Hidden Discrimination in Frictional Labor Markets

Elisa Macchi* and Claude Raisaro†

Version: September 6, 2026.

Abstract

We study how workplace frictions shape gender discrimination in hiring in three male-dominated sectors in Kampala, Uganda. Employers associate gender with several attributes whose returns depend on these frictions. They perceive women as less technically skilled and more trustworthy, but also as more exposed to workplace harassment. In an experiment, 903 employers select vocational trainees for probationary jobs and select men 10 percentage points more often than women. We experimentally relax monitoring and safety frictions while holding information about workers fixed. Auditing worker behavior to prevent stealing widens the gender gap by 56%, consistent with monitoring substituting for women's perceived trustworthiness. By contrast, auditing interactions among workers, which alleviates concerns about harassment, closes the gender gap. Differences in signaled technical performance do not explain the remaining gap. In a separate experiment, male and female trainees perform similarly on technical exams and in an honesty game, contrary to employers' beliefs. The status-quo gender gap thus hides two offsetting forms of discrimination: inaccurate statistical discrimination against men on trustworthiness and a larger residual penalty against women. Because identity signals multiple attributes, residual measures of bias may either understate or overstate discrimination, depending on which attributes researchers classify as productive. Because gender advantages need not align across attributes, workplace frictions create a second-best problem. Alleviating monitoring constraints removes the penalty men face on perceived trustworthiness but magnifies the hiring consequences of the remaining penalty against women.

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

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

‡We thank Siwan Anderson, Abhijit Banerjee, Justus Bamert, Arielle Bernhardt, Aditi Bhowmick, Laura Boudreau, Nina Buchmann, Lorenzo Casaburi, Ada González-Torres, Peter Hull, Alex Imas, Rocco Macchiavello, Kristina Manysheva, Rohini Pande, Supreet Kaur, Frank Schilbach, Anna Vitali, Roberto Weber for helpful discussions. Josh Bwiira and Carlotta Riva provided outstanding field management and research assistance. Claude Code provided additional 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. The experiments were approved by Mildmay Uganda (0408-2023) and the Uganda National Council for Science and Technology (SS2574ES), and pre-registered at the AER Registry (AEARCTR-0013698).

1 Introduction

Labor-market frictions shape discrimination through two standard channels. They limit information about workers, leading employers to infer productivity from group identity (Phelps, 1972; Arrow, 1973), or allow biased preferences to persist by limiting competition (Becker, 1957; Black, 1995). As is common in models of hiring, both channels take the structure of labor demand as given: the production function pins down which worker attributes employers value.

We study a different channel through which workplace frictions shape discrimination: they change which worker attributes employers value and by how much.[1] Because such attributes are difficult to signal, identity becomes a signal for a bundle of traits. When beliefs about this bundle differ across groups, changes in workplace frictions can change hiring gaps even if beliefs remain fixed. Interpreting these gaps depends on whether beliefs are accurate and which attributes count as productive. We ask: how do workplace frictions shape demand across identity groups, and what does allowing for multidimensional workers imply for measuring discrimination?

Our focus is employer demand for male and female workers in male-dominated manufacturing sectors in Uganda: motor mechanics, welding, and carpentry, among the highest-return vocational occupations in the country. The context, like that of many poorer countries, is prone to workplace frictions (Breza and Kaur, 2025) and has strong gender norms (Jayachandran, 2021). While female labor-force participation is high (76.5%, in line with the sub-Saharan Africa average; International Labour Organization, 2025), most occupations remain nearly single-gendered. Policies promoting women's entry focus on the supply side, e.g., by training women for these sectors.[2]

The paper has four components. We measure preferences for male and female workers in a large incentivized hiring experiment: 903 employers evaluate hypothetical trainee profiles, knowing that their ratings determine real referrals from partner vocational training institutes (VTIs). We estimate the hiring gender gap: the difference in selection rates between profiles cross-randomized on candidate gender and a performance signal. To understand how frictions shape demand by gender, we randomize employers to audits that each mitigate one of two binding workplace frictions, and compare the hiring gender gap

[1] Firms facing moral hazard, for example, often rely on family labor and network hiring, consistent with trustworthiness carrying greater value (Bloom et al., 2013; Heath, 2018).

[2] Occupational segregation drives the earnings gender gap (Blau and Kahn, 2017; Goldstein et al., 2019). Training programs show mixed results for women (McKenzie, 2017; Bandiera et al., 2022).

under business-as-usual to the gap under each audit. The interventions are gender-neutral in wording; any differential effect by gender should operate through the targeted friction, a mechanism we confirm using employers' posterior beliefs. We benchmark employers' beliefs about workers' skills by gender against actual trainee behavior in a complementary lab-in-the-field experiment. Finally, we quantify the residual bias under alternative assumptions about which worker attributes are productive.

Two features of the setting are central to our question. First, employers value multiple worker attributes, and frictions shape which of them matter most. When asked to identify the most important trait in a worker, employers in our sample cite trustworthiness (good behavior/no stealing), effort, technical skills, learning, and cooperation (Figure 1). Trustworthiness and cooperation are the first and second hardest-to-find-and-monitor traits, both tied to well-documented workplace frictions in low-income settings: monitoring constraints over worker behavior and teamwork (Hjort, 2014).[3] Second, employers hold gendered beliefs about worker traits. Men are perceived as stronger on technical skills, learning interest, and effort, which we denote jointly as technical performance. Women are seen as more trustworthy and less likely to steal. While employers see men and women as equally cooperative, their main worry when hiring women is disruption to teamwork: harassment or distraction.

Most employers (85% of 903) state a preference for hiring at least one woman.[4] To measure revealed preferences for hiring women, the hiring experiment has employers choose trainees to hire on probation, the margin at which all hiring in this setting begins. We offer employers access to candidates from VTIs: these trainees are better skilled than their usual applicants, so they may carry higher experimentation costs (Groh et al., 2016). Demand for female workers is hard to observe because employers' usual applicant pools contain almost no women; the experiment removes this constraint by simplifying access to a pool of vocationally trained candidates where the share of women is much higher.

One-third of the employers evaluate profiles, cross-randomized on candidate gender and performance, under business-as-usual conditions. In this group, male selection rates exceed female selection rates by about 10 percentage points. Nevertheless, employers reveal substantial demand for female workers: 89.6% select at least one female trainee profile, and women are 42% of the average employer's selected candidates. Stated and revealed

[3]Employers can discipline effort through piece-rate pay (Foster and Rosenzweig, 1994) and address skills gaps through on-the-job training.

[4]The average stated ideal gender mix is 21.7% women, about seven times the actual gender mix (3.2%).

demand also align. The hiring gender gap narrows as stated demand for women rises and disappears in the top quintile.

We investigate to what extent workplace frictions shape the demand for male and female workers. Two-thirds of employers evaluate the same trainee profiles knowing that new hires will come with one of two external workplace audits: weekly visits by trained staff of our partner VTIs, coordinated by our team, during the one-month probation period. Visits are unannounced in timing but known to employers and employees. The audits mimic support that staff at some VTIs provide during internships, and add monitoring capacity at no cost to the firm. The two treatments use quasi-identical scripts, differing only in the friction they target: Behavior audits monitor new workers, focusing on stealing; Safety audits focus on ensuring new workers' safe and respectful treatment. Neither friction is intrinsically gendered and both audits are framed as gender-neutral.[5] Treatments are randomized and employers are explicitly told that audits do not reflect firm or worker characteristics; trainees learn about visits only if they accept a job, to avoid changes in supply. Neither audit affects the production technology itself: absent frictions, worker productivity is the same across the three arms. Nor do the audits differentially change the information employers have about candidates: both provide the same additional monitoring and reduce the cost of experimenting with new workers.[6] We compare the gender gap under business-as-usual and under each audit.

Our hypothesis is that each friction changes the set of attributes that employers value. Because beliefs about these attributes are gendered, the two audits, each mitigating one friction, should move the gender gap in opposite directions. We test two preregistered hypotheses. Behavior audits should lower the relative demand for women by reducing the value of trustworthiness, widening the hiring gender gap. Safety audits should raise the relative demand for women by addressing harassment concerns, narrowing the gap. In contrast, under a standard framework where workers differ along a single productivity dimension, the two audits would be the same intervention: a gender-neutral reduction in the cost of experimenting with new workers. Such an intervention may leave the gap unchanged or narrow it, depending on whether the gap reflects taste-based or statistical discrimination, but it should move the gap in the same direction in both arms.[7]

[5]For example, in Hjort (2014), teamwork frictions stem from worker ethnicity.

[6]Technologies that change the production function can shift the comparative advantage of one gender in production (Alesina et al., 2013).

[7]Under pure taste-based discrimination, the audits leave the disutility of employing women unchanged, and with it the gap (Becker, 1957). If animus is contact-based, audits that substitute for employer interaction reduce the effective taste cost; under statistical discrimination, cheaper experimentation reduces reliance on gendered priors about productivity (Phelps, 1972; Arrow, 1973).

We find that the two audit treatments have opposite impacts on the gender gap. Both behavior and safety audits increase total selection relative to the control group, consistent with lower experimentation costs. However, as predicted, they do so through opposite margins: Behavior audits increase male selection by 5.3 percentage points ($p < 0.001$) with no significant effect on women ($p = 0.685$), while safety audits increase female selection by 7.9 percentage points ($p < 0.001$) with no significant effect on men ($p = 0.473$). The average hiring gender gap under behavior audits is 16.7 percentage points, a 56% increase relative to business-as-usual. Behavior audits increase the gender gap most among the employers who state the strongest preferences for hiring women: in the top quintile, the increase is nearly three times that in the full sample. The gap under safety audits instead falls to 1.7 percentage points, statistically indistinguishable from zero, a six-fold reduction relative to business-as-usual.

We confirm that each audit operates by mitigating the targeted friction. We measure post-experimental worries about theft and workers' safety risk, as well as about two placebo dimensions the scripts do not mention, effort and turnover. Each audit shifts worries only about its targeted friction, with no effect on the placebos. This limits concerns that audits affect hiring through channels unrelated to the targeted friction. Confounders that beliefs do not capture remain possible, such as scrutiny: employers may select fewer women under behavior audits to avoid being observed around female workers. However, such confounders predict effects in the same direction under safety audits too, so the between-arm comparison serves as an additional mechanism test.

Behavior audits thus reveal a hidden gap: women's perceived trustworthiness generates demand that masks part of the penalty against them under business-as-usual. Safety concerns account for at most half of the uncovered penalty. The remainder may reflect taste against hiring women or statistical discrimination on technical skills, based on accurate or inaccurate beliefs. Under statistical discrimination, employers use gender as a proxy for expected performance and hold less favorable priors about women.[8] If so, providing more information about candidates should help women (Bertrand and Duflo, 2017). The profiles vary this information on two margins: its content (the rank in the class) and its amount (whether the candidate has taken a certification test). We find that employers reward both signals, but better or more information does not narrow the gender gap. If anything, returns to the signals are slightly lower for women, consistent with a mild undervaluation of the signals they carry (Phelps, 1972). These patterns, however,

[8]Employers report less favorable beliefs about women's technical performance in general (Figure 1). Whether they apply these priors to VTI trainees is unclear ex ante.

could also be consistent with moderate gendered priors or beliefs about unmeasured skills. We therefore test for accurate statistical discrimination directly by comparing employers' incentivized beliefs with behavior in the trainee pool (Bohren et al., 2025a).

The evidence is consistent with hidden gender discrimination. For the hidden gap to instead reflect accurate screening, employers' beliefs would need to be correct. We run a lab-in-the-field experiment with 182 vocational trainees. Employers expect female trainees to perform worse on sector-specific theory exams and practical tasks, the skills summarized by the profiles' performance signal. These beliefs are inaccurate: female trainees perform as well as male trainees on practical tasks and, if anything, better on theory. Employers' beliefs about trustworthiness are also inaccurate. In a novel behavioral task, trainees can misreport their output for pay. We find that cheating is rare and does not differ by gender. Yet employers expect nearly half of male trainees to cheat when unsupervised, about seven times the observed rate, and expect female trainees to cheat less than men. Thus, the gendered beliefs underlying both the technical-skill penalty and the trust premium are inaccurate.

How much bias do women face?[9] Our results show that estimates of bias are sensitive to which attributes the researcher measures and whether they are treated as productive. We quantify this sensitivity within a discrete-choice model of hiring (Appendix A.3). In the standard omitted-variable-bias problem, adding attributes shrinks the residual; here, it can move the estimate in either direction (Lang and Spitzer, 2020). A model with only technical performance attributes the full 10.7 percentage point gap to bias; adding trustworthiness raises the estimate to 16.7 percentage points; not accounting for trustworthiness while treating harassment concerns as productive lowers it to 1.7. This implies that residual measures of bias understate it under business-as-usual.

This paper provides experimental evidence on how workplace frictions shape the demand for workers by identity, operating through the demand for worker attributes. A defining feature of labor markets in poor countries is the coexistence of multiple frictions (see Breza and Kaur, 2025 for a review). Within the firm, frictions shape hiring practices (Heath, 2018) and technology adoption (Atkin et al., 2017; Houeix, 2025). We show experimentally that mitigating a friction directly changes the demand for worker attributes and the composition of hires: employers stereotypically link personality traits and soft skills to

[9]Our results also imply hidden discrimination against men: the trust premium that favors women under business-as-usual rests on an inaccurate belief. Because the net penalty runs against women, we focus on discrimination against them.

identity groups, in line with supply-side evidence on the importance of identity for labor-market outcomes in poor countries (Cassan et al., 2022; Oh, 2023).

Mitigating frictions is a central goal of development policy; indeed, certification and application support help workers signal their skills (Abebe et al., 2021; Bassi and Nansamba, 2022; Carranza et al., 2022), and enforcing safety standards improves working conditions (Boudreau et al., 2023). However, these results suggest that when multiple frictions coexist, solving one can move a target outcome, such as the gender gap, in opposite directions depending on which friction is addressed. This is the logic of the second best (Lipsey and Lancaster, 1956; Amodio et al., 2025): removing one distortion need not bring the allocation closer to the first best while others remain.

Our paper contributes to the literature identifying the sources of discrimination, and in particular separating belief-based discrimination from taste (Phelps, 1972; Arrow, 1973; Bohren et al., 2019, 2025a). Standard models consider one latent skill on which the disadvantaged group is, or is perceived to be, worse. Inspired by recent multidimensional models of workers (Deming, 2017; Woessmann, 2024; Deming and Silliman, 2025; Chambers and Echenique, 2021), we allow employers to value several unobserved attributes, uncorrelated with one another yet all inferable from identity. Gendered priors can then cross, with men believed better on some attributes and women on others.[10] We show that this simple yet realistic departure from the standard framework has two implications.

First, it creates an identification problem: outcome gaps can hide discrimination, and residuals can mismeasure bias. Favorable beliefs on one attribute, whether accurate or not, can mask unfavorable beliefs on another or a taste-based penalty. In our setting, the small net penalty against women masks at the same time inaccurate statistical discrimination against men, based on honesty, and a larger bias against women. Residual estimates can therefore understate discrimination; in standard models, by contrast, omitted productive attributes inflate the residual (Lang and Spitzer, 2020). Our results provide another reason not to interpret small outcome gaps as evidence of little discrimination. Bohren et al. (2019) show that outcome gaps are unreliable measures of discrimination because evaluations embed the history of past, possibly biased, evaluations; we add that gaps can be unreliable because gender signals multiple traits whose gendered beliefs can compensate.

[10]The stereotypes literature also generates crossing beliefs across skills (Bordalo et al., 2016, 2019), but each skill enters a separate decision. Macchiavello et al. (2026) measure gendered beliefs across eight supervisory attributes in Bangladeshi garment factories; there, beliefs uniformly favor men.

Second, it introduces a new source of disparate treatment beyond information and preferences: the set of attributes employers care about, which responds to features of the production environment such as workplace frictions. In the literature, taste-based and belief-based discrimination are identified through variation in the information environment (Goldin and Rouse, 2000; Carranza et al., 2022; Bassi and Nansamba, 2022; Bohren et al., 2019, 2025a). Our treatments keep beliefs and preferences fixed, and the two audits provide the same information; yet they move the gender gap in opposite directions, through the returns to attributes on which beliefs are gendered. This mechanism can help explain why gender discrimination varies across firms and sectors (Kline et al., 2022). It also rationalizes why information about soft skills can reduce a hiring gap even when credentials do not (Kaas and Manger, 2012).

This new source also blurs the line between statistical discrimination and bias. These attributes resemble traits more than skills: workers have little scope to invest in or signal them, and their productivity is ambiguous. Safety is the clearest example. Employers may act on the belief that integration costs reduce profits, or respond to altruistic or reputational concerns, akin to the paternalism documented by Buchmann et al. (2023). Whether researchers treat safety as productive therefore determines whether the same hiring response is read as statistical discrimination or bias. The problem connects to systemic discrimination: exposure to harassment is imposed on the disadvantaged group rather than chosen (Bohren et al., 2025b).

2 Setting, Sample, and Motivating Evidence

2.1 Setting and Sample

We study motor mechanics, welding, and carpentry firms in Kampala, Uganda. These are among the higher-return vocational sectors in the country (Alfonsi et al., 2020), and are extremely male-dominated.

Our sample consists of 903 employers, one per firm: 300 from motor mechanics, 300 from carpentry, and 303 from welding (see Table 1). We collected the data in two waves. The first covered only motor mechanics; the second added welding and carpentry to improve external validity across male-dominated sectors and increase statistical power (Appendix A.4). We selected firms in the neighborhoods of the greater Kampala area

where these sectors are clustered, as shown in Appendix Figure A3. We conducted a listing of all firms in these neighborhoods and interviewed all those who consented to participate, conditional on meeting the following criteria: (i) working in small or medium-sized firms (fewer than 100 employees); (ii) aged 18 or older; (iii) expressing interest in recruiting new workers on probation; and (iv) proficiency in either English or Luganda.

Our sample covers some of the most common male-dominated sectors in urban Uganda: women make up 2.3% of the workforce in our sample, in line with the 2.0% reported in Alfonsi et al. (2020) and Bassi et al. (2022).[11] We assess representativeness by comparing our firms to the Census of Business Establishments conducted by the Uganda Bureau of Statistics in 2020 (Uganda Bureau of Statistics, 2020) and in 2010 (Uganda Bureau of Statistics, 2011), restricted to Kampala and to our three sectors (Table A21). The censuses covers all establishments with a fixed location: our 903 firms are equivalent to 29.7% of the establishments censused in 2020 and 27.5% of those censused in 2010. Male ownership is near-universal throughout, at 96.3% in our sample against 96.1% in 2020 and 95.1% in 2010.[12] Firm age is comparable, at 10.8 years against 11.5 in 2020. Revenue is distributed similarly to the 2020 census, with 88.9% of our firms reporting annual turnover above 10 million UGX against 80.2%. Our firms are larger than the census average, at 9.3 persons engaged (median 6) against 8.4 in 2020 and 5.5 in 2010, since we required employers with workers to evaluate while the census includes owner-only establishments.

85.5% of respondents are business owners; 12% hold managerial positions. The firms are well established, serving approximately 4.1 clients per day and report mean monthly profits of USD 471 on revenues of USD 1,104.[13] 96.3% of employers are men, and only 170 out of 7,434 employees are women. Fewer than 14% of firms employ any women. This extreme occupational segregation is not specific to our sample. Alfonsi et al. (2024) document that when COVID-19 disrupted the Ugandan labor markets, occupational segregation by gender intensified.

We also collaborate with seven vocational training institutes (VTIs) in Kampala to collect data on trainees and gain access to candidates seeking employment. Our primary VTI partner is the Smart Girls Foundation, a nonprofit that delivers nationally recognized vocational training through its "Girls With Tools" program, preparing women for male-

[11] The Alfonsi et al. (2020) firm census covers motor mechanics and welding but contains no carpentry firms, while Bassi et al. (2022) cover carpentry and metal fabrication alongside grain milling.

[12] Male ownership is similar in the two firm surveys, at 92.6% in the Alfonsi et al. (2020) sectors that overlap with ours and at 97.9% and 99.2% in carpentry and metal fabrication in Bassi et al. (2022).

[13] Uganda's 2024 GDP per capita was USD 1,023.

dominated sectors including automotive mechanics, electrical installation, welding, and carpentry. Although the program primarily targets women, it also trains male students (see Appendix Figure A4, panel B).

Through these partnerships, we survey 182 trainees enrolled in automotive mechanics, carpentry, or welding programs. Of these, 33% specialize in automotive mechanics, 19.8% in carpentry, and 28% in welding (see Appendix Table A17).[14] Trainees are 20.8 years old on average and enrolled in programs lasting 6 to 24 months, with those longer than one year accredited by the Uganda Directorate of Industrial Training (DIT). Women make up 38.5% of our trainee sample; this share falls to 19.5% when excluding trainees from the Girls With Tools program. This nonetheless appears to be an upward trend compared to the 11% share of female vocational training applicants in welding and motor mechanics documented by Alfonsi et al., 2020.

2.2 The Multitasking Firm and Monitoring Constraints

Firms in this setting are organized around bundles of tasks rather than specialized production activities and exhibit minimal labor specialization (Bassi et al., 2023). Employers in our sample report spending an average of 10.4 hours at the firm each day, dedicating 2.06 hours to monitoring, the second most time-consuming activity after technical work, and more than they spend on training workers. In this setting, employers can access skilled workers through vocational training or train them on the job: firm-provided training raises sector-specific skills as much as formal vocational training (Alfonsi et al., 2020), at the cost of the owner's time.

Monitoring constraints naturally emerge from this production environment. The physical layout of workshops compounds the challenge: firms typically combine indoor workspaces with large outdoor areas (see Figure A4, panel A), making it difficult to observe all workers at once. Over 90% of employers report wanting to increase monitoring, and 83% have experienced theft by an employee. Piece-rate pay offers a partial solution (four out of five employers report paying piece rates) to moral hazard problems related to effort but it does not help with theft risk.

[14] An additional 19.2% are enrolled in broader fields such as electrical work but report job aspirations aligned with our target sectors. About 9.5% of the trainees are enrolled in multiple programs.

Employers respond to this residual moral hazard through network hiring: most current workers in our sample (86%) were recruited through personal networks, in line with other work in developing countries (Heath, 2018; Chandrasekhar et al., 2020). Network hiring restricts the talent pool, both in terms of skills and gender. Evidence from wage subsidy experiments suggests that employers perceive an experimentation cost from hiring outside their usual channels (Groh et al., 2016), although Alfonsi et al. (2020) show that vocationally trained workers (the alternative path to network hiring) have better skills. In male-dominated sectors, network-based recruitment also produces overwhelmingly male candidate pools (Beaman et al., 2018).

In our sample, employers estimate that on average 3.6% of their applicants are women; 80.5% report none at all.[15] The scarcity is in the channel rather than in the stock of trained women. Women are 3.6% of the applicants employers see, against 19.5% of trainees at the vocational institutes we partner with, excluding the program that trains women specifically, and 38% including it. Networks in these sectors are male, so they screen women out before employers form any view about them. This makes it difficult to assess whether employers are at all interested in hiring women. In particular, we know little about employers' preferences for gender and skills.

2.3 Demand for Multiple Worker Attributes

The multidimensional skills literature emphasizes that workers are bundles of cognitive, social, personality, and behavioral attributes, and that the labor-market value of these attributes depends on the tasks firms need performed and the production environment (Deming, 2017; Woessmann, 2024; Deming and Silliman, 2025). The descriptive patterns in our data align with this view. Employers demand multiple worker attributes, and the attributes they value most are closely tied to the workplace frictions they face.

The categories come from the employers themselves. In the pilot we asked an open-ended question, "What is the single most important trait you look for when hiring a worker?", and coded the answers with local enumerators. We ask the same question in the main survey and recover the same categories: fewer than one percent of answers fall outside them. We use these categories to classify the open-ended answers to the same question in

[15] We lack data on the actual gender composition of the applicant pool. For comparison, Ashraf et al. (2025) find that firms in India and Turkey draw from applicant pools where men outnumber women four to one.

the main survey, and as the fixed options in the closed-choice questions that follow. When asked to state the most important trait in a worker, employers most frequently cite good behavior, described as trustworthiness, honesty, and not stealing (Figure 1, panel A).[16] Good behavior is also the trait employers report as the hardest to find and the hardest to monitor. Effort and hard work rank second in importance, but only 34% of employers cite effort as hard to find or monitor, consistent with the prevalence of piece-rate pay, which makes effort contractible (Foster and Rosenzweig, 1994). Employers also mention learning interest (14.5%). Few employers mention cooperation as the single most important trait (4.8%), yet it is the second-hardest trait to find or monitor, cited by 45.7% of employers.

Beyond technical performance, defined as the combination of technical skills, learning interest, and effort, employers place distinctive emphasis on trustworthiness and cooperation. The emphasis on trustworthiness is consistent with monitoring frictions being pervasive in low-income firms (Houeix, 2025): evidence suggests that asymmetric information leads to hiring family labor, hiring from networks, reducing delegation, with implications for firm growth (Bloom et al., 2013; Kotia, 2025; Heath, 2018). The importance of cooperation is consistent with evidence that frictions in workplace interactions and team production can arise from worker identity and social preferences, including ethnicity (Hjort, 2014), gender (Ashraf et al., 2024), and fairness concerns (Breza et al., 2018).

The descriptive evidence also highlights a second feature: the attributes employers value most are difficult to assess and monitor, especially trustworthiness and cooperation. This aligns with evidence from Sub-Saharan Africa that asymmetric information makes it hard for workers to credibly signal their skills to employers, especially soft skills and other hard-to-observe attributes (Carranza et al., 2022; Bassi and Nansamba, 2022).[17]

[16] We classify answers into categories using GPT-5 mini and benchmark the classification to human coding, with a 96.6% success rate. Demographic characteristics, such as age or personal background, are coded for Panel A but excluded from Panels B–D because they are fixed worker traits. Appendix A.1 describes the procedure.

[17] Monitoring frictions are themselves a form of asymmetric information. We use the terms separately to emphasize that asymmetric information matters through two channels in our setting: it makes worker attributes difficult to observe before hiring, and it raises the value of trustworthiness once workers are hired.

2.4 Gendered Beliefs about Worker Attributes

Relative to the literature, the novel pattern in our data is that employers hold gendered perceptions of these attributes. For each of the five worker-trait categories in Figure 1, we ask whether men are better, women are better, or there is no difference, so that a perceived advantage for one gender implies a corresponding liability for the other.[18] Men are perceived as stronger in technical domains, effort, and learning, while women are perceived as more trustworthy and less likely to steal (Figure 1, panels C and D). Employers do not perceive women as less cooperative. If anything, in the Business-as-usual arm, more employers report that women are better than men at cooperation and interactions with clients, managers, and colleagues.[19] We also ask employers open-endedly about their concerns when hiring women (Figure 1, panel E). This question captures gender-specific concerns that may not appear in the trait battery.[20]

Several responses echo the gendered beliefs above, but the open-ended question clarifies their content. When employers mention skills, they mostly refer to physical strength; when they mention interest in the job, they mostly worry that women will leave. Stealing is never mentioned, consistent with women's perceived trustworthiness. The most notable finding concerns cooperation. Harassment from men to women, or distraction from women to men, is the most frequently cited concern when hiring women, mentioned by about 40% of employers. Thus, mixed-gender workplaces appear to introduce cooperation frictions similar to those documented in inter-ethnic contexts (Hjort, 2014).

[18] We elicit these gendered-belief questions after the hiring experiment, near the end of the survey, to avoid priming gender before employers evaluate candidate profiles. Three checks suggest that the timing does not drive the patterns. First, the same qualitative patterns appear in pilot data collected without the hiring experiment. Second, the patterns are unchanged when we restrict to employers in the control group. Third, treatment assignment does not predict responses to the gendered-belief battery (Appendix Table A11, Panel A).

[19] The battery of questions is inspired by Macchiavello et al. (2026), who also find that managers hold inaccurate gendered beliefs about women's attributes and skills. In their setting, these beliefs are mostly negative or neutral, rather than offsetting across attributes.

[20] The battery asks about the attributes employers consider most important in workers, a set likely shaped by their predominantly male workforce. In focus groups, asking employers what worried them about hiring male workers largely repeated the general worker-attribute questions, with responses focused on stealing, effort, and cooperation. Asking about women directly allows employers to mention concerns that arise specifically when women enter male-dominated workplaces.

3 Hiring Experiment

We measure employers' preferences for the gender mix at the firm. To motivate the experiment, we first elicit stated preferences with a simple survey question.

3.1 Stated Preferences for Worker Gender Mix

We ask employers, "What is the best gender composition of the workers in your firm?" Employers answer on an out-of-10 scale, from 10 men out of 10 workers to 0 men out of 10 workers. We then recode the response as the preferred number of women out of 10 workers. We ask for a composition rather than a number of women because firms vary in size, and the normalized scale gives small firms, which are the most common in our sample, room to express intermediate preferences. We frame the question as the share of men to make the women aspect less salient.

Using this recoded measure, 85% of employers report preferring at least one woman, and 70.5% prefer two or more (Figure A2). 14.4% state a preference for a roughly gender-balanced workforce of four to six women out of 10. The average preferred female share is 21.7%, about seven times the firm-level female worker share of 3.2%. Most employers attribute the gap between preferred and actual gender composition to supply constraints: there are few women in the applicant pool, which is mostly network-based. The data are consistent with this account, given that the perceived share of women in the applicant pool (3.6%) is similar to actual share of female workers.

3.2 Revealed Preferences: Incentivized Ratings Design

We measure employer demand for trained male and female workers through an incentivized hiring experiment with 903 employers in Kampala, following Kessler et al. (2019). Employers evaluate hypothetical trainee profiles knowing their ratings determine real probationary referrals from a pool of candidates at partner VTIs. Gender and a VTI performance signal are cross-randomized within otherwise identical profiles. Across firms, we manipulate hiring frictions by assigning employers to randomly receive monitoring support via external audits that focus on either theft risk or workplace integration.

The design has three features. First, choices are consequential without deception. Employers know that profiles are hypothetical, but their ratings determine referrals from partner VTIs, giving them access to high-quality candidates outside their usual hiring networks. Second, the experiment measures hiring preferences from a gender-balanced candidate pool. Women are rare in employers' usual applicant flows, so the choice between a male and a comparable female candidate is seldom observed; the balanced pool generates this comparison directly. Holding access to male and female candidates fixed allows us to separate employer preferences from search costs, which are typically higher for women in male-dominated sectors. Third, the design combines two light-touch sources of variation, which are gender-neutral. The scripts do not mention gender and do not affect the selection of profiles employers evaluate.

To measure the relative demand for female workers, we follow an incentivized ratings design (Kessler et al., 2019). Employers evaluate hypothetical trainee profiles, knowing that their ratings determine referrals from a real pool of candidates at partner VTIs. Before rating profiles, employers are told that more accurate ratings lead to better matches. The design therefore makes choices consequential without deception. We measure the share of selected candidates who are women and the hiring gender gap, defined as the male-female difference in selection rates.[21]

The pool from which employers can hire differs from firms' usual applicants in two ways. First, VTI trainees are more skilled: they have completed vocational training, while firms otherwise train workers on the job.[22] Second, women are more represented: 38% of VTI trainees are women, compared to 3.6% of firms' usual applicants. Given that employers could otherwise fill vacancies through standard hiring channels, access to this non-standard hiring pool is part of the value of the incentivized ratings design. The experimental profiles are gender-balanced, so the design measures preferences under a supply environment more favorable to women than the one employers usually face. Estimating demand in this counterfactual supply environment is policy relevant. Programs increasingly train women for higher-return, male-dominated occupations, yet evidence that training translates into employment is mixed (McKenzie, 2017; Bandiera et al., 2022; Goldstein et al., 2019). Whether, and under what conditions, employers would hire qual-

[21] We focus on the hiring margin rather than the wage gap. In this setting, wages are difficult to measure: many apprentices are unpaid, payment structures vary widely across firms (piece rates, fees, stipends), and probationary workers often also receive in-kind compensation which is hard to monetize.

[22] Existing evidence from Uganda shows that vocational training raises sector-specific skills and that firms demand such workers, partly because certification makes their technical skills easier to verify (Alfonsi et al., 2020).

ified female candidates is an open question, and one relevant for understanding when supply-side investments translate into employment.[23]

Profiles. We collect administrative data on 472 trainees enrolled with partner VTIs in Kampala since 2016. Using these data, we create 36 base profiles by randomizing information that VTIs routinely collect from trainees and share with employers: age, marital status, possession of a driver's license, language spoken, motivation for career choice, education, training background, and certification status. In addition, we add the referral contact of the VTI director.[24]

We generate four versions of each profile by cross-randomizing two attributes: gender, signaled by an avatar, and trainee performance, indicated by class rank and visualized as a star rating. Performance signals are anchored to the actual distribution within each VTI cohort: median performers receive three stars, while top performers, above the 95th percentile, receive five stars. This yields 144 unique profiles (4 versions $\times$ 36 base profiles; see Figure 2 for an example). We use the performance signal to test whether the hiring gender gap responds to information about technical performance. The signal is broad: it is an instructor assessment that bundles the three productive attributes employers name as most important apart from trustworthiness and cooperation—technical skills, effort, and learning interest (Figure 1, panel A)—and reflects anything that helps trainees do the practical work on which VTIs assess them. Insofar as it captures a measure of technical skill, we use it to test whether employers use gender as a proxy for technical performance. Of course, employers may value dimensions of skill the signal does not capture, or may not trust the VTI assessment; we return to both in the results (Section 5.1, Section 5.3).

Applicant Pool. Each employer evaluates 24 of the 36 base profiles, each shown in one randomly assigned version, during regular work hours.[25] This creates within-employer variation in candidate gender and performance. The design allows us to estimate the hiring

[23]Employers' preferences for gender and skills are understudied, especially in sub-Saharan Africa (Heath et al., 2024). Existing work documents demand-side discrimination in wages and promotion within sectors where women already work in South Asia (e.g., Brown 2023; Macchiavello et al. 2026).

[24]The information shown in the profiles is designed to mirror the information employers typically request from applicants.

[25]We calibrate the number of evaluations to reflect the average monthly inflow of jobseekers reported by employers during scoping activities.

gender gap conditional on observed performance signals and to test whether stronger signals about productive attributes narrow this gap.

We keep the profile pool gender balanced. This design identifies employer demand for trained VTI candidates when access to male and female applicants is held constant. The design measures hiring preferences with search costs equalized across male and female candidates. This matters because network-based hiring raises the cost of finding female workers in male-dominated sectors (Beaman et al., 2018). Moreover, a pool matching firms' usual applicant flows, where women are nearly absent, would leave little power to test for discrimination.

The gender-balanced profile pool is closer to the VTI referral pool than to firms' usual applicant flows. Across the VTIs we consider, women account for approximately 38% of trainees. Gender composition varies substantially across centers: some centers train mostly women, while others train mostly men. This variation makes it difficult to define a single representative gender composition for the referral pool. It also means that firms' access to female trainees depends on whether they are connected to the newer set of VTIs that train women.

Estimating demand for trained female workers is policy-relevant. Training programs increasingly aim to move women into higher-return, male-dominated sectors (Goldstein et al., 2019). Their returns depend on whether employers are willing to hire trained women once supply constraints are relaxed.

Outcomes. The primary outcome is whether the employer wants to initiate a probationary hire (hereafter, *Hire*). The question was administered in Luganda: "*Wandyagadde tukuweeyo omukozi eyefananyirizako ono okutandiika okugezesebwa mu business yo?*", which translates to the English: "Do you want us to refer you a similar worker to start the probation period at your firm?" (Yes/No). The verb *tukuweeyo* ("send/place with you") and *okugezesebwa* ("to be placed on probation") make explicit that the outcome is the start of a probationary job (as opposed to an interview or another selection stage). In the setting all hiring begins with a probationary period, so this is the relevant margin. Referred candidates are trainees in the firm's own sector: mechanics, welders, or carpenters.[26]

[26]As discussed in Appendix A.4, we preregistered a second outcome: an indicator for whether the employer expected to retain the worker after the probation period. Due to a programming error, this outcome was not elicited in the first wave of data collection (motor mechanics), and we therefore did

Secondary outcomes elicit employers' perceptions of candidates' work quality, trustworthiness, and expected monthly earnings one year after the evaluation. The questions are administered in Luganda and translate into English as follows. *Work quality*: "How would you rate the worker profile's technical 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." *Behavior*: "How would you rate the worker profile's trustworthiness and honesty (good behavior)? Please rate on a scale from 0 to 10, where 0 is not at all trustworthy/honest and 10 is very trustworthy/honest." *Earnings*: "What is your best guess of the monthly earnings of this worker profile a year from now?"

Referrals. At the end of each wave, the research team facilitated matches between interested employers and partner VTIs to fulfill the experimental incentives. To do so, we referred employers to VTI managers (focal points), and provided the focal points with information about employers' worker preferences to facilitate the match. Each focal point received a list of interested employers together with their stated preferences—the type of trainee sought and a flag for whether the employer was seeking female workers. We did not place trainees directly—with far fewer graduating trainees than interested firms, direct placement was in any case infeasible; our role was to make it easier for interested employers and VTIs to match.

Each of the eight partner VTI focal points received information on a subset of employers, together with each employer's stated preferences: the type of trainee sought and a flag for whether the employer was seeking female workers. We flagged an employer as seeking women if their stated optimal number of female workers, elicited at baseline, exceeded the number of women employed at the firm at the time—that is, if the employer had unmet stated demand for female workers. Employers and VTIs were matched by geographic proximity and the share of female workers at the VTI. We communicated the information by dedicated visits.

On the employer side, VTI information was communicated via an SMS in both English and Luganda. The message reminded employers of the study and of the hiring interest they had expressed, named the VTI and the contact details, announced that a successful hire would come with audit support during the probation period, upon verification with both the employer and the VTI. For employers with a stated unmet demand for female

not collect it in the second wave (welding and carpentry). Its omission does not affect referrals: the incentivized hiring choice in the experiment is the probationary hire, whereas retention would have captured beliefs about future match quality.

workers, the message explicitly mentioned female trainees, making sure they knew they could ask for female workers.[27] We tracked referral, matching, and audit outcomes through December 2025. In August 2026, we conducted a follow-up phone survey of all firms in the sample. Implementation details and outcomes are reported with the results (Section 4).

3.3 Mitigating Workplace Frictions: The Audit Treatments

Employers are randomly assigned to one of three experimental conditions: Business-as-usual, Monitoring-Behavior, or Monitoring-Safety. Treatment assignment is stratified by enumerator, survey wave, and sector. The Business-as-usual arm serves as the control group (C). In both treatment arms, trained staff of our partner VTIs, working with our team, conduct weekly visits during the one-month probation period, unannounced in timing. The visits focus on the new worker and follow a familiar format: training institutes routinely visit their trainees during internships, so employers know the practice, and such visits carry weight with new workers, whose standing with the training institute matters for future placements. Two features make the visits meaningful relative to the friction they target. First, they concentrate on a single new worker rather than the whole workforce, so the added capacity is large for the monitoring problem at hand. Second, the program is announced once the new worker is hired: new workers know from the outset that visits may occur, though not when, so deterrence can operate through expectations rather than detection alone. Consistent with this, each audit reduces posterior worry only about the dimension it targets (Section efsec:results), evidence against both spillovers across frictions and demand effects, which would move all reported worries together. Employers are told that visits are randomly assigned and do not reflect characteristics of the firm or the worker. Workers learn about the visits only if they accept the job, so employers cannot anticipate supply-side responses when making hiring choices; anticipated monitoring can otherwise change who applies and how workers fare (Cavounidis et al., 2024).

The two monitoring regimes differ only in the friction they target. Monitoring-Behavior visits are framed as focusing on discouraging new workers' misbehavior, including stealing

[27]The message read: "Hello from Clientele Research Consulting. We previously interviewed you about hiring a vocational trainee on a 1-month probation period [including a female trainee]. At the time, you indicated that you were looking to hire. Are you still interested? If so, please contact VTI [name] to find trainees who match your preferences [including female trainees]. Contact: [name, phone number]. If you successfully hire a trainee, you will receive audit support from the VTI during the 1-month probation period. The hire will be verified with both you and the VTI focal point. If you have any questions, please call us at [number]. You will receive UGX 1,000 in airtime today as a thank-you, please confirm if you received it.".

and dishonesty. Monitoring-Safety visits are framed as ensuring that new workers are safe and treated with respect; we highlight that these visits do not target stealing and dishonesty. The scripts are otherwise identical and are read before employers evaluate the profiles:

- Business-as-usual script: "We are committed to ensuring that managers and workers will have a positive experience if they end up being matched.";
- MB script: "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.";
- MS script: "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. The visits received do not reflect the 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 interventions are designed to target these frictions without priming gender. Neither script mentions gender, women, or men. The Monitoring-Safety script refers to harassment, but never links it to gender: visits ensure that new workers are safe and treated with respect, with no reference to who poses or faces the risk. Employers in both active arms then answer the same comprehension check, whose answer options name safety and harassment alongside stealing and dishonesty, so each concern is mentioned in both arms. This limits the scope for treatment effects to arise from gender salience induced by the treatment scripts.

The design also limits experimenter demand (de Quidt et al., 2026). Employers make consequential choices tied to real probationary referrals, and the exact rating-to-referral

rule is not disclosed. This leaves no simple way to tailor choices to a perceived hypothesis. The treatment contrast is between employers: each employer hears only one gender-neutral script and cannot compare conditions. Candidate gender is visible on each profile, but gender is never named as the object of study and appears alongside many randomized attributes. We elicit beliefs about gendered worker attributes only after the rating task, so these questions cannot prime hiring choices; the only gender-related question preceding the profiles is the stated ideal gender mix, asked more than 30 questions earlier.[28] At the end of the survey, we directly ask employers whether they thought the research team preferred them to hire men, women, or had no preference. Most employers report that the research team was indifferent, consistent with a setting where gender-based affirmative action in private firms is not salient.

Our hypothesis is that our treatments, by reducing workplace frictions that are perceived to bind differently by gender, change the cost of experimenting with male or female workers. Theft concerns raise the cost of hiring men, whom employers perceive as less trustworthy. Cooperation and safety concerns raise the cost of hiring women, whom employers perceive as creating harassment or distraction risks in male-dominated workplaces.

We expect the two audits to move demand for women in opposite directions. Behavior audits reduce the value of trustworthiness and should widen the hiring gender gap. Safety audits reduce workplace integration costs and should narrow it. To verify that each audit shifts the perceived friction it targets, we measure posterior beliefs immediately after the rating task. We ask employers: "How worried are you about workers' lack of trustworthiness, dishonesty, stealing?", "How worried are you about workers' safety and wellbeing?", and "How worried are you about workers' effort and concentration?". Employers respond on a scale from 0 to 10, where 0 is not worried at all and 10 is very worried. The questions directly follow the profile evaluations and precede the referral request, so employers answer them in the context of hiring one of our workers.[29] The first two questions map onto the frictions targeted by the audits. We also ask: "How worried are you about the workers' turnover / leaving the firm?" (asked in the 2024 survey waves

[28] The survey proceeds in this order: stated gender-mix preference and (non-gendered) trait-importance items; the treatment script and comprehension check; the profile-rating task; and only then the gendered items—beliefs about gendered skills, harassment concerns, and the perceived research-team preference. Employers articulate gendered concerns only after all hiring choices are made. Treatment assignment does not predict these stated beliefs (Appendix Table A11, Panel A).

[29] Employers in the within-subject design (Appendix A.5) answer these questions after rating profiles under their first condition only, before the second condition is introduced. Posterior beliefs therefore always reflect exposure to a single script.

only). The effort and turnover questions are placebos: neither script mentions effort or turnover, so neither audit should move them.

The test has power under three conditions. First, employers must base hiring decisions on attributes that are relevant under workplace frictions; otherwise the audits should have no effect on selection. Second, employers must perceive men and women as differing along these attributes; otherwise the audits should not have differential effects by gender. Third, the audits must be strong enough to relax the targeted frictions. Audit-based interventions have affected behavior in other developing-country settings, including reducing corruption (Olken, 2007) and pollution (Duflo et al., 2013).

3.4 Interpreting the Experimental Variation

The audits increase monitoring capacity for the new workers, at no cost to the firm. Over 90% of employers report that they would monitor more hours if they could. Because the visits focus on the new workers, even limited additional hours are not diluted across the workforce. The audit covers the monitoring a new worker would receive under business-as-usual: the median worker gets about two hours of employer monitoring a week, aligned with the audit duration expected by the employer and delivered by the VTIs. They mirror the visits VTIs routinely use to supervise trainees during internships, and are conducted by VTI staff with whom trainees have reputational stakes. Workers learn about the visits at job acceptance and know they can occur at any time, so the audits can deter misbehavior between visits and raise expected monitoring throughout probation. Experimentation with new workers is costly in this market, and temporary audits can produce lasting changes in hiring by lowering the cost of learning about unfamiliar workers, consistent with models of hiring as experimentation (Pallais, 2014; Macchiavello et al., 2026). Other interventions that match firms with trainees use large wage subsidies and more handholding support to induce firms to hire from their pool (Alfonsi et al., 2020). We do not subsidize wages: a subsidy reduces the cost of experimenting with every candidate in the same way, independently of which friction binds, whereas each audit lowers the cost of experimentation by reducing one friction at a time. Still, the audits are a relatively light-touch approach: whether they relax the targeted frictions enough to shift hiring is an empirical question.

Most treatments in the literature show that discrimination depends on information: they vary the availability or precision of information about workers, so that employers update

their beliefs (Goldin and Rouse, 2000; Carranza et al., 2022; Bassi and Nansamba, 2022). Our goal is to test for a different channel: whether workplace frictions are a determinant of discrimination. The audits change the production environment but not the production function, as a new technology would (Alesina et al., 2013). In other words, absent frictions, we expect worker productivity would be identical across arms. The audits also leave the information available to employers unchanged: both audits increase monitoring capacity by the same extent and provide no information about workers. To verify this, we measure beliefs about workers directly and test that they do not vary across treatments: the profile ratings elicit employers' perceptions of candidates' quality and trustworthiness.

Because both treatments involve external audits, spillovers across frictions are plausible. Employers may expect behavior audits to improve workplace integration and safety audits to deter worker misbehavior. Importantly for identification, such spillovers would make the two arms more similar and attenuate the differential effects we predict. Two exceptions are particularly worrisome. First, scrutiny. Under a scrutiny channel, employers respond to being observed rather than to the relaxed friction, and may hire fewer women if they prefer their conduct toward female workers to go unobserved. This is problematic because it offers an alternative explanation for the behavior-audit effect; the safety-audit arm provides the benchmark to test it, as scrutiny would predict lower effects under safety audits too. The second concern is priming. Mentioning theft may direct employers' attention to it, regardless of whether the audits address it. Indeed, to keep the two scripts as close as possible, we mention theft in both, saying in one that visits focus on it and in the other that they do not.[30] This could explain the differential patterns. To test for it, after the experiment we elicit posterior worries directly, including two placebo dimensions, turnover and effort (Section 3.3).

Strategic behavior, meaning that employers tailor ratings to increase the chance of receiving a referral, is unlikely to explain the treatment effects. Employers do not know the exact rating-to-referral rule, but strategic behavior matters for our estimates only if it differs by candidate gender (Litwin and Low, forthcoming). One possibility is perceived scarcity: employers may believe female referrals are harder to obtain and adjust their selections downward accordingly. The design limits this concern. Treatment assignment is random, employers are told that visits do not reflect firm or worker characteristics, and all arms evaluate the same gender-balanced profile pool. Trainees learn about visits only

[30]Employers in the Monitoring-Safety arm are explicitly told that safety visits cannot discourage stealing, dishonesty, or disrespect.

after they are matched, so the treatments cannot change the supply of male or female candidates at the rating stage.

Experimenter demand is also unlikely to explain the treatment effects. The natural concern is that employers, thinking the research team wants them to hire women, may select more women in every arm. This would attenuate the gender gap against women and make differences across arms harder to detect. To explain our treatment effects, demand effects would instead need to have opposite effects across arms. Existing evidence suggests that experimenter demand is unlikely to reverse this kind of comparative-static inference (de Quidt et al., 2026). Our direct demand measure points in the same direction. Most employers report that the research team was indifferent between male and female hires; only 7.1% report believing that the team preferred women. Appendix Table A11 reports these checks: treatment assignment does not predict perceived research-team preferences or post-rating gendered skill beliefs (Panel A), and the treatment effects do not vary with this direct demand item (Panel C).

4 Empirical Strategy and Results

4.1 Empirical Strategy

We restrict the sample to the 888 employers (out of 903) who pass attention checks and exhibit variation in profile evaluations, following our pre-registered main sample protocol.[31] Table 2 shows covariate balance across the three experimental arms, including employers' characteristics and firm-level attributes.[32]

We estimate the hiring gender gap, defined as the difference in selection rates between male and female candidates with otherwise identical trainee profiles, using the following regression model:

$$\text{Hire}_{ijs} = \beta_0 + \beta_1 \text{Female}_{ij} + \delta_i + \sigma_s + u_{ijs}. \tag{1}$$

[31] Sample selection is uncorrelated with treatment assignment (Appendix Table A5).

[32] We 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 (Appendix Table A13).

Hire_{ijs} denotes whether employer j wants to hire the candidate i on probation. Female_{ij} equals 1 if employer j is randomly assigned to the female version of profile i. The coefficient β_1 captures the gender gap, measured as the differential effect for female relative to male candidates. δ_i denote profile fixed effects interacted with the randomized performance signal. σ_s stands for sector, survey-wave, and enumerator fixed effects. Enumerator fixed effects absorb differences in survey administration, including enumerator gender, which may lead to stronger Hawthorne effects. Standard errors are clustered at the employer level.

To identify the average treatment effect of monitoring regimes on the gender gap, we estimate:

$$\begin{aligned}\text{Hire}_{ijs} = \beta_0 + \beta_1 \text{Female}_{ij} + \beta_2 \text{MB}_j + \beta_3 \text{MS}_j \\ + \beta_4 \text{Female}_{ij} \cdot \text{MB}_j + \beta_5 \text{Female}_{ij} \cdot \text{MS}_j + \delta_i + \sigma_s + u_{ijs},\end{aligned} \tag{2}$$

which mirrors the baseline specification in equation 1. MB_j and MS_j are indicators for assignment to the Monitoring-Behavior and Monitoring-Safety arms, respectively; the Business-as-usual arm is the omitted category. The main coefficients of interest are $\hat{\beta}_4$ and $\hat{\beta}_5$, which capture the differential effect of each monitoring regime on the probability of selecting a female profile, that is, the change in the gender gap relative to business-as-usual conditions.

Employers' stated preferences for workforce gender composition are a natural source of heterogeneity in treatment effects. Employers with no stated preference for women are five times more likely to select no female trainees at baseline, leaving limited scope for monitoring to shift their behavior. Therefore, we also estimate a fully interacted version of this model to assess heterogeneity in treatment effects by employers' stated preferences for the gender composition of their workforce.

Employers select a meaningful share of profiles, which makes variation in the gender gap meaningful. If employers selected nearly all profiles or almost none, differences by candidate gender would be mechanically limited and less informative about demand. This is not the case. In the main sample, employers select 11.1 profiles on average out of 24 (standard deviation 3.8).[33]

[33]By construction, no employer selects zero or all profiles. The pre-registered main sample excludes employers who fail comprehension checks or exhibit no variation in profile evaluations, a restriction that excludes only 15 of 903 employers.

We first characterize demand for female workers under business-as-usual (control group). We then estimate how behavior and safety audits affect the hiring gender gap and discuss evidence on the mechanism behind these effects. Finally, we test whether beliefs about technical performance explain the gender gap under the different monitoring regimes.

4.2 Demand for Female Workers and Hiring Gender Gap under Business-As-Usual

Employers show substantial willingness to hire trained female workers under business-as-usual. When given access to a trained applicant pool, 89.6% of employers in the Business-as-usual arm ($N = 308$) select at least one female trainee. Among selected candidates, the average share of women is 42.0% (median 43.8%).

Given the experimental setup, we interpret this level descriptively: it shows that most employers include a meaningful share of women in the set of candidates they want to be referred to when trained female workers are available.[34] The demand we elicit in the experiment is consistent with the gender composition employers say they want. Women are 42% of the candidates employers select. Revealed demand lines up with stated preferences for workforce gender composition they state at baseline. Employers ask us to refer a median of four trained workers (mean 4.6). Applying the 42% selection rate to these requests and adding them to a firm's current workforce gives a post-hire female share of 20.2%, close to the 22.2% ideal share these employers state at baseline.[35]

The main object of interest is relative demand for female versus male trainees, the experimental hiring gender gap, which we estimate using the randomized comparison between otherwise identical male and female profiles. Employers, on average, prefer hiring male over female vocational trainees, all else equal. As shown in column 1 of Table A23, women are 10.7 percentage points less likely to be selected than men with otherwise identical profiles ($p < 0.001$). This corresponds to a 21% gap relative to the mean selection rate of men, and to 1.34 standard deviations of the trainee exam score distribution. As a sanity check, employers who state some preference for hiring women exhibit a smaller gender gap. Excluding employers who state a preference for an all-male workforce reduces the hiring gender gap to 8.5 percentage points ($p < 0.001$; Table A23, column 2).

[34] In the incentivized ratings design, more selections do not map mechanically into more referrals.

[35] We compute this among employers who stated how many workers they wished to recruit under business-as-usual.

A few features of the experiment help interpret the gap between current employment and demand for female workers. Realized employment combines employer preferences with supply-side constraints and search frictions. The effect of making trained women available was ambiguous ex ante. Training may increase employers' willingness to consider women if it makes female candidates observable, signals interest in the sector, and lowers the cost of experimenting with an unfamiliar type of worker. It may reduce the gender gap if employers mainly worry about women's technical skills, especially for high-performance candidates. Access to trained female candidates could even lead employers to prefer women over men if training addresses skills concerns while employers value other traits they associate with women, such as trustworthiness. But training may have little effect if employers, customers, or coworkers are strongly biased against women.

The results point to latent demand for female workers, but not to gender-neutral hiring preferences. Thus, lowering search and matching frictions does not eliminate the role of employer preferences as a barrier. The average hiring penalty against women under business-as-usual is economically meaningful and masks substantial heterogeneity. The gender gap narrows significantly across the distribution of stated preferences for workforce gender mix, as shown in Figure A6, panel A. A 10-percentage-point increase in employers' preferred female share in the workforce is associated with a 4.95 percentage point reduction in the hiring gap against female candidates ($p < 0.001$) (Appendix Table A15, column 1). To summarize this gradient, we compare relative demand for female workers between the bottom and top quintiles of the stated preferences for workforce gender mix distribution. Employers in the bottom quintile select female candidates 23.3 percentage points less often than otherwise identical male candidates ($p < 0.001$). In the top quintile, where employers state an ideal workforce between 40% to 60% women, the hiring gender gap is not statistically significant.

Follow-through. We follow employers' choices beyond the experiment through the referral and audit process described in Section 3.2. The referral SMS—inviting employers to contact their matched VTI and offering audit support in case of a hire—reached 99.2% of the employers contacted, and 81% of employers confirmed having received the message—that is, the information needed to act on their stated hiring interest. Nonetheless, as expected, realized matches were limited by trainee supply: the partner VTIs' Summer 2024 graduating cohort comprised 182 trainees.

Only firms whose match we could verify directly with both the VTI focal point and the employer received audits. When a VTI reported a match, we called the firm and confirmed whether the match was real; if so, we paid the VTI to implement audit visits with their staff during the 1-month probation period. For ethical reasons, given our ex-ante concern for the safety of newly placed workers, especially women, the audit staff was instructed to monitor and report on both safety and behavior, regardless of treatment arm. The audit format otherwise mirrored the one described during the experiment, and during the visits the audit team recorded whether the firm had hired a woman thanks to the experiment.

In total, VTI focal points reported 62 matches; we could verify 57 by phone with the employers, and these firms received the VTI audits. These placements amount to roughly a third of the available trainee cohort. Because counting a match required both the VTI focal point to report it to us and our verification with the employer, some matches are likely to have gone undetected; the 57 audited placements are therefore a lower bound. Table A19 reports the results. During the audit visits the VTI auditor recorded whether the firm had hired a woman: 28% employed at least one woman at the time of the audit, against 11% of the same firms at baseline.[36]

Two years after the experiment we conducted a follow-up phone survey of all firms in the experimental sample, administered by our own team.[37] Table A20 reports the results.

Female employment is substantially higher than at baseline. The share of firms employing at least one woman roughly doubles, from 13.0% to 23.8%, and the female share of the workforce rises from 2.6% to 5.5%; both changes are significant at the 1% level. The average firm employs 0.21 more women than at baseline, against a decline of 0.95 in total employment. Hiring is more frequent than these stocks suggest: 59.7% of firms hired at least one woman at some point since the experiment ended, at a rate of 0.94 women per firm-year, but for an average tenure of 7.0 months, so that at any moment the typical firm employs 0.50 women, or 9.2% of its workforce on average over the follow-up period.

Employers also turned to the training institutes more often. The share naming VTIs as a channel for finding workers rose by 70%. In all, 57% of firms are connected to a

[36]The audited firms that hired women are a highly selected sample; we do not use these data to draw causal or arm-by-arm comparisons. Beyond this observation, we collected no additional survey data during the audits.

[37]We reached 79.4% of the firms in the sample; 97.6% of those we called consented. Panel C of Table A20 compares the baseline characteristics of firms we reached with those we did not: they are statistically indistinguishable on firm size, the female share of the workforce, and the stated ideal composition.

VTI today, either having been in touch since the experiment or currently employing a VTI-trained worker. Firms connected to a VTI are far more likely to employ at least one woman today (41.7% against 10.0%), and hired more than twice as many women per year over the follow-up period (0.98 against 0.44).

Employers' stated ideal composition also shifted, from 21.8% to 26.9% female. The shift is concentrated among firms that actually hired women: those that hired at least one woman since the experiment raised their stated ideal by 6.8 percentage points, against 2.8 points for those that did not, a difference that survives controlling for baseline characteristics. Experimenting with female hires is associated with an increase in stated demand for women in the workforce.

Firms that took on a trainee through the referral of the experiment and received an audit are more likely to have hired at least one woman since the experiment (69.4% against 59.0%).[38]

4.3 The Effect of Audits on the Gender Gap

We compare the selection rate of male and female candidates between employers randomly assigned to receive behavior audits (MB) and those in the business-as-usual condition (C). Our hypothesis is that behavior audits reduce monitoring constraints, lowering the value of trustworthiness in hiring decisions. Because employers perceive women as more trustworthy, this should widen the hiring gender gap.

As a sanity check, we confirm that behavior audits reduce the experimentation cost of hiring from a pool of unknown workers: employers are 4.9% more likely to select a trainee ($p = 0.055$), consistent with audits reducing monitoring constraints (Appendix Table A8).

The higher selection rate under MB operates entirely through a higher likelihood of selecting male trainees, consistent with our hypothesis. Figure 3 shows selection rates by gender across arms: male selection under MB is higher, while female selection rates are comparable across the business-as-usual and MB arms. Monitoring-Behavior is associated with a 5.3 percentage point higher male selection rate relative to the Business-as-usual

[38]This comparison is descriptive: firms that matched with a trainee are self-selected as the matching happened organically between the firm and the VTIs, and the comparison group includes firms that sought a trainee but could not obtain one. Reassuringly, audited firms were no more female-intensive at baseline than the rest of the sample.

arm (Table A23, column 1; $p < 0.001$), with no statistically significant effect on female candidates ($p = 0.685$).

Providing support in monitoring workers' behavior widens the hiring gender gap against women relative to business-as-usual conditions. Table A23 estimates equation 2 on the full sample. The gender gap under behavior audits is 16.7 percentage points, compared to 10.7 percentage points under business-as-usual, a 56% increase. The difference is statistically significant at the 5% level ($p = 0.011$). As an alternative benchmark, the widening of the gender gap induced by behavior audits is equivalent to 53% of the hiring premium of a top- relative to an average-performer male profile (11.3 percentage points; $p < 0.001$).

The effects of behavior audits on the gender gap are stronger among employers with stronger stated preferences for hiring women. Figure 4 reports the gender gap for three subsamples: the full sample, employers with any stated preference for hiring women, and the top quintile of the stated preferred gender mix distribution (40%+ desired share women in the workplace). The first split was pre-registered: we expected that employers with no stated preference for hiring women would be less likely to hire women under business-as-usual conditions, leaving limited scope for behavior audits to widen the gap further. Indeed, employers with no stated preference are nearly four times more likely to select no female trainees at baseline.[39]

The effects are larger with employers' stated preference for hiring women. Among employers who state any preference for hiring women, the gap is 8.5 percentage points under business-as-usual and 15.6 under MB. In the top quintile, it is statistically indistinguishable from zero under business-as-usual and 15.0 percentage points under MB ($p < 0.001$) — a widening of 17.1 percentage points, against 6.0 in the main sample ($p = 0.011$; Table A23), nearly three times larger. Employers who hire without regard to gender under business-as-usual end up, under MB, with a gap close to the 16.7 percentage points observed on average. Appendix Figure A6 shows the same patterns in the raw data.

Stated preferences for gender diversity thus lose most of their predictive power once monitoring frictions are relaxed under MB. Under business-as-usual, employers in the top quintile of stated preferences for hiring women are 1.57 times more likely to select women than those in the bottom quintile; once monitoring frictions are relaxed, this gradient flat-

[39] 31 employers in the Business-as-usual arm who currently employ no women and express a preference for a 100% male workforce nonetheless select at least one female profile in the experiment. Consistent with our mechanism, when asked to explain their apparently inconsistent choices, over 30% cite honesty and trustworthiness as the reason for selecting that female candidate; the remainder state they chose female applicants to give them a chance to prove themselves.

tens. The pattern holds across the full preference distribution: estimating the gap as a continuous function of stated preferences, the preference–gap relationship under business-as-usual is significantly weaker under MB (Appendix Table A15, column 4).[40]

The results are consistent with our hypothesis: support in monitoring workers' behavior and preventing theft widens the hiring gender gap, revealing latent gender gaps especially among employers who are predicted to behave as gender-neutral under business-as-usual conditions. This widening operates through a higher likelihood of selecting men, which is consistent with a perceived reduced risk of theft.

Appendix Table A10 reports the remaining pre-registered dimensions: whether the employer hires from VTIs at baseline, and the employer's gender. As expected, the widening under behavior audits is larger for employers who do not recruit from vocational training institutes ($p = 0.074$). These employers report filling vacancies mostly through personal networks, which typically screen for trustworthiness; audits substitute for that screening, revealing a large gender gap against women. Only 32 of 888 employers are women, so we interpret heterogeneity qualitatively. Their baseline gap is indistinguishable from zero, but behavior audits open a gap of twenty percentage points among them ($p = 0.002$), the same direction as in the full sample.

4.4 Do Safety Audits Affect the Gender Gap?

We compare the selection rate of male and female candidates between employers randomly assigned to receive safety audits (MS) and those in the business-as-usual condition. We expect safety audits to reduce the cost of integrating women into male-dominated workplaces, narrowing the gap. External oversight, however, introduces scrutiny of employers' own conduct, which could deter them from hiring women, widening it. As with behavior audits, we first test how safety audits affect the experimentation cost of hiring from a pool of unknown workers. We find that employers are 7.1% more likely to select a trainee ($p = 0.005$), consistent with audits reducing monitoring constraints rather than deterring hiring through scrutiny (Appendix Table A8).

[40] The linearity assumption does not drive this pattern: allowing a quadratic interaction leaves the pattern unchanged, with none of the quadratic terms statistically significant (column 5). Also, Splitting employers at the median stated preference reproduces the pattern above as presented in Appendix Table A10.

While behavior audits increase overall selection by increasing preferences for hiring men, safety audits affect selection rates only by increasing the preferences for hiring women. In the Monitoring-Safety arm, the likelihood of selecting a female trainee is 7.9 percentage points higher relative to the Business-as-usual arm (Table A23, column 1; $p < 0.001$), with no effect on male candidates ($p = 0.473$). Figure 3 illustrates this pattern: female selection rises from 41.5% to 49.3%, while male selection stays flat.

Safety audits close the hiring gender gap. Table A23 estimates equation 2 on the full sample. The gender gap under safety audits is 1.7 percentage points, compared to 10.7 percentage points under business-as-usual, an 84.1% reduction. The difference is statistically significant at the 1% level ($p < 0.001$). We cannot reject the null that the hiring gender gap under safety audits is zero. These results are inconsistent with scrutiny effects dominating the safety channel. Indeed, safety audits neither deter hiring overall, nor reduce the selection rate of female trainees.

Unlike behavior audits, safety audits do not have larger effects among employers with stronger stated preferences for hiring women. The effect is similar across the preference distribution, except in the bottom quintile, where employers who prefer not to hire women do not respond to safety audits ($p = 0.871$; Table A23, columns 3–7; Appendix Figure A6). Consistent with this pattern, neither the linear nor the quadratic interaction with stated preferences is statistically significant.[41]

4.5 Mechanism Discussion

The results are consistent with our proposed channel: workplace frictions affect the gender gap through the value employers place on gendered attributes. Reducing monitoring frictions lowers the value of trustworthiness, the dimension on which employers perceive women more favorably, lowering the relative demand for women and widening the penalty against them. Addressing cooperation frictions related to the integration of new workers, the dimension on which women are perceived as a liability because of harassment or distraction, increases the relative demand for women and narrows the gap.

The audits reduce the perceived constraints they are designed to address. As shown in Appendix Table A6, the posterior worry measures described in Section 3.3 move in the

[41] Appendix Table A15 reports $p = 0.681$ for the linear interaction (column 4) and $p = 0.801$ for the quadratic interaction (column 5).

predicted directions. Monitoring-Behavior reduces theft worry by 0.62 points on a 0–10 scale (Holm-adjusted $p = 0.004$), consistent with a reduction in perceived monitoring constraints over worker behavior, and does not reduce safety worry. Monitoring-Safety reduces worry about worker safety by 0.47 points (Holm-adjusted $p = 0.069$). The estimated effect of Monitoring-Safety on theft worry is negative but does not survive the multiple-testing correction (Holm-adjusted $p = 0.213$); in particular, safety audits do not *increase* theft worry, evidence against the salience channel. Neither treatment affects the placebo worries, effort and turnover, which neither script mentions. Under a salience account, theft worry should instead move in opposite directions across the two arms: down when audits target behavior, and up when employers hear that safety visits leave misbehavior unaddressed. Selection margins provide a second test. If the text accompanying safety audits primed employers primarily to think about theft, the audits should reduce the selection of men, whom employers view as more likely to steal. Instead, safety audits increase female selection by 7.9 percentage points and leave male selection unchanged ($p = 0.473$).

The magnitude of the audit effects is plausible given the intervention and the hiring margin we study (Section 3.4). Trustworthiness may be especially important in this sample. Employers normally learn about worker behavior through their networks (Heath, 2018), whereas our experiment offers skilled VTI candidates from outside those networks. Employers therefore have better information about candidates' skills than about their trustworthiness.

The preregistered secondary outcomes show that gendered beliefs are reflected in employers' profile evaluations, and that the audits leave those beliefs unchanged. Before each hiring decision, employers rate the candidate on the attributes they say matter, and these ratings reproduce the pattern in Figure 1 on the specific profiles they see: female profiles are rated higher on trustworthiness and lower on quality and effort, holding other profile attributes constant (Appendix Table A7). All four gaps survive correction for multiple testing.[42] The gaps are the same in every arm: no interaction between candidate gender and monitoring regime approaches significance once we account for the eight tests, including trustworthiness, where the largest estimate has an adjusted p-value of 0.40. Employers therefore use gender as a signal of worker attributes, and the audits do not change it.

[42]The ratings show which gender employers expect to be better on each attribute, with women ahead on trustworthiness and men on quality and effort. They do not show what each attribute is worth in hiring, which depends on the cost of the risk and on how much of it the employer can prevent (Appendix A.2). The audits identify that variation.

We next consider alternative explanations for the treatment pattern. First, Monitoring-Safety helps rule out scrutiny as the driver of the gender gap increase under Monitoring-Behavior. The concern is that women's safety may depend partly on employers' own conduct. If some employers anticipate mistreating female workers, or prefer not to have their behavior toward women observed, audits could reduce their willingness to hire women. This scrutiny channel should be strongest in the Monitoring-Safety arm, which explicitly monitors new workers' safety. However, the data show a pattern inconsistent with scrutiny as the main explanation. Monitoring-Safety increases the relative demand for women. This average effect does not mask offsetting responses across employers. The Monitoring-Safety gap distribution shifts left of the control distribution throughout: for each employer subgroup, Monitoring-Safety increases hiring of women relative to men and reduces the hiring gender gap (Appendix Figure A8).

Second, experimenter demand cannot generate the treatment pattern (Section 3.4). Demand favors female hiring and so attenuates the gap: excluding the 7.1% of employers who believe the research team preferred women widens the business-as-usual gender gap from 10.7 to 12.3 percentage points and leaves the treatment effects unchanged (Appendix Table A11, Panel B). Treatment assignment does not shift perceived research-team preferences or gendered skill beliefs (Panel A), and the treatment effects do not vary with the demand belief (Panel C).

Third, the audits may signal that our partner institutes screen candidates carefully, or that we can source and support female candidates. Employers would then select more candidates because they expect better workers, not because monitoring is cheaper. Both audits do raise overall selection. But the two scripts differ in a single clause, so any signal about candidate quality is the same in both arms. It cannot raise male selection under behavior audits (5.3 percentage points, $p < 0.001$, with no effect on women) while raising female selection under safety audits (7.9 percentage points, $p < 0.001$, with no effect on men). A signal about candidate quality would also change how employers rate the candidates, and their ratings of quality and trustworthiness are the same in every arm (Appendix Table A7).

Finally, results are robust across empirical strategies. Within-subject estimates are consistent in sign and larger in magnitude: Monitoring-Behavior roughly doubles the gender gap (12.2 percentage points; $p < 0.001$), while Monitoring-Safety reduces it by 14.9 percentage points ($p < 0.001$; Appendix Table A16). The between-subject estimates are also unchanged when we control for profile order or allow the gender gap and treatment

effects to vary linearly with profile order (Appendix Table A12). They are similar when the sample is restricted by registry timeline, including to observations collected only after the last amendment (Appendix Table A3).

The results show that workplace frictions determine the size of hiring disparities, holding information and preferences fixed. Both treatments carry the same information about candidates and neither mentions gender. Employers rate the candidates no differently across arms, on trustworthiness as on quality and effort (Appendix Table A7), and their stated beliefs about gendered worker traits do not differ either (Appendix Table A11, Panel A). What changes is how traits are valued: behavior audits widen the gender gap by 6.0 percentage points and safety audits close it, and each audit moves employers' worries only on the friction it targets (Appendix Table A6). The mechanism is therefore distinct from belief updating. Monitoring changes the set of attributes that carry value in hiring, and because beliefs about those attributes are gendered, the gap moves with it.

5 Hidden Gender Discrimination

Comparing gender gaps across the three arms isolates two components bundled into the business-as-usual gap: women's trust premium and the penalty from harassment concerns. The remaining female penalty may reflect taste or inaccurate beliefs about other attributes. Under business-as-usual, we observe only the sum of these offsetting components. The identification problem arises because gender signals multiple attributes whose effects on hiring may offset. We use "hidden discrimination" to describe a component concealed by the net gap. Hidden does not refer to a friction-free or true level of discrimination. Each environment produces a different gap because frictions determine which attributes matter for hiring. As frictions multiply, more gendered attributes enter the hiring decision, creating greater scope for discrimination to remain hidden.

If beliefs are accurate, there is no taste-based bias, and researchers observe all relevant components, the net hiring gap reflects the weighted sum of actual differences in job-relevant attributes. If any condition fails, however, the net gap no longer identifies the direction or extent of discrimination because actual differences and distortions may offset. Because a relative premium for women is also a relative penalty against men, the net gap may conceal disparate treatment in both directions. For example, if women are more trustworthy and equally skilled, equal hiring conceals a penalty against women.

Conversely, if men are more technically productive and women's perceived trust advantage is inaccurate, a small net gap conceals inaccurate discrimination against men.

Interpreting these hidden components requires determining whether they reflect accurate group differences. First, high VTI performance does not attenuate the female penalty, and trainee performance rejects accurate screening on technical skills. The remaining penalty is therefore consistent with bias, whether from taste or inaccurate beliefs. Second, trainee behavior rejects women's perceived advantage in trustworthiness. The trust premium thus reflects inaccurate statistical discrimination against men. The business-as-usual gap conceals discrimination in both directions. Finally, we quantify how the estimated bias depends on which attributes researchers measure and treat as productive.

We proceed in three steps. Section 5.1 asks whether better information about skills closes the gap. Section 5.2 asks whether employers' beliefs about women are accurate. Section 5.3 asks how much bias remains once trust and safety are accounted for.

5.1 Technical Performance Signals Do Not Reduce the Gender Gap

The hidden gender penalty could reflect skills-based screening rather than bias. Employers choose among workers with multiple relevant attributes: trustworthiness, integration risk, and production skills. If employers who appear gender-neutral under business-as-usual are hiring women partly for perceived trustworthiness, then Monitoring-Behavior may reduce the value of that attribute and shift hiring weight toward production skills, where employers perceive men as stronger. Under this interpretation, the wider gender gap under Monitoring-Behavior reflects statistical discrimination on production skills rather than bias. If employers use gender as a proxy for unobserved production skills, providing more information about candidates should intuitively help women, if employers hold high beliefs about the quality of men and are more uncertain about the quality of women (Bertrand and Duflo, 2017).

We test this prediction using two sources of variation in the skill information the profiles carry: a cross-randomized performance signal, which contrasts average and top class rank, and the display of a DIT certification, which we also combine with the other credentials on the profile in a linear index. The two margins are complementary. The cleanest test would experimentally vary the availability or precision of skill information, comparing profiles

with and without it, as in Macchi (2023); we did not include such a cell because the experiment is powered to detect the effects of the two audits. The class rank is the sharper margin, cross-randomized within profile; certification, a fixed profile attribute, comes closer to varying how much information employers have. Profiles are randomly assigned a top or average performance ranking within their vocational training cohort, based on evaluations of both theory and practical tasks. We estimate a fully interacted model of hiring on gender, technical performance, and monitoring regimes (Appendix Table 4).[43] As a sanity check, employers place substantial weight on this signal—high-performance profiles are 9.2 percentage points more likely to be selected (21.9%; $p < 0.001$)—indicating that they trust the VTI assessment and view the skills needed to succeed at the VTI as useful on the job.

Returns to the performance signal are slightly lower for women: the interaction between technical performance and gender is negative and marginally significant (-4.1 pp; $p < 0.10$; Appendix Table 4). Results are similar using an alternative profile quality index that equally weights secondary education, DIT certification, enrollment in a 12-month program, and GPA: a one-standard-deviation increase raises the probability of selection by 11.8 percentage points ($p < 0.001$) but does not reduce the gender gap (-1.0 pp; $p = 0.454$; Appendix Table 4).

Certification provides a second source of variation that raises the precision of the skill signal rather than its level. Seven of the twenty-four profiles display a DIT certificate; the remaining profiles show that the certification has not yet been taken. Employers value the externally certified signal—certified profiles are 15.0 percentage points more likely to be selected ($p < 0.001$)—yet the penalty is unchanged: the interaction between gender and certification is small, negative, and statistically insignificant (-0.016; $p = 0.557$), and certification alters neither the widening under Monitoring-Behavior ($p = 0.713$) nor the narrowing under Monitoring-Safety ($p = 0.782$; Appendix Table 4). Making skills externally verifiable, rather than merely raising the reported level, leaves the female penalty intact.

Employers reward these signals, by 9.2 percentage points for a top ranking and 15.0 for a certificate, yet the gap does not narrow at higher performance. If employers used gender to infer the skills the signal conveys, a stronger signal would resolve more uncertainty about women and close the gap. It does not: the interaction runs the other way, and the treatments do not operate through the performance margin either. Solving safety or

[43] We also present sample-split estimates by arm in Figure 5.

monitoring frictions does not shift employers toward higher-performing women, or away from lower-performing ones. Under Monitoring-Behavior, the demand for male trainees rises while the demand for female trainees remains unchanged at both performance levels ($p = 0.871$). Under Monitoring-Safety, the pattern is symmetric: female selection rates rise uniformly across performance levels ($p = 0.758$).

These results reject a story in which gender is a strong proxy for the skills the profiles signal and better information compresses the gap. They are instead consistent with a mild undervaluation of the signals women carry (Phelps, 1972), and with moderate gendered priors that the information does not narrow, a natural case when employers are uncertain about candidate quality. Employers could also statistically discriminate on a productive attribute they believe the signal does not capture—for example, physical strength, the main skill-related concern employers raise about women (Figure 1, panel E). This interpretation, however, requires that employers view strength as unnecessary to perform well at the VTI, whose assessment reflects hands-on practical work. Because none of these beliefs can be identified from the response of the gap to information (Bertrand and Mullainathan, 2004), we test for accurate statistical discrimination directly, benchmarking employers' incentivized beliefs against the trainee pool (Section 5.2), and return to the residual in Section 5.3.

We find no evidence of customer or coworker discrimination: 45% of employers expect profits to increase with more female workers and only 9% expect a decrease, with 84% of the former citing either that customers value women's trustworthiness or that customers would be attracted by women. The business-as-usual gender gap does not differ between employers who believe female workers attract customers and those who do not (Appendix Table A14).

5.2 Beliefs Accuracy

Whether the trust advantage that masks discrimination under business-as-usual reflects efficient screening or a distortion depends on the accuracy of employers' beliefs. For screening to be efficient, employers' beliefs would need to be accurate: women would need to actually be less technically skilled and more trustworthy. We cannot determine whether trust or safety are productive attributes; we can, however, assess whether employers' beliefs about these attributes are accurate. We do so by comparing employers' gendered beliefs about worker attributes to survey data from 182 vocational trainees across seven

training centers, incentivizing employer responses with the true survey results, as proposed by Bohren et al. (2025a).[44] The questions concern the pool we measure—vocationally trained workers seeking jobs in the employer's sector—and accuracy is rewarded against our survey of that pool. Our sample of trainees is described in Appendix Table A17.

We find that employers' beliefs are inaccurate on both dimensions, and in opposite directions: employers understate women's technical skill and overstate their honesty. In the language of Bohren et al. (2025a), both the penalty and the premium women carry are inaccurate statistical discrimination. A natural concern is that this finding reflects sample selection rather than genuine belief inaccuracy: women who self-select into vocational training in male-dominated sectors may be positively selected on skills or other attributes, so the null gender gap in technical performance could reflect selection rather than parity in the broader population. However, the relevant comparison is precisely the pool of vocationally trained candidates that employers in our experiment would hire from.

Technical skills. We assess two measures of technical skills. 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).[45] We also assess theoretical knowledge with a theoretical exam, administered at the vocational training center, with questions specific to their specialization. The theoretical questions were designed by VTI instructors; an exam example is shown in Figure A9. The results of both trainee outcomes and employers' beliefs by gender are summarized in Table 5.[46]

Employers perceive female workers as less skilled, both on practical and theoretical tasks. On average, employers expect women to be 11.8 and 15.2 percentage points less likely

[44]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 can receive a bonus" (up to UGX 10,000). The elicitations ask, for each gender separately: for practical skills, "Out of 10 vocationally trained male [female] workers looking for a job in your sector, how many can perform [the sector-specific task]?"; for theory, after displaying the sector exam question, "Out of 10 vocationally trained male [female] workers, how many could answer correctly?"; for effort, "what was the average number of tasks completed when unmonitored [monitored]?"; and for cheating, "Out of 10 vocationally trained male [female] workers, how many misreported?," where the unsupervised version notes "we had no way to know who lied" and the supervised version "we paid anyway, but we could see who lied."

[45]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 stainless steel surface?"; for carpenters: "Can you perform a wood sanding using a coated abrasive?".

[46]Skill beliefs are incentivized against the trainees' exams, so they were elicited only once the exams for the employer's sector had been scored. We detect no differences on observables relative to the full sample (Table A22).

than men to answer correctly on theory and practical exams, respectively, or 20.4% and 25.1% below the rate they expect for men (both $p < 0.001$). These beliefs are inaccurate: we find no evidence of gender differences in trainee performance. In self-assessments, female and male trainees are equally likely to report being able to perform a standard sector task ($p = 0.511$). On the theory exam, women perform at least as well as men, if anything outperforming them, though the difference is not statistically significant (8.0 percentage points; $p = 0.075$).[47]

Trustworthiness. We design a real-effort over-reporting task to measure workplace dishonesty under varying supervision. Employers define trustworthiness in these terms: the absence of theft and other workplace dishonesty, such as embezzling clients' money.[48]

Our task measures whether trainees misreport their output for pay in a work-like setting. Each trainee completes two five-minute blocks, checking twelve registration papers from the training program for mismatched identifiers. Five of the twelve papers contain a planted mistake. Trainees mark which registrations they check and record the number of mistakes they find. After each block, trainees answer a yes/no question asking whether they found six or more mistakes. An affirmative answer earns UGX 1,000. Because only five mistakes exist, answering yes claims payment for output that does not exist.

The task also tests whether supervision reduces dishonesty. The two blocks differ in whether a false claim can be attributed to the trainee. The unsupervised sheet is anonymous, whereas the supervised sheet displays the trainee's name and ID. The UGX 1,000 bonus for over-reporting is identical across blocks, although the base-payment schedule creates stronger effort incentives under supervision. Following standard practice, we randomize block order and select one block at random for payment.

The design separates effort from over-reporting, two outcomes that misreporting-for-pay tasks often confound. We measure effort by the number of registrations checked. Separately, we observe the number of mistakes recorded and whether the trainee claims the

[47] 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 23.0 percentage points more likely to answer correctly ($p = 0.020$); see Table 5.

[48] Existing experimental and survey measures capture related but distinct concepts. Survey scales measure generalized trust (Falk et al., 2018); trust games measure reciprocity after being trusted or the choice to trust someone else (Berg et al., 1995); and die-rolling tasks measure lying in an abstract environment (Fischbacher and Föllmi-Heusi, 2013). The closest production-setting task studies moral hazard through effort rather than theft (Caria and Falco, 2024). Our task is similar in spirit to the "claim your output" design of Mazar et al. (2008).

bonus for finding more than five. We can therefore condition over-reporting on effort and test whether reports bunch at six, the minimum report that earns the bonus. We can also check whether trainees claim the bonus after recording no more than five mistakes. The supervision contrast is also informative: honest mistakes should not systematically respond to whether the work is observed, whereas cheating should.

Cheating is rare, and we find no evidence that women cheat less than men in our behavioral task. This conclusion is not driven by low engagement: 93% of trainees check all twelve registrations, and 71% identify exactly the five planted mistakes (Figure A10). It is also consistent across our two measures of cheating: over-reporting on the tracking sheet and claiming the bonus. Across both measures, raw gender differences are either negligible or show more cheating among women. In the unsupervised block, 15.7% of women claim the bonus, compared with 7.1% of men. This difference is marginally significant (Fisher's exact $p = 0.08$) and is no longer significant after controlling for VTI and effort. Supervision provides a further check: bonus claims fall from 10.4% to 3.85%, a 63% decline ($p = 0.017$). Because honest mistakes should not respond to observation, this decline supports deliberate misreporting. The response to supervision does not differ significantly by gender ($p = 0.44$). The small number of cheating events, however, limits the precision of this interaction estimate. We therefore interpret the evidence as providing no support for the view that female trainees are more honest than male trainees.

When asked to predict cheating separately for male and female trainees, employers substantially overestimate its prevalence, especially among men, and predict a female honesty advantage not supported by our behavioral data (Table 5).[49] They expect 46.3% of unsupervised men to cheat, compared with an actual rate of 7.1%, and believe women are 11.4 percentage points less likely to cheat ($p < 0.001$). This perceived advantage persists under supervision: employers expect 25.3% of men to cheat and women to be 5.8 percentage points less likely to cheat ($p < 0.001$).

Employers also overstate the effect of supervision. They expect it to reduce cheating by 21.0 percentage points and raise effort by 23.3 percentage points. In the data, bonus claims fall by about 7 percentage points, while effort remains unchanged at 97% of registrations checked. Thus, employers expect frequent cheating among unsupervised workers,

[49]During the employer experiment, we describe the task and the sample of vocational trainees who completed it. We then ask employers to predict average effort and the share claiming the bonus, separately for male and female trainees under supervised and unsupervised conditions. Predictions are incentivized against observed trainee behavior. To avoid deception, the unsupervised tracking sheets are anonymous and cannot be linked to individual trainees; we observe effort and gender on each sheet, but not trainee identity.

especially men, and believe that supervision prevents much of it. These beliefs are consistent with the female trustworthiness premium and the strong hiring response to the behavior audits. These gaps are of comparable size and cross in sign: employers expect women to be worse on skills and better on honesty by similar margins.

Taken together, these results are consistent with employers holding inaccurate gender beliefs about trainee trustworthiness. Our task captures one specific behavior: claiming payment for output that does not exist. However, it combines a standard monetary incentive to misreport with employment-related reputational stakes. We introduced the task as checking Girls With Tools registration records for payment, and the consent form stated that survey information could be used to match trainees with employers. Other survey and behavioral measures point in the same direction. Across measures of trust and social preferences, we find no consistent evidence that women are more prosocial than men, although employers hold pronounced gendered beliefs on almost every measure (Appendix Table A18).[50]

5.3 How Much Bias Do Women Face?

We measure bias within a discrete-choice model of hiring (Appendix A.3), in which an employer's willingness to hire a candidate is a latent value. Relative to an identical man, a woman carries a penalty (or a premium) on this value, which sums the contributions of the valuable attributes on which employers perceive the two to differ and a residual we read as bias. Which attributes carry value depends on the frictions employers face, and each treatment switches one off — behavior audits the value of trustworthiness, safety audits the cost of harassment. This variation lets us measure each contribution and shows that the residual depends on which attributes we account for. In our setting, residual estimates range from 1.7 to 16.7 percentage points.

We decompose the business-as-usual gender gap into a trust premium, a harassment penalty, and a residual we interpret as bias, as standard in the literature. We do not separate a contribution for technical performance: the gap does not respond to the performance signal (Section 5.1), and trainee data show no gender differences in measured skills (Section 5.2). Any belief component insensitive to the information we vary remains

[50] Appendix A.6 describes the measures. The only exception is the decision to trust another player in the trust game: women are less likely than men to do so, consistent with Ashraf et al. (2024). Employers do not anticipate this gender difference.

in the residual, which therefore bundles taste and inaccurate beliefs; the two are not separately identified from hiring behavior (Bohren et al., 2025a). We formalize the decomposition with a discrete-choice model (Appendix A.3), in which the female penalty on the latent value of hiring is the sum of these contributions, and the treatments identify each: every arm removes one friction, and with it the value of the attribute it activates. Identification rests on the assumption that the contributions are additively separable on the latent value — the trust premium does not depend on whether harassment concerns are present, and conversely. While we cannot test directly this assumption, lacking an arm that combines both treatments, we provide correlational support, as, for example, the effect of behavior audits on the gender gap does not vary with employers' perceived harassment risk for women ($p = 0.139$).

We estimate the three components jointly in the full model, then compare the residual bias estimate to simpler specifications. If technical performance were the only productive skill, the full gap of 10.7 percentage points would be interpreted as bias. Adding women's trustworthiness premium of 6.0 p.p. raises the residual to 16.7 p.p. Adding the harassment penalty instead lowers the residual to 1.7 p.p., explaining almost the entire gap. The full model leaves a residual bias of 7.7 p.p. The estimated bias ranges from 1.7 to 16.7 p.p., depending on which skills and attributes are observed and treated as productive.

These estimates make three points. First, residual bias is not monotone in the number of attributes included. Unlike the classic omitted-variable bias logic in Lang and Spitzer (2020), adding attributes need not shrink the residual. It can move the estimate in either direction.

Second, residual estimates of bias depend on what the researcher (observes and) treats as productive. Harassment risk is the clearest case. If harassment lowers women's expected productivity, then avoiding women may look like statistical discrimination. But harassment is imposed on women, not chosen by them, so the same behavior can also be read as discrimination (Buchmann et al., 2023). In a multidimensional model, attributes that affect measured productivity therefore blur the line between statistical discrimination and bias. We take no stance on whether safety concerns reflect altruism, paternalism, or genuine costs to the firm, and we show that this distinction matters for the measurement of bias.

Third, discrimination is measured in a given production environment. Frictions determine which attributes gender signals, so each arm produces a different gap: 10.7 percentage points under business-as-usual, 16.7 under behavior audits, and 1.7 under safety audits.

Each measures the penalty women face under a different set of frictions, and none recovers a friction-free level of bias. The more frictions coexist, the more forms of discrimination a single net gap can hide.

6 Conclusions

We study employers' hiring preferences for women in three male-dominated manufacturing sectors in Uganda. Employers report substantial unmet demand for women. Yet under business-as-usual, they are 10.6 percentage points less likely to select an otherwise identical female profile than a male profile. Behavior audits widen this gap to 16.5 percentage points. Safety audits reduce it to 1.6 percentage points, statistically indistinguishable from zero. These changes match employers' gendered beliefs about technical performance, trustworthiness, and harassment risk.

We interpret the results as showing that workplace frictions can hide discrimination. Under business-as-usual, monitoring constraints raise the value of trustworthiness, which employers mistakenly associate with women. This trust premium masks part of the penalty women face on other dimensions. Behavior audits reduce this premium, widening the gender gap most among employers with the strongest stated preferences for women. The net status-quo gap therefore combines inaccurate statistical discrimination against men on honesty with a larger residual bias against women. A complementary experiment with trainees confirms that employers' beliefs about gender differences in both technical performance and honesty are inaccurate.

Our results also show that allowing for multidimensional worker attributes complicate the identification of statistical discrimination from bias. Gender signals several attributes, some favoring women and others favoring men. Residual estimates can therefore overstate or understate bias, depending on which attributes the researcher treats as productive. Harassment risk illustrates this ambiguity. If harassment lowers a woman's expected match value, avoiding her may appear to be statistical discrimination. Yet because others impose this risk on women, the same decision may also reflect paternalistic discrimination or bias (Buchmann et al., 2023). Each experimental arm measures the penalty women face under a different set of frictions. None identifies a friction-free level of bias.

The mechanism is general but requires three conditions: workers have multiple attributes, employers hold group-specific beliefs about those attributes, and workplace frictions de-

termine their relative weight in hiring. These conditions extend beyond our setting. Soft skills increasingly matter for labor-market outcomes and are difficult to signal, particularly in developing countries (Heckman and Kautz, 2012; Deming, 2017; Carranza et al., 2022). Gendered beliefs about social preferences are widespread and often inaccurate (Exley et al., 2025). The same mechanism may apply to race, ethnicity, and caste.

One limitation is that our experiment measures hiring preferences rather than realized employment. We did not systematically track placements or retention and therefore cannot estimate the experiment's effects on longer-run employment. Audit records confirm that some selections led to hires, but incomplete tracking likely misses matches outside our referral channel. At the two-year follow-up, women's share of the workforce among firms we reached had risen from 2.6% to 5.6% ($p < 0.001$; Table A20).

One final implication of our results is an application of the theory of the second best. When workplace frictions interact, removing one friction need not reduce unequal treatment. Monitoring constraints partly offset the penalty women face by raising the value of trustworthiness, which employers associate with women. Behavior monitoring removes this offset and widens the hiring gap. Gender-neutral improvements in firm organization may therefore worsen women's hiring prospects when other distortions remain.

References

Abebe, Girum, A. Stefano Caria, Marcel Fafchamps, Paolo Falco, Simon Franklin, and Simon Quinn, "Anonymity or Distance? Job Search and Labour Market Exclusion in a Growing African City," *The Review of Economic Studies*, 2021, *88* (3), 1279–1310.

Alesina, Alberto, Paola Giuliano, and Nathan Nunn, "On the Origins of Gender Roles: Women and the Plough," *Quarterly Journal of Economics*, 2013, *128* (2), 469–530.

Alfonsi, Livia, 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.

Amodio, Francesco, Pamela Medina, and Monica Morlacco, "Labor Market Power, Self-Employment, and Development," *American Economic Review*, 2025. Forthcoming.

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*.

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

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

Bandiera, Oriana, Ahmed Elsayed, Andrea Smurra, and Céline Zipfel, "Young adults and labor markets in Africa," *Journal of Economic Perspectives*, 2022, *36* (1), 81–100.

Bassi, Vittorio and Aisha Nansamba, "Screening and Signaling 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.

_ , **Raffaela Muoio, Tommaso Porzio, Ritwika Sen, and Esau Tugume**, "Achieving Scale Collectively," *Econometrica*, 2022, *90* (6), 2937–2978.

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.

Black, Dan A, "Discrimination in an equilibrium search model," *Journal of Labor Economics*, 1995, *13* (2), 309–333.

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, "Stereotypes," *The Quarterly Journal of Economics*, 2016, *131* (4), 1753–1794.

_ , _ , _ , **and** _ , "Beliefs about gender," *American Economic Review*, 2019, *109* (3), 739–773.

Boudreau, Laura E, Sylvain Chassang, Ada Gonzalez-Torres, Rachel Heath, and National Bureau of Economic Research, "Monitoring harassment in organizations," Technical Report, National Bureau of Economic Research Cambridge, MA 2023.

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

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

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

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

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.

Cassan, Guilhem, Daniel Keniston, and Tatjana Kleineberg, "A Division of Laborers: Identity and Efficiency in India," *American Economic Review*, 2022, *112* (10), 3423–3460.

Cavounidis, Costas, Kevin Lang, and Russell Weinstein, "The Boss is Watching: How Monitoring Decisions Hurt Black Workers," *The Economic Journal*, 2024, *134* (658), 485–514.

Chambers, Christopher P and Federico Echenique, "A Characterisation of 'Phelpsian' Statistical Discrimination," *The Economic Journal*, 2021, *131* (637), 2018–2032.

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

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.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan, "Truth-telling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India," *Quarterly Journal of Economics*, 2013, *128* (4), 1499–1545.

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.

Fischbacher, Urs and Franziska Föllmi-Heusi, "Lies in Disguise: An Experimental Study on Cheating," *Journal of the European Economic Association*, 2013, *11* (3), 525–547.

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," *World Bank Policy Research Working Paper*, 2019, (8865).

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.

Heckman, James J. and Tim Kautz, "Hard Evidence on Soft Skills," *Labour Economics*, 2012, *19* (4), 451–464.

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," *Growth*, 2025.

International Labour Organization, "ILO Modelled Estimates and Projections database (ILOEST)," https://ilostat.ilo.org/data/ 2025. Accessed January 7, 2025.

Jayachandran, Seema, "Social norms as a barrier to women's employment in developing countries," *IMF Economic Review*, 2021, *69* (3), 576–595.

Kaas, Leo and Christian Manger, "Ethnic Discrimination in Germany's Labour Market: A Field Experiment," *German Economic Review*, 2012, *13* (1), 1–20.

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.

Litwin, Ashley and Corinne Low, "Measuring Discrimination with Experiments," in Alex Rees-Jones, ed., *Handbook of Experimental Methods in the Social Sciences*, Cheltenham, UK: Edward Elgar Publishing, forthcoming.

Macchi, Elisa, "Worth your weight: experimental evidence on the benefits of obesity in low-income countries," *American Economic Review*, 2023, *113* (9), 2287–2322.

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.

Oh, Suanna, "Does Identity Affect Labor Supply?," *American Economic Review*, August 2023, *113* (8), 2055–83.

Olken, Benjamin A, "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," *Journal of Political Economy*, 2007, *115* (2), 200–249.

Pallais, Amanda, "Inefficient Hiring in Entry-Level Labor Markets," *American Economic Review*, 2014, *104* (11), 3565–3599.

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

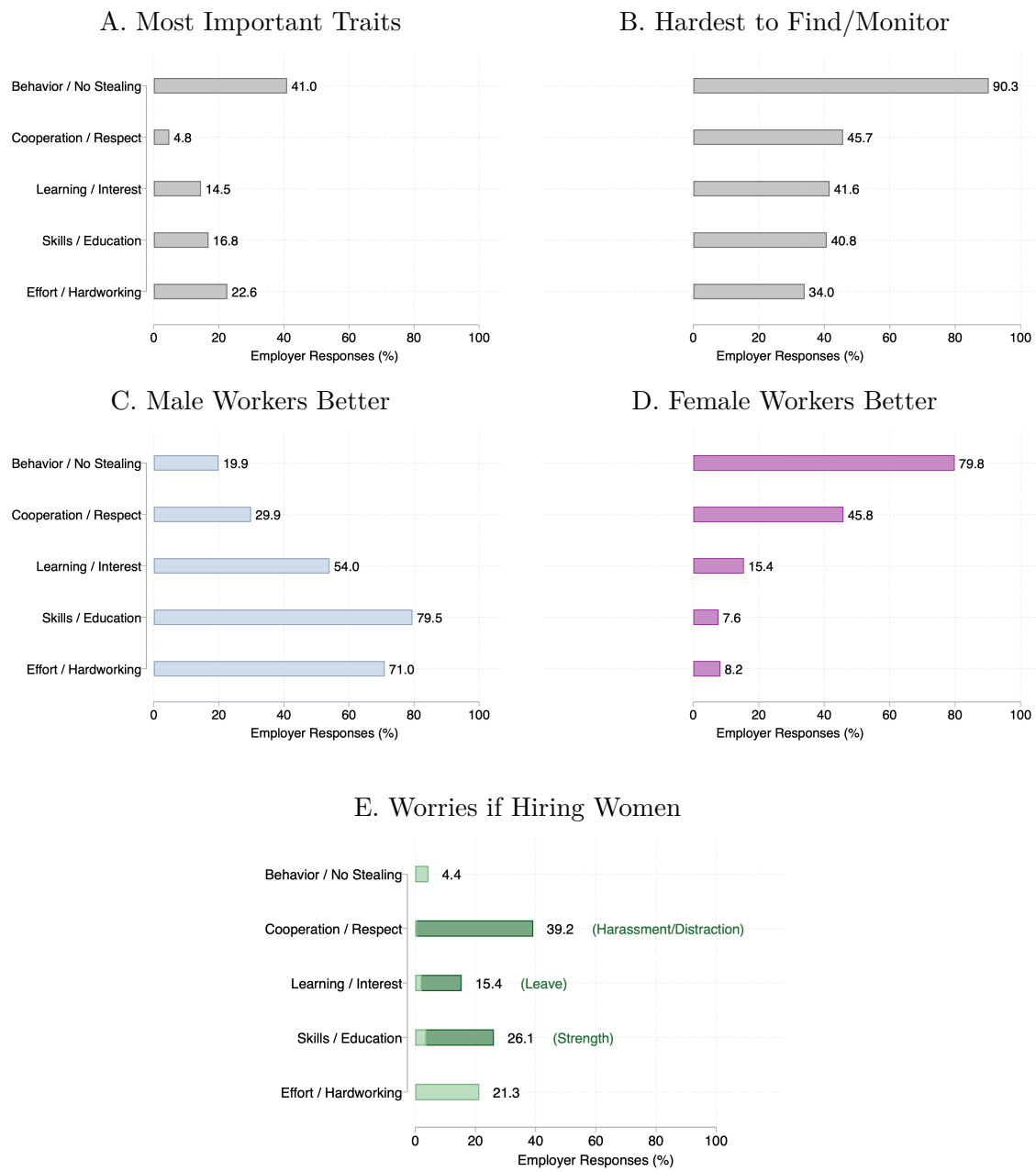
Uganda Bureau of Statistics, "Census of Business Establishments, 2010/11," Census Report, Uganda Bureau of Statistics, Kampala, Uganda 2011.

_ , "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*, 2024, *17*.

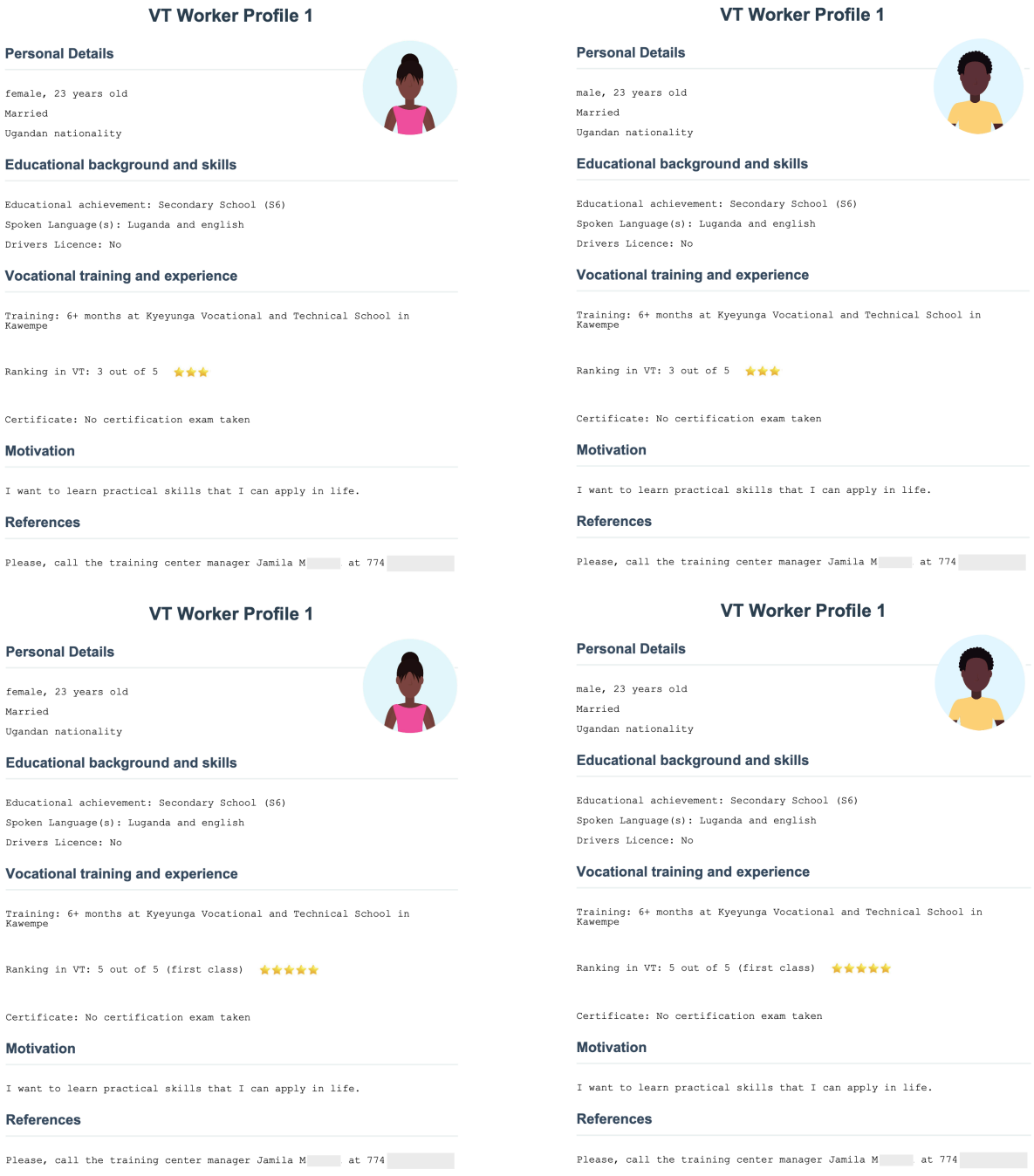
Figures

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



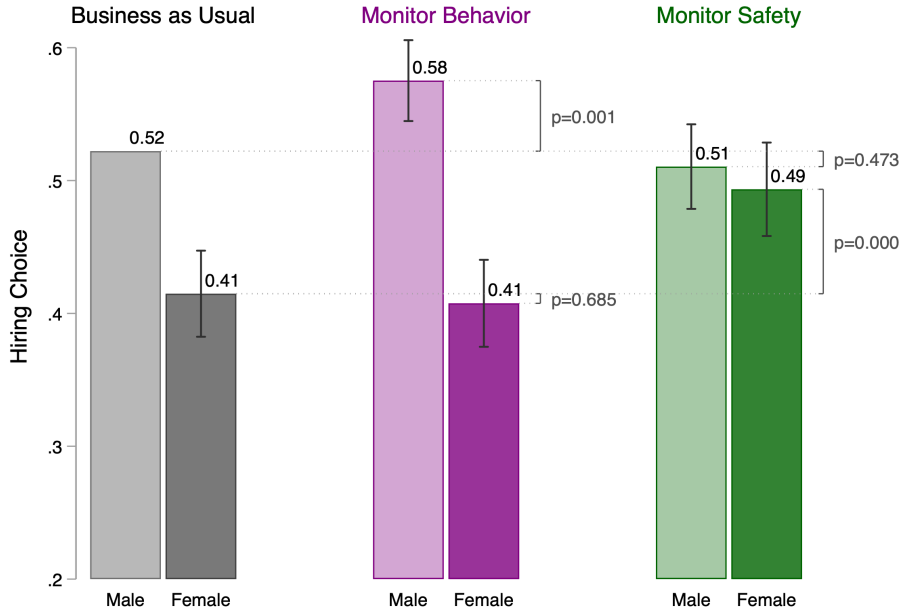
Notes: Panels A–D use the employer sample (N=903); Panel E uses employers with nonmissing responses to the open-ended worry question (N=597). Panel A reports the worker traits employers cite as most important. Panels B–D use closed-choice questions based on the categories derived from the open-ended answers in Panel A. Panel E 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" (10.72% and 1.51%). Open-ended answers in Panels A and 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 hypothetical profile 1 in both female and male versions and for high and low performance levels. In total, there are 24 profiles, with four variations for each profile that reflect the 2×2 design based on gender and performance.

Figure 3: Hiring Choice by Gender and Monitoring Arm



Notes: Bars report the likelihood of the employer to hire the candidate on probation for male and female candidates across the three monitoring arms: Business as Usual (Control), Monitor Behavior, and Monitor Safety. Levels are estimated from a fully interacted specification of candidate gender and monitoring condition as in Equation 2, 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. Monitor Behavior significantly increases hiring of male candidates ($p = 0.001$) with no effect on female candidates ($p = 0.685$), while Monitor Safety significantly increases hiring of female candidates ($p < 0.001$) with no effect on male candidates ($p = 0.473$). $N = 20{,}304$ evaluations from 888 employers.

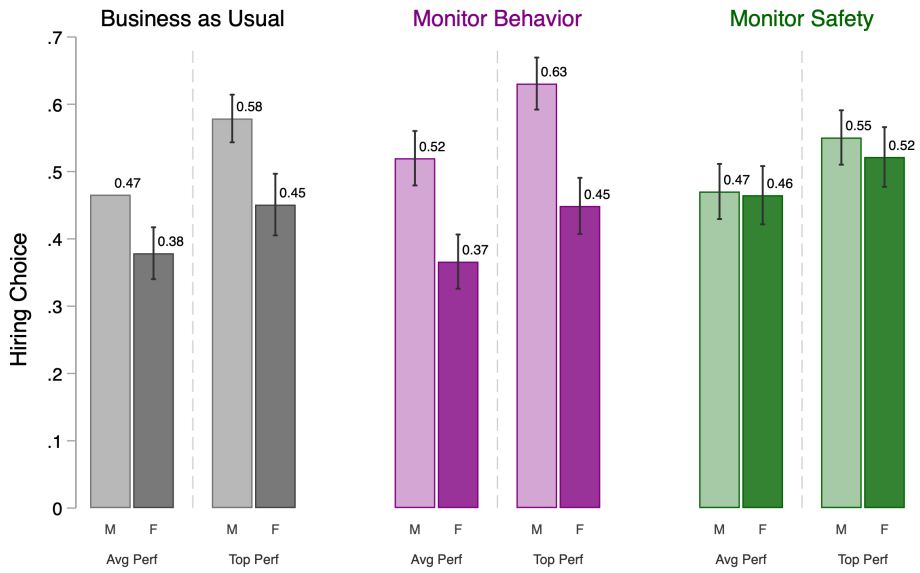
Figure 4: Hiring Gender Gap by Monitoring Arm and Preferences for Workforce Gender-Mix



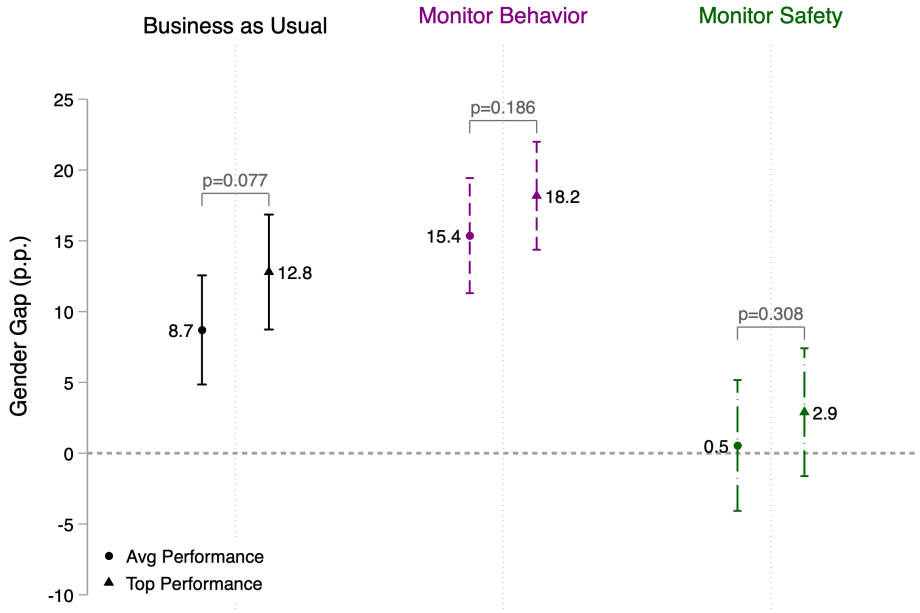
Notes: Data from the experiment. Point estimates and 95% confidence intervals for the gender gap against women are estimated using the regression in Equation 2 and reported by treatment arm. The figure presents results across three levels of employers' stated preferences for workforce gender mix. Circles denote estimates for the full sample (N = 20,304). Squares denote estimates for employers with any stated preference for workforce gender mix (N = 17,421), and triangles for employers in the top quintile of stated preferences (i.e., 40% or more female workers; N = 2,967). 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

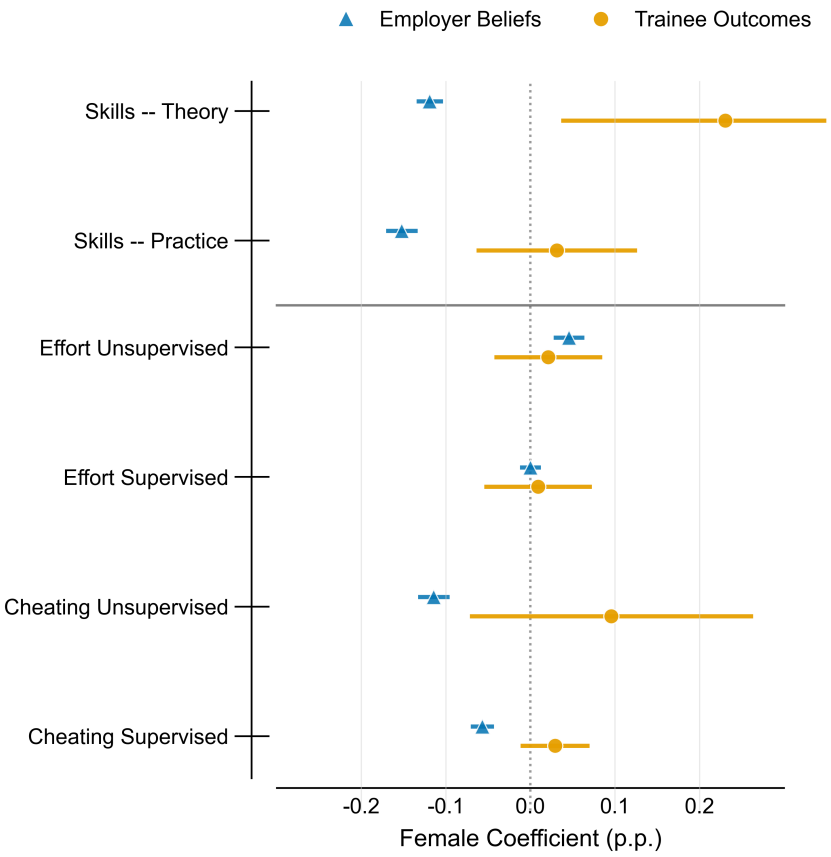


(B) Hiring gender gap



Notes: Both panels come from a single fully interacted model of hiring on candidate gender, technical performance signal, and monitoring regime, 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 respondent level.

Figure 6: Gender Gap in Technical Skills and Trustworthiness



Notes: Data from the baseline employer survey and the trainee survey. Each marker is the coefficient on a female indicator from a separate regression, in percentage points; bars are 95% confidence intervals. For trainees the indicator is the trainee's own gender and the regression includes training-institute fixed effects, with robust standard errors. For employers it indicates whether the belief was elicited about a female or a male trainee; the regression includes sector fixed effects and standard errors are clustered at the employer level. All outcomes are measured as shares: whether the sector theory question was answered correctly, whether the practical task can be performed (self-assessed), the share of twelve tasks completed, and the share lying. Practical technical skills are self-assessed; theoretical skills are elicited via an exam. The Cheating task is described in Section 5.2.

Tables

Table 1: Descriptive Statistics by Sector

	All	Mechanics		Carpentry		Welding	
	Mean	Mean	[Q25;Q75]	Mean	[Q25;Q75]	Mean	[Q25;Q75]
Panel A: Firms							
Workers (N)	8.26	15.33	[7.0 ; 20.0]	4.55	[2.5 ; 5.0]	5.00	[3.0 ; 6.0]
Female workers (N)	0.19	0.19	[0.0 ; 0.0]	0.26	[0.0 ; 0.0]	0.12	[0.0 ; 0.0]
Trainees (N)	2.64	5.54	[2.0 ; 7.0]	0.97	[0.0 ; 2.0]	1.41	[0.0 ; 2.0]
Family workers (share)	0.14	0.07	[0.0 ; 0.1]	0.16	[0.0 ; 0.2]	0.20	[0.0 ; 0.3]
Age (years)	10.77	12.31	[6.0 ; 16.0]	10.01	[5.0 ; 14.0]	9.99	[5.0 ; 13.0]
Firm capital (USD)	1,920	2,917	[741 ; 2,963]	1,385	[370 ; 1,481]	1,613	[741 ; 1,852]
Firm revenue (USD)	1,165	1,101	[556 ; 1,407]	1,016	[370 ; 1,481]	1,388	[556 ; 1,852]
Firm profits (USD)	471	473	[185 ; 556]	400	[185 ; 556]	550	[185 ; 741]
Customers (N/day)	4.07	6.38	[4.0 ; 8.0]	2.87	[1.2 ; 4.0]	2.95	[1.0 ; 4.0]
Any stealing	0.83	0.91	[1.0 ; 1.0]	0.81	[1.0 ; 1.0]	0.77	[1.0 ; 1.0]
Monitor constrained	0.91	0.90	[1.0 ; 1.0]	0.93	[1.0 ; 1.0]	0.89	[1.0 ; 1.0]
Panel B: Employers							
Male	0.96	0.96	[1.0 ; 1.0]	0.96	[1.0 ; 1.0]	0.97	[1.0 ; 1.0]
Age (years)	37.14	40.45	[33.0 ; 47.0]	35.78	[28.0 ; 42.0]	35.24	[29.0 ; 40.0]
Experience (years)	15.25	18.98	[10.0 ; 25.0]	13.19	[6.0 ; 19.0]	13.60	[8.0 ; 20.0]
Higher education	0.63	0.64	[0.0 ; 1.0]	0.62	[0.0 ; 1.0]	0.65	[0.0 ; 1.0]
VTI trained	0.31	0.37	[0.0 ; 1.0]	0.27	[0.0 ; 1.0]	0.29	[0.0 ; 1.0]
Hours worked (per day)	10.42	10.45	[10.0 ; 12.0]	10.43	[9.0 ; 12.0]	10.38	[10.0 ; 11.0]
Hours monitor (per day)	2.06	2.02	[1.0 ; 2.0]	2.10	[1.0 ; 2.0]	2.07	[1.0 ; 2.0]
Observations	903	300		300		303	

Notes: The table reports summary statistics for the main sample, with means and interquartile ranges by sector. Panel A presents firm characteristics and Panel B employer characteristics. Monetary values are in USD. *Family workers (share)* is the share of workers who are family members or close friends. *Any stealing*, *Monitoring constrained*, *Higher education*, and *VTI trained* are binary indicators for having experienced worker theft, desiring more monitoring, completing secondary education, and having received VTI training, respectively.

Table 2: Balance Table

	C (1)		MB (2)		MS (3)		*p*-value	*p*-value	N
	Mean	SD	Mean	SD	Mean	SD	(1)-(2)	(1)-(3)	
Panel A: Firms									
Workers (N)	8.00	(8.52)	8.67	(8.69)	8.08	(9.03)	0.139	0.566	903
Female workers (N)	0.19	(0.55)	0.21	(0.56)	0.17	(0.48)	0.523	0.596	903
Trainees (N)	2.48	(3.48)	2.57	(3.24)	2.86	(5.41)	0.675	0.161	903
Family workers (share)	0.15	(0.25)	0.15	(0.24)	0.13	(0.22)	0.805	0.226	901
Age (years)	10.73	(7.88)	11.31	(8.21)	10.27	(7.17)	0.341	0.478	903
Firm capital (USD)	1915.4	(3305.4)	2362.4	(4304.9)	1515.9	(1570.6)	0.125	0.215	543
Firm revenue (USD)	1085.4	(894)	1151.2	(884.5)	1255.7	(1080.4)	0.631	0.236	531
Firm profits (USD)	438.3	(437)	481.2	(449.2)	494.2	(495.7)	0.393	0.369	526
Customers (N/day)	4.13	(4.38)	4.43	(4.27)	3.64	(2.98)	0.344	0.054	903
Any stealing	0.84	(0.37)	0.85	(0.36)	0.81	(0.39)	0.884	0.247	903
Monitor constrained	0.90	(0.30)	0.92	(0.28)	0.90	(0.30)	0.421	0.817	903
Prefers any women	0.85	(0.36)	0.85	(0.36)	0.86	(0.35)	0.856	0.990	903
Panel B: Employers									
Male	0.97	(0.18)	0.95	(0.23)	0.98	(0.15)	0.181	0.424	903
Age (years)	37.31	(9.89)	36.41	(9.26)	37.7	(10.56)	0.218	0.506	900
Experience (years)	15.22	(9)	15.35	(9.1)	15.17	(9.25)	0.878	0.715	901
Higher education	0.62	(0.49)	0.62	(0.49)	0.66	(0.48)	0.929	0.407	903
VTI trained	0.26	(0.44)	0.33	(0.47)	0.34	(0.47)	0.045	0.047	903
Hours worked (per day)	10.53	(1.53)	10.44	(1.7)	10.29	(1.63)	0.645	0.097	902
Hours monitor (per day)	2.08	(1.60)	2.01	(1.31)	2.10	(1.52)	0.546	0.794	899

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 Behavior (MB), and Monitoring Safety (MS)—with standard deviations shown in parentheses in adjacent columns. Columns (1)–(2) and (1)–(3) report *p*-values from regressions of each outcome on a treatment indicator, including 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. Significance levels are based on randomization inference *p*-values from 10,000 permutations of treatment assignment within strata: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Hiring Gender Gap Across Treatment Arms and by Stated Preferences

			Stated Preferences for Workforce Gender Mix					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Hire (0-1)	Main Sample LPM	Gender-Mix Pref. > 0	Bottom QT	2nd QT	3rd QT	4th QT	Top QT	Main Sample Probit
Female	-0.107***	-0.085***	-0.228***	-0.160***	-0.107***	-0.094***	0.021	-0.108***
	[0.017]	[0.016]	[0.055]	[0.046]	[0.028]	[0.027]	[0.033]	[0.016]
Monitor Behavior	0.053***	0.062***	-0.014	0.039	0.051*	0.060**	0.105***	0.052***
	[0.016]	[0.016]	[0.046]	[0.042]	[0.027]	[0.029]	[0.040]	[0.016]
Monitor Safety	-0.012	-0.016	0.005	-0.057	-0.017	-0.021	-0.005	-0.012
	[0.016]	[0.017]	[0.047]	[0.041]	[0.030]	[0.029]	[0.046]	[0.017]
Female x Monitor Behavior	-0.060**	-0.071***	-0.001	-0.048	-0.053	-0.032	-0.171***	-0.060**
	[0.024]	[0.024]	[0.075]	[0.067]	[0.038]	[0.045]	[0.049]	[0.024]
Female x Monitor Safety	0.090***	0.097***	0.015	0.138**	0.112**	0.099**	0.077	0.091***
	[0.026]	[0.026]	[0.094]	[0.062]	[0.043]	[0.044]	[0.071]	[0.026]
Observations	20,304	17,421	2,883	2,978	6,399	5,077	2,967	20,304
R-squared	0.130	0.129	0.198	0.178	0.149	0.125	0.125	
Control Mean	0.522	0.517	0.549	0.500	0.525	0.535	0.490	0.522

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) 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. Column (8) re-estimates the column (1) specification by probit and reports average marginal effects on the probability scale. 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 Behavior* is a binary indicator equal to 1 if the firm was randomly assigned to audit visits to monitor trainees' behavior. *Monitor Safety* is a binary indicator equal to 1 if the firm was randomly assigned to audit 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 4: Hiring Gender Gap by Skill Signal and Monitoring

	Top Performance		Performance Index		Certificate	
Hire (0-1)	(1)	(2)	(3)	(4)	(5)	(6)
Female		-0.087***		-0.108***		-0.103***
		[0.020]		[0.017]		[0.018]
Monitor Behavior	0.020	0.054***	0.023*	0.053***	0.034**	0.065***
	[0.014]	[0.021]	[0.012]	[0.015]	[0.014]	[0.019]
Monitor Safety	0.045***	0.005	0.033***	-0.012	0.040***	-0.003
	[0.015]	[0.021]	[0.012]	[0.016]	[0.015]	[0.020]
Female × Monitor Behavior		-0.067**		-0.060**		-0.062**
		[0.029]		[0.024]		[0.026]
Female × Monitor Safety		0.082***		0.090***		0.087***
		[0.031]		[0.026]		[0.029]
Ability Signal	0.092***	0.113***	0.118***	0.122***	0.150***	0.157***
	[0.013]	[0.018]	[0.012]	[0.013]	[0.016]	[0.022]
Female × Ability Signal		-0.041*		-0.010		-0.016
		[0.023]		[0.012]		[0.026]
Monitor Behavior × Ability Signal	0.004	-0.002	-0.011	-0.017	-0.040*	-0.046
	[0.019]	[0.025]	[0.012]	[0.014]	[0.023]	[0.031]
Monitor Safety × Ability Signal	-0.024	-0.033	-0.016	-0.012	-0.026	-0.030
	[0.019]	[0.025]	[0.012]	[0.015]	[0.023]	[0.031]
Female × Monitor Behavior × Ability Signal		0.013		0.012		0.014
		[0.031]		[0.017]		[0.038]
Female × Monitor Safety × Ability Signal		0.017		-0.007		0.010
		[0.033]		[0.018]		[0.037]
Observations	20,304	20,304	20,304	20,304	20,304	20,304
R-squared	0.115	0.129	0.115	0.128	0.022	0.036
Control Mean	0.421	0.466	0.431	0.475	0.475	0.475

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 Behavior* is a binary indicator equal to 1 if the firm was randomly assigned to audit visits to monitor trainees' behavior. *Monitor Safety* is a binary indicator equal to 1 if the firm was randomly assigned to audit visits to monitor trainees' safety. *Ability Signal* denotes a different measure in each column pair. 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 are clustered at the respondent level and reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Gender Gap in Skills and Trustworthiness: Trainee Behavior and Employer Beliefs

	Trainee		Employer	
	Actual		Beliefs	
Panel A. Technical skills				
	Theory	Practical	Theory	Practical
Female	0.230**	0.031	-0.118***	-0.152***
	[0.098]	[0.047]	[0.007]	[0.008]
Mean, men	0.759	0.920	0.579	0.605
Observations	182	182	584	585
Panel B. Cheating task				
	Effort	Claimed (≥ 6)	Effort	Claimed (≥ 6)
Pooled across supervision conditions				
Female	0.015	0.051	0.023***	-0.086***
	[0.022]	[0.039]	[0.005]	[0.006]
By supervision condition				
Female	0.024	0.074	0.046***	-0.114***
	[0.027]	[0.060]	[0.008]	[0.008]
Supervised	-0.001	-0.054*	0.233***	-0.210***
	[0.007]	[0.030]	[0.006]	[0.009]
Female × Supervised	-0.018	-0.046	-0.046***	0.056***
	[0.033]	[0.058]	[0.008]	[0.008]
Mean, unsupervised men	0.989	0.071	0.650	0.463
Observations	364	364	898	896

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

A Appendix

A.1 Classifying Open-Ended Answers

Some of the descriptive evidence in Figure 1 comes from open-ended survey questions. Panel A uses answers to "What is the single most important trait you look for when hiring a worker?" 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 Panel A, enumerators recorded open-ended answers for 618 employers. 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 remaining 303 employers, enumerators assigned each answer directly to the closest category during the survey. Enumerators did not read the categories aloud. Employers answered in their own words, and enumerators coded the response using the categories derived from earlier open-ended answers: behavior/respect/trust/no stealing, effort/hardworking/discipline, skills/education, learning/flexibility/motivation/interest, background/history, age, cooperation, and other.

Although we could use the manually coded categories directly, Figure 1, panel A reports model-assisted coding for replicability. 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 618 model-coded open-ended answers with the 303 mechanically mapped responses yields 903 observations for Figure 1, panel A.

For Panel E, we classify the 615 nonmissing open-ended answers 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 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 Conceptual Framework

We formalize the descriptive evidence in a simple model of hiring with multidimensional workers. Employers value three attributes in a worker: technical skill, honesty, and safety from harassment. Each attribute is high (one) or low (zero), $a = (a^S, a^T, a^H) \in \{0,1\}^3$, and none is visible at hiring. The employer sees the candidate's gender $g \in \{m, f\}$ and a performance signal q, and holds a gender-conditional guess about each attribute,

$$p_g^S \equiv \Pr\big(a^S{=}1 \mid g\big), \qquad p_g^T \equiv \Pr\big(a^T{=}1 \mid g\big), \qquad p_g^H \equiv \Pr\big(a^H{=}1 \mid g\big).$$

Assumption 1: differential signalability. The attributes differ in how visible they are before hiring. The performance signal updates the skill guess by Bayes' rule, $p_g^S \to \hat{p}_g^S(q)$. Honesty and safety have no pre-hire signal: those guesses stay at the priors.

Assumption 2: crossing priors. The priors are about candidates in the pool: what the employer guesses about this profile, a VTI-certified trainee, given everything on it except gender. Employers see women as more honest, with prior gap $\Delta^T \equiv p_f^T - p_m^T > 0$, and less safe, $\Delta^H \equiv p_m^H - p_f^H > 0$, consistent with their stated beliefs and concerns (Figure 1). The skill prior is left free. The sign of $\Delta^S \equiv p_m^S - p_f^S$ is open: employers may think men are stronger, as in the general population; they may think men and women are equally good once in VTI training; or they may think women are better, because the few women who enter these sectors are highly selected. The model takes no stance: skill has a signal, so this prior can be measured; honesty and safety do not, so theirs must be assumed.

The value of a hire. Firm j values hiring candidate i at

$$V_{ij} = y_L + \Delta y\, \hat{p}_g^S(q) - (1-\mu_j)\,\theta\,(1-p_g^T) - (1-\sigma_j)\,\kappa\,(1-p_g^H) - d\,\mathbf{1}\{g=f\}, \tag{3}$$

and hires when V_{ij} beats its outside option, the candidates it can reach through its network, plus idiosyncratic noise, so the chance of hiring rises with V_{ij}. Reading the terms: y_L is the output of a low-skill worker and Δy the premium a skilled one adds. A dishonest worker costs θ, and the firm bears the share $1-\mu_j$ it cannot prevent: μ_j is its capacity to prevent theft, $1-\mu_j$ its monitoring friction (Section 2.2). A harassed worker costs κ, and the firm bears the share $1-\sigma_j$ it cannot prevent: σ_j is its capacity to prevent harassment, $1-\sigma_j$ its cooperation friction. The residual d is a penalty on women. A firm with $\mu_j = 1$ does not care who is honest; a firm with $\sigma_j = 1$ does not care about safety. The two risk costs enter the payoff separately and the attributes are independent conditional on gender; this separability is what the decomposition below rests on.

The gender gap. Between male and female profiles identical in everything the employer sees except gender,

$$G_j(q) = \Delta y\, \Delta^S(q) - (1-\mu_j)\,\theta\,\Delta^T + (1-\sigma_j)\,\kappa\,\Delta^H + d, \tag{4}$$

where $\Delta^S(q) \equiv \hat{p}_m^S(q) - \hat{p}_f^S(q)$ is the belief gap after the signal. If a skill prior against women survives conditional on VTI certification, the first term penalizes them; if $\Delta^S = 0$ in the pool, it is zero at every signal. Men carry the bigger expected theft cost, so the honesty term is a subsidy to women—the one piece working in their favor, worth $(1-\mu_j)\theta\Delta^T$ and nothing at $\mu_j = 1$. Women carry the bigger expected harassment cost, a penalty of $(1-\sigma_j)\kappa\Delta^H$. The observed gap is penalties net of one subsidy: it understates what works against women, and it can be small while every force in it is large. This is the hidden gap.

Prediction 1: mitigating the two frictions moves the gap in opposite directions. From (4), $\partial G/\partial \mu = \theta\, \Delta^T > 0$ and $\partial G/\partial \sigma = -\kappa\, \Delta^H < 0$. Mitigating the monitoring friction removes the honesty subsidy and widens the gap; mitigating the cooperation friction removes the harassment penalty and narrows it. An intervention that mitigates a friction changes no belief about any candidate: only the price of a risk moves. The opposite signs exist only because the priors cross.

Prediction 2: how the gap responds to the performance signal. The capacities and d do not vary with the signal, so every term of (4) cancels across signal levels except the skill term: $G(\text{high}) - G(\text{low}) = \Delta y\, [\Delta^S(\text{high}) - \Delta^S(\text{low})]$. If employers hold a strong skill prior against women and the signal is informative, a favorable signal destroys doubt, women have more doubt to destroy, and the gap narrows at high signals. If $\Delta^S = 0$, the gap does not vary with the signal. The converse is not exact: with moderate priors, the posterior gap can be similar at the two signal levels, so a flat gap rejects the strong-prior case but not a moderate prior that the signal leaves in place; any such component loads on the residual. The level response is the companion check: hiring must rise with the signal overall, which certifies that employers read and believe it, and returns that differ by gender indicate that employers discount or mark up the signal when a woman carries it.

In the experiment, the two audits mitigate the monitoring and the cooperation friction, one at a time (Section 3.3), and the profiles carry the cross-randomized performance signal (Section 3.2). The decomposition below uses the model to split the observed gap into a trust premium, a harassment penalty, and a residual.

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

This appendix estimates bias against women residually, in a multidimensional skill framework. Workers are bundles of attributes whose value depends on the frictions employers face, so the hiring gender gap reflects gender differences in each attribute employers value; bias is what remains once these attributes are accounted for. Three attributes are at play in our setting—technical skills, trustworthiness, and safety—and they differ in how we identify their contribution. Technical skill is signaled directly: we cross-randomize a performance signal on each profile, so it can be conditioned on as a control would be in a standard wage regression. Trustworthiness and safety cannot be randomized or read off a profile; they are latent, identity-linked attributes whose value depends on the production environment. We therefore back them out from the experimental conditions—the monitoring treatments remove the frictions that make each attribute matter, and the induced movement in the gap identifies its contribution. As the framework implies, the resulting estimate of bias depends on which attributes the researcher measures and treats as productive.

Discrete-choice model. Consider employer e evaluating profile i with gender $g_i \in \{m, f\}$ and a cross-randomized technical-performance signal $q_i \in \{\text{low}, \text{high}\}$, in arm $a \in \{\text{C}, \text{MB}, \text{MS}\}$. Let U^*_{ei} be the employer's latent net value of hiring the candidate,

$$U^*_{ei} \;=\; \mathbf{x}'_{ei}\boldsymbol{\beta} \;+\; \mathbf{1}[g_i = f]\, \pi(a, q_i) \;+\; \varepsilon_{ei}, \qquad \varepsilon_{ei} \sim \mathcal{N}(0,1), \tag{5}$$

and the employer selects the candidate, $\mathit{Hire}_{ei} = 1$, whenever $U^*_{ei} > 0$. The controls $\mathbf{x}_{ei}$ (profile, performance, and strata fixed effects) are common to men and women; $\pi(a, q)$, the female penalty on the latent index, is the object we decompose.

The penalty follows from the model. Firm j's gap in hiring value between identical male and female profiles is $G_j(q)$ in equation (4). The arms change only the capacities: business-as-usual leaves (μ_j, σ_j) at baseline, the behavior audit sets $\mu_j = 1$, the safety audit sets $\sigma_j = 1$. Averaging $G_j(q)$ within arm—randomization makes baseline capacities comparable across arms, with means $(\bar{\mu}_0, \bar{\sigma}_0)$—gives the penalty

$$\pi(a,q) \;=\; \underbrace{B}_{\text{residual}} \;-\; \underbrace{T}_{\text{trust}}\,\mathbf{1}[\text{trust friction in } a] \;+\; \underbrace{H}_{\text{safety}}\,\mathbf{1}[\text{safety friction in } a] \;+\; \underbrace{S}_{\text{skill}}\,\mathbf{1}[q = \text{low}], \quad (6)$$

with $T = (1-\bar{\mu}_0)\,\theta\,\Delta^T$, $H = (1-\bar{\sigma}_0)\,\kappa\,\Delta^H$, $S = \Delta y\,[\Delta^S(\text{low}) - \Delta^S(\text{high})]$, and $B = \Delta y\,\Delta^S(\text{high}) + d$. The additive form is inherited from the model, where the two risk costs enter the payoff separately and the attributes are independent conditional on gender: the trust premium does not depend on whether the safety friction is present, and conversely. The data support this restriction: the behavior-audit effect on the gender gap does not vary with employers' perceived harassment risk ($p = 0.139$). The mapping also assumes the audits set the capacities to one and change nothing else; the caveats below discuss partial removal and spillovers. Differencing across cells identifies the components:

$$S = \pi_{\text{low}} - \pi_{\text{high}}, \qquad T = \pi_{\text{MB}} - \pi_{\text{C}}, \qquad H = \pi_{\text{C}} - \pi_{\text{MS}}, \qquad B = \pi_{\text{MB}} + \pi_{\text{MS}} - \pi_{\text{C}}, \quad (7)$$

where each π is the male–female gap in the relevant cell.

Testing each channel. We estimate (5) by probit and test each channel against zero. The trust and safety channels are statistically significant (Table A1); the skill channel is the only one we cannot reject: the gender gap does not respond to the performance signal ($S = -2.9$ p.p., $p = 0.17$; the analogous control-arm test in Section 5.1 likewise fails to reject), and male and female trainees show no difference in measured skills (Section 5.2). We therefore impose $S = 0$. Imposing it leaves the just-identified decomposition $T = \pi_{\text{MB}} - \pi_{\text{C}}$, $H = \pi_{\text{C}} - \pi_{\text{MS}}$, and the residual $B = \pi_{\text{MB}} + \pi_{\text{MS}} - \pi_{\text{C}}$.

We recover the components on the latent index, where the independent channels add up, and let the nonlinear link map the index to selection probabilities, so the channels are not constrained to add on the probability scale.

Table A1, column 1, reports the components on the latent index, in probit (standard-normal) units, on the main between-subjects sample and with the same fixed effects and clustering as Table A23.

Column 2 expresses the same components on the probability scale, as average male–female differences in predicted selection probability (in percentage points), so they can be read as shares of the hiring gender gap rather than in the latent model's standard-deviation units. These figures coincide with a linear probability model to within a tenth of a percentage point, because selection rates lie near one half — where the link is approximately linear — and the model is saturated in gender × arm; floor and ceiling effects are therefore not a concern.

The residual depends on which attributes are treated as productive. What the researcher attributes to the residual depends on which friction-linked attributes she measures and treats as productive, and our arms trace out the cases (Table A1, Panel C). Conditioning on technical performance alone, the whole business-as-usual gap is residual (π_{C}), because the gap does not respond to the performance signal. Treating trustworthiness as productive removes the offsetting trust premium and *raises* the residual (π_{MB}); treating safety as productive removes the harassment penalty and *lowers* it (π_{MS}); treating both leaves B. The residual thus ranges from 1.7 to almost 17 percentage points depending on the researcher's assumptions. Two points follow. Omitting an attribute on which women are *favored*—trustworthiness—*understates* the residual, reversing the usual direction of omitted-variable

bias (Lang and Spitzer, 2020). And the sign of the correction depends on a stance the researcher must take—whether a friction-linked attribute is productive or itself a form of discrimination—which is first-order for the estimate. Whether the residual reflects bias is the interpretation we defend in Section 5.3, drawing on the absence of skill differences, the inaccuracy of employers' beliefs, and the lack of customer or coworker discrimination.

How much of the hidden gap is bias. Relative to the business-as-usual residual, separating the trust premium reveals $T = 6.0$ p.p. of hidden discrimination. How much is bias depends on whether harassment is treated as productive: if it is, the $\pi_{\text{MB}} = 16.7$ p.p. residual is 46% bias ($B = 7.7$) and 54% harassment ($H = 9.0$), and the hidden gap splits in the same proportion, $T \cdot B/(B + H) \approx 2.8$ p.p. of bias and $T \cdot H/(B + H) \approx 3.2$ p.p. of harassment. Under the most conservative reading, the hidden bias amounts to 26% of the observed business-as-usual gap.

Caveats. First, the decomposition assumes no spillovers across audits: each treatment switches off only its targeted friction. Because the design is just-identified—three arms recover exactly three components—this assumption is observationally equivalent to additivity, the restriction that the trust premium and harassment penalty do not interact. Any spillovers would push the two arms toward each other and work against the opposite-signed effects we find; consistent with treatments operating on their targeted friction, each audit shifts employers' post-experiment worries only on its own dimension (Appendix Table A6). If a treatment only partially removes its friction, $|T|$ and $|H|$ are understated and the estimates should be read as bounds.

Second, the decomposition treats each attribute as either fully productive or fully non-productive. In practice each likely combines the two: the trust premium may partly reflect genuine productivity and partly inaccurate beliefs, and harassment concerns may partly reduce women's output and partly reflect paternalism. The extreme cases bound the bias estimate.

Third, employers may value productive attributes beyond technical performance, trustworthiness, and safety that we do not observe. Any such attribute would constitute a fourth dimension that our design does not separately identify, and its effect remains in the residual. A natural candidate is physical strength, the main skill-related concern employers raise about women (Figure 1, Panel E). If employers expect strength to be correlated with the instructor performance signal, then the cross-randomized signal partly captures it. If not, strength remains in the residual: under a four-dimensional framework that treats strength as productive, our residual overestimates bias, and should be read as an upper bound rather than as bias itself. More generally, this is central to our measurement contribution: with multidimensional skills, the residual can be interpreted as bias only relative to the attributes the researcher measures. What remains conflates bias with any productive dimension not separately identified.

Table A1: A Discrete-Choice Decomposition of the Hiring Gender Gap

	(1) Latent index	(2) Probability (pp)
Panel A. Hiring gender gap by arm		
Business-as-usual	0.30	10.81
	[0.05]	[1.65]
Monitoring-Safety	0.05	1.74
	[0.06]	[2.01]
Monitoring-Behavior	0.47	16.83
	[0.05]	[1.70]
Panel B. Decomposition components		
Trust premium (T)	0.17	6.01
	[0.07]	[2.37]
Safety penalty (H)	0.25	9.07
	[0.07]	[2.60]
Residual (B)	0.22	7.75
	[0.09]	[3.11]
Panel C. Residual by attributes treated as productive		
Technical performance only	0.30	10.81
	[0.05]	[1.65]
+ trustworthiness	0.47	16.83
	[0.05]	[1.70]
+ safety	0.05	1.74
	[0.06]	[2.01]
+ trustworthiness and safety	0.22	7.75
	[0.09]	[3.11]
Employers	888	888
Observations	20,304	20,304

Notes: The table reports the discrete-choice (probit) decomposition of the business-as-usual hiring gender gap into a trust premium (T), a safety 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_{MB}, π_{MS}, and B). Under the assumption that the two treatments do not interact, $T = \mathrm{MB} - \mathrm{C}$, $H = \mathrm{C} - \mathrm{MS}$, and $B = \mathrm{MB} + \mathrm{MS} - \mathrm{C}$, where C, MB, and MS denote the gaps under business-as-usual, Monitoring-Behavior, 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 A23.

A.4 Preregistration

We note three main deviations from the first pre-registration. 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.

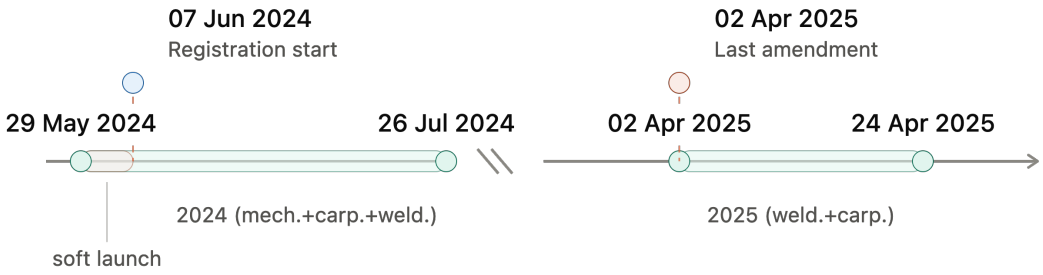
Second, we expanded the study to include two additional sectors—carpentry and welding—in addition to mechanics. As preregistered, we planned to expand data collection to additional sectors if power proved insufficient. We observed substantial heterogeneity in stated preferences for hiring women, and the additional sectors provide sufficient power to study treatment effect heterogeneity by stated preferences.

Third, because the number of firms per new sector in the second wave was uncertain, we added a *within-subject* component. After completing the initial 24 evaluations under the main treatment condition, each employer rated 12 additional profiles under a second monitoring condition.

Fourth, although the safety audit arm was preregistered as an active control or benchmark, we report its results as an additional treatment of interest given their policy relevance.

We amended the preregistration before expanding the sample to reflect these changes (except the last one) and to document the motivations. Table A3 shows that the main effects are similar when the sample is restricted to observations collected after the registration start date or after the last amendment. Table A2 lists each outcome and analysis with its pre-registration status.

Figure A1: Timeline of IRR 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 (02–24 April 2025) covers welding and carpentry. The period between 29 May and 07 June 2024, prior to the registration start date, denotes the *soft launch*, during which the data collection at scale was tested. The two markers above the timeline indicate the registration start date (07 June 2024) and the last registry amendment (02 April 2025), which coincides with the start of the 2025 wave.

Table A2: Pre-Registration Status of Outcomes and Analyses

	Status	Date	Note
Panel A. Primary outcomes			
Hire (interest in candidate)	Pre-registered	07 Jun 2024	
Retention (likelihood of retention)	Pre-registered; not collected	07 Jun 2024	Programming error in wave 1; dropped in wave 2.
Panel B. Secondary outcomes			
Quality (perceived skill)	Pre-registered	07 Jun 2024	
Behavior (trustworthiness, honesty)	Pre-registered	07 Jun 2024	
Effort (hardworking, punctual)	Pre-registered	07 Jun 2024	
Earnings (expected monthly, 1 yr)	Pre-registered	07 Jun 2024	
Panel C. Treatment arms and comparisons			
BAU vs. Monitor Behavior	Pre-registered	07 Jun 2024	
BAU vs. Monitor Safety	Pre-registered (active control)	07 Jun 2024	
Within-subject second monitoring condition	Amended	02 Apr 2025	Excludes mechanics (wave 1).
Gender × ability interaction	Pre-registered	07 Jun 2024	
Panel D. Heterogeneity and mechanism analyses			
By preference for workplace gender diversity	Amended	02 Apr 2025	Diversity-preference measure and heterogeneity added in amendment.
By preference quintile (specific cutoff)	Exploratory	—	Cutoff not pre-specified. Results are robust to different cutoffs and continuous specifications.
By VTC hiring and manager gender	Pre-registered	07 Jun 2024	
Beliefs vs. actual trainee behavior	Pre-registered	07 Jun 2024	

Notes: "Pre-registered" = specified in the original preregistration (AEA RCT Registry AEARCTR-0013698, registration date 07 June 2024); "Amended" = added in the 02 April 2025 amendment, filed before the corresponding fieldwork began (Fig. A1); "Exploratory" = not pre-specified. The amendment expanded the study from motor mechanics to welding and carpentry, added the within-subject monitoring variation, and refocused the heterogeneity analysis on stated preferences for workplace gender diversity.

Table A3: Robustness to the registration and amendment timeline

	W/ soft launch	≥ 07jun2024 (reg. start)	≥ 02apr2025 (last amend.)
Female	-0.103***	-0.107***	-0.136***
	[0.016]	[0.017]	[0.025]
Female × Monitor Safety	0.086***	0.090***	0.282***
	[0.026]	[0.026]	[0.045]
Female × Monitor Behavior	-0.064***	-0.060**	-0.157***
	[0.023]	[0.024]	[0.042]
Firms	906	888	266
Observations	20,726	20,304	6,185
Control mean	0.522	0.522	0.568
Share dropped	—	2.0%	70.2%

Notes: The table re-estimates column (1) of Table A23 on the subsample of observations collected on or after the indicated registry date: the registration start date (07jun2024) and the last amendment (02apr2025). The first column retains the soft-launch observations collected before the registration start and applies only the preregistered attention screen. *Share dropped* is the share of the first column's observations that the cutoff removes. The dependent variable *Hire* is a binary indicator equal to 1 if the employer wants to hire the candidate on probation. 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: Hiring Gender Gap Across Treatment Arms and by Stated Preferences (Including Soft-Launch Firms)

			Stated Preferences for Workforce Gender Mix					
Hire (0-1)	(1) Main Sample LPM	(2) Gender-Mix Pref. > 0	(3) Bottom QT	(4) 2nd QT	(5) 3rd QT	(6) 4th QT	(7) Top QT	(8) Main Sample Probit
Female	-0.103***	-0.078***	-0.235***	-0.144***	-0.099***	-0.092***	0.029	-0.104***
	[0.016]	[0.016]	[0.053]	[0.045]	[0.028]	[0.025]	[0.032]	[0.016]
Monitor Behavior	0.052***	0.061***	-0.014	0.037	0.052*	0.062**	0.104**	0.050***
	[0.015]	[0.016]	[0.045]	[0.041]	[0.027]	[0.028]	[0.040]	[0.016]
Monitor Safety	-0.013	-0.017	0.001	-0.059	-0.016	-0.019	-0.005	-0.013
	[0.016]	[0.017]	[0.046]	[0.040]	[0.029]	[0.028]	[0.046]	[0.016]
Female x Monitor Behavior	-0.064***	-0.077***	0.003	-0.065	-0.061	-0.034	-0.178***	-0.064***
	[0.023]	[0.024]	[0.074]	[0.066]	[0.038]	[0.044]	[0.048]	[0.023]
Female x Monitor Safety	0.086***	0.091***	0.024	0.122*	0.104**	0.096**	0.070	0.087***
	[0.026]	[0.025]	[0.090]	[0.062]	[0.043]	[0.043]	[0.071]	[0.026]
Observations	20,726	17,748	2,978	3,026	6,471	5,220	3,031	20,726
R-squared	0.127	0.126	0.177	0.173	0.147	0.120	0.124	
Control Mean	0.522	0.517	0.552	0.501	0.523	0.534	0.491	0.522

Notes: The table reproduces the main specifications on the full preregistration sample, which adds the 18 soft-launch firms surveyed before the preregistration start date. Column (1) shows the effects from the main sample; column (2) excludes employers with preferences for an all-male workforce. Columns (3) to (7) restrict the sample to each quintile of the distribution of preferred gender workforce composition. Column (8) re-estimates the main-sample specification of column (1) by probit and reports average marginal effects on the probability scale. 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. *Monitoring Behavior* is a binary indicator equal to 1 if the firm was randomly assigned to audit visits to monitor trainees' behavior. *Monitoring Safety* is a binary indicator equal to 1 if the firm was randomly assigned to audit visits to ensure trainees' safety. Profile and strata fixed effects are included (not reported in the table). Standard errors are clustered at the business level. Standard errors in brackets * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A.5 Within-Subject Design Results

The within-subject design is not our primary specification, but it provides greater statistical power. Employers assigned to the Business-as-usual arm were randomly assigned a second condition of either Monitoring-Behavior or Monitoring-Safety; employers assigned to an active monitoring arm rated profiles under the other active arm as their second condition. The Business-as-usual condition was never assigned as the second condition, ensuring that control-group responses are uncontaminated by prior exposure to an active monitoring regime. The order of the two conditions was randomized. The results, summarized in Table A16, are consistent in sign with those from the between-subjects design and larger in magnitude. Monitoring-Behavior increases the gender gap by 12.2 percentage points ($p < 0.001$) and Monitoring-Safety closes it by 14.9 percentage points ($p < 0.001$; Table A16, column (1)). The larger magnitudes are consistent with the within-subject design removing noise from heterogeneity in employers' baseline preferences, which attenuates the between-subject estimates. In particular, the between-subject Monitoring-Behavior estimate is a lower bound: employers with no baseline demand for female workers cannot exhibit a further reduction, compressing the average effect in the between-subject design.

A.6 Social Preference Measures and Behavioral Games

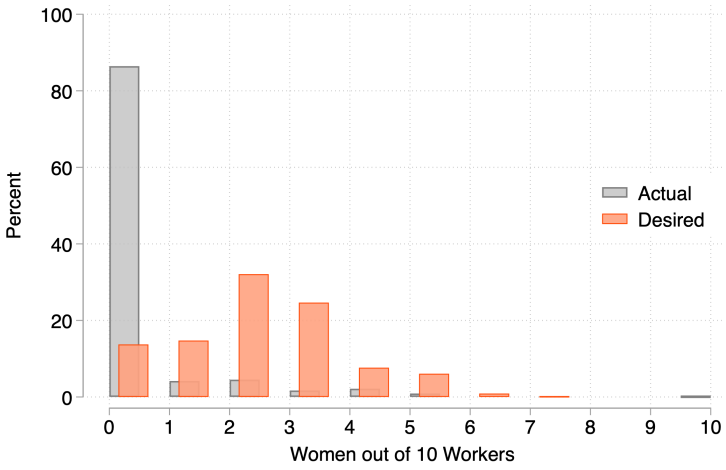
This appendix describes the measures behind Appendix Table A18. All behavior is measured in the trainee survey; the employer survey elicits beliefs about male and female trainees separately.

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 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 35 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.

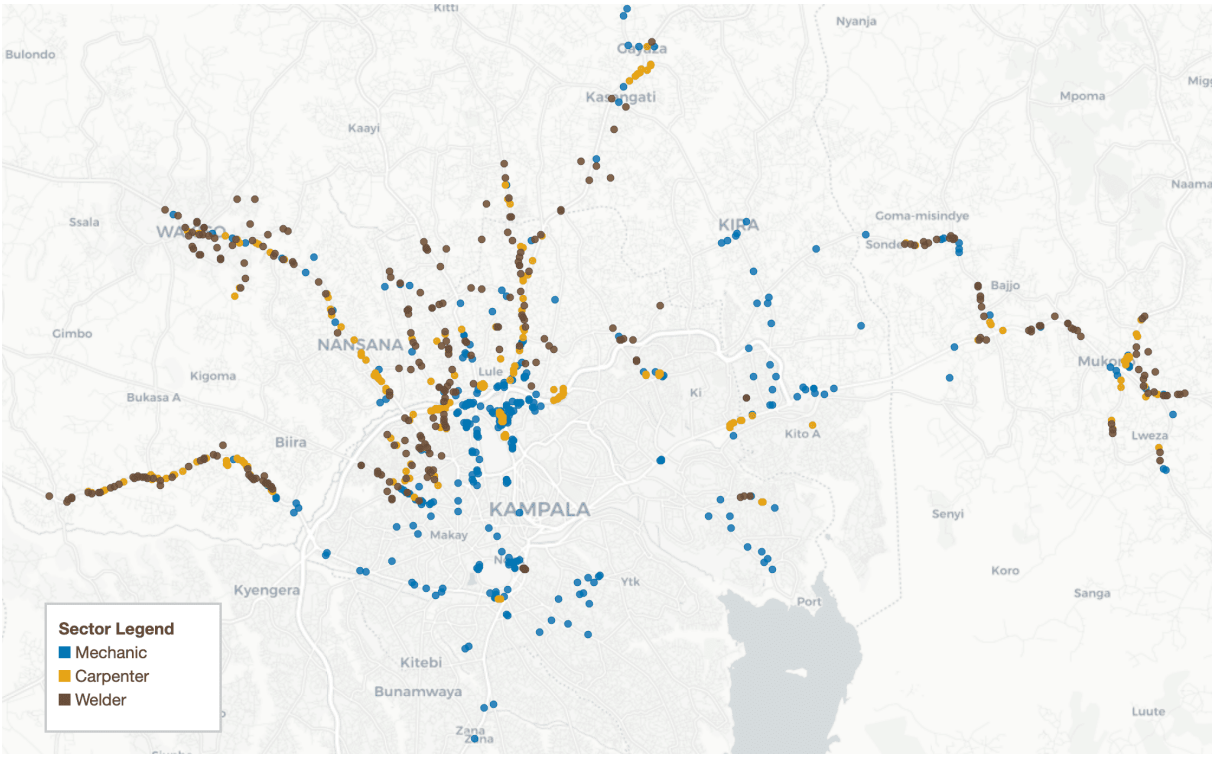
A.7 Appendix Figures

Figure A2: Employers' Desired and Actual Gender Mix Among Workforce



Notes: The figure is based on the full sample of employers (N=903). 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: This figure reports the locations of firms in the study sample in the metropolitan area of Kampala, Uganda. For privacy, location accuracy has been reduced by showing random positions within a 100m radius of the actual location.

Figure A4: Firm and Vocational Training Environment

(A) Firm Layout



(B) Practical VTI Class



(C) Internship



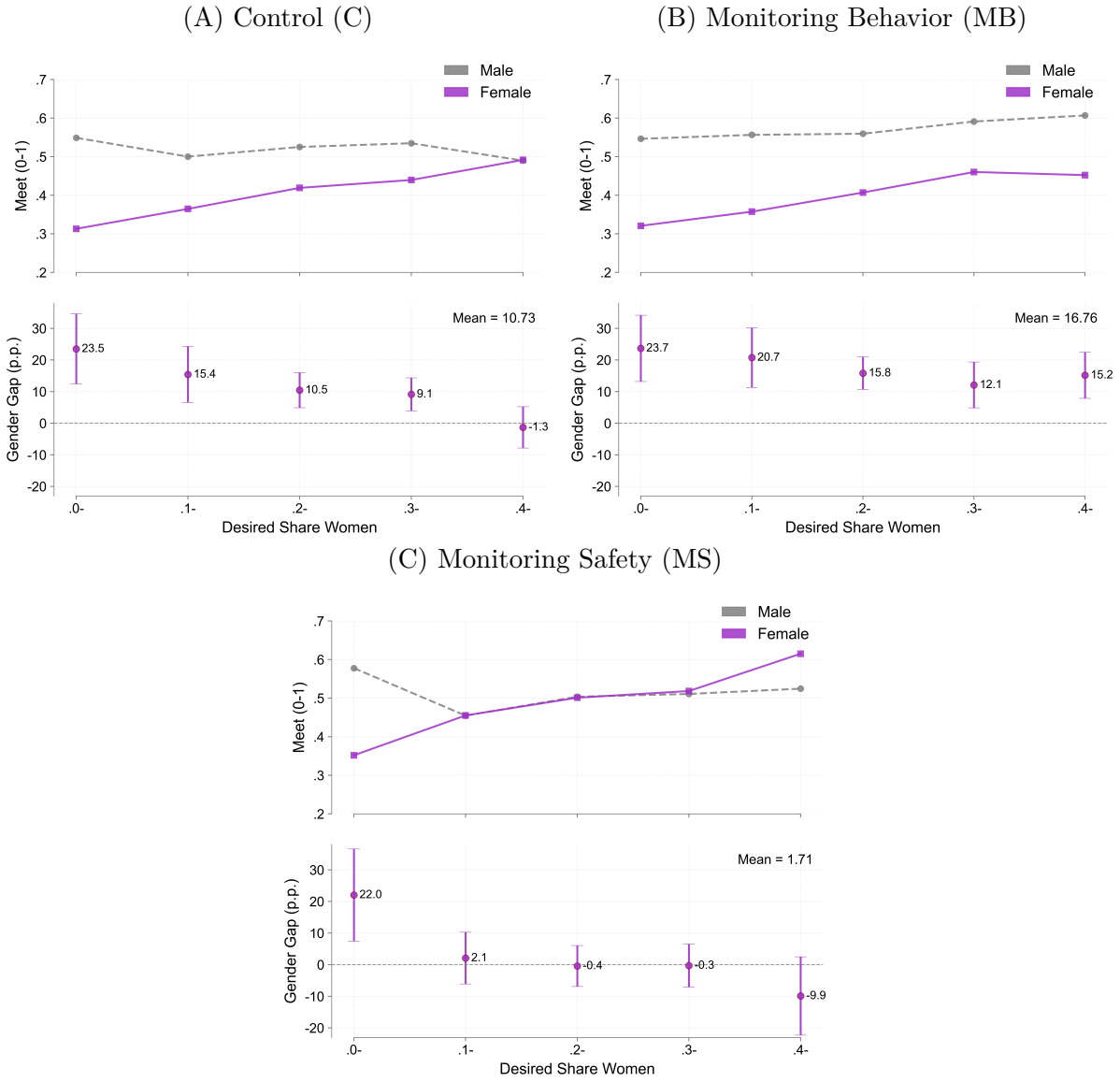
Photo credit: © Mariajose Silva-Vargas.

Figure A5: Timeline of the Experiment, Audits, and Follow-up



Notes: The figure plots the sequence of data collection. Teal blocks denote the two experimental waves, purple blocks the audit rounds implemented by the partner VTIs among verified matches, and the coral block the employer phone follow-up administered by our team. The horizontal axis is broken between blocks and is not drawn to scale; block widths do not encode duration. Audit dates are approximate.

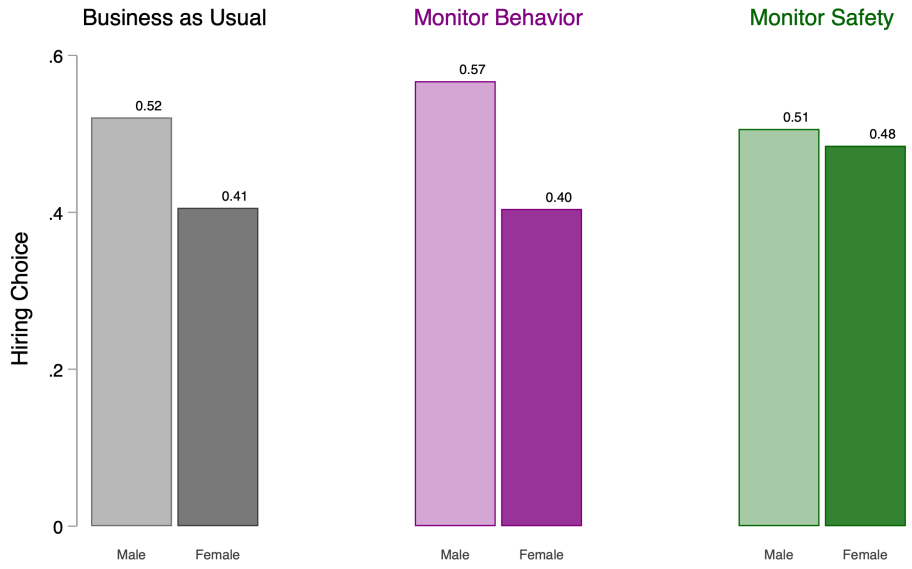
Figure A6: Hiring Gender Gap by Treatment Arm and by Stated Preferences for Hiring Women (Raw Data)



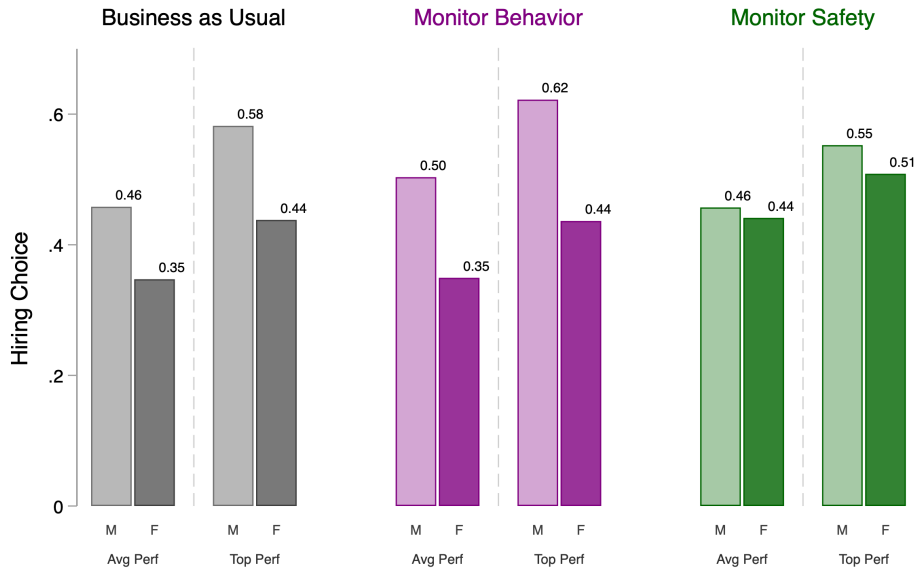
Notes: The figures show the heterogeneity in the gender gap by stated preferences for workforce gender mix, by treatment arm. Panel A presents descriptive statistics for the Control group, while panels B and C present the corresponding statistics for the Monitoring Behavior and Monitoring Safety groups, 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 2. Confidence intervals are at the 95% level.

Figure A7: Hiring Rates by Gender, Trainee Performance, and Treatment Arm (Raw Data)

(A) By gender

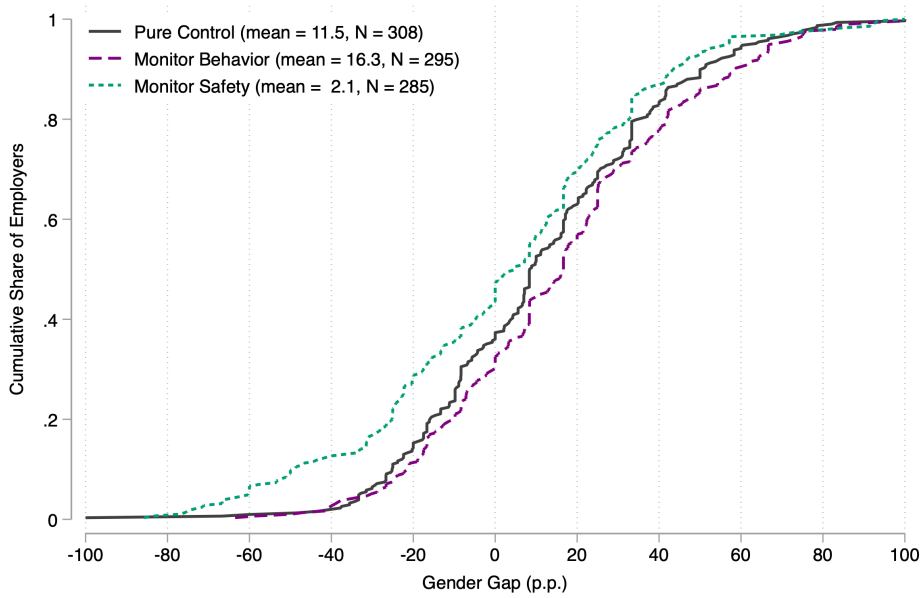


(B) By gender and performance signal



Notes: Raw means, with no regression adjustment. For each employer we compute the share of profiles selected within a cell; each bar is the mean of those shares across employers in that cell. Panel A splits by candidate gender (M, F) and treatment arm; Panel B adds the technical performance signal (Avg, Top).

Figure A8: 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), Monitor Behavior (dashed purple), and Monitor Safety (dotted green). Behavior audits shift the distribution rightward relative to Control; safety audits shift it leftward. The Monitoring-Safety CDF lies strictly above the Monitoring-Behavior CDF at every point and does not cross the Control CDF, confirming that safety dominates scrutiny for every employer subgroup under MS.

Figure A9: Theory Technical Exam — An Example (Motor Mechanics)

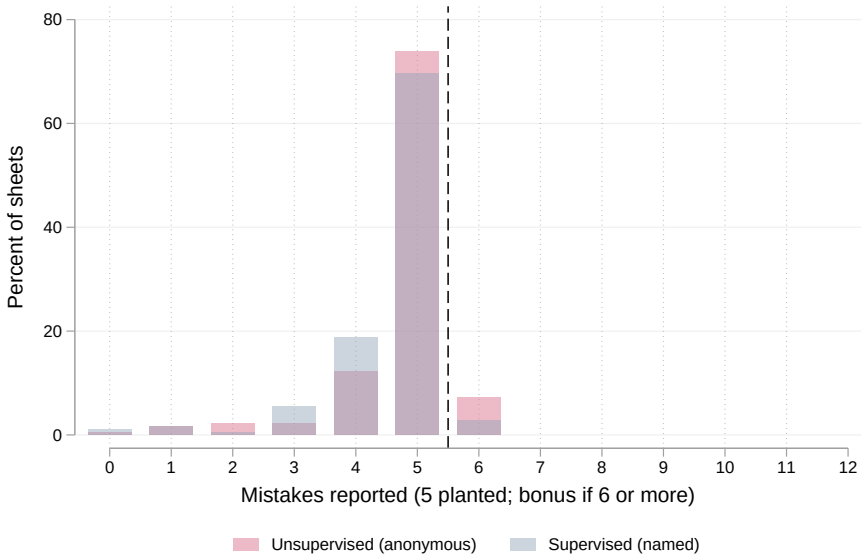
FULL NAME:_____ ☐ MAN

ID:_____ ☐ WOMAN

MOTOR-MECHANICS

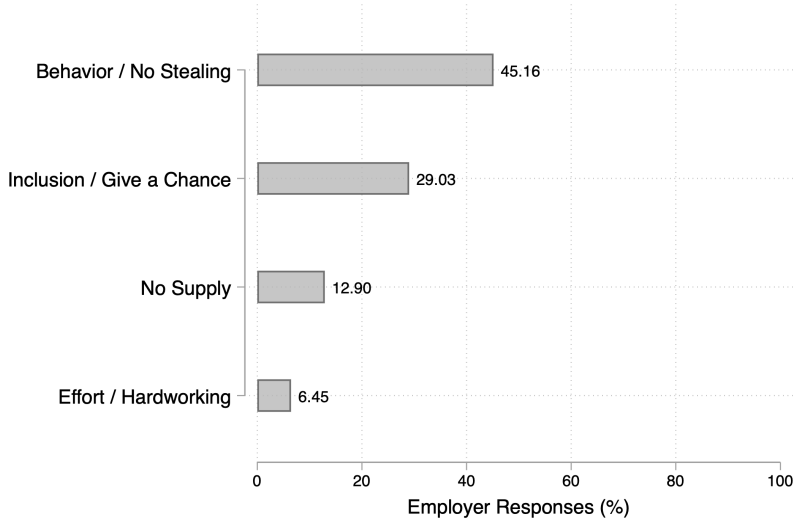
	Question	Answers
1	*multiple-choice* What are you advised to do when servicing the engine by changing oil?	A. Top up lubricating oil B. Replace oil filter C. Over hand engine D. Over hand cylinder head
2	*multiple-choice* What immediate remedy can you give to a vehicle with a problem of excessive tyre 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 A10: Reported Mistakes in the Cheating Task



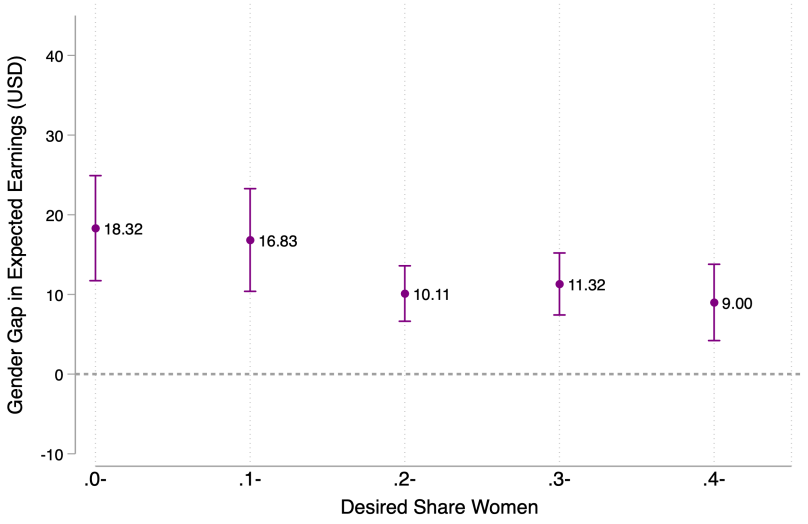
Notes: Data from the cheating task, at the sheet level ($N = 364$). Each trainee proofreads two blocks of twelve mock job registrations, five of which contain a planted mistake; one block is supervised (carries the trainee's name and ID) and one is unsupervised (anonymous). Bars show the distribution of the number of mistakes reported, by supervision condition. The dashed line marks the bonus threshold: a bonus is paid for reporting six or more mistakes, and because only five exist, any report of six or more flags a correct registration. All over-reports record exactly six—the minimum that pays the bonus—and none exceed six, although ninety-three percent of trainees check all twelve registrations.

Figure A11: Motivation for Selecting Female Workers in Experiment if No Stated Preferences



Notes: Data from employers who express a preference for male workers only and currently do not employ any women, yet selected at least one female CV in the experiment ($N = 33$). We ask an open-ended question: "Can you explain why you were choosing women in the exercise with the profiles before?". Responses were re-coded into standardized categories, with 9.68% falling under "Other". "No Supply" indicates question misunderstanding (e.g.: "There are no women workers that have come to ask for jobs").

Figure A12: Beliefs About Gender Gap in Labor Market Outside Options



Notes: Data from the experiment, focusing on employers in the Control group. The figure reports the predicted gender gap in monthly earnings (USD) of the candidates a year from the hire by stated preferences for workforce gender mix. The average gender gap is USD 14, with men expected to earn an average of USD 117.9 monthly (12.3%).

A.8 Appendix Tables

Table A5: Attention Checks and Attrition

	Monitor Treat		Female Treat
	(1)	(2)	(3)
	PC vs. MB Sample	PC vs. MS Sample	PC Sample
Main Sample	0.080	-0.082	-0.032
	[0.063]	[0.058]	[0.023]
Observations	14,524	14,487	7,382
R-squared	0.001	0.002	0.000
p-value	0.204	0.160	0.154

Notes: The table tests whether attrition due to attention checks is correlated with treatment assignment. The dependent variable is an indicator for treatment assignment: Monitoring Behavior (MB) in column (1), Monitoring Safety (MS) in column (2), and a female profile in column (3). The independent variable is *Main Sample*, a binary indicator equal to 1 if the evaluation of a given CV is included in the main analysis sample. An evaluation is excluded if preregistered attention checks are not met: the respondent understood the IRR exercise, and a given block of eight profiles has some variation in the evaluations. The flag is defined at the level of a block of profiles rather than the employer, so a firm can lose some blocks and keep others. Samples used to estimate the regressions are restricted to pure control (PC) and MB, PC and MS, and PC only in columns (1)–(3), respectively. Standard errors in brackets are clustered at the firm level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: Post-Experiment Worries by Treatment Arm

	(1) Stealing worry	(2) Safety worry	(3) Effort worry	(4) Turnover worry
Monitor Behavior	-0.624	-0.129	-0.167	-0.376
	[0.177]	[0.177]	[0.159]	[0.212]
Holm p-value	0.003	0.936	0.877	0.381
WY p-value	0.002	0.715	0.628	0.278
Monitor Safety	-0.377	-0.473	-0.227	0.077
	[0.179]	[0.183]	[0.171]	[0.215]
Holm p-value	0.213	0.069	0.736	0.936
WY p-value	0.166	0.062	0.522	0.715
Observations	878	884	887	585
R-squared	0.421	0.458	0.399	0.364
Control mean	5.500	4.646	4.968	4.019

Notes: The table reports treatment effects on post-experiment worry measures, elicited on a 0–10 scale after the profile evaluations (Section 3.3). Each column reports a regression of the indicated worry measure on indicators for Monitoring-Behavior 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. 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 A7: Hiring Gender Gap by Treatment Arm: Secondary Outcomes

	(1) Quality	(2) Trustworthiness	(3) Effort	(4) Monthly earnings
Female	-0.174	0.053	-0.211	-12.355
	[0.022]	[0.025]	[0.026]	[1.069]
Holm p-value	0.000	0.033	0.000	0.000
WY p-value	0.000	0.030	0.000	0.000
Monitor Behavior	-0.019	-0.032	-0.010	1.253
	[0.044]	[0.051]	[0.045]	[3.091]
Monitor Safety	-0.034	-0.032	-0.013	-2.457
	[0.046]	[0.052]	[0.045]	[3.095]
Female × Monitor Behavior	-0.040	-0.007	-0.055	-1.004
	[0.034]	[0.038]	[0.038]	[1.550]
Holm p-value	1.000	1.000	1.000	1.000
WY p-value	0.775	0.956	0.605	0.942
Female × Monitor Safety	0.038	0.064	0.019	-0.624
	[0.033]	[0.036]	[0.039]	[1.543]
Holm p-value	1.000	0.589	1.000	1.000
WY p-value	0.775	0.396	0.956	0.956
Observations	20,303	20,230	20,279	20,304
R-squared	0.228	0.084	0.174	0.211
BAU male mean	0.088	-0.025	0.107	115.663

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 USD: "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 respondent 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 A8: Effect of Monitoring Treatments on Number of Selected Profiles

	(1)
	Hire (0–1)
Monitor Behavior	0.023*
	[0.012]
Monitor Safety	0.033***
	[0.012]
Observations	20,304
R-squared	0.116
Pure Control Mean	0.468

Notes: The table reports the treatment effects for the main sample. We estimate the following regression model: $\text{Hire}_{ijs} = \beta_0 + \beta_1 \text{MB}_j + \beta_2 \text{MS}_j + \delta_i + \sigma_s + u_{ijs}$. *Hire* is a binary indicator equal to 1 if the employer wants to hire the candidate on probation. *Monitoring Behavior* is a binary indicator equal to 1 if the firm was randomly assigned to behavior audits. *Monitoring Safety* is a binary indicator equal to 1 if the firm was randomly assigned to audit visits to monitor trainees' safety. The control mean corresponds to the *Pure Control* arm. Profile and strata fixed effects are included (not reported). Standard errors are clustered at the respondent level and reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A9: Hiring Gender Gap by Treatment Arm and by Resume Characteristics

	(1)	(2)	(3)	(4)	(5)
Hire (0-1)	Age ≥ Median	Married	Second. Educ.	English	DIT Certified
CHRS	-0.097***	-0.219***	0.009	0.072***	0.150***
	[0.011]	[0.019]	[0.014]	[0.015]	[0.016]
CHRS × Monitor Behavior	0.014	0.030	-0.012	0.012	-0.040*
	[0.017]	[0.029]	[0.020]	[0.022]	[0.023]
CHRS × Monitor Safety	-0.005	0.029	-0.005	-0.015	-0.026
	[0.017]	[0.027]	[0.020]	[0.023]	[0.023]
Observations	20,304	20,304	20,304	20,304	20,304
R-squared	0.017	0.049	0.008	0.011	0.022
Control Mean	0.529	0.577	0.462	0.459	0.424

Notes: The table reports the treatment effects for the main sample, by resume characteristics (CHRS). Estimates are shown in Columns (1)–(5), respectively, for whether the candidate: is above the median age (22.5 years), is married, has attended secondary education, speaks English, holds a Directorate of Industrial Training (DIT) certification. Strata fixed effects and monitoring arm main effects are included but not reported in the table. Standard errors are clustered at the employer level and reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A10: Hiring Gender Gap by Treatment Arm – Heterogeneous Treatment Effects (Preregistered)

	Preferred Share of Women		Employer Gender		VTI Search	
	(1) Fewer women	(2) More women	(3) Female	(4) Male	(5) Yes	(6) No
Female	-0.148***	-0.047**	0.001	-0.109***	-0.183***	-0.090***
	[0.023]	[0.021]	[0.108]	[0.017]	[0.042]	[0.018]
Monitor Behavior	0.035*	0.083***	0.149	0.050***	-0.025	0.068***
	[0.020]	[0.023]	[0.097]	[0.016]	[0.034]	[0.017]
Monitor Safety	-0.021	-0.005	0.280***	-0.017	-0.007	-0.012
	[0.021]	[0.024]	[0.087]	[0.016]	[0.036]	[0.018]
Female × Monitor Behavior	-0.042	-0.087**	-0.206*	-0.057**	0.036	-0.081***
	[0.032]	[0.034]	[0.120]	[0.024]	[0.059]	[0.026]
Female × Monitor Safety	0.099***	0.081**	-0.218	0.096***	0.107*	0.085***
	[0.035]	[0.038]	[0.139]	[0.026]	[0.063]	[0.029]
Observations	12,260	8,044	727	19,577	3,449	16,855
R-squared	0.150	0.115	0.173	0.133	0.175	0.129
Control Mean	0.526	0.516	0.465	0.524	0.566	0.513
p-value, difference across groups	0.615		0.211		0.163	
Female × Monitor Behavior	0.324		0.519		0.074	
Female × Monitor Safety	0.687		0.091		0.802	

Notes: The table reports the treatment effects for the main sample, by pre-registered heterogeneity dimensions: Columns (1)–(2) split employers by the share of women they state would be optimal in their own workforce, at the sample median: column (1) covers employers who would prefer fewer women and column (2) those who would prefer more; Columns (3)–(4) split the sample by the employers' gender; Columns (5)–(6) split the sample by whether employers typically hire from vocational training institutes or not. 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. *Monitoring Behavior* is a binary indicator equal to 1 if the firm was randomly assigned to audit visits to monitor trainees' behavior. *Monitoring Safety* is a binary indicator equal to 1 if the firm was randomly assigned to audit visits to monitor trainees' safety. The final three rows report p-values from the pooled model with the full set of triple interactions, estimated on the whole sample: the joint test that both Female by arm interactions differ across the two groups, and the two interactions separately. Each spans the pair of columns it refers to. Profile and strata fixed effects are included. Standard errors are clustered at the employer level and reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A11: Robustness to Experimenter Demand

Panel A. Effect of treatment on beliefs

	Demand belief	Prefers women	Skill beliefs
Monitor Safety	-0.036	-0.024	-0.027
	[0.044]	[0.020]	[0.044]
Monitor Behavior	0.035	0.008	0.019
	[0.044]	[0.022]	[0.044]
Joint p (MS=MB=0)	0.26	0.26	0.59
Observations	879	879	880
Control mean	-0.167	0.077	-0.001

Panel B. Hiring gap, excluding demand-effects prone employers

	(1) Baseline	(2) Excl. prefers-women
Female	-0.107***	-0.123***
	[0.017]	[0.017]
Monitor Safety	-0.012	-0.015
	[0.016]	[0.017]
Monitor Behavior	0.053***	0.059***
	[0.016]	[0.016]
Female × Monitor Safety	0.090***	0.102***
	[0.026]	[0.027]
Female × Monitor Behavior	-0.060**	-0.066***
	[0.024]	[0.025]
Observations	20,304	18,881
Control mean	0.522	0.529

Panel C. Treatment effect heterogeneity by demand belief

	Coef.	p-value
Female × Demand belief	0.177***	0.00
	[0.032]	
Female × Monitor Behavior × Demand belief	-0.015	0.74
	[0.044]	
Female × Monitor Safety × Demand belief	-0.016	0.77
	[0.055]	
Observations	20,088	

Notes: The table reports three checks that experimenter-demand effects do not drive the results. Panel A regresses each stated belief on the monitoring arms (Business-as-usual omitted) at the firm level, with robust standard errors, and reports the joint test that both arm coefficients are zero. The outcome is the belief shown in each column: the demand-effect beliefs *Choose women* (-1 prefer men, 0 indifferent, +1 prefers women), an indicator that the team prefers women (*Prefers women*), and a standardized index of the six post-experiment gendered skill items (*Skill beliefs*; men vs. women better at each trait). Panel B estimates the main treatment-effect regression; the outcome is *Hire*. Column (2) drops the employers who report believing the research team prefers they hire women, as defined in Panel A. Employers who declined to answer that item are retained. Panel C interacts the gender gap and the arms with the direct demand item; the outcome is *Hire*. *Female* × *Demand belief* is the baseline relationship between the demand item and the gap, and the triple interactions test whether the arms change it. Panel C excludes the employers with a missing demand item, since it interacts on that item. Interacting additionally with the stated preference gradient leaves the four-way arm interactions insignificant ($p = 0.57$ for Monitor Behavior, $p = 0.90$ for Monitor Safety). Panels B and C are 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 business level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A12: Robustness to Profile Order

	(1) Baseline	(2) Order control	(3) Order interactions
Female	-0.107***	-0.107***	-0.107***
	[0.017]	[0.017]	[0.017]
Monitor Safety	-0.012	-0.012	-0.012
	[0.016]	[0.016]	[0.016]
Monitor Behavior	0.053***	0.053***	0.053***
	[0.016]	[0.016]	[0.016]
Female × Monitor Safety	0.090***	0.090***	0.091***
	[0.026]	[0.026]	[0.026]
Female × Monitor Behavior	-0.060**	-0.060**	-0.060**
	[0.024]	[0.024]	[0.024]
Profile order		-0.001**	-0.002*
		[0.000]	[0.001]
Female × Order			0.002
			[0.002]
Monitor Safety × Order			-0.000
			[0.002]
Monitor Behavior × Order			0.001
			[0.002]
Female × Monitor Safety × Order			-0.001
			[0.002]
Female × Monitor Behavior × Order			-0.003
			[0.002]
Observations	20,304	20,304	20,304
R-squared	0.130	0.130	0.130
Control mean	0.522	0.522	0.522
Order control	No	Yes	Yes
Order × treatment	No	No	Yes

Notes: The table tests robustness of the main between-subject treatment effects to profile order. The dependent variable is *Hire*, an indicator equal to 1 if the employer wants to hire the candidate on probation. Column (1) reports the baseline specification. Column (2) controls linearly for the profile's order in the task. Column (3) interacts profile order with candidate gender and monitoring treatment. Profile order is the exact position of the profile in the employer's task, running from 1 to 24, centered at 12.5, so the uninteracted coefficients are evaluated at the middle of the task. Blank cells indicate terms not included in that specification. All columns include CV profile fixed effects interacted with trainee performance, sector, survey-wave, and enumerator fixed effects. Standard errors are clustered at the employer level and reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A13: Hiring Gender Gap Across Treatment Arms, Controlling for Imbalanced Covariates

			Stated Preferences for Workforce Gender Mix					
Hire (0-1)	(1) Main Sample LPM	(2) Gender-Mix Pref. > 0	(3) Bottom QT	(4) 2nd QT	(5) 3rd QT	(6) 4th QT	(7) Top QT	(8) Main Sample Probit
Female	-0.107***	-0.085***	-0.227***	-0.162***	-0.107***	-0.093***	0.023	-0.108***
	[0.016]	[0.016]	[0.056]	[0.046]	[0.028]	[0.027]	[0.033]	[0.016]
Monitor Behavior	0.051***	0.060***	-0.021	0.036	0.046*	0.059**	0.098**	0.050***
	[0.015]	[0.016]	[0.044]	[0.042]	[0.027]	[0.029]	[0.040]	[0.016]
Monitor Safety	-0.012	-0.017	-0.003	-0.060	-0.020	-0.016	-0.007	-0.013
	[0.016]	[0.017]	[0.048]	[0.042]	[0.030]	[0.029]	[0.046]	[0.017]
Female × Monitor Behavior	-0.060**	-0.071***	-0.000	-0.046	-0.054	-0.033	-0.170***	-0.059**
	[0.024]	[0.024]	[0.075]	[0.067]	[0.038]	[0.045]	[0.049]	[0.024]
Female × Monitor Safety	0.090***	0.097***	0.014	0.140**	0.112***	0.097**	0.077	0.091***
	[0.026]	[0.026]	[0.094]	[0.062]	[0.043]	[0.044]	[0.071]	[0.026]
Observations	20,304	17,421	2,883	2,978	6,399	5,077	2,967	20,304
R-squared	0.130	0.130	0.201	0.180	0.150	0.127	0.128	
Control Mean	0.522	0.517	0.549	0.500	0.525	0.535	0.490	0.522
Imbalanced covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table repeats the specification of Table A23, adding as controls the baseline characteristics that came out imbalanced under randomization inference. Columns are defined as in that table. The controls are a dummy for whether the employer has any vocational training, the monthly number of clients and the employer's daily hours worked, each entered together with an indicator for a missing value, which is median-imputed. Column (8) re-estimates the main-sample specification by probit and reports average marginal effects on the probability scale. Fixed effects are as in the main table and are not reported. Standard errors are clustered at the employer level and reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A14: Customer Discrimination Channel Test

	(1) Full sample	(2) Restricted sample
Female	-0.126***	-0.130*
	[0.023]	[0.072]
Female × Monitor Behavior	-0.065**	-0.092
	[0.031]	[0.102]
Female × Monitor Safety	0.050	0.102
	[0.035]	[0.096]
Female × *Customer channel*	0.044	0.048
	[0.033]	[0.076]
Female × MB × *Customer channel*	0.027	0.053
	[0.048]	[0.108]
Female × MS × *Customer channel*	0.104**	0.050
	[0.052]	[0.103]
Observations	20,280	9,118
Employers	887	399
Control Mean (male)	0.522	0.546
Employer FE × Performance	Yes	Yes
Sector, Year, Enumerator FE	Yes	Yes

Notes: The dependent variable *Hire* is a binary indicator equal to one if the employer wants to hire the candidate on probation. *Customer channel* is coded from open-text responses to the survey question: if the share of female workers at your firm were to increase, what do you expect to happen to the profits of your firm, and why? It equals one if the employer cited customer attraction, customer care or customer trust as the reason profits would rise, which is 86.5 percent of coded responses among employers who expected profits to rise. Column (1) uses the full analysis sample, with employers who did not expect profits to rise and employers citing non-customer reasons both assigned *Customer channel* = 0. Column (2) restricts to employers who expected profits to rise from hiring more women. Both columns include 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 A15: Gender Gap by Treatment Assignment and Interaction with Gender Preferences

	(1)	(2)	(3)	(4)	(5)
Hire (0-1)	Pure Control	Monitor Behavior	Monitor Safety	Main Sample Linear	Main Sample Quadratic
Female	-0.216***	-0.216***	-0.142***	-0.216***	-0.231***
	[0.036]	[0.033]	[0.044]	[0.036]	[0.049]
Monitor Behavior				-0.001	0.013
				[0.031]	[0.040]
Monitor Safety				-0.025	-0.001
				[0.033]	[0.043]
Diversity	-0.120	0.135	-0.054	-0.118	0.110
	[0.088]	[0.089]	[0.106]	[0.083]	[0.225]
Diversity^2					-0.466
					[0.431]
Female × Monitor Behavior				0.000	-0.009
				[0.049]	[0.067]
Female × Monitor Safety				0.074	0.065
				[0.057]	[0.079]
Female × Diversity	0.488***	0.221*	0.575***	0.489***	0.676*
	[0.127]	[0.114]	[0.175]	[0.127]	[0.377]
Female × Diversity^2					-0.386
					[0.690]
Monitor Behavior × Diversity				0.244**	0.058
				[0.116]	[0.317]
Monitor Safety × Diversity				0.061	-0.257
				[0.128]	[0.364]
Monitor Behavior × Diversity^2					0.395
					[0.597]
Monitor Safety × Diversity^2					0.694
					[0.733]
Female × Monitor Behavior × Diversity				-0.270	-0.155
				[0.171]	[0.505]
Female × Monitor Safety × Diversity				0.089	0.217
				[0.216]	[0.640]
Female × Monitor Behavior × Diversity^2					-0.217
					[0.874]
Female × Monitor Safety × Diversity^2					-0.314
					[1.246]
Observations	6,943	6,830	6,531	20,304	20,304
R-squared	0.151	0.157	0.112	0.136	0.137
Baseline Mean	0.522	0.572	0.510	0.549	0.549

Notes: The table reports treatment effects on employer interest to hiring candidates on probation, estimated as a function of stated preferences for workforce gender diversity. Columns (1)–(3) report results for employers randomly assigned to Control (C), Monitor Behavior (MB), and Monitor Safety (MS) arms in the between-subject design (main specification). Column (4) includes all respondents in the main sample and estimates a fully interacted model with candidate gender, monitoring treatment, and employer stated gender diversity preferences. Column (5) adds a quadratic interaction with diversity preferences; none of the quadratic terms is statistically significant. 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 resume corresponds to a woman. *Monitor Behavior* is a binary indicator equal to 1 if the firm was assigned to audit visits monitoring trainee behavior. *Monitor Safety* is a binary indicator equal to 1 if the firm was assigned to audit visits monitoring trainee safety. *Diversity* is a continuous variable (ranging from 0 to 1) measuring the share of female workers the employer prefers in their ideal workforce composition. All models include profile and strata fixed effects (not reported). Standard errors are clustered at the employer and profile levels (reported in brackets: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

Table A16: Hiring Gender Gap by Treatment Arm: Within-Subject Design

			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	-0.115***	-0.096***	-0.236***	-0.122**	-0.100***	-0.122***	-0.050*
	[0.015]	[0.015]	[0.058]	[0.053]	[0.026]	[0.026]	[0.028]
Monitor Behavior	0.092***	0.103***	0.037	0.076**	0.121***	0.083***	0.106***
	[0.013]	[0.013]	[0.041]	[0.035]	[0.022]	[0.026]	[0.029]
Monitor Safety	-0.046***	-0.045***	-0.016	-0.092**	-0.041*	-0.050**	-0.027
	[0.013]	[0.013]	[0.048]	[0.036]	[0.024]	[0.024]	[0.027]
Female x Monitor Behavior	-0.122***	-0.125***	-0.119	-0.149**	-0.129***	-0.096**	-0.123***
	[0.020]	[0.020]	[0.072]	[0.066]	[0.034]	[0.039]	[0.039]
Female x Monitor Safety	0.149***	0.163***	-0.020	0.179***	0.163***	0.174***	0.166***
	[0.023]	[0.023]	[0.089]	[0.062]	[0.040]	[0.040]	[0.050]
Observations	20,801	18,409	2,392	2,447	6,521	5,714	3,727
Employers	597	527	70	70	189	161	107
R-squared	0.221	0.214	0.314	0.249	0.237	0.208	0.191
Control Mean	0.529	0.528	0.536	0.522	0.513	0.550	0.526
p-value MB vs. MS	0.000	0.000	0.268	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 assigned to the business-as-usual arm rated a second set of profiles under either Monitoring-Behavior or Monitoring-Safety; employers assigned to an active arm rated their second set under the other active arm. Business-as-usual was never assigned second, so control responses are uncontaminated by prior exposure to an active monitoring regime. 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 Behavior* is a binary indicator equal to 1 if the firm was randomly assigned to audit visits to monitor trainees' behavior. *Monitor Safety* is a binary indicator equal to 1 if the firm was randomly assigned to audit 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 A17: Descriptive Statistics: Trainee Sample

	Mean	SD	P25	P75	Min	Max	*N*
Panel A: Demographics							
Female (share)	0.38	0.49	0.00	1.00	0.00	1.00	182
Age (years)	20.82	2.23	19.00	22.00	18.00	27.00	182
Married (share)	0.02	0.13	0.00	0.00	0.00	1.00	182
Completed secondary, S4+ (share)	0.96	0.19	1.00	1.00	0.00	1.00	182
Panel B: Training							
Mechanics (share)	0.33	0.47	0.00	1.00	0.00	1.00	182
Carpentry (share)	0.20	0.40	0.00	0.00	0.00	1.00	182
Welding (share)	0.28	0.45	0.00	1.00	0.00	1.00	182
Other programs (share)	0.19	0.40	0.00	0.00	0.00	1.00	182
Girls With Tools VTI (share)	0.30	0.46	0.00	1.00	0.00	1.00	182
Training duration (months)	24.81	7.26	24.00	24.00	3.00	60.00	182
24-month program (share)	0.76	0.43	1.00	1.00	0.00	1.00	182
Panel C: Skills and Work							
Theory exam score (share correct)	0.53	0.20	0.43	0.71	0.14	1.00	182
Practical skill, self-reported (share)	0.92	0.27	1.00	1.00	0.00	1.00	182
Worked last month (share)	0.21	0.41	0.00	0.00	0.00	1.00	182
Expected monthly income, 1yr (USD)	268.0	268.6	135.1	270.3	40.54	1891.9	182

Notes: The table reports summary statistics for the 182 vocational trainees surveyed across seven training centers in the Kampala metropolitan area. *Female* is a binary indicator equal to 1 for female trainees. *Completed secondary, S4+* indicates completion of at least Senior 4. Sector shares reflect the trainee's primary training program. *Other programs* includes hairdressing/barbering, tailoring, electrical installation, ICT, and plumbing. The latter three also report being interested in job in the mechanics and welding sectors. *Girls With Tools VTI* indicates enrollment at the Smart Girls Foundation (GWT), the primary provider of female candidates in male-dominated sectors. *Training duration* is the total program length in months; 76% of trainees are enrolled in 24-month programs. *Theory exam score* is the share of correct answers on a sector-specific technical exam. *Practical skill* is a binary self-reported indicator. *Worked last month* equals 1 if the trainee performed any work in the month before the survey. *Expected monthly income* is the trainee's expected earnings in USD one year after training.

Table A19: Women Hired at Audited Firms

	Mean	SD	Median	Max	Total
Panel A. Workforce and women hired					
Workers today	7.40	5.25	6	23	422
Women today	0.35	0.61	0	2	20
Women in the past year	1.21	1.69	1	8	69
Panel B. Share of firms					
Any woman in the past year	52.6%				30
Any woman today	28.1%				16
Any hire from a VTI, past six months	47.4%				27
Panel C. Benchmarks					
Female share of the workforce today			4.7%		
Stated ideal female share, baseline			0.193		
Firms audited			62		
Firms reporting			57		

Notes: The table reports the hiring module collected during the audit visits described in Section 3.2. Of the 62 matched firms that received an audit, 57 completed the module; the remaining five could not be reached by phone. Panel A reports counts of workers, including trainees and excluding the owner. *Women in the past year* counts every woman the employer agreed to hire over the preceding twelve months, whether or not she was still employed at the time of the visit, and so weakly exceeds *Women today* for all but one firm. Panel B reports the share of firms with at least one such worker; the final column gives the corresponding number of firms. In Panel C, the female share of the workforce is the ratio of total women to total workers across reporting firms (20/422), and the stated ideal female share is the employer's preferred female share of the workforce elicited at baseline. Twenty-seven of the 57 reporting firms report no woman over the past year. Because only the four hiring questions were collected during the visit, a given worker cannot always be attributed to our referral.

Table A18: Trainee Outcomes and Employer's Beliefs

Panel A: **Trainees**

	Social Preferences				Dictator Game	Trust Game		Cooperation Game
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Patience	Altruism	Pos. Recipr.	Trusting	Share	Trust Pl. 1	Reciprocate Pl. 2	Cooperate
Female	0.089	0.067	0.064	0.041	0.177*	-0.375***	-0.005	0.131
	[0.073]	[0.071]	[0.045]	[0.059]	[0.106]	[0.093]	[0.099]	[0.099]
Observations	171	171	171	171	182	182	182	182
R-squared	0.028	0.038	0.064	0.034	0.133	0.096	0.088	0.059
Mean, men	0.696	0.681	0.768	0.450	0.723	0.679	0.875	0.830

Panel B: **Employers**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	-0.123**	-0.117**	0.057	0.490***	-0.087***	-0.011	-0.034***	-0.017**
	[0.050]	[0.052]	[0.046]	[0.042]	[0.012]	[0.009]	[0.009]	[0.008]
Observations	901	901	902	903	1,802	1,799	1,793	1,795
R-squared	0.011	0.009	0.046	0.317	0.031	0.002	0.007	0.020
Mean, men	0.433	0.447	0.253	0.127	0.573	0.536	0.604	0.572

Notes: The table reports gender differences in trainees' outcomes (Panel A) and in employer beliefs about trainees (Panel B); Appendix A.6 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* coefficient is therefore 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); Panel B reports the believed share of trainees choosing each action. Technical skills and the cheating task are reported in Table 5. In Panel A the *Female* coefficient is the difference between female and male trainees, and regressions include vocational-training-institute fixed effects, with robust standard errors; in Panel B it is the difference in beliefs about female versus male trainees, and regressions include sector fixed effects. Standard errors are robust in Panel B columns (1)–(4), which contain one observation per employer, and clustered at the employer level in columns (5)–(8), which contain one observation per employer and trainee gender. Except in Panel B columns (1)–(4), the *Mean, men* row reports the male (or believed male) mean. Standard errors in brackets * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A20: Employer follow-up

	Baseline (1)	Follow-up (2)	Difference (3)	p-value (4)	N (5)
Panel A. Workforce at the time of the survey					
Female share of workforce	0.026	0.056	0.030	0.000	717
Share of firms with at least one woman	0.132	0.241	0.108	0.000	748
Women employed	0.175	0.389	0.214	0.000	748
Workers (excl. owners)	8.22	7.20	-1.02	0.000	746
Panel B. Hiring since the end of the experimental wave					
Share of firms hiring at least one woman		0.596			748
Women hired		1.66			748
Women hired per firm-year		0.96			748
Average women employed at any point		0.52			748
Average tenure (months)		7.1			446
Share of firms contacting a VTI		0.338			517
Panel C. Selection into the follow-up (baseline characteristics)					
	Reached	Not reached	Difference	p-value	N
Workers (excl. owners)	8.22	8.45	-0.23	0.770	900
Women employed	0.175	0.252	-0.076	0.102	903
Female share of workforce	0.027	0.030	-0.002	0.769	900
Stated ideal female share	0.216	0.222	-0.006	0.642	903

Notes: The table reports outcomes from the employer phone follow-up, covering all firms in the experimental sample that we reached and that consented. In Panels A and B, column (1) reports the baseline value of the same quantity for the same firms, measured in the employer survey at the start of the firm's experimental wave; column (2) reports the follow-up value; column (3) the within-firm difference; column (4) the p-value of a Wilcoxon signed-rank test of the null of no within-firm change; and column (5) the number of firms contributing to the comparison. Panel A measures the workforce at the moment of each survey. *Female share of workforce* is the number of women divided by the number of workers, defined for firms with at least one worker. *Workers* excludes owners and includes trainees, matching the baseline question wording. Panel B measures hiring over the interval between the end of the firm's experimental wave and the follow-up call, a window of 24 months for wave-1 firms and 15 months for wave-2 firms; it has no baseline counterpart, so columns (1), (3) and (4) are empty. *Women hired* counts all women who worked at the firm at any point in the window, whether or not they are still employed. *Women hired per firm-year* normalises that count by the window length so that the two waves are comparable. *Average women employed at any point* is the product of the number of women hired and their average tenure, divided by the window length; because spells still running are right-censored, it is a lower bound. Panel C compares the baseline characteristics of firms we reached in the follow-up with those we did not, where the difference in column (3) is reached minus not reached and column (4) reports the p-value of a two-sample t-test; firms are classified as not reached if they were never called or did not consent, and among firms called the consent rate was XX%. All statistics are unweighted firm-level means unless stated otherwise.

Table A21: Representativeness of the Study Sample

	Our sample (1)	2010 Census (2)	2020 Census (3)
Firm size	9.26	5.54	8.40
	(8.75)	(10.00)	(7.22)
Owner is male	0.963	0.951	0.961
	(0.188)	(0.204)	(0.193)
Firm age (years)	10.77	8.17	11.50
	(7.77)	(7.82)	(9.18)
Annual revenue (nominal UGX)			
Below 5m	0.013	0.280	0.037
	(0.114)	(0.449)	(0.190)
5m to 10m	0.098	0.418	0.160
	(0.297)	(0.493)	(0.367)
Above 10m	0.889	0.302	0.802
	(0.315)	(0.459)	(0.398)
Observations	903	3287	3045

Notes: The table compares our employer sample to the Census of Business Establishments conducted by the Uganda Bureau of Statistics in 2010 (Uganda Bureau of Statistics, 2011) and in 2020 (Uganda Bureau of Statistics, 2020). All three columns cover Kampala and our three study sectors (ISIC 4520 motor mechanics, 2511 welding, and carpentry through ISIC 3100 in 2010 and ISIC 3110 in 2020). Our sample was drawn from employers with at least one worker to evaluate, whereas the census columns include owner-only establishments; this accounts for the larger mean firm size in column 1. Standard deviations are in parentheses. Firm size is the number of persons engaged *including* the working owner. Owner gender is the establishment-level share of owners who are male in 2010 and an indicator for a male owner in 2020. Firm age is computed from the reported year of establishment relative to each census reference year. Revenue is reported as the share of firms in each of three annual turnover bands, the form in which the census records it; our own revenue is binned into the same bands. At an exchange rate of 2,700 UGX per USD, the two thresholds correspond to roughly USD 1,850 and USD 3,700 of annual turnover. The bands are nominal and are held fixed across the fifteen years separating the sources. The revenue rows use the 531 of 903 firms in our sample reporting revenue, and all firms in both census waves.

Table A22: Representativeness of the Employer Skills-Beliefs Sample

	Full sample (1)	Beliefs recorded (2)	Difference (3)	p (4)
Pure Control	0.341	0.349	0.008	0.479
Monitor Safety	0.329	0.322	-0.007	0.548
Monitor Behavior	0.330	0.329	-0.001	0.915
Stated ideal female share	0.217	0.212	-0.006	0.119
VTI trained	0.309	0.315	0.006	0.590
Hours worked (per day)	10.420	10.489	0.069	0.085
Joint test, F			1.266	0.277
Firms	903	584		

Notes: The table compares the full sample of employers with those whose skills beliefs are recorded, the sample used in Panel A of Table 5. Employers' predictions of trainee performance are incentivized against the trainees' own exam results, so the module could be asked only once the exams for the employer's own sector had been administered and scored. Motor mechanics were surveyed first and their trainees had already sat the exam, so employers in that sector were asked throughout and 299 of 300 answered. Carpentry and welding trainees were examined later, so employers in those sectors were asked only from April 2025: 108 of 300 carpentry employers and 177 of 303 welding employers. Whether an employer was asked is therefore determined by sector and interview date, both fixed before the employer was contacted, and only one employer who was asked declined to answer. The belief sample is accordingly weighted toward motor mechanics. Firm size measured by customers served per day is therefore not reported here, because it differs sharply across sectors and so proxies the coverage rule rather than testing representativeness; it is compared within sector in Table **??**. Column (3) is the difference between columns (2) and (1). Column (4) reports the p-value from a regression of each characteristic on an indicator for having beliefs recorded, with robust standard errors; that contrast is proportional to the difference in column (3), so the two carry the same p-value. The joint test is of all rows simultaneously in a linear probability model for that indicator, omitting one treatment arm. Treatment was assigned within strata and is orthogonal to the interview date by construction. All estimates are at the employer level, one observation per firm.

Table A23: Hiring Gender Gap Across Treatment Arms and by Stated Preferences

			Stated Preferences for Workforce Gender Mix					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Hire (0-1)	Main Sample LPM	Gender-Mix Pref. > 0	Bottom QT	2nd QT	3rd QT	4th QT	Top QT	Main Sample Probit
Female	-0.107***	-0.085***	-0.228***	-0.160***	-0.107***	-0.094***	0.021	-0.108***
	[0.017]	[0.016]	[0.055]	[0.046]	[0.028]	[0.027]	[0.033]	[0.016]
Monitor Behavior	0.053***	0.062***	-0.014	0.039	0.051*	0.060**	0.105***	0.052***
	[0.016]	[0.016]	[0.046]	[0.042]	[0.027]	[0.029]	[0.040]	[0.016]
Monitor Safety	-0.012	-0.016	0.005	-0.057	-0.017	-0.021	-0.005	-0.012
	[0.016]	[0.017]	[0.047]	[0.041]	[0.030]	[0.029]	[0.046]	[0.017]
Female x Monitor Behavior	-0.060**	-0.071***	-0.001	-0.048	-0.053	-0.032	-0.171***	-0.060**
	[0.024]	[0.024]	[0.075]	[0.067]	[0.038]	[0.045]	[0.049]	[0.024]
Female x Monitor Safety	0.090***	0.097***	0.015	0.138**	0.112**	0.099**	0.077	0.091***
	[0.026]	[0.026]	[0.094]	[0.062]	[0.043]	[0.044]	[0.071]	[0.026]
Observations	20,304	17,421	2,883	2,978	6,399	5,077	2,967	20,304
R-squared	0.130	0.129	0.198	0.178	0.149	0.125	0.125	
Control Mean	0.522	0.517	0.549	0.500	0.525	0.535	0.490	0.522

Notes: The table reports the treatment effects for the main sample, and by quintiles distribution of stated preferred gender workforce composition of the employer. Column (1) shows the effects from the main sample; column (2) excludes employers with preferences for an all-male workforce. Columns (3) to (7) restrict the sample to each quintiles distribution of preferred gender workforce composition. Column (8) re-estimates the main-sample specification of column (1) by probit and reports average marginal effects on the probability scale. 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. *Monitoring Behavior* is a binary indicator equal to 1 if the firm was randomly assigned to audit visits to monitor trainees' behavior. *Monitoring Safety* is a binary indicator equal to 1 if the firm was randomly assigned to audit visits to ensure trainees' safety. Profile and strata fixed effects are included (not reported in the table). Standard errors are clustered at the business level. Standard errors in brackets * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.