, but need not eliminate, these risks without changing beliefs about underlying worker characteristics or preferences over gender. Beyond these beliefs, the framework rests on two assumptions. We model the performance signal as leaving gender differences in theft and harassment risk unchanged, and each type of monitoring as reducing only the risk it targets. We discuss the evidence on both assumptions in Section 4.

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

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 $T(\mu) = \theta[\tilde{s}_m(\mu) - \tilde{s}_f(\mu)]$ favors women and falls as theft monitoring capacity μ increases. The harassment penalty $H(\eta) = \tau[\tilde{h}_f(\eta) - \tilde{h}_m(\eta)]$ works against women. It falls as safety monitoring capacity η increases.

The framework yields two implications. First, changes in the production environment affect disparate treatment even when beliefs and preferences over gender remain unchanged. Greater capacity to monitor theft reduces women's trust premium and widens the gap in expected hiring value. Greater capacity to monitor harassment reduces the harassment penalty and narrows the gap. These predictions hold whether employers' concerns reflect productive or nonproductive motives.

Second, women's trust premium offsets penalties arising from perceived technical performance differences, harassment risk, or residual preferences against women. A small or zero gender gap can therefore coexist with substantial penalties against women. Greater capacity to monitor theft reduces this offsetting advantage and reveals penalties concealed in the status quo. Interpreting these components as bias requires assessing the accuracy of employers' beliefs and the productive relevance of the attributes they value.

The opposite predictions distinguish our mechanism from channels common to both treatments, including greater scrutiny, experimenter demand, and a general reduction in the cost of experimenting with new workers. Each would predict effects on the hiring gender gap in the same direction across the two arms. Scrutiny could discourage hiring women if employers anticipate more problems when doing so. Experimenter demand could encourage hiring women. Both types of visits could make employers more willing to try unfamiliar workers, i.e., women; this experimentation channel would predict increased relative demand for women under both treatments.

The empirical strategy below tests whether theft and safety monitoring move the hiring gender gap in opposite directions. Section 5.2 combines the three experimental arms in a discrete-choice model to estimate the trust premium, the harassment penalty, and, under complete mitigation, the penalty that remains with both frictions removed.

3.4 Empirical Strategy

Analysis Sample. Our main analysis includes profile evaluations from 886 of the 915 interviewed employers after applying two preregistered screens. The first is a comprehension test for employers in the treatment arms. To pass, they must select only the intended purpose of their assigned visits: theft monitoring under MT or safety monitoring under MS. We exclude those who select an incorrect or additional purpose, as well as six treated employers interviewed before the comprehension question was introduced. The second screen is an attention check applied to all employers at the level of eight-profile blocks. We exclude any block in which the employer selects either all eight profiles or none. Employers remain in the sample if they pass the comprehension test, where applicable, and retain at least one block with variation in hiring choices. Only retained blocks enter the analysis. Results are robust to including the excluded employers and profile blocks (Appendix Table A3).

Table 2 shows covariate balance across the three experimental arms, including employers' characteristics and firm-level attributes. We only find mild imbalances in three baseline variables: whether the employer has formal vocational training, the number of customers served daily, and the employer's daily hours worked. Results are robust to including these controls.

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:

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

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–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 hiring of women relative to men. A positive interaction narrows the hiring gender gap; a negative interaction widens it.

We include base-profile-by-performance-signal fixed effects, δ_{iq}, so 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 term u_{ij} is the error term. 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 ten-worker scale. Using a standardized workforce of ten lets employers express gender-mix preferences that would be difficult to capture with only a few workers.[20]

We first ask 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.

Second, we 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

[20]Employers answer on a scale ranging from ten men out of ten workers to zero men out of ten workers. We recode responses as the preferred female share. We frame the scale in terms of men to make its focus on women less salient.

female share, measured continuously or split at the median. We also report preregistered heterogeneity by employer gender and recruitment from VTIs.

Technical Performance Signals. We examine whether two signals, VTI class rank and DIT certification, narrow the hiring gender gap. We extend equation (3) to interact each signal with candidate gender and monitoring assignment, including their three-way interaction. For class rank, we replace base-profile-by-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 two out of ten 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 vacancies, hiring women in the same proportion as in its 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.[21] These results show that employers are willing to hire trained women when given access to qualified candidates.

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

We next ask 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 1.34 standard deviations of the trainee exam score distribution.[22]

The average hiring gender gap masks substantial heterogeneity that tracks employers' stated preferences. The gender gap under business-as-usual 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 4.83-percentage-point reduction in the gender gap ($p < 0.001$; Appendix Table A4, column 1). In the top quintile, where employers state an ideal workforce between 40% and 60% women, the hiring gender gap is not statistically significant.[23]

We also confirm that employers' assessments of otherwise identical profiles reflect the gendered beliefs documented in the motivating survey. Under business-as-usual, female profiles receive lower ratings for technical performance and effort (both Holm-adjusted $p < 0.001$), but higher trustworthiness ratings (Holm-adjusted $p = 0.020$; Appendix Table A5). 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.

[22]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 statistically significant (Table 3, column 2).

[23]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).

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 4.9% under theft monitoring visits ($p = 0.052$) 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 16.5 percentage points, a 56% increase relative to business-as-usual ($p = 0.013$). This increase comes from increased hiring of men: male selection rises by 5.3 percentage points ($p = 0.001$), while female selection does not detectably change ($p = 0.714$).

Safety monitoring reduces the hiring gender gap by 84%, from 10.6 to 1.7 percentage points ($p < 0.001$), leaving it statistically indistinguishable from zero. This reduction comes from increased hiring of women: female selection rises by 7.6 percentage points ($p < 0.001$), while male selection does not detectably change ($p = 0.404$).

The opposite effects are consistent with our prediction that monitoring support changes the value employers place on gendered worker attributes: theft monitoring may reduce women's trust premium, while safety monitoring may ease 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 0.61 points on a 0–10 scale (Holm-adjusted $p = 0.004$), while safety monitoring reduces worries about worker safety by 0.48 points (Holm-adjusted $p = 0.056$; Table 4). Worries about the other targeted friction and the placebo concerns of effort and turnover do not detectably change. Second, the safety monitoring treatment does not detectably increase theft worries, despite mentioning that the visits cannot discourage theft, weighing against theft-risk priming as an explanation for the effect on female hiring.

Employers' ratings of candidate attributes allow us to test for belief updating. Neither treatment detectably changes the gender gaps in technical performance, effort, trust-

worthiness, or expected earnings (Appendix Table A5). 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 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 7.1% 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.

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 A6). Results are similar using probit (Table 3, column 2) or excluding employers who prefer an all-male workforce (column 3). They also persist when reinstating observations excluded by the attention and comprehension screens or including soft-launch firms (Appendix Table A3).

Hidden Hiring Penalties. Women's perceived trustworthiness offsets other hiring penalties when monitoring theft and other worker misconduct is difficult. 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 17.1 percentage points ($p < 0.001$), from a level statistically indistinguishable from zero to 14.9 percentage points. 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 4).

More generally, the relationship between stated preferences and the hiring gender gap is less than half as strong under theft monitoring as under business-as-usual, although

the difference is imprecisely estimated.[24] 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, columns 1–2).[25] 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.

4.3 Effects of Technical Performance Signals

We next examine the role of technical performance in the remaining hiring penalty. Following the standard approach in the discrimination literature, we test whether stronger or additional performance signals narrow the hiring gender gap. We consider two signals: VTI class rank (median or top) and whether the profile mentions a DIT certification.

Employers respond positively to both signals, but neither detectably narrows the hiring gender gap. Under business-as-usual, a top VTI ranking increases selection by 9.3 percentage points (22.0%; $p < 0.001$). If anything, women benefit less: their hiring response is 4.2 percentage points smaller ($p = 0.066$; Table 5, column 2). Top-ranked candidates also receive higher work-quality ratings, but the gender gap in these ratings does not detectably change. Reporting DIT certification increases selection by 15.0 percentage points ($p < 0.001$) and raises quality ratings by 0.46 points on the 0–10 scale ($p < 0.001$). Again, certification does not detectably narrow the gender gap in hiring ($p = 0.463$) or quality ratings ($p = 0.283$).[26] The hiring effects of either signal do not detectably vary across monitoring arms (Table 5, columns 2 and 6).

[24] The coefficient falls from 0.483 to 0.216, a change of −0.267 (standard error 0.171; Appendix Table A4, column 1).

[25] Columns 3–4 report the other preregistered heterogeneities, by employer gender and recruitment from VTIs.

[26] Results also hold using an index combining secondary education, DIT certification, training duration, and GPA. A one-standard-deviation increase raises selection by 10.6 percentage points ($p < 0.001$), but does not detectably narrow the hiring gender gap (Table 5, columns 3–4).

Under statistical discrimination, more informative signals reduce employers' reliance on group beliefs and can narrow the hiring gender gap (Phelps, 1972; Aigner and Cain, 1977). The persistent gap in our setting could reflect bias or more complex forms of statistical discrimination. Employers may be risk-averse and regard the signals as less informative for women (Aigner and Cain, 1977), or infer differences in attributes the signals do not fully reveal, such as physical strength or effort. However, candidates come from a highly selected pool of vocational trainees, and the VTI ranking provides a broad signal of job performance, covering theoretical knowledge, practical ability, effort, and interest in learning. As noted above, no employer mentions strength as an attribute that matters when hiring, yet it is a common concern about hiring women. These findings could suggest a role for bias in the criteria employers apply or how they weigh the same information (Bertrand and Duflo, 2017). In the next section, we directly assess whether employers' gendered beliefs match differences in technical skills and trustworthiness among the trainees they could hire.

5 Hidden Gender Discrimination

We next examine how employers' gendered beliefs contribute to the extent of discrimination. We first compare employers' incentivized predictions with measured gender differences in technical skills and trustworthiness among vocational trainees. If beliefs are accurate and concern productive attributes, the hiring responses may reflect efficient screening. If beliefs are inaccurate or the attributes are not productive, the responses may instead indicate bias.

In the second part of this section, we return to the distinction between productive and nonproductive attributes. We use a discrete-choice model of hiring to estimate bias under different assumptions about which attributes researchers observe and treat as productive.

5.1 Accuracy of Employers' Beliefs

We compare employers' gendered beliefs with the skills and behavior of 173 vocational trainees across seven partner institutes (Table 1, panel C). These trainees represent the relevant hiring pool: trained men and women seeking work in the study sectors. Technical skills are measured through sector-specific theory exams and the ability to perform

practical tasks. These skills are assessed at the VTIs, in the same training context as the theoretical and practical assessments underlying the performance rankings shown in the profiles. We measure trustworthiness and effort through a task designed to offer financial incentives to over-report output. Both exercises carry reputational stakes because trainees know that results may be used for job referrals.

We describe the trainee pool and each measure to employers, then ask them to predict outcomes separately for men and women in their own sector. Predictions refer to the same measures collected from trainees, allowing a direct comparison between beliefs and outcomes. Employers receive a bonus for accuracy.[27]

Technical Skills. We consider two measures of technical skills. The first is a theory exam, administered at the VTIs, with questions specific to the trainees' specialization. VTI instructors designed the questions; Figure A7 shows an example. To measure technical skills on practical tasks, we elicit self-reported ability to perform sector-specific technical tasks, following the approach developed by Alfonsi et al. (2020).[28] The results of both trainee outcomes and employers' beliefs by gender are summarized in Table 6. Employers perceive female workers as less skilled, both on practical and theoretical tasks. On average, employers expect women to be 12.0 and 15.3 percentage points less likely than men to answer correctly on theory and practical tasks, respectively, or 20.7% and 25.2% below the rate they expect for men (both $p < 0.001$).

These beliefs are inaccurate: we find no evidence that female trainees perform worse than male trainees. In self-assessments, female and male trainees are equally likely to report being able to perform a standard sector task ($p = 0.124$). On the theory exam, women perform at least as well as men (6.2 percentage points; $p = 0.198$).[29]

These results are consistent with previous evidence of similar technical skills among vocationally trained men and women in this setting (Alfonsi and Ferreira, 2026). They suggest

[27] 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).

[28] For trainees in motor mechanics: "Can you perform an oil change on a vehicle?"; for welders: "Can you perform a clean and strong weld on a stainless steel surface?"; for carpenters: "Can you perform a wood sanding using a coated abrasive?".

[29] Because of time constraints, we elicited employers' beliefs on one randomly selected theory question and matched these to actual trainee outcomes. On this question, female trainees are 21.4 percentage points more likely to answer correctly ($p = 0.022$); see Table 6.

that employers' perceptions of women as less technically skilled reflect inaccurate gender beliefs rather than actual skill differences in the trainee pool.

Trustworthiness. We design a record-checking task to measure effort and dishonesty. Trainees check VTI registration forms for errors and can earn a bonus by reporting more errors than are actually present. Each trainee completes two five-minute rounds, one supervised and one unsupervised. In each round, they check twelve VTI registration forms for mismatched identifiers. Five forms contain a planted error. Trainees mark the forms they check and record how many errors they find on a tracking sheet. The sheet is anonymous in the unsupervised round, while it carries their name and ID in the supervised round. After each round, we directly ask whether they found six or more errors. Answering yes earns an additional UGX 1,000, creating an incentive to over-report. Trainees receive UGX 2,000 regardless of performance in the unsupervised round. In the supervised round, payment depends on correctly checked forms, up to UGX 2,000. We randomize round order and select one round at random for payment.

Our task draws on the production-setting task of Caria and Falco (2024), which measures shirking. We instead measure cheating through over-reporting output for pay, similar to the "claim your output" design of Mazar et al. (2008). Our design separates effort from over-reporting: we observe how many forms trainees check, how many errors they record, and whether they claim the bonus. We can therefore compare over-reporting conditional on effort, measured by the number of forms checked. By varying supervision, we also test whether employers expect monitoring to reduce dishonesty more among men, consistent with their hiring responses to theft monitoring.

Importantly, the task carries stakes beyond the financial bonus. Trainees work with VTI records and know that survey information may be used for job referrals, so dishonest reporting could put their reputation and employment prospects at risk. These reputational stakes are relevant in a labor market where recruitment through networks makes reputation important.

The trainee data show that effort is high and bonus claims are uncommon. Trainees check about 97% of forms, and 23 of 173 trainees (13.3%) claim the bonus in at least one round. Separately, 18 of the 352 tracking sheets with recorded error counts report more than five errors (Figure A8). We use responses to the incentivized bonus question as our main measure of over-reporting. Supervision reduces over-reporting without detectably changing effort. Bonus claims fall from 10.4% to 4.1%, a 61% decline ($p = 0.016$). Effort

remains high, consistent with a ceiling effect because most trainees already check nearly all forms. The decline in claims therefore appears to reflect less over-reporting rather than less work.

We find no evidence that women over-report less than men. In the unsupervised round, 15.9% of women claim the bonus, compared with 7.3% of men—more than twice the rate. However, the estimate is imprecise because bonus claims are rare, and the difference is not statistically significant (Fisher's exact test, $p = 0.12$). Pooling both rounds and controlling for effort, supervision, and VTI fixed effects, women are 5.0 percentage points more likely to claim a bonus, although the effect is still imprecisely estimated ($p = 0.249$; Table 6).[30] Results are similar when we measure over-reporting as recording more than five errors on the tracking sheets: women are not detectably less likely to over-report than men. The response to supervision also does not detectably differ by gender ($p = 0.68$), though the small number of claims limits precision.

Employers substantially overestimate cheating, especially among men, and expect women to cheat less. They predict that 46.3% of unsupervised male trainees will claim the bonus, compared with the observed rate of 7.3%. They expect women to be 11.3 percentage points less likely to claim it ($p < 0.001$). Under supervision, expected male cheating falls to 25.4%, but women are still expected to cheat 5.7 percentage points less ($p < 0.001$; Table 6). Employers also expect supervision to reduce cheating more among men, consistent with their hiring responses to theft monitoring. The predicted reduction is 20.9 percentage points for men and 15.2 for women (both $p < 0.001$), a difference of 5.6 percentage points ($p < 0.001$).[31]

These results suggest that employers hold inaccurate gender beliefs about trainee trustworthiness. The other measures of prosociality we collect, including generalized trust (Falk et al., 2018) and behavior in dictator and trust games (Berg et al., 1995), show no consistent evidence that female trainees are more prosocial than male trainees, despite pronounced gender differences in employers' predictions (Appendix Table A7).[32] These findings align with the mixed evidence on actual gender differences in prosocial behavior in the literature (Niederle, 2016), alongside widespread beliefs that women are more prosocial than men (Exley et al., 2025).

[30] 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.

[31] Employers also expect supervision to increase the share of forms checked by male trainees by 23.3 percentage points, despite little observed change in effort.

[32] Appendix A.5 describes the measures.

5.2 How Much Bias Do Women Face?

We next ask how much of the hiring penalty against women can be attributed to bias. Researchers commonly estimate bias as the residual hiring gap after controlling for observed qualifications or providing additional information about workers. A familiar concern is that unmeasured productive attributes explain part of this residual, leading researchers to overstate bias (Lang and Spitzer, 2020).

We use the two monitoring treatments to decompose the hiring gender gap into a trust premium, a harassment penalty, and a residual. Following the conceptual framework (Section 3.3), we estimate these components with a probit model (Appendix A.3). The decomposition assumes that each treatment affects only its targeted friction, leaving beliefs and preferences unchanged, and that the components enter employers' valuations additively. The mechanism tests support the first assumption: each treatment reduces worries about its targeted risk without detectably changing other worries or assessments of candidates (Table 4; Appendix Table A5). To recover the full components, we also assume that each treatment completely removes its targeted friction.[33]

We estimate a trust premium of 5.9 percentage points, a harassment penalty of 9.0 percentage points, and a residual penalty against women of 7.6 percentage points. Treating trustworthiness and harassment risk as productive attributes, we interpret the residual as bias under the assumption that no other productive differences explain the gap. Two findings support this assumption: the attributes employers mention fall within the three dimensions covered by the experiment, and the trainee evidence does not reveal a female disadvantage in measured technical skills or trustworthiness.

We compare this estimate with those from models that account for fewer attributes. If we assume employers care only about technical performance, broadly defined, estimated bias is 10.6 percentage points. In a two-dimensional model that also accounts for trustworthiness, it rises to 16.6 percentage points. If the second dimension is instead harassment risk, it falls to 1.7 percentage points. The three-dimensional model lies between these estimates, at 7.6 percentage points (Table 7).

This exercise yields two conclusions. First, residual estimates of bias depend on which attributes researchers observe. Second, omitting attributes employers value can understate bias as well as overstate it relative to the full model. Holding the treatment of harassment

[33] If monitoring only partly mitigates the targeted frictions, the model implies larger underlying trust and harassment components than those recovered under complete removal.

risk constant, omitting trustworthiness lowers estimated bias by 5.9 percentage points because women's trust premium offsets other hiring penalties.

Even when researchers observe all the attributes employers value, measuring bias requires assumptions about which count as productive. Harassment risk illustrates this distinction. Violence and harassment can impose substantial economic costs on workers and firms (Adams and Nix, 2026), but exposure to these risks reflects coworkers' conduct and workplace conditions rather than the candidate's own decisions. Employers may also care about safety out of a desire to protect women from perceived harm or because of personal reputational concerns. Such concerns can impose an additional hiring penalty on women (Buchmann et al., 2023). Treating employers' concerns about harassment risk as accurate beliefs about productivity removes the associated hiring penalty from the residual, potentially excluding part of the discrimination we seek to measure.

Our treatments therefore do not uncover a single "true" level of bias independently of assumptions about which attributes count as productive. Our results illustrate how the distinction between statistical discrimination and bias becomes less clear when employers' judgments of productivity may themselves reflect preferences or stereotypes. In our setting, residual estimates range from 1.7 to 16.6 percentage points, depending on which attributes we account for and treat as productive (Table 7).

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, 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.[34] 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, giving a sense of the scale relative to the trainee pool, although 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.

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 original sample. These firms are similar to the full sample (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 A8, 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 A8, panel B).

Stated demand for women also increased: the average preferred female share rose 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

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

3.2 percentage points among the other firms ($p = 0.011$). Employers who hired women therefore did not update negatively about female workers.

7 Conclusions

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, given the same beliefs and preferences over identity.

The results also show why observed gender gaps can understate discrimination. 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 status-quo 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.

Disparate treatment can itself be a response to the production environment. Monitoring frictions make women's perceived trustworthiness valuable, increasing the cost to employers of acting on bias against them. Easing these frictions reduces that cost and can amplify the hiring consequences of bias, consistent with Becker's insight that discrimination responds to its economic cost (Becker, 1957). Addressing one friction can therefore widen hiring disparities when other sources of disadvantage remain.

When employers value attributes beyond technical skills, measuring bias requires accounting for offsetting advantages and penalties and making assumptions about which attributes count as productive. Residual estimates depend on which attributes researchers observe: omitting a perceived advantage can understate bias, while omitting a perceived disadvantage can overstate it. Even when all attributes are observed, their productive

relevance may be ambiguous. Harassment concerns, for example, may reflect production costs, paternalism, or reputational motives.

These findings have two 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.

Our experiment measures hiring preferences and provides a controlled test of how workplace frictions affect demand for workers by gender. At the two-year follow-up, firms observed at both dates employ more women than at baseline, consistent with the demand for female workers revealed in the experiment. An important next step is to test whether changes in workplace frictions translate into sustained gains in minority employment, 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," August 2026. Working paper.

_ , Mary Namubiru, and Sara Spaziani, "Gender gaps: back and here to stay? Evidence from skilled Ugandan workers during COVID-19," *Review of Economics of the Household*, 2024, *22* (3), 999–1046.

_ , 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, Edward Glaeser, and Kim Sarnoff, "Female Entrepreneurship and Trust in the Market," 2024. Revise and Resubmit at *Journal of Political Economy*.

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.

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.

Bassier, Ihsaan and Leila Gautham, "The Firm-Pay Gender Gap and Formal Sector Churn over the Life Cycle," *Journal of Development Economics*, 2025, *176*, 103498.

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

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

Berg, Joyce, John Dickhaut, and Kevin McCabe, "Trust, reciprocity, and social history," *Games and economic behavior*, 1995, *10* (1), 122–142.

Bertrand, Marianne and Esther Duflo, "Field experiments on discrimination," *Handbook of economic field experiments*, 2017, *1*, 309–393.

__ **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.

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

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," Technical Report, National Bureau of Economic Research 2020.

Chiplunkar, Gaurav and Pinelopi Koujianou Goldberg, "Aggregate Implications of Barriers to Female Entrepreneurship," *Econometrica*, 2024, *92* (6), 1801–1835.

de Quidt, Jonathan, 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.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde, "Global evidence on economic preferences," *The quarterly journal of economics*, 2018, *133* (4), 1645–1692.

Field, Erica and Kate Vyborny, "Women's mobility and labor supply: experimental evidence from Pakistan," *Asian Development Bank Economics Working Paper Series*, 2022, (655).

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, pp. 213–227.

Goldin, Claudia and Cecilia Rouse, "Orchestrating impartiality: The impact of "blind" auditions on female musicians," *American economic review*, 2000, *90* (4), 715–741.

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, "Do Wage Subsidies Provide a Stepping-Stone to Employment for Recent College Graduates? Evidence from a Randomized Experiment in Jordan," *Review of Economics and Statistics*, 2016, *98* (3), 488–502.

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.

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.

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," Technical Report, Working Paper 2026.

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.

McKenzie, David, "How effective are active labor market policies in developing countries? a critical review of recent evidence," *The World Bank Research Observer*, 2017, *32* (2), 127–154.

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.

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

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.

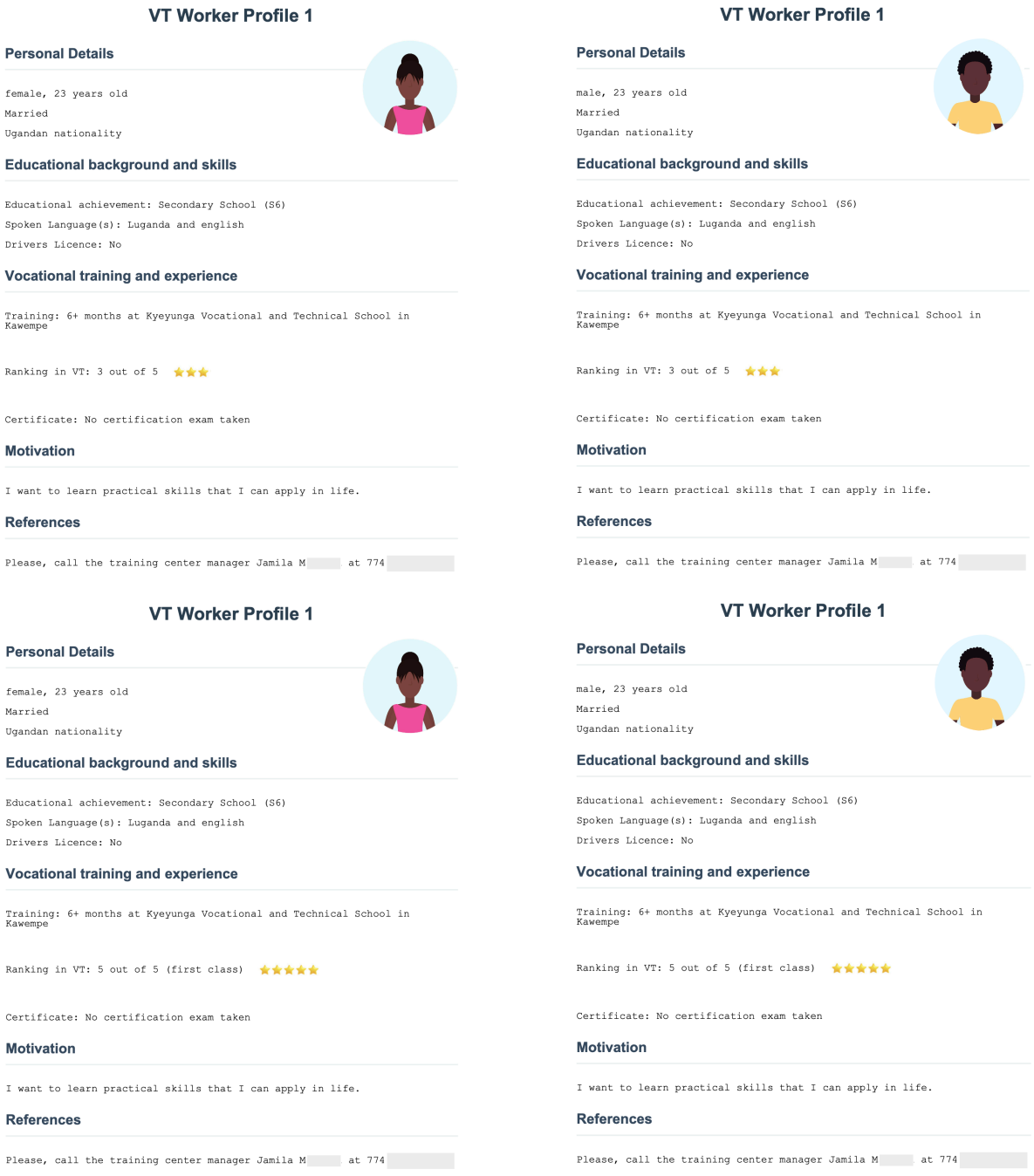
Figures

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



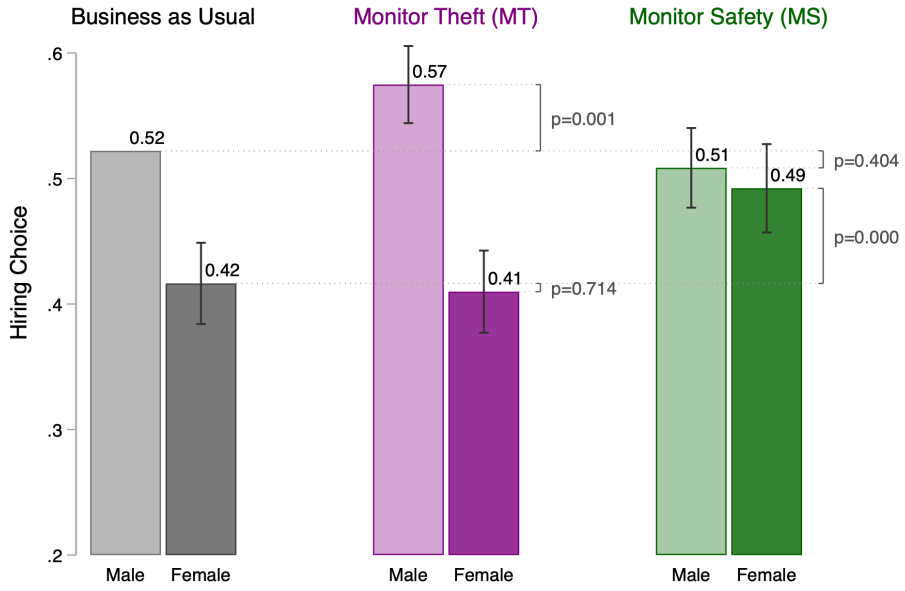
Notes: Panels A–D use the employer sample (N=915). Panel A reports the share of employers rating each trait 10 on a 0–10 importance scale; it excludes 73 employers who gave the same rating to all six traits (N=842). Behavior / No Stealing combines the honesty item and the handling-money item. Panels B–D use closed-choice questions with the same categories. Panel E uses employers with coded responses to the open-ended worry question (N=589) and classifies worries about hiring women into the same categories. Darker portions in panel E show women-specific sub-worries: strength, leaving, and harassment/distraction. Panel E omits responses categorized as "Nothing" and "Other" (10.70% and 1.53%). Open-ended answers in panel E are classified using GPT-5 mini; Appendix A.1 describes the procedure and the benchmark with human classification.

Figure 2: CV Examples



Notes: Examples of one base profile in its four versions, varying gender and performance. We create 24 base profiles, each with four versions, yielding 96 profiles. Each employer evaluates all 24 base profiles, each shown in one randomly assigned version.

Figure 3: Hiring Choice by Gender and Monitoring Arm



Notes: Bars report the likelihood that the employer hires the candidate on probation for male and female candidates across the three monitoring arms: Business-as-usual (Control), Monitoring-Theft, and Monitoring-Safety. Levels are estimated from a fully interacted specification of candidate gender and monitoring condition as in equation (3), with 95% confidence intervals. Brackets and p-values report within-gender, across-arm comparisons testing whether each monitoring treatment shifts hiring relative to the Control, separately for male and female candidates. Monitoring-Theft significantly increases hiring of male candidates ($p = 0.001$) with no statistically significant effect on female candidates ($p = 0.714$), while Monitoring-Safety significantly increases hiring of female candidates ($p < 0.001$) with no statistically significant effect on male candidates ($p = 0.404$). $N = 20{,}170$ evaluations from 886 employers.

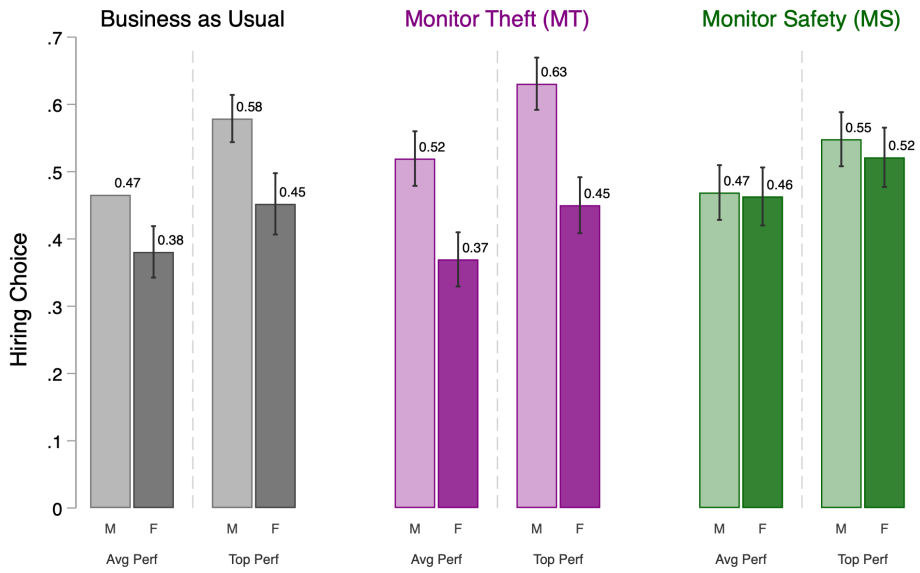
Figure 4: Hiring Gender Gap by Monitoring Arm and Preferences for Workforce Gender Mix



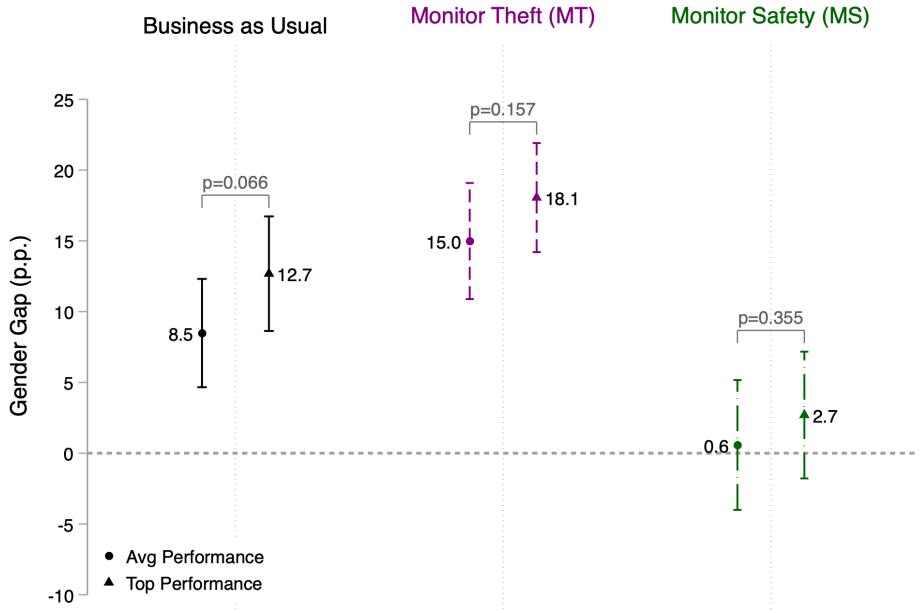
Notes: Data from the experiment. Point estimates and 95% confidence intervals for the gender gap against women are estimated using the regression in equation (3) and reported by treatment arm. The figure presents results across three levels of employers' stated preferences for workforce gender mix. Circles denote estimates for the full sample (N = 20,170). Squares denote estimates for employers with any stated preference for workforce gender mix (N = 17,296), and triangles for employers in the top quintile of stated preferences (i.e., 40% or more female workers; N = 2,957). Each set of estimates comes from a separate regression on the corresponding sample. Estimates from a fully saturated model are quantitatively equivalent.

Figure 5: Hiring Outcomes by Gender, Trainee Performance, and Treatment Arm

(A) Hiring rates



(B) Hiring gender gap



Notes: Both panels come from a single fully interacted model of hiring on candidate gender, technical performance signal, and visit arm, with profile and strata fixed effects. Panel A reports regression-adjusted hiring rates by candidate gender (M, F) and performance signal (Avg, Top) in each treatment arm; the reference cell (male, average performance, business-as-usual) is shown without a confidence interval. Panel B reports the implied hiring gender gap in percentage points, separately for average- and top-performance profiles, where positive values indicate a gap against women. Bars are 95% confidence intervals. Standard errors are clustered at the employer level.

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	—
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	—
Higher 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: The table reports sample means, overall and by sector; monetary values are in millions of UGX. Panel A presents firm characteristics and panel B employer characteristics. Employees include managers and trainees but exclude owners. *Any stealing*, *Monitoring constrained*, *Higher education*, and *VTI trained* are binary indicators for worker theft experienced, desiring more monitoring, secondary education, and VTI training. The *UBOS 2020* column reports the 2020 Census of Business Establishments (Uganda Bureau of Statistics, 2020), restricted to Kampala and the three study sectors, for the rows the census measures. Census employees are total employment minus one working owner. *Annual revenue > UGX 10m* is the top band in which the census records turnover. Panel C reports the 173 surveyed trainees in the study sectors' labor pools, assigned by training sector; the mechanics pool also includes electrician trainees, who are typically hired by mechanics firms, and trainees in related programs are assigned by their stated desired occupation. *Theory exam* is the share correct on a sector-specific exam, *Practical skill* is self-assessed, and *Expected income* is expected monthly earnings one year after training. The *Follow-up* column reports baseline means for the firms reached and consenting 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)	8.67	(8.71)	8.27	(9.44)	0.112	0.535	912
Female workers (N)	0.19	(0.55)	0.21	(0.56)	0.17	(0.49)	0.489	0.613	909
Trainees (N)	2.49	(3.47)	2.56	(3.24)	2.40	(3.38)	0.662	0.902	910
Family workers (share)	0.15	(0.25)	0.15	(0.23)	0.13	(0.22)	0.605	0.197	910
Age (years)	10.06	(7.86)	10.66	(8.16)	9.66	(7.11)	0.326	0.515	915
Capital (UGX m)	6.19	(14.29)	5.78	(11.21)	3.93	(4.30)	0.962	0.060	599
Revenue (UGX m)	3.10	(2.99)	3.18	(2.87)	3.40	(2.91)	0.775	0.466	548
Profits (UGX m)	1.16	(1.18)	1.27	(1.21)	1.34	(1.35)	0.381	0.233	545
Customers (N/day)	3.51	(3.51)	3.74	(3.39)	3.39	(3.02)	0.206	0.473	909
Any stealing	0.84	(0.37)	0.85	(0.36)	0.81	(0.39)	1.000	0.220	915
Monitor constrained	0.90	(0.30)	0.92	(0.27)	0.90	(0.29)	0.454	0.780	908
Panel B: Employers									
Male	0.97	(0.18)	0.95	(0.23)	0.98	(0.15)	0.233	0.464	915
Age (years)	37.31	(9.85)	36.45	(9.26)	37.68	(10.57)	0.234	0.599	915
Experience (years)	15.24	(9.00)	15.37	(9.09)	15.15	(9.27)	0.839	0.877	913
Higher education	0.62	(0.49)	0.62	(0.49)	0.66	(0.47)	1.000	0.391	915
VTI trained	0.26	(0.44)	0.33	(0.47)	0.33	(0.47)	0.035	0.046	915
Hours worked (/day)	10.53	(1.53)	10.43	(1.70)	10.30	(1.62)	0.600	0.090	914
Hours monitor (/day)	2.08	(1.59)	2.01	(1.30)	2.10	(1.51)	0.523	0.824	911

Notes: The table presents baseline balance for the full sample. Columns (1), (2), and (3) report means by treatment arm—Business-as-usual (C), Monitoring-Theft (MT), and Monitoring-Safety (MS)—with standard deviations shown in parentheses in adjacent columns. Columns (1)–(2) and (1)–(3) report randomization-inference p-values, from 10,000 permutations of treatment assignment within strata, on regressions of each outcome on a treatment indicator with strata fixed effects and robust standard errors. The final column reports the sample size used in each balance test. Some observations are missing due to respondent 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]	[0.055]	[0.045]	[0.028]	[0.027]	[0.034]
Monitor Theft	0.053***	0.053***	0.061***	-0.014	0.053	0.051*	0.050	0.106***
	[0.016]	[0.016]	[0.017]	[0.045]	[0.045]	[0.027]	[0.030]	[0.040]
Monitor Safety	-0.014	-0.014	-0.018	0.007	-0.044	-0.017	-0.031	-0.002
	[0.016]	[0.016]	[0.017]	[0.047]	[0.041]	[0.030]	[0.029]	[0.046]
Female x Monitor Theft	-0.059**	-0.059**	-0.069***	-0.001	-0.043	-0.054	-0.030	-0.171***
	[0.024]	[0.024]	[0.024]	[0.075]	[0.066]	[0.038]	[0.046]	[0.049]
Female x Monitor Safety	0.089***	0.090***	0.097***	0.013	0.135**	0.112**	0.098**	0.076
	[0.026]	[0.026]	[0.026]	[0.094]	[0.060]	[0.043]	[0.044]	[0.071]
Monitor Theft, pooled	0.023* [0.012]							
Monitor Safety, pooled	0.031*** [0.012]							
Observations	20,170	20,170	17,296	2,874	2,972	6,389	4,978	2,957
R-squared	0.129		0.130	0.197	0.178	0.149	0.127	0.125
Control Mean	0.522	0.522	0.517	0.548	0.493	0.525	0.540	0.487

Notes: The table reports treatment effects for the main sample and by quintile of the employer's stated preferred workforce gender composition. Column (1) reports the main sample; column (2) re-estimates the same specification by probit and reports average marginal effects on the probability scale. Column (3) excludes employers who state a preference for an all-male workforce. Columns (4) to (8) restrict the sample to each quintile of the stated preference distribution. The dependent variable *Hire* is a binary indicator equal to 1 if the employer wants to hire the candidate on probation. *Female* is a binary variable indicating whether the profile corresponds to a woman. *Monitor Theft* is a binary indicator equal to 1 if the firm was randomly assigned to visits to monitor trainees' behavior. *Monitor Safety* is a binary indicator equal to 1 if the firm was randomly assigned to visits to ensure trainees' safety. The bottom panel reports the monitoring effects on selection pooled across candidate genders, computed from the column (1) regression as the arm coefficient plus one half the interaction (profiles are half female by design); standard errors in brackets, and the pooled business-as-usual selection rate is 0.468. Profile and strata fixed effects are included and not reported. Standard errors, clustered at the employer level, are reported 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	-0.614	-0.128	-0.138	-0.420
	[0.177]	[0.178]	[0.159]	[0.211]
Holm p-value	0.004	1.000	1.000	0.237
WY p-value	0.004	0.743	0.743	0.182
Monitor Safety	-0.389	-0.484	-0.222	0.029
	[0.177]	[0.182]	[0.169]	[0.214]
Holm p-value	0.168	0.056	0.761	1.000
WY p-value	0.141	0.047	0.526	0.892
Observations	876	882	885	585
R-squared	0.419	0.452	0.400	0.366
Control mean	5.511	4.656	4.965	4.052

Notes: The table reports treatment effects on post-experiment worry measures, elicited on a 0–10 scale after the profile evaluations (Section 3.2). Each column reports a regression of the indicated worry measure on indicators for Monitoring-Theft and Monitoring-Safety assignment, with the Business-as-usual arm omitted and sector, survey-wave, and enumerator fixed effects absorbed. Robust standard errors in brackets. Below each coefficient we report p-values adjusted for multiple hypothesis testing across the family of eight tests (two treatment arms $\times$ four outcomes): Holm step-down and Westfall–Young (WY; 2,000 bootstrap replications). Stars are omitted: inference should be read from the adjusted p-values. Effort and turnover worries are placebo outcomes: neither visit type targets them. The turnover worry question was asked in the 2024 survey waves only, hence the smaller sample in column (4). Results are robust to adding baseline covariates (vocational training of the respondent, firm revenues, and monthly number of clients).

Table 5: Hiring and the Technical Performance Signal

Signal:	Top Performance		Performance Index		Certificate	
Hire (0-1)	(1)	(2)	(3)	(4)	(5)	(6)
Female		-0.085***		-0.106***		-0.100***
		[0.019]		[0.017]		[0.018]
Monitor Theft	0.021	0.054***	0.022*	0.053***	0.032**	0.065***
	[0.015]	[0.021]	[0.012]	[0.016]	[0.014]	[0.019]
Monitor Safety	0.042***	0.003	0.031***	-0.013	0.038***	-0.003
	[0.015]	[0.021]	[0.012]	[0.016]	[0.015]	[0.020]
Female × Monitor Theft		-0.065**		-0.060**		-0.064**
		[0.029]		[0.024]		[0.026]
Female × Monitor Safety		0.079***		0.089***		0.084***
		[0.030]		[0.026]		[0.029]
Signal	0.093***	0.113***	0.106***	0.113***	0.150***	0.160***
	[0.013]	[0.018]	[0.011]	[0.013]	[0.016]	[0.022]
Female × Signal		-0.042*		-0.016		-0.019
		[0.023]		[0.012]		[0.026]
Monitor Theft × Signal	0.003	-0.002	-0.013	-0.022	-0.035	-0.045
	[0.019]	[0.025]	[0.012]	[0.015]	[0.023]	[0.031]
Monitor Safety × Signal	-0.024	-0.034	-0.018	-0.018	-0.026	-0.034
	[0.019]	[0.025]	[0.012]	[0.015]	[0.023]	[0.030]
Female × Monitor Theft × Signal		0.011		0.020		0.020
		[0.031]		[0.017]		[0.039]
Female × Monitor Safety × Signal		0.021		0.000		0.016
		[0.032]		[0.017]		[0.037]
Observations	20,197	20,170	20,268	20,199	20,268	20,199
R-squared	0.115	0.128	0.115	0.128	0.023	0.035
Control Mean	0.422	0.466	0.491	0.538	0.474	0.474

Notes: The table reports treatment effects for the main sample. The dependent variable *Hire* is a binary indicator equal to 1 if the employer wants to hire the candidate on probation. *Female* is a binary variable indicating whether the profile corresponds to a woman. *Monitor Theft* is a binary indicator equal to 1 if the firm was randomly assigned to visits to monitor trainees' behavior. *Monitor Safety* is a binary indicator equal to 1 if the firm was randomly assigned to visits to monitor trainees' safety. *Signal* denotes a different measure in each column pair, named in the column-group header. In columns (1) and (2) it is a binary indicator equal to 1 if the profile is ranked in the top of its training cohort on the combined theory and practical assessment. In columns (3) and (4) it is a continuous index of profile quality—equally weighting completed secondary education, certification, enrollment in a twelve-month rather than six-month programme, and grade point average—standardised on the *Control* arm distribution, so its coefficient is the effect of a one standard deviation increase. In columns (5) and (6) it is a binary indicator equal to 1 if the profile displays a certificate. Coefficients in columns (3) and (4) are therefore per standard deviation and are not directly comparable in magnitude with the binary contrasts in the remaining columns. Odd-numbered columns pool male and female profiles; even-numbered columns are fully saturated in gender, signal, and monitoring regime. Columns (1) to (4) include profile and strata fixed effects. The control mean is that of *Control*-arm candidates without the signal: average-performance candidates in columns (1) and (2), candidates within 0.3 standard deviations of the index mean in columns (3) and (4), and non-certified male candidates in columns (5) and (6). Standard errors clustered by employer are reported in brackets; inference is conditional on the experimental profile set. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Trainee Attributes and Employers' Beliefs

	Trainee		Employer	
	Actual		Beliefs	
Panel A: Technical skills				
	Theory	Practical	Theory	Practical
Female	0.214**	0.070	-0.120***	-0.153***
	[0.092]	[0.045]	[0.007]	[0.008]
Mean, men	0.773	0.918	0.579	0.607
Observations	173	173	597	598
Panel B: Cheating task				
	Effort	Claimed (≥ 6)	Effort	Claimed (≥ 6)
Pooled across supervision conditions				
Female	0.017	0.050	0.023***	-0.085***
	[0.023]	[0.043]	[0.005]	[0.006]
By supervision condition				
Female	0.029	0.072	0.046***	-0.113***
	[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	-0.046***	0.056***
	[0.035]	[0.062]	[0.008]	[0.007]
Mean, unsupervised men	0.989	0.073	0.649	0.463
Observations	355	346	910	907

Notes: Trainee columns report actual trainee behavior; *Employer* columns report employer beliefs about trainees, elicited within employer by trainee gender and, in Panel B, by supervision condition. Trainee estimates in Panel A are at the trainee level, with vocational-training-institute fixed effects and robust standard errors. In Panel B, effort is measured at the sheet level: the unsupervised sheet is anonymous and identified only by gender, so it cannot be linked to a trainee and standard errors are robust; the bonus claim is at the trainee level, with two observations per trainee, and standard errors are clustered at the trainee. Employer estimates are within-employer contrasts collapsed to one observation per employer, so standard errors are robust and there is no remaining within-employer dimension to cluster on. *Female* is a dummy for whether the trainee is female or the employer was asked to guess the outcome for female trainees. *Supervised* is a dummy for whether the trainee performed the task supervised or the employer was asked to guess the outcome for a trainee under the supervised condition. Panel A outcomes are theory knowledge and self-reported ability to perform a standard sector task; employer columns report believed probabilities of each. The trainee theory outcome is the exam question on which employer beliefs were elicited. Panel B outcomes are effort, measured as the share of twelve registrations checked, and the bonus claim, defined as reporting six or more mistakes when only five are planted. Employer beliefs are rescaled to shares. In the unsupervised condition, effort is measured on anonymous sheets identified only by gender; the claim controls for effort. Mean rows report the omitted-category mean: male trainees in Panel A and unsupervised male trainees in Panel B. Standard errors in brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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	0.05	1.68
	[0.06]	[2.00]
Monitoring-Theft	0.46	16.57
	[0.05]	[1.71]
Panel B: Decomposition components		
Trust premium (T)	0.17	5.94
	[0.07]	[2.37]
Harassment penalty (H)	0.25	8.95
	[0.07]	[2.59]
Residual (B)	0.21	7.62
	[0.09]	[3.10]
Panel C: Residual by attributes treated as productive		
Technical performance only	0.30	10.64
	[0.05]	[1.65]
+ trustworthiness	0.46	16.57
	[0.05]	[1.71]
+ safety	0.05	1.68
	[0.06]	[2.00]
+ trustworthiness and safety	0.21	7.62
	[0.09]	[3.10]
Employers	886	886
Observations	20,170	20,170

Notes: The table reports the discrete-choice (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) reports the decomposition on the latent index of willingness to hire, in probit (standard-normal) units; column (2) translates the same quantities to the probability scale, as average male–female differences in predicted selection probability, in percentage points (pp). Panel A reports the gender gap in each experimental arm, Panel B the implied components, and Panel C the residual a researcher would attribute to bias as further attributes are treated as productive (equal to π_{C}, π_{MT}, π_{MS}, and B). Under the assumption that the two treatments do not interact, $T = \mathrm{MT} - \mathrm{C}$, $H = \mathrm{C} - \mathrm{MS}$, and $B = \mathrm{MT} + \mathrm{MS} - \mathrm{C}$, where C, MT, and MS denote the gaps under business-as-usual, Monitoring-Theft, and Monitoring-Safety. All specifications include profile-by-ability and strata fixed effects and use the main between-subjects sample. Standard errors, in brackets, are obtained by the delta method and clustered at the employer level, as in Table 3. The decomposition coincides with a linear probability model on the same sample (logged in the generating do-file).

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 analysis 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 in the analysis sample, 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 analysis sample with the 303 mechanically mapped responses yields 895 of the 915 analysis-sample observations for Figure 1, panel A.

For panel E, we classify the 615 nonmissing open-ended answers, 589 of them in the analysis 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/distraction 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/distraction 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%.

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 support 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 $m \in (0, 1]$.

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.

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. 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.[35]

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. A theft costs him $\theta > 0$ and a harassment incident costs him $\tau > 0$. 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

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

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

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.

[35] 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 male selection by 5.2 percentage points in firms that already employ women and lower it by 2.3 points in all-male firms (difference $p = 0.077$). This pattern is consistent with the omitted term, but imprecise.

Assumptions. We derive the monitoring predictions under two assumptions.

Assumption 1 (targeted effects of monitoring). Theft monitoring raises μ without changing η, while safety monitoring raises η without changing μ. Neither changes employers' beliefs about underlying worker characteristics, the costs they attach to theft and harassment, or their preferences over gender.

The mechanism tests support this separation: each treatment reduces worries about its targeted risk without detectable changes in other worries or assessments of candidate attributes (Table 4; Appendix Table A5).

Assumption 2 (gendered beliefs). Employers believe that 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 assumptions 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).

Prediction: effects of monitoring visits. Under Assumptions 1 and 2:

(i) theft monitoring visits widen the gender gap:

$$\frac{\partial \Delta}{\partial \mu} = \theta\, \pi \left(\frac{1}{\kappa_m} - \frac{1}{\kappa_f} \right) > 0; \tag{9}$$

(ii) safety monitoring visits narrow the gender gap:

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

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 examines its identifying 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. By Assumption 1, theft monitoring visits remove the trust premium and safety monitoring visits remove the harassment penalty, when each raises its monitoring to one. Averaging over employers before the visits, 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 he 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.[36]

Assumptions. The decomposition maintains two assumptions from the model: each type of visit affects only its targeted friction without changing beliefs or preferences, and the trust premium and harassment penalty enter employers' valuations additively.

We also assume that each type of visit fully removes its targeted friction: theft monitoring removes women's trust premium, and safety monitoring removes their harassment penalty.

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.

[36] 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 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

Employers 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.6). Second, it serves as a robustness check, estimating the visit effects within the same employer. After 24 evaluations under his first condition, each employer rates 12 more profiles, randomly assigned from the same base profiles in another version, under the second condition. We estimate

$$\begin{aligned}\text{Hire}_{ijb} = \beta_1 \text{Female}_{ij} + \beta_2 \text{MT}_{jb} + \beta_3 \text{MS}_{jb} \\ + \beta_4 \text{Female}_{ij} \times \text{MT}_{jb} + \beta_5 \text{Female}_{ij} \times \text{MS}_{jb} + \delta_{iq} + \alpha_j + u_{ijb},\end{aligned} \tag{15}$$

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 A6, 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).[37]

A.5 Social Preference Measures and Behavioral Games

This appendix describes the measures behind Appendix Table A7. All behavior is measured in the trainee survey; the employer survey elicits beliefs about male and female trainees separately.

Survey scales. Following Falk et al. (2018), trainees report on 0–10 scales their willingness to give to good causes without expecting anything in return (altruism), to give up something today to benefit in the future (patience), to return a favor (positive reciprocity), and to trust others (trust). We rescale each to $[0, 1]$. Employers answer a comparative version of each item—whether men are more willing, women are more willing, or there is no difference—coded -1, $+1$, and 0; the *Female* coefficient in panel B is the net balance of these responses.

One-shot games. Trainees make binary choices in three games with payoffs in Ugandan shillings. In the *dictator game*, Player 1 keeps (UGX 10,000 for self, 0 for Player 2) or splits (5,000 each); the outcome is splitting. In the *trust game*, Player 1 distrusts (10,000, 0) or

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

trusts; if trusted, Player 2 rewards (10,000 each) or punishes (0, 20,000); the outcomes are Player 1 trusting and Player 2 rewarding. The *cooperation game* is a prisoner's dilemma (mutual cooperation 10,000 each; mutual defection 5,000 each; unilateral defection 15,000 against 0); the outcome is cooperating. Trainees also play an ultimatum game, omitted from the table because employer beliefs were not elicited for it; there, female trainees are 32 percentage points more likely than men to propose an equal split ($p < 0.01$). For each game, employers state how many of ten male and ten female trainees would choose each action; beliefs are rescaled to shares.

A.6 Deviations from Preregistration

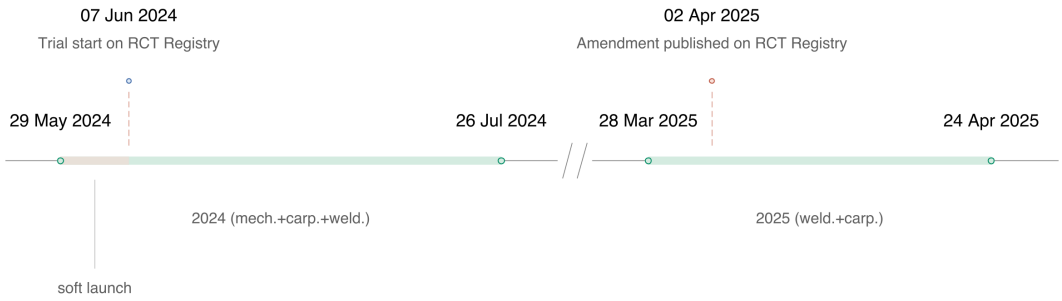
We note three main deviations from the first preregistration. First, the second primary outcome was not recorded as intended due to a coding error in the survey. The variable *Retain* — elicited using the question, "Okumuwa: Kisooboka kyenkanaki gwe okuwa omukozi ono omulimu nga [job] ku bizinensi yo? 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.

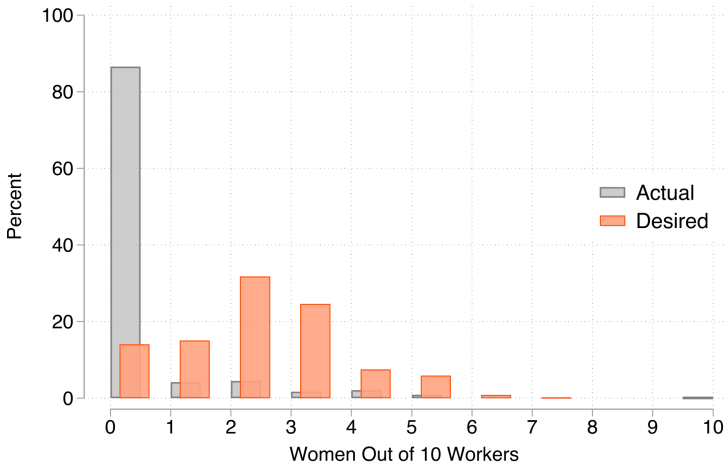
Figure A1: Timeline of Data Collection and Registry Milestones



Notes: The figure plots the data-collection timeline by survey date, with a broken horizontal axis that compresses the gap between collection 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 period between 29 May and 07 June 2024, prior to the trial start, denotes the *soft launch*, during which the data collection at scale was tested. The two markers above the timeline indicate the registered trial start (07 June 2024) and the only amendment to the preregistration (02 April 2025), published shortly after the start of the 2025 wave.

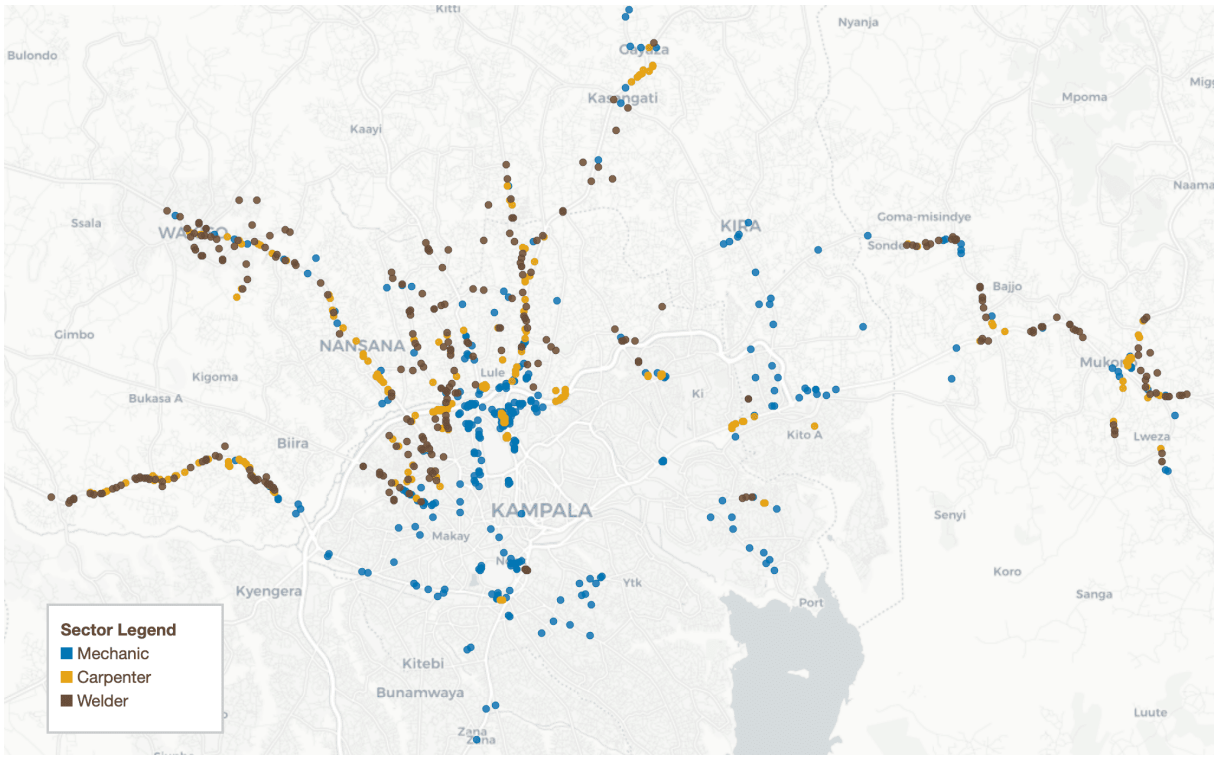
A.7 Appendix Figures

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



Notes: The figure is based on the full sample of employers (N=915; the actual share is available for 906). It presents the distributions of the desired gender composition (on a scale from 0 women out of 10 (all men) to 10 women out of 10 (all women)) versus the actual share of women among current workers, rescaled to 10.

Figure A3: Firm Spatial Distribution



Notes: The figure shows the locations of the 915 study firms in the Kampala metropolitan area, by sector. Each location is randomly displaced by up to 100 meters to preserve firm anonymity.

Figure A4: Firm and Vocational Training Environment

(A) Firm Layout



(B) Practical VTI Class



(C) Internship

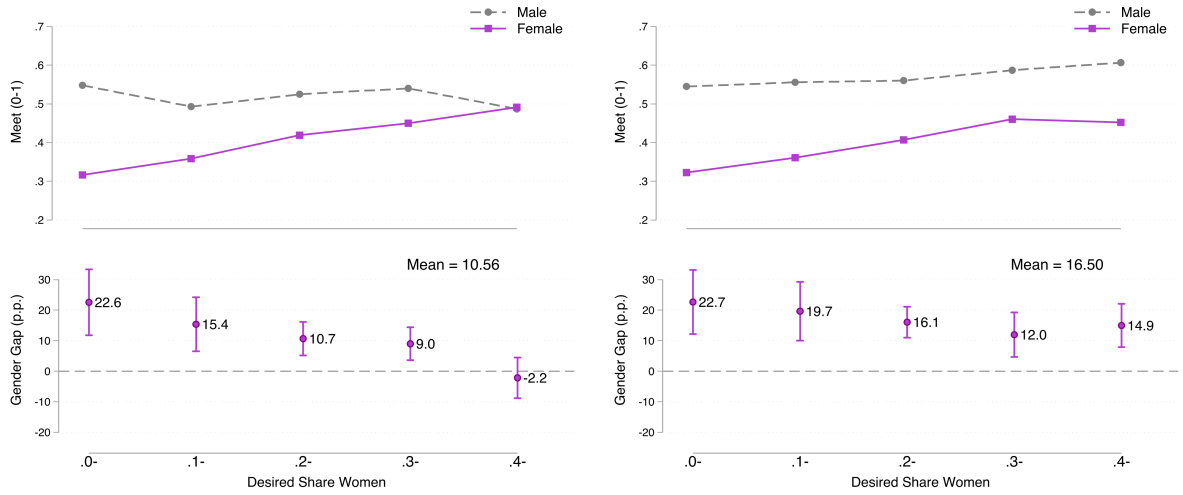


Photo credit: © Mariajose Silva-Vargas.

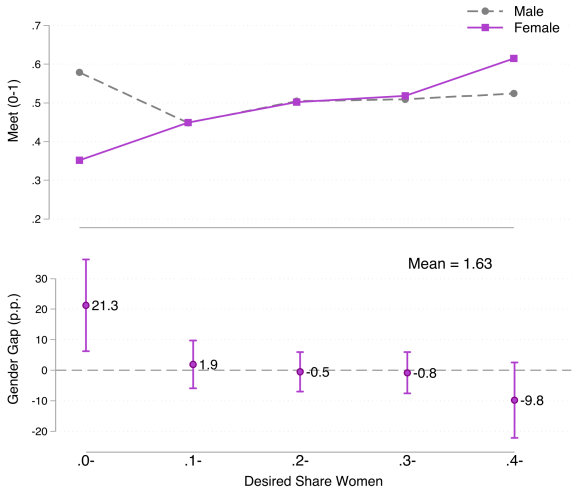
Figure A5: Hiring Gender Gap by Treatment Arm and by Stated Preferences for Hiring Women (Raw Data)

(A) Control (C)

(B) Monitoring-Theft (MT)

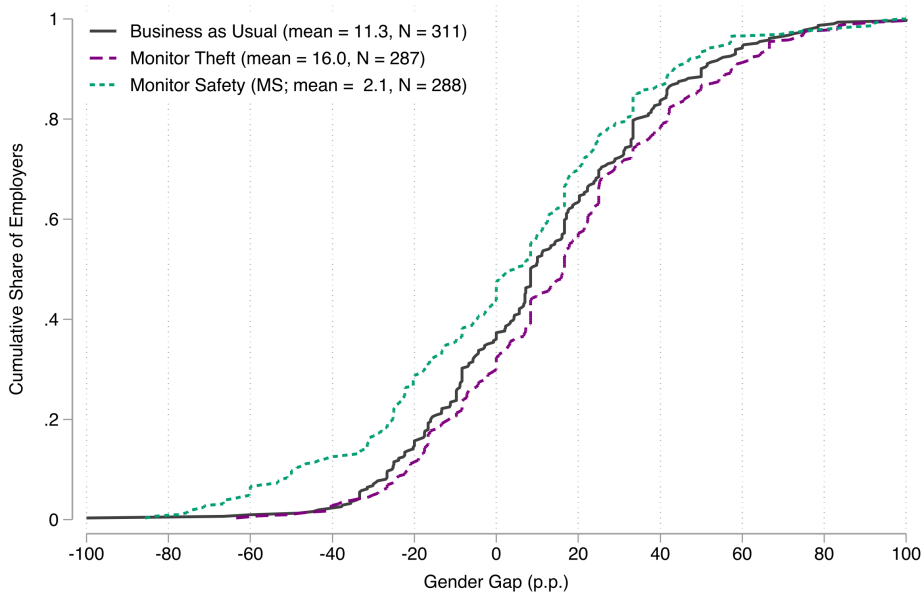


(C) Monitoring-Safety (MS)



Notes: The figures show the heterogeneity in the gender gap by stated preferences for workforce gender mix, by treatment arm. Panel A presents descriptive statistics for the Control group, while panels B and C present the corresponding statistics for the Monitoring-Theft and Monitoring-Safety groups, respectively. The top graph of each panel presents the raw data average of the employer's interest in hiring the candidate on probation by profile gender and by the share of female workers preferred at baseline, while the bottom graph displays the heterogeneous treatment effect of a profile being assigned to a female candidate, estimated with our main specification from equation (3). Confidence intervals are at the 95% level.

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



Notes: Lines show the empirical CDF of the within-employer gender gap in hiring (male minus female selection rate, in percentage points) for employers in Control (solid grey), Monitoring-Theft (dashed purple), and Monitoring-Safety (dotted green). Legend entries report each arm's mean gap and number of employers.

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

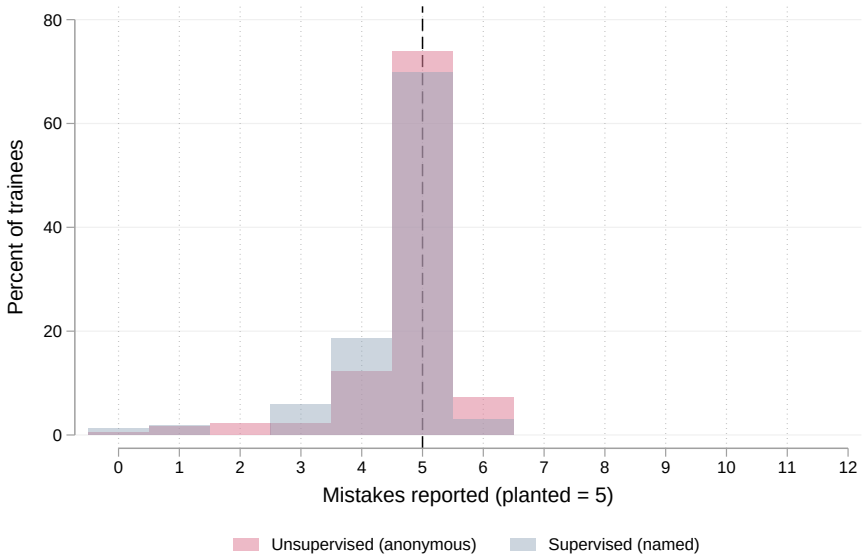
FULL NAME:_____ ☐ MAN

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: Reported Errors in the Record-Checking Task



Notes: Data from the record-checking task, at the tracking-sheet level: the 352 of 355 sheets with a recorded error count. Each trainee checks two rounds of twelve VTI registration forms, five of which contain a planted error; one round is supervised (the sheet carries the trainee's name and ID) and one is unsupervised (anonymous). Bars show the distribution of the number of errors reported, by supervision condition. The dashed line marks the bonus threshold: a bonus is paid for reporting six or more errors, and because only five exist, any report of six or more flags a correct form as mistaken. All over-reports record exactly six—the minimum that pays the bonus—and none exceed six, although 92% of sheets record all twelve forms checked.

A.8 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	0.241	0.015 [0.036]	-0.029 [0.035]	0.44
No preference	0.689	0.009 [0.038]	0.022 [0.038]	0.85
Women	0.071	-0.024 [0.020]	0.007 [0.022]	0.26

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	-0.014	-0.016	0.68
	[0.016]	[0.017]	
Monitor Theft	0.053***	0.059***	0.22
	[0.016]	[0.016]	
Female × Monitor Safety	0.089***	0.100***	0.35
	[0.026]	[0.026]	
Female × Monitor Theft	-0.059**	-0.063**	0.67
	[0.024]	[0.025]	
Observations	20,170	18,771	
Control mean	0.522	0.529	
Share female among selected	0.448	0.441	

Notes: The table reports checks that experimenter-demand effects do not drive the results, following de Quidt et al. (2026). The measure: at the end of the survey, employers state whether they thought the research team preferred them to hire men, women, or had no preference; 877 of the 886 main-sample employers answer. Demand effects could explain the treatment effects only if treatment shifted this belief (Panel A) or if the estimates depended on the employers who believe the team preferred them to hire women (Panel B). Panel A reports, for each answer, the share of employers and the coefficients from regressing the answer's indicator on the monitoring arms (business-as-usual omitted) at the firm level, with robust standard errors inline in brackets; *Joint p* tests that both arm coefficients are zero. Panel B estimates the main treatment-effect regression; the outcome is *Hire*. Column (2) drops the believers – the employers who report believing the research team preferred them to hire women; employers who declined to answer the item are retained. *Share female among selected* is the share of the selected profiles that correspond to women, in each column's sample. Column (3) reports, for each coefficient, the p-value of the test that it differs for the excluded employers, from the model fully interacted with the believer indicator. The treatment effects also do not vary with the item: interacting the gender gap and the arms with it leaves the triple interactions insignificant ($p = 0.60$ for Monitor Theft, $p = 0.72$ for Monitor Safety), as does adding the stated preference gradient ($p = 0.61$, $p = 0.86$). Panel B is at the profile level with CV profile fixed effects interacted with trainee performance, plus sector, survey-wave, and enumerator fixed effects; 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 check	(4) Soft launch	(5) Post-amendment
Female	-0.106***	-0.103***	-0.106***	-0.102***	-0.136***
	[0.016]	[0.016]	[0.016]	[0.016]	[0.025]
Female × Monitor Safety	0.089***	0.080***	0.085***	0.080***	0.281***
	[0.026]	[0.025]	[0.025]	[0.025]	[0.045]
Female × Monitor Theft	-0.059**	-0.051**	-0.061***	-0.060**	-0.151***
	[0.024]	[0.023]	[0.024]	[0.023]	[0.042]
Employers	886	889	912	925	265
Observations	20,170	21,189	20,759	21,079	6,162
Control mean	0.522	0.513	0.522	0.522	0.568

Notes: The table re-estimates column (1) of Table 3 on alternative samples, one per column. Column (1) reports the main analysis sample. Column (2) adds back the evaluation blocks excluded by the attention check, where the employer rated every profile in the block identically. Column (3) adds back the treated firms that failed, or were never asked, the comprehension check about their assigned treatment. Column (4) adds the soft-launch firms interviewed before the registered trial start on 07 June 2024; the comprehension check, introduced with the final questionnaire, could not be applied to them. Column (5) restricts the sample to interviews on or after the only amendment to the preregistration, published 02 April 2025. The dependent variable *Hire* is a binary indicator equal to 1 if the employer wants to hire the candidate on probation. *Female* is a binary variable indicating whether the profile corresponds to a woman. *Monitor Theft* is a binary indicator equal to 1 if the firm was randomly assigned to visits to monitor trainees' behavior. *Monitor Safety* is a binary indicator equal to 1 if the firm was randomly assigned to visits to ensure trainees' safety. The monitoring main effects are estimated but not reported. Profile and strata fixed effects are included. Standard errors, clustered at the employer level, are reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Heterogeneity: Preregistered Dimensions

Hire (0-1)	(1)	(2)	(3)	(4)
	Heterogeneity variable:			
	Preferred female share		Female	Recruits
	Continuous (0–1)	High (> median)	employer	from VTIs
Female	-0.213***	-0.146***	-0.107***	-0.090***
	[0.036]	[0.023]	[0.017]	[0.018]
Monitor Theft	0.003	0.037*	0.049***	0.069***
	[0.031]	[0.021]	[0.016]	[0.018]
Monitor Safety	-0.025	-0.020	-0.019	-0.013
	[0.033]	[0.021]	[0.016]	[0.018]
Het. var	-0.104	-0.012	-0.040	0.080***
	[0.083]	[0.022]	[0.054]	[0.027]
Female × Monitor Theft	-0.000	-0.041	-0.056**	-0.080***
	[0.049]	[0.032]	[0.024]	[0.026]
Female × Monitor Safety	0.071	0.097***	0.095***	0.086***
	[0.056]	[0.035]	[0.026]	[0.028]
Female × Het. var	0.483***	0.103***	0.052	-0.082*
	[0.126]	[0.031]	[0.106]	[0.045]
Monitor Theft × Het. var	0.228**	0.041	0.091	-0.089**
	[0.116]	[0.032]	[0.071]	[0.037]
Monitor Safety × Het. var	0.053	0.018	0.235***	0.011
	[0.128]	[0.033]	[0.071]	[0.040]
Female × Monitor Theft × Het. var	-0.267	-0.045	-0.089	0.114*
	[0.171]	[0.047]	[0.126]	[0.066]
Female × Monitor Safety × Het. var	0.098	-0.019	-0.244*	0.003
	[0.215]	[0.051]	[0.144]	[0.068]
Joint p (Female × arm × Het. var)	0.13	0.63	0.22	0.16
Observations	20,170	20,170	20,170	20,170
R-squared	0.136	0.133	0.130	0.131

Notes: The table reports the preregistered heterogeneity dimensions of the treatment effects, one column per dimension. Each column estimates column (1) of Table 3 fully interacted with the column's heterogeneity variable (*Het. var*); evaluations with a missing heterogeneity variable drop from that column. Columns (1) and (2) are two versions of the same underlying measure, the share of female workers the employer prefers in their ideal workforce composition: column (1) enters it continuously (0 to 1), and column (2) as a binary indicator for a share above the median – the preregistered split, equal to 1 for shares of 0.3 or more; the mass point at 0.2, the median, falls in the low group. *Female employer* is a binary indicator for a female respondent. *Recruits from VTIs* is a binary indicator for the firm recruiting workers from vocational training institutes. *Joint p* tests that the two Female × arm × heterogeneity terms are jointly zero. The dependent variable *Hire* is a binary indicator equal to 1 if the employer wants to hire the candidate on probation. *Female* is a binary variable indicating whether the profile corresponds to a woman. *Monitor Theft* is a binary indicator equal to 1 if the firm was randomly assigned to visits to monitor trainees' behavior. *Monitor Safety* is a binary indicator equal to 1 if the firm was randomly assigned to visits to ensure trainees' safety. Profile and strata fixed effects are included and not reported. Standard errors, clustered at the employer level, are reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Effects of Monitoring Visits on Employers' Beliefs About Profile Attributes

	(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	0.027	0.000	0.000
Monitor Theft	-0.019	-0.026	-0.002	0.008
	[0.044]	[0.052]	[0.045]	[0.011]
Monitor Safety	-0.033	-0.031	-0.006	-0.010
	[0.046]	[0.051]	[0.045]	[0.011]
Female × Monitor Theft	-0.029	0.001	-0.051	-0.003
	[0.034]	[0.038]	[0.039]	[0.006]
Holm p-value	1.000	1.000	1.000	1.000
WY p-value	0.919	0.998	0.685	0.956
Female × Monitor Safety	0.029	0.055	0.002	-0.002
	[0.033]	[0.036]	[0.040]	[0.006]
Holm p-value	1.000	1.000	1.000	1.000
WY p-value	0.919	0.601	0.998	0.966
Observations	20,196	20,123	20,172	20,197
R-squared	0.230	0.084	0.175	0.215
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 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-by-performance and strata fixed effects are absorbed. 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 A6: 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	-0.116***	-0.097***	-0.243***	-0.125**	-0.100***	-0.122***	-0.050*
	[0.015]	[0.015]	[0.060]	[0.054]	[0.026]	[0.026]	[0.028]
Monitor Theft	0.092***	0.102***	0.034	0.074**	0.122***	0.081***	0.107***
	[0.013]	[0.014]	[0.043]	[0.036]	[0.022]	[0.026]	[0.029]
Monitor Safety	-0.047***	-0.046***	-0.016	-0.100***	-0.041*	-0.051**	-0.027
	[0.013]	[0.014]	[0.049]	[0.036]	[0.024]	[0.024]	[0.028]
Female x Monitor Theft	-0.119***	-0.122***	-0.111	-0.147**	-0.130***	-0.087**	-0.123***
	[0.021]	[0.021]	[0.074]	[0.067]	[0.034]	[0.040]	[0.039]
Female x Monitor Safety	0.154***	0.168***	-0.017	0.188***	0.163***	0.184***	0.170***
	[0.023]	[0.023]	[0.091]	[0.063]	[0.041]	[0.040]	[0.050]
Observations	20,518	18,199	2,319	2,411	6,501	5,582	3,705
R-squared	0.220	0.214	0.313	0.250	0.236	0.208	0.191
Control Mean	0.530	0.528	0.540	0.525	0.514	0.549	0.523
p-value MT vs. MS	0.000	0.000	0.302	0.000	0.000	0.000	0.000

Notes: The table reports treatment effects on the subsample of employers for whom the within-subject randomization was implemented, and by quintile of the employer's stated preferred workforce gender composition. Employers rated a second set of profiles under a different condition; the two conditions and their order are randomized across all six possible pairs. Column (1) reports the full within sample; column (2) excludes employers who state a preference for an all-male workforce. Columns (3) to (7) restrict the sample to each quintile of the stated preference distribution. The dependent variable *Hire* is a binary indicator equal to 1 if the employer wants to hire the candidate on probation. *Female* is a binary variable indicating whether the profile corresponds to a woman. *Monitor Theft* is a binary indicator equal to 1 if the firm was randomly assigned to visits to monitor trainees' behavior. *Monitor Safety* is a binary indicator equal to 1 if the firm was randomly assigned to visits to ensure trainees' safety. The final row reports the p-value for the test that the two *Female* by arm interactions are equal. Employer and profile-by-ability fixed effects are included and not reported. Standard errors, clustered at the employer level, are reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: Other Survey and Behavioral Prosociality Measures

	Social preferences				Dictator	Trust game		Cooperation
	(1) Patience	(2) Altruism	(3) Pos. recipr.	(4) Trust	(5) Share	(6) Trust (Pl. 1)	(7) Reciprocate (Pl. 2)	(8) Cooperate
Panel A: Trainees								
Female	0.072	0.012	0.057	0.011	0.131	-0.343***	-0.019	0.180*
	[0.055]	[0.055]	[0.043]	[0.054]	[0.111]	[0.101]	[0.098]	[0.103]
Observations	157	152	160	154	173	173	173	173
R-squared	0.050	0.032	0.056	0.018	0.126	0.086	0.073	0.071
Mean, men	0.731	0.758	0.800	0.486	0.736	0.673	0.882	0.827
Panel B: Employers' beliefs								
Female	-0.061**	-0.085***	0.163***	0.485***	-0.092***	-0.012	-0.043***	-0.019
	[0.030]	[0.030]	[0.026]	[0.023]	[0.020]	[0.016]	[0.015]	[0.013]
Observations	913	913	914	915	913	911	908	909
R-squared	0.000	0.000	0.000	0.000	0.001	0.000	0.002	0.000
Mean, men	0.436	0.451	0.251	0.127	0.572	0.533	0.604	0.571

Notes: The table reports gender differences in trainees' prosociality measures (Panel A) and in employer beliefs about trainees (Panel B); Appendix A.5 describes the measures in detail. Columns (1)–(4) are survey measures of social preferences following Falk et al. (2018)—patience, altruism, positive reciprocity, and trust—elicited from trainees on 0–10 willingness scales and rescaled to the unit interval. The corresponding employer beliefs are comparative: employers state whether men are more willing (−1), women are more willing (+1), or there is no difference (0); in Panel B columns (1)–(4) the *Female* cell is the net balance of opinion, and the *Mean, men* row reports the share of employers answering that men rank higher. Columns (5)–(8) report one-shot behavioral games: sharing in a dictator game (5), player 1 (trusting) and player 2 (reciprocating) behavior in a trust game (6)–(7), and cooperation in a prisoner's-dilemma game (8); in Panel B the *Female* cell is the within-employer contrast of the believed shares of female versus male trainees choosing the prosocial action. Technical skills and the cheating task are reported in Table 6. Panel A covers the 173 surveyed trainees in the study sectors' labor pools; regressions include vocational-training-institute fixed effects, with robust standard errors. Panel B covers the interviewed firms; columns (1)–(4) report means with robust standard errors, and columns (5)–(8) within-employer contrasts with sector fixed effects and robust standard errors. Except in Panel B columns (1)–(4), the *Mean, men* row reports the male (or believed male) mean. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A8: 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: The table reports the employer phone follow-up, conducted about two years after the experiment; we reached 776 firms, a consent rate of 97.5% among those called. Columns report the baseline value for the same firms, the follow-up value, the within-firm difference, the p-value of a Wilcoxon signed-rank test of no within-firm change, and the number of firms; Panel B quantities have no baseline counterpart. Panel A measures the workforce at each survey. *Workers* excludes owners and includes trainees, matching the baseline question wording, and *Female workers (share)* requires at least one worker at both surveys. *Ideal female share* is the employer's stated ideal share of women in the workforce. Panel B measures hiring between the end of the firm's data-collection round and the follow-up call: 24 months for firms surveyed in 2024, 15 for 2025, 21.0 on average. *Women hired* counts all women who worked at the firm at any point in the window, harmonized by the window length; *Female worker tenure* is defined only for firms that hired a woman. *Women employed at a given time* multiplies women hired by their average tenure and divides by the window length – the average number of women employed at a given time over the window; spells still running are right-censored. *Female workers at a given time (share)* divides that number by the workforce, for firms with at least one worker. Selection into the follow-up is examined in the *Follow-up* column of Table 1. All statistics are unweighted firm-level means.