Hidden Discrimination in Frictional Labor Markets

Elisa Macchi* and Claude Raisaro†

Version: April 3, 2026.

Abstract

This paper shows that frictions shape the set of worker attributes employers value. When those attributes carry identity connotations, changes in the friction environment affect hiring disparities. Through this channel, frictions can hide discrimination. We test for gender discrimination among 921 employers in male-dominated Ugandan manufacturing sectors. Employers perceive women as more trustworthy but worry about harassment risk, two attributes related to common monitoring and cooperation frictions. In an experiment where employers select from a pool of trainees for probationary hire, we randomly assign monitoring support targeting either worker misbehavior or workplace safety. Under business-as-usual conditions, men are selected at rates 10 percentage points higher than women, yet 42% of selected candidates are women, revealing substantial unmet demand. Audits to monitor new workers and prevent stealing widen the hiring gender gap by 62%, concentrated among employers with the strongest stated preferences for hiring women. Audits to prevent harassment, instead, close the gender gap. Monitoring frictions mask discrimination: the gender penalty is not driven by statistical discrimination on technical performance, we find no technical performance or trustworthiness differences by gender, and safety concerns can explain at most half of the uncovered penalty. The results provide proof of concept that when multiple distortions coexist, solving one friction can worsen allocative efficiency, consistent with second-best logic. Residual measures of bias depend on which worker attributes are measured and treated as productive, and can both understate and overstate discrimination.

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

†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 markets in developing countries are characterized by multiple frictions. Standard theories predict more discrimination in such markets, emphasizing two channels: frictions allow prejudice to persist by reducing competition (Becker, 1957; Black, 1995) and generate statistical discrimination by limiting information (Phelps, 1972; Arrow, 1973).

This paper highlights a third channel: frictions expand the set of worker attributes that matter for hiring beyond technical skills. When these friction-relevant attributes are associated with identity groups, solving frictions affects hiring gaps. Unlike the first two, this channel can both generate and dampen hiring disparities. If a group faces bias but is perceived as superior in at least one friction-relevant attribute, the two forces partially offset, producing a smaller observed gap. Resolving that friction removes the offsetting advantage and reveals hidden discrimination.

We investigate this channel by testing for gender discrimination in hiring in three Ugandan male-dominated manufacturing sectors, through a hiring experiment that holds supply fixed and manipulates two workplace frictions: monitoring constraints over workers' behavior (stealing) and cooperation (harassment/distraction). This allows us to isolate how the friction environment affects the demand for female workers. Our design also measures the role of beliefs about technical performance, by cross-randomizing gender and performance signals in candidate profiles, and the accuracy of employers' beliefs, through a complementary lab-in-the-field experiment with trainees.

Female labor force participation in Uganda is relatively high (76.5%; ILO, 2025), as is common in sub-Saharan Africa, yet most sectors remain nearly single-gendered. Occupational segregation is a driver of the gender gap (Blau and Kahn, 2017; Goldstein et al., 2019). Mechanics, welding, and carpentry (our focus) are among the higher-return vocational sectors in the country and a target of policies promoting women's entry. Demand-side barriers remain understudied, particularly in sub-Saharan Africa (Heath et al., 2024), in part because supply constraints limit the pool of female applicants.[1] Whether employers in these sectors would hire women given access to female candidates is an open question.

[1]Labor supply decisions are shaped by intra-household bargaining (Bursztyn et al., 2020; Kala and McKelway, 2025), social norms (Fernández et al., 2004; Jayachandran, 2015; Agte and Bernhardt, 2023), and constraints such as transport, safety, and childcare (Field and Vyborny, 2022; Amaral et al., 2023; Halim et al., 2023; Ajayi et al., 2023).

Most employers in our sample (84.9% of 921) state a preference for hiring more women than they currently employ.[2] Employers value multiple worker attributes and hold gendered beliefs about most. When asked to identify the most important trait in a worker, employers cite trustworthiness (good behavior/no stealing), effort, technical skills, learning, and cooperation.[3] Men are perceived as stronger on technical performance and effort, but employers can discipline effort through piece-rate pay (Foster and Rosenzweig, 1994) and address skills gaps through on-the-job training. The attributes that are hardest to find and monitor are trustworthiness (90.4%) and cooperation (46%), both tied to well-documented frictions in low-income settings: monitoring constraints over worker behavior (Heath, 2018) and teamwork (Hjort, 2014). Women are perceived as more trustworthy, the attribute most valued under monitoring frictions. Employers do not perceive women as less cooperative, but over 40% worry about the risk of harassment or distraction when hiring women. The gendered perception of friction-relevant attributes creates the conditions to test for our channel.

To measure revealed preferences for hiring women, we design an incentivized hiring experiment that holds supply fixed: employers evaluate and select from a gender-balanced pool of vocationally trained candidates. In the status quo, employers mostly hire through networks, which mitigate adverse selection and moral hazard (Chandrasekhar et al., 2020) but constrain the talent pool and, in male-dominated sectors, exclude women (Beaman et al., 2018).[4] In an incentivized hiring experiment following Kessler et al., 2019, we randomly assign a third of the employers to evaluate trainee profiles cross-randomized by gender and technical performance under business-as-usual conditions. Revealed choices broadly confirm stated preferences: 89.6% select at least one female CV, and 42% of selected candidates are women. Employers are not gender-neutral on average: the hiring gender gap, the male-female difference in hiring rates, is 10 percentage points, but the gap narrows with stronger stated preferences and disappears in the top quintile. Women face a hiring penalty but positive returns to entry.

We investigate whether the demand for female workers is shaped by workplace frictions, as opposed to technical performance or bias. Our main experimental variation relaxes the two frictions that bind when hiring new workers: monitoring constraints over theft risk and concerns about integrating new workers into the workplace. We test whether the hiring gender gap responds consistently with the gendered beliefs about friction-relevant

[2]The stated ideal gender mix is 20% women, about 10 times higher than the actual gender mix.

[3]We recode these categories from open-ended answers, and elicit beliefs for each of the 5 categories.

[4]Even when vocationally trained workers are available and have better skills, employers perceive an experimentation cost from hiring outside their networks (Breza and Kaur, 2025; Groh et al., 2016).

worker attributes we document. The design has three conditions. One-third of the employers, described above, receive no audits and measure preferences under business-as-usual (control group). The remaining two-thirds also evaluate candidates, but are randomized to one of two monitoring support technologies, implemented through workplace audits.[5]

Employers in the Monitoring-Behavior (MB) arm receive audits focused on deterring new workers from stealing and dishonesty. This addresses a principal-agent problem. Our prediction is that reducing the cost of monitoring new workers increases the gender gap. Monitoring-Safety (MS) provides audits to ensure new workers' safe and respectful treatment, testing how reducing concerns related to the integration of new workers in the workplace affects the gender gap. We think of these two frictions as separate, although in our context the same technology (audits) can address both, which allows for a cleaner experimental design. We are careful to frame both audits as gender-neutral, as one may be concerned that safety audits are themselves a gendered treatment.[6] Notably, unlike technologies that directly shift the comparative advantage of one gender over another in production (Alesina et al., 2013), neither audit changes the production technology of the firm. Our interventions change the cost of experimenting with new workers, whether through freeing employer time from monitoring, reducing expected losses from theft, or potentially social concerns.[7]

We test two interconnected hypotheses. Monitoring support for workers' behavior should lower the relative demand for women by reducing the value of trustworthiness, widening the hiring gender gap. Monitoring support for workplace safety should increase demand for women by addressing harassment concerns, narrowing the gap. Because both audits involve sending an external visitor, each arm may partly generate effects intended by the other; these spillovers work against us finding opposite effects on the gender gap, making our treatment effects conservative estimates. The main identification concern is that behavior audits widen the gender gap not through the monitoring channel but through employer scrutiny: employers may select fewer women to avoid being observed around female workers. Under behavior audits, both channels predict a wider gender gap, so we cannot separate them. Safety audits help distinguish the two: scrutiny predicts a

[5]The audits mimic support provided by staff at some vocational training institutes during internships and consist of weekly visits of our enumerators during the 1-month probation period. Audit visits are unannounced in timing but known to occur by trainees and employers. We highlight to employers that treatments are randomized; there is no selection into treatment on either firm or worker side.

[6]In similar contexts, as in Hjort (2014), these integration frictions stem from ethnicity and exist within same-sex workers.

[7]While the interventions may also affect worker productivity, we think of this as second order relative to the change in employer costs. Our design does not rely on distinguishing between these channels.

wider gap, while reducing integration concerns predicts a narrower one. The comparison between the effect of safety and behavior audits therefore provides a mechanism test of our hypothesis. The cross-randomization of gender and technical performance allows a separate test for statistical discrimination on performance.

As predicted, the two treatments have opposite impacts on the gender gap. Both behavior and safety audits increase total selection relative to the control group, but through different margins: behavior audits increase male selection by 5.2 percentage points ($p < 0.001$) with no significant effect on women ($p = 0.480$), while safety audits increase female selection by 7.4 percentage points ($p < 0.001$) with no significant effect on men ($p = 0.399$). Employers assigned to behavior audits show a gender gap of 16.7 percentage points, 62% larger than under business-as-usual. This widening is stronger among employers with stronger stated preferences for hiring women: among those expressing any such preference, the gap doubles; among those above the median, it more than triples. Safety audits increase the relative demand for women. The gender gap under safety audits is 1.7 percentage points, statistically indistinguishable from zero, a six-fold reduction relative to business-as-usual. That safety audits increase rather than decrease demand for women rules out that audit visits deter hiring through scrutiny.

We show that solving monitoring frictions reveals hidden discrimination. One interpretation of the treatment effects is that frictions change the costs of hiring different workers, and employers reshuffle the workforce accordingly. Under this interpretation, the shift in the gender composition of hiring reflects profit-maximizing responses to changing costs, with implications for allocative efficiency but not for discrimination. This interpretation breaks down under two conditions: if any part of the gender gap reflects bias rather than efficient screening, then the trust premium masks discrimination rather than offsetting a productivity difference; and if employers' beliefs about gendered attributes are inaccurate, then the screening itself is not efficient. We find evidence for both. First, while high-performance profiles are more likely to be selected (10.1 percentage points; $p < 0.001$), the effect does not differ by gender ($p = 0.307$), ruling out that employers use gender as a proxy for technical performance. Second, a lab-in-the-field experiment with 182 trainees shows that employers' beliefs are inaccurate on both dimensions. On technical performance, we find no gender differences in practical skills or in a sector-specific theory exam; yet employers expect women to perform worse. On trustworthiness, a novel behavioral task we design measuring cheating in a workplace-like setting finds no gender differences in misreporting output for pay; yet employers expect women to cheat less when unsupervised. Finally, benchmarking the extent of harassment concerns with the Safety

treatment arm, we show that at most half of the hidden penalty can be attributed to safety concerns.[8] The remainder is inconsistent with efficient screening and reflects bias.

How much bias do women face? Our results show that residual estimates underestimate bias in the status quo. Leveraging our experimental data, we can also show that bias estimates are sensitive to which attributes the researcher measures as determinants of hiring and whether they are assumed to reflect productivity or bias. A one-dimensional framework that considers only technical performance would attribute the full 10.3 percentage point gap to bias (since this does not respond to performance signals). Accounting for trustworthiness increases the estimate to 16.7 pp. Treating harassment concerns as productive reduces it; the bias estimate ranges from 1.7 to 16.7 percentage points depending on which attributes are accounted for and assumed productive (Appendix A.1).

Our paper contributes to the literature testing for discrimination in the labor market (Goldin and Rouse, 2000; Bertrand and Mullainathan, 2004; Kline et al., 2022). Recent work takes a systemic view, showing that discrimination in one domain feeds into others (Lang and Spitzer, 2020) and that the structure of the evaluation process shapes discrimination (Bohren et al., 2025). We highlight a different systemic channel: not how employers evaluate workers, but which attributes matter for hiring. When these attributes carry identity connotations, changes in the production environment widen or narrow hiring disparities, even if the production technology remains unchanged.

Modeling workers as multidimensional, a perspective increasingly emphasized in the labor literature (Deming, 2017; Woessmann, 2024; Deming and Silliman, 2025), has implications for measuring discrimination. Failing to account for determinants of hiring beyond standard productivity measures—such as trustworthiness or cooperation—creates an omitted variable bias problem. If the omitted attribute favors the discriminated group, as in our setting, the observed gap understates bias. This reverses the standard direction of omitted variable bias highlighted by Lang and Spitzer (2020), and adds to a growing literature cautioning against interpreting outcome gaps as evidence for or against discrimination (Bohren et al., 2019, 2025; Macchiavello et al., 2026).

Our findings contribute to a growing literature at the intersection of labor and development economics (see Breza and Kaur, 2025 for a recent review). This literature identifies multiple coexisting frictions in poor countries' labor markets. Relevant to this paper, there is evidence of information frictions that make it hard for workers to signal skills

[8] Whether safety concerns constitute bias (via paternalism or reputational concerns) or should be considered as statistical discrimination is debatable.

(Abebe et al., 2021; Bassi and Nansamba, 2022; Carranza et al., 2022), on the importance of identity (Cassan et al., 2022; Oh, 2023), and that monitoring constraints shape hiring composition in that firms facing monitoring constraints rely more heavily on referral hiring (Heath, 2018). We differ from this literature in that we experimentally manipulate the workplace environment to solve frictions, and vary two frictions simultaneously, allowing us to compare how each shapes hiring composition along identity lines. We make two contributions. First, we show that frictions raise the importance of soft skills such as trustworthiness relative to technical performance, and that employers use identity as a proxy for both. Second, when multiple frictions coexist, solving one does not necessarily improve allocative efficiency. In our setting, two frictions have opposite effects on the hiring gender gap. This provides micro-level evidence of the second-best concern (Lipsey and Lancaster, 1956; Amodio et al., 2025): solving frictions can unintentionally worsen outcomes when it removes a distortion that was offsetting another.[9]

Finally, a rich literature studies barriers to female labor force participation (see Heath et al., 2024 for a review). 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). On hiring, Buchmann et al. (2023) show that paternalistic safety concerns lead employers in Bangladesh not to hire women for a night-shift job, even when female applicants are willing to take it. Our results are consistent: alleviating workplace integration concerns related to harassment increases demand for women in the male-dominated sectors we study in sub-Saharan Africa. We show that frictions can also push in the opposite direction. Monitoring frictions generate demand for women through the trust premium, making gender discrimination—whether taste-based or paternalistic—costly for employers in a Beckerian sense. Which frictions bind determines whether demand-side distortions favor or penalize women, which may help explain why supply-side interventions such as vocational training yield mixed effects (McKenzie, 2017; Bandiera et al., 2022).

[9]This does not require assumptions that men and women are equally productive; regardless of productivity differences, the two interventions cannot both be efficiency-improving.

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 921 employers, one per firm: 318 from motor mechanics, 300 from carpentry, and 303 from welding (see Table 1).[10] We selected firms in the neighborhoods of the greater Kampala area where these sectors are clustered, as shown in Appendix Figure A1. 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 hiring a trainee within nine months; and (iv) proficiency in either English or Luganda.

In our sample, 84.9% of respondents are business owners; 15.1% hold managerial positions. The firms are well established, employing on average 8.3 workers (median 5). They serve approximately 11.8 clients per day, operate for over 10 years, and report mean monthly profits of USD 286 on revenues of USD 718.[11] 96.4% of employers are men, and only 2.26% of all workers are women (173 out of 7,657 employees). 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-dominated sectors including automotive mechanics, electrical installation, welding, and

[10] Of these, 906 are retained in the main analysis after applying preregistered attention checks. We initially focused on a single sector, motor mechanics. In the preregistration, we noted that we might expand to additional sectors for power and external validity, depending on the number of mechanics we were able to interview. After the first wave, we decided to expand the study to carpentry and welding in the second wave. The preregistration was amended before expanding data collection and includes a document that summarizes the motivations (see Appendix A.2 for details).

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

carpentry. Although the program primarily targets women, it also trains male students (see Appendix Figure A2, 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 A13).[12] 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.

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 A2, 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.

Employers respond to this residual moral hazard through network hiring: around 80% of current workers were recruited through personal networks (Heath, 2018). 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

[12] 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.

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% report none at all.[13] 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 Stated Preferences for Worker Gender Mix

When asked, "What is the best gender composition of the workers in your firm?" (on a scale from 0 to 10 women out of 10), 84.9% of employers report preferring at least one woman out of ten workers, and 70.5% prefer two or more (Figure 2). 14.6% state a preference for a roughly gender-balanced workforce of four to six women out of 10. These stated diversity preferences are notable given the well-documented role of gender norms (Fernández and Fogli, 2009; Jayachandran, 2021), male status concerns (Bernhardt et al., 2018), and male backlash (Abou Daher et al., 2025) in limiting women's labor force participation in low-income settings.

The stated preferences show little indication of experimenter demand effects, with no bunching at focal values such as 1 or 5 women out of 10. Most employers attribute the gap between desired and actual gender mix to supply constraints: there are few women in the applicant pool, which, as noted above, is mostly network-based. The data are consistent with this account. The firm-level share of female workers (2.7%) roughly tracks the perceived share of female applicants (3.6%). These stated preferences may also be consistent with evidence that gender norms are perceived as more binding than they actually are (Bursztyn et al., 2020).

2.4 Stated Preferences for Worker Attributes and Gender Beliefs

When asked to identify the most important trait in a worker, employers most frequently cite good behavior, described as trustworthiness, honesty, and not stealing (Figure 1,

[13] 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.

panel A). 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 monitor, consistent with the prevalence of piece-rate pay, which makes effort contractible (Foster and Rosenzweig, 1994). Employers also mention learning interest (14.24%) and, in a few cases, cooperation. Physical strength is never mentioned.

Employers hold gendered perceptions of these attributes. Following Macchiavello et al. (2026), for each of the five traits mentioned, 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. Men are perceived as stronger in technical domains (technical skills, interest, and effort), while women are perceived as more trustworthy (Figure 1, panels C and D). Employers perceive men and women as similarly cooperative.

We also ask employers directly about their concerns when hiring women (Figure 1, panel E).[14] Many responses echo the trait data: employers worry about skills (mostly strength) and interest in the job (mostly leaving).[15] Stealing is never mentioned, consistent with women's perceived trustworthiness. The most notable finding is about cooperation. Harassment (from men to women) or, less frequently, distraction (from women to men) are the most frequently cited concerns when hiring women, mentioned by over 40% of employers. Thus, mixed-gender workplaces appear to introduce frictions similar to those documented in inter-ethnic contexts (Hjort, 2014).

Cooperation frictions raise the cost of integrating women into male-dominated workplaces. They can operate through several channels. They may impose costs on employers outside the production function: reputational concerns, time spent managing interpersonal conflicts, or perceived liability for workers' safety. Harassment or distraction may also reduce the productivity of female workers directly, for instance through mental health effects, or lower teamwork productivity, as documented in mixed-gender teams in other settings (Ronchi and Salvestrini, 2025).

[14] Gendered perceptions are elicited only for the traits employers mention as most important, a set likely shaped by their predominantly male workforce. Asking about women directly captures additional concerns this approach may miss.

[15] Interestingly, no employer mentions strength as a relevant attribute for their business.

3 Experimental Design

We measure revealed preferences for hiring female and male trainees in an incentive-compatible experiment with 921 employers in Kampala. Employers evaluate and select from a gender-mixed pool of vocationally trained candidates, providing access to female workers outside their usual networks. We manipulate the constraints under which employers make hiring decisions by randomly assigning monitoring support that targets the two frictions that bind when hiring new workers: theft risk and workplace integration. This allows us to identify how workplace frictions shape demand for women and the hiring penalty against them.

We think of our treatments as lowering the costs of experimenting with new hires along specific dimensions. Matching interventions of this type typically offer a wage subsidy to reduce experimentation costs (Alfonsi et al., 2020). We deliberately do not: if a wage subsidy already compensates employers for the cost of experimenting with new workers, the frictions our treatments target become less binding, reducing their power to detect which friction drives the gender gap. The test has power under the joint assumption that employers base selection on friction-relevant attributes — otherwise we would expect no treatment effect — and that they perceive men and women as differing along these attributes — otherwise we would expect no differential effect by gender.

3.1 Incentivized Ratings Experiment

To measure the relative demand for female workers, our experiment follows an incentivized ratings design (Kessler et al., 2019). We provide employers with a gender-balanced pool of hypothetical trainee profiles. Based on their ratings of these profiles, employers receive referrals of real candidates for probationary hire. Our primary outcomes are the share of female candidates selected and the gender gap in hiring.[16]

Profiles. We collect administrative data on 472 trainees enrolled with partner VTIs in Kampala since 2016. From this information, we create 36 base profiles by cross-

[16]We focus on the hiring margin rather than the wage gap. In this setting, as noted by Alfonsi et al. (2020), 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.

randomizing attributes such as age, marital status, possession of a driver's license, language spoken, motivation for career choice, education, training background, and references.[17]

We then generate four versions of each base profile by cross-randomizing two additional attributes: gender (signaled by an avatar) and technical performance (indicated by class rank, visualized as a star rating). Performance signals are anchored to the actual distribution within each VTI cohort: median performers receive three stars, top performers (above the 95th percentile) receive five stars. This yields 144 unique profiles (4 versions $\times$ 36 base profiles; see Figure 3 for an example). Each employer reviews one randomly assigned version of each profile, creating within-employer variation in both gender and performance. This design allows us to estimate the hiring gender gap, the male-female difference in hiring rates, conditional on performance signals, and to test whether performance signals affect the gender gap, as in the standard statistical discrimination test.

Each employer evaluates 24 of the 36 base profiles, each shown in one randomly assigned version, during regular work hours.[18] Our primary outcome is *Meet*: whether the employer wants to initiate a probationary hire ("Do you want us to refer you a similar worker to start the probation period at your firm?"; Yes/No). Secondary outcomes probe employers' perceptions of work quality (rated 0–10), behavior and trustworthiness (rated 0–10), and expected monthly earnings one year from now.[19]

Incentives. The incentivized ratings design provides incentives for decision-makers to select candidate profiles according to their true preferences, while avoiding deception. Employers are told that profiles are hypothetical but that referrals will be made from a pool of real candidates at partner VTIs based on their choices in the experiment. Before evaluating profiles, each employer is informed that more accurate ratings lead to better matches.

[17] We pilot-test these profiles with 25 employers out of sample.

[18] We calibrate the number of evaluations to reflect the average monthly inflow of jobseekers (24.6) reported by employers during scoping activities.

[19] *Work Quality*: "How would you rate the worker'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 workers' behavior (trustworthiness and honesty)? Please, rate on a scale from 0 to 10, where 0 is not at all trustworthy/honest and 10 is very trustworthy/honest." *Earnings*: "What is your best guess of the monthly earnings of this worker a year from now?". As discussed in Appendix A.2, we initially preregistered two primary outcomes, the second being an indicator of the desire to make an offer. Due to a programming error, this outcome was not elicited in the first wave (motor mechanics); consequently, we chose not to collect it in the second wave (welding and carpentry) either.

Because the mapping from ratings to referrals is not fully disclosed, selection may partly reflect strategic behavior rather than preferences. However, this is only problematic if strategic behavior differs by gender (Litwin and Low, forthcoming). Since candidate characteristics are independently and randomly assigned conditional on gender, employers do not have to manipulate the gender composition of referrals to avoid the risk or maximize the likelihood of being referred to certain candidates.

Competition for female candidates induces employers to avoid selecting women; however, this mechanism would bias measured demand downward. Upward bias would require employers to strategically avoid men, which is implausible given that vocationally trained men are perceived as high quality and dominate the pool. Referrals are also stratified geographically, limiting competition for specific candidate types.

Social incentives, such as the desire not to look discriminatory, may also attenuate the referral incentives. Existence or awareness of affirmative action policies is also limited.[20] Nonetheless, inspired by De Quidt et al. (2018), we assess the problem by asking employers: "In your opinion, would it be better for us if you hired men or women or it does not matter?" The majority (66.7%) report it does not matter, and only 6.8% believe the research team prefers they hire women, suggesting that the scope for social desirability bias is limited. We test that our results are robust to excluding the subsample of employers that show social desirability bias.

3.2 Treatment Assignment

Employers are randomly assigned to make hiring decisions under one of three experimental conditions: a Business-as-usual arm and two monitoring support regimes. Treatment assignment is stratified by enumerator, survey wave, and sector. The Business-as-usual arm, beyond measuring the relative demand for trained female workers under status-quo conditions, serves as the control group (C).

In both treatment arms, monitoring support is implemented via weekly unannounced visits from our enumerator team during the probation period. These external audits mirror the support provided by VTI staff during internships. Employers are explicitly told that receiving audits is random and does not reflect any characteristics of the firm or

[20] Only 18.7% of employers had heard of the United Nations Development Programme's Gender Equality Seal, the only such policy we could identify in this context.

workers. The script also explains that any selected worker will be aware of the monitoring regime upon starting the probationary period, so that treatment effects operate through expectations of monitoring on top of ex-post detection. However, workers will only learn of the audits after accepting the job, meaning that employers should not expect supply responses. We think of these audits as a monitoring technology randomly made available for free to the firm.[21]

The two monitoring support regimes differ only in the dimension they target. Employers assigned to the Monitoring-Behavior (MB) arm receive unannounced visits aimed at discouraging new workers' misbehavior and theft, targeting the cost of experimenting with new workers. Employers in the Monitoring-Safety (MS) arm receive unannounced visits focused on ensuring that new workers are safe and treated with respect, targeting how new workers are received in the workplace. Both arms receive identical scripts except for the targeted dimension; treatment assignment is conveyed in the script read to employers before they evaluate the profiles.[22] Both arms are framed gender-neutrally to avoid priming effects on demand for female workers. For example, we do not mention sexual harassment to avoid signaling a gendered concern.

Behavior audits target the primary monitoring friction in these sectors: deterring misbehavior and theft. These frictions make trustworthiness a valued attribute, and the dimension on which employers perceive women more favorably. We predict that reducing monitoring constraints on behavior lowers the relative value of trustworthiness in hiring, thereby widening the gender gap compared to business-as-usual conditions. Safety audits target cooperation frictions: the integration costs of new workers in the workplace (e.g., how they are received and treated). These frictions may apply to any diversity dimension,

[21] Audit-based interventions have been shown to affect behavior in developing-country settings, including reducing corruption (Olken, 2007) and pollution (Duflo et al., 2013).

[22] 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.

including religion or ethnicity, but in these male-dominated sectors, gender is the binding diversity dimension. The same logic applies to any context where cooperation frictions deter hiring from an underrepresented group: reducing them should narrow the gap.

Because both audits involve sending an external visitor to the firm, each arm may partly generate effects intended by the other: behavior audits could reduce harassment risk through the visitor's presence, and safety audits could deter theft. These spillovers attenuate the difference between the two arms, making our treatment effects conservative estimates. We nevertheless discuss three specific identification concerns.

First, the Monitoring-Behavior script explicitly mentions theft, which could prime associations between female workers and dishonesty or make theft risk more salient in the safety arm. We address this by deliberately including the same theft wording in the Monitoring-Safety script: employers in the safety arm are told that visits cannot deter stealing or dishonesty. Any priming effect is therefore present in both arms, and the comparison between them differences it out. This sentence also limits the scope for monitoring spillovers from safety audits, since employers are explicitly told that safety visits do not serve a monitoring function.

Second, safety audits may affect demand for women through scrutiny rather than safety: if employers themselves are a source of harassment, monitoring their conduct could deter them from hiring women. This channel predicts the opposite of our primary mechanism— scrutiny would widen the gender gap, while reducing integration concerns would narrow it. Because the effect could go either way, understanding the effect of safety audits is also relevant for policy: workplace safety audits are increasingly used to monitor compliance with labor standards (ILO, 2019), yet whether they encourage or deter the hiring of women is an open question.

Third, the same scrutiny concern applies to behavior audits, as a spillover effect. Behavior audits may deter employers from hiring women not because monitoring reduces the value of trustworthiness, but because employers want to avoid being observed around female workers. Both channels predict the same direction— a wider gender gap — so we cannot separate them within the MB arm alone. The comparison with Monitoring-Safety helps to identify whether scrutiny effects are relevant. If audits generate scrutiny effects, they should be stronger under Monitoring-Safety, which explicitly monitors workplace interactions to prevent harassment. Thus, Monitoring-Safety allows us to bound potential scrutiny spillovers in Monitoring-Behavior from above. If we observe significantly higher demand for women under Monitoring-Safety, where scrutiny is strongest, it

is unlikely that scrutiny spillovers are the main driver of lower demand for women under Monitoring-Behavior.

Beyond these identification concerns, a separate question is whether employers' hiring preferences should at all be affected by audits that only apply during the probation period. Employers report an expected audit duration of approximately 1.7 hours per visit, suggesting that interventions could meaningfully relax monitoring constraints during this period. The audits, however, do not permanently reduce the cost of hiring a worker; rather, they reduce the cost of experimenting with an unfamiliar one. As Macchiavello et al. (2026) show, even short-term reductions in experimentation costs can shift outcomes: in their setting, temporarily reducing the cost of promoting women into managerial positions has lasting effects. The probation period is precisely when monitoring and integration frictions bind hardest. In focus groups, employers report that they feel they can screen workers once they spend some time at the firm, suggesting that the initial period is where support matters most.

3.3 Empirical Strategy

We restrict the sample to the 906 employers (out of 921) who pass attention checks and exhibit variation in profile evaluations, following our pre-registered main sample protocol.[23] Table 2 shows covariate balance across the three experimental arms, including employers' characteristics and firm-level attributes.[24]

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:

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

Meet_{ijs} denotes employer j's interest in meeting candidate profile i to hire them 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

[23] Sample selection is uncorrelated with treatment assignment (Appendix Table A1).

[24] We find mild imbalances in one baseline variable: employer having formal vocational training. Results are robust to including this control.

for female relative to male candidates. δ_i and σ_s denote profile and strata fixed effects. Standard errors are clustered at the employer and profile levels.

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

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

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.

4 Results

4.1 Experimental Demand for Female Workers

We examine hiring choices in the Business-as-usual arm ($N = 326$). Employers' revealed hiring choices are broadly consistent with their stated preferences for gender mix. Women face positive returns to entry: when given access to a trained applicant pool, 89.6% of employers select at least one female trainee (Figure 4). Of the profiles shown to employers, 47% are female; among all selected candidates, the average share of women is 42.2% (median 44.1%). We refer to these selection choices as relative demand for female workers, since the experimental design identifies them against an identical male candidate.

Yet employers are not gender-neutral on average. As shown in column 1 of Table 3, Panel A, women are 10.3 percentage points less likely to be selected than men with otherwise identical profiles (equation 1). The hiring gender gap is equivalent to 87% of the hiring premium of a top- relative to an average-performer profile, and is statistically significant ($p < 0.001$).[25] The gap is moderate relative to the extreme gender segregation of the Ugandan labor market, where only 2.3% of workers in these sectors are women. Relative to the mean selection rate of men, this corresponds to a gap of approximately 22%. The gap is present across all preference groups, though smaller among employers with some stated preference for hiring women.

Stated preferences predict the experimental gender gap, as shown in Figure A3, panel A: the gender gap narrows significantly across the preference distribution. A 10-percentage-point increase in stated gender-mix preferences is associated with a 4.95 percentage point narrowing of the gap against female candidates ($p < 0.001$), equivalent to a 23.4% reduction relative to the gender gap among employers who report a preference for all-male ideal workforce composition.[26] Employers in the top quintile, those with a stated ideal gender mix of at least 40% women, behave as gender-neutral.

4.2 Do Behavior Audits Increase 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 4.1% more likely to select a trainee ($p = 0.090$), consistent with audits reducing monitoring constraints (Appendix Table A6).

The higher selection rate under MB operates entirely through a higher likelihood of selecting male trainees, consistent with our hypothesis. Figure 5 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

[25] Equivalently, the gap corresponds to 1.4 standard deviations of the trainee exam score distribution (approximately 1.4 GPA points on a five-point scale).

[26] See Appendix Table A11, Column (1).

with a 5.2 percentage point higher male selection rate relative to the Business-as-usual arm (Table 4, column 1; $p < 0.001$), with no statistically significant effect on female candidates ($p = 0.480$).

Providing support in monitoring workers' behavior widens the hiring gender gap against women relative to business-as-usual conditions. Table 4 estimates equation 2 on the full sample. The gender gap under behavior audits is 16.7 percentage points, compared to 10.3 percentage points under business-as-usual, a 62% increase. The difference is statistically significant at the 1% level ($p = 0.008$). As an alternative benchmark, the widening of the gender gap induced by behavior audits is equivalent to 54% of the hiring premium of a top- relative to an average-performer male profile (11.8 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 6 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 five times more likely to select no female trainees at baseline.[27]

The effects are stronger among employers with any stated preference for hiring women: the gender gap under MB is 15.5 percentage points, about twice the 7.8 percentage points under business-as-usual. The effect is even stronger among employers in the top quintile of the stated preferences for hiring women, who behave as gender-neutral under business-as-usual. Among employers with the strongest stated preferences for hiring women, those assigned to business-as-usual show a gender gap statistically indistinguishable from zero, while those assigned to MB show a gap of 15.2 percentage points ($p < 0.001$) — an effect 2.7 times larger than the full-sample estimate (Table 4; see also Appendix Figure A3 and Table A11).

Stated preferences for gender diversity lose most of their predictive power for hiring choices once monitoring frictions are relaxed under Monitoring-Behavior. Under business-as-

[27] 33 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.

usual, employers in the top quintile of stated preferences for hiring women are 1.68 times more likely to select women than employers in the bottom quintile. Once monitoring frictions are relaxed, employers in the top quintile of the stated preferred gender mix distribution exhibit gender gaps comparable to those of employers with no stated preference for hiring women.

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.

4.3 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 6.4% more likely to select a trainee ($p = 0.012$), consistent with audits reducing monitoring constraints rather than deterring hiring through scrutiny (Appendix Table A6).

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 7.4 percentage points higher relative to the Business-as-usual arm (Table 4, column 1; $p < 0.001$), with no effect on male candidates ($p = 0.399$). Figure 5 illustrates this pattern: female selection rises from 41.9% to 49.3%, while male selection stays flat.

Safety audits close the hiring gender gap. Table 4 estimates equation 2 on the full sample. The gender gap under safety audits is 1.7 percentage points, compared to 10.3 percentage points under business-as-usual, an 83.5% 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 dominat-

ing the safety channel. Indeed, safety audits neither deter hiring overall, nor reduce the selection rate of female trainees.

We also find that, unlike behavior audits, the effect of safety audits does not concentrate among employers with the strongest stated preferences for hiring women. The effect is broadly uniform across the preference distribution, except for the bottom quintile, where employers who prefer not to hire women do not respond to safety audits ($p = 0.770$; Table 4, columns 3–7; Appendix Figure A3). If anything, point estimates are strongest in the mid quintiles. Neither linear nor quadratic interactions with stated preferences pick up this difference.[28] This pattern is consistent with two types of employers: those with a preference for not hiring women, for whom safety audits have limited scope, and those with some preference for hiring women, for whom safety constraints are more binding.

4.4 Mechanism Discussion and Robustness

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 harassment of new workers, the dimension on which women are perceived as a liability, increases the relative demand for women and narrows the gap.

An important result to pin down the mechanism is that Monitoring-Safety increases the relative demand for women in the experiment, implying that scrutiny spillovers are not first order. One may worry that safety audits generate both scrutiny and safety effects that partially offset, so that the net positive effect on the demand for women understates the scrutiny component. This would requires employers to simultaneously plan to mistreat female workers (making scrutiny costly) and worry about other workers harassing them (making safety audits valuable). These motivations are not impossible to reconcile, but are rather contradictory. A subtler version of this concern is that effects are heterogeneous: some employers respond to scrutiny and some to safety. The distributional evidence rules this out. If scrutiny effects were concentrated among a subset of employers, some would increase and others decrease the demand for women under safety

[28] Appendix Table A11 reports $p = 0.767$ for the linear interaction (column 4) and $p = 0.541$ for the quadratic (column 5).

audits, generating crossings in the distribution. Instead, the Monitoring-Safety gap distribution shifts entirely to the left of the control with no crossing, narrowing the gap for every employer subgroup (Appendix Figure A5). Also, to explain the difference between the Monitoring-Behavior and the Control arms through scrutiny, the safety effect would need to be nearly twice as large as what we observe—large enough to both overcome the scrutiny deterrent and still produce the increase in female hiring we measure.

The mechanism we propose requires that stated gendered perceptions of worker attributes are also reflected in how employers actually evaluate candidates. Pre-registered secondary outcomes confirm this: holding other attributes constant, female profiles are rated as more trustworthy but as having lower technical performance (Appendix Table A5). These patterns suggest that the same attributes employers state to care about also shape baseline demand for female workers and the treatment effects.

Results are robust across empirical strategies. Within-subject results are consistent in sign and stronger in magnitude: Monitoring-Behavior roughly doubles the gender gap (12.2 percentage points, $p < 0.001$) and Monitoring-Safety reduces it by 14.9 percentage points ($p < 0.001$; Appendix Table A12). The larger magnitudes likely reflect heterogeneity in employer baseline preferences muting the between-subject estimates.

Finally, experimenter demands are unlikely to drive demand for female workers or explain the treatment effects. As discussed in Section 3.1, we find little evidence of social desirability bias: only 6.8% of employers believe we prefer they hire women. Dropping these respondents from the sample, if anything, strengthens the results (Appendix Table A9).

5 Hidden Gender Discrimination

Our results reveal an underlying hiring penalty against women in the status quo, masked by a trustworthiness premium sustained by monitoring constraints. Two questions follow. Does the hidden penalty reflect efficient discrimination or bias? And how much bias do women face?

To answer the first question, we leverage the cross-randomization of gender and performance signals to test for statistical discrimination on technical performance, and we assess the accuracy of employers' beliefs using the trainee data. If employers hold accurate beliefs about women's skills and use gender to identify the most productive workers

under each production technology, the hidden discrimination reflects efficient screening. If the gender gap does not respond to performance signals, or if beliefs about women's skills or trustworthiness are inaccurate, then solving frictions reveals bias. Moreover, if beliefs about trustworthiness are wrong, the premium that offsets discrimination is itself a distortion. To answer the second question, we use the comparison between the three experimental conditions to decompose the gender gap into its components and estimate the extent of the hidden gap that can be attributed to bias, versus safety concerns or productivity. We also show that the residual estimate of bias depends on which attributes the researcher measures and treats as productive.

5.1 Statistical Discrimination on Technical Performance

The hidden gender penalty could reflect efficient screening rather than bias. If employers who appeared gender-neutral under business-as-usual were hiring women for their perceived trustworthiness, removing that advantage would shift hiring toward technical performance, where men are perceived as stronger. Under this interpretation, the widening gender gap under Monitoring-Behavior may reflect a efficient statistical discrimination based on technical performance. If so, this means that employers use gender as a proxy for technical performance, and in turn, the hiring gender gap should be lower among candidates with high-performance signals (Bertrand and Duflo, 2017).

We test this prediction using the cross-randomization of gender and technical performance signals in the worker profiles. 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 A2).[29] As a sanity check, employers place substantial weight on this signal: high-performance profiles are 10.1 percentage points more likely to be selected (22.9%; $p < 0.001$).

We find that female and male candidates benefit equally from the performance signal under business-as-usual conditions. The interaction between technical performance and gender is small, negative, and statistically insignificant ($p = 0.312$): the gender gap does not narrow with higher observed technical performance. 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

[29] We also present sample-split estimates in Appendix Table A3 and Figure A4.

the probability of selection by 13.3 percentage points ($p < 0.001$) but does not reduce the gender gap ($p = 0.454$; Appendix Table A4).

Across all three arms, the gender gap is orthogonal to technical performance signals, ruling out statistical discrimination on this dimension. The treatment effects are inconsistent with statistical discrimination on technical performance: solving safety or 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 for both performance levels ($p = 0.871$). Under Monitoring-Safety, the pattern is symmetric: female selection rates rise uniformly across performance levels ($p = 0.758$).

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 gender gap does not differ between employers who believe female workers attract customers and those who do not (Appendix Table A10).

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. Our sample of trainees is described in Appendix Table A13.

We find that employers' beliefs are inaccurate on both dimensions. 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).[30] 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 A6. The results of both trainee outcomes and employers' beliefs by gender are summarized in Figure 7 and in Appendix Table A14.

Employers perceive female workers as less skilled, both on practical and theoretical tasks. On average, employers expect women to be 20.8% and 25.4% less likely than men to answer correctly on theory and practical exams, respectively (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 ($p = 0.831$). On the theory exam, women perform at least as well as men, if anything outperforming them, though the difference is not statistically significant (8.3 percentage points; $p = 0.068$).[31]

Trustworthiness. The literature typically measures trust using the investment game (Berg et al., 1995), but this game captures beliefs about how much another party will reciprocate trust, conditional on having been trusted. Similarly, survey-based measures, such as those in the World Values Survey, elicit generalized trust or prosociality. The die-rolling paradigm (Fischbacher and Föllmi-Heusi, 2013) detects dishonesty statistically by comparing self-reported outcomes to known distributions, but this is very distant from the type of stealing or moral hazard that employers are concerned about in our setting. In a production setting closer to ours, Caria and Falco (2024) design an incentive-compatible game to measure moral hazard, but they focus on unsupervised effort in productive tasks. None of these definitions captures trustworthiness as honesty and non-stealing. Finding that employers' beliefs are inaccurate on these outcomes would not be informative. We

[30] 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?".

[31] 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 23.4 percentage points more likely to answer correctly ($p = 0.017$); see Table A14.

therefore design an incentivized task measuring both effort and cheating, supervised and unsupervised.

Our "Cheating" task is as follows. We randomly assign trainees to supervised or unsupervised conditions. In the unsupervised condition, trainees receive a flat wage to complete up to 12 proofreading tasks, with no output verification. In the supervised condition, trainees are paid piece rate, with supervisors verifying output and paying only for completed tasks. The number of completed tasks is our measure of output/effort. We introduce an incentive to cheat: trainees can increase their pay by 50% by claiming to find more than 10 proofreading mistakes, although by design there are fewer than 10, making this a direct measure of cheating. The sample is stratified by treatment and gender to estimate mean outcomes separately for each group.[32] Employers predict average task completion and likelihood of cheating for male and female trainees separately, under both supervised and unsupervised conditions. Results are presented in columns (7)–(11) of Appendix Table A14.

Most relevant to our question, employers expect unsupervised female trainees to misreport less than unsupervised men (11.5 percentage points relative to a perceived average likelihood of 46.6% for men, $p < 0.001$).[33] The effect is driven by both men and women: 25.6% of men cheat when supervised, and women are 5.7 percentage points less likely than men to cheat. Thus, overall employers expect women to exert more effort and cheat less than men, especially when unsupervised. Employers' beliefs are not in line with actual trainee behavior. Task completion is nearly universal (97%) across gender and supervision. Cheating is low on average (10.4%), with no statistically significant gender differences. Supervision substantially reduces misreporting (63%, $p = 0.014$), with no differential effect by gender ($p = 0.593$).

5.3 How Much Bias Do Women Face?

Failing to account for women's trustworthiness premium leads to underestimating bias against women. Building on the results above, in a standard framework where workers

[32] To avoid deception, individual output is not verified in the unsupervised arm. Effort can only be assessed at the treatment-by-gender cell level. As a sanity check, employers expect supervision to increase task completion by 21.2 percentage points ($p < 0.001$) from a baseline rate of 66.9%; this effect is driven by male trainees increasing effort under supervision, with the gender gap in completion disappearing under supervision ($p = 0.985$).

[33] Employers overall expect supervision to reduce the overall likelihood of cheating by 18.1 percentage points.

differ only in technical performance, the hiring gender gap (10.3 percentage points) would be attributed entirely to bias, since it does not respond to performance signals (and male and female trainees appear to have similar performance). But our experimental evidence reveals that, under monitoring constraints, trustworthiness is an attribute that employers base hiring on. Because employers perceive women as more trustworthy, women receive a trust premium under business-as-usual that partially offsets the penalty against them. Removing this friction under Monitoring-Behavior removes the trust premium, revealing 6.4 percentage points of hidden discrimination against women. The residual estimate of bias rises to 16.7 percentage points, 62% larger than under a one-dimensional model.[34] Given the evidence that employers' beliefs are inaccurate and that the gender gap is inconsistent with statistical discrimination on performance, the hidden 6.4 percentage points represents bias, not efficient screening.

Of course, our second treatment suggests that hiring decisions depend on more than two dimensions. Harassment concerns also affect demand for women, and may explain part of the penalty. We use the comparison between the three experimental conditions to estimate each component. Under an additivity assumption, the trust premium equals the difference between the Monitoring-Behavior and business-as-usual gaps (6.4 percentage points), and the harassment penalty equals the difference between the business-as-usual and Monitoring-Safety gaps (8.6 percentage points). The residual bias is 8.1 percentage points (Appendix A.1). Under the same assumption, the share of the gap explained by harassment concerns under business-as-usual applies proportionally to the gap revealed under Monitoring-Behavior. This allows us to split the 6.4 percentage points of hidden discrimination into hidden bias (approximately 3 percentage points) and hidden harassment concerns (approximately 3 percentage points).[35] Whether the harassment component should be counted as bias is not conceptually obvious: if harassment risk reduces women's productivity, these concerns reflect statistical discrimination; if instead they reflect paternalistic preferences, as in Buchmann et al. (2023), they are themselves a form of discrimination. Even under the most conservative interpretation, business-as-usual conditions underestimate bias against women by a meaningful extent: at least 30% of the hidden discrimination reflects bias.

[34] If monitoring frictions are not fully eliminated by our treatment, the true trust premium exceeds 6.4 pp and the true hidden discrimination exceeds what we measure. We note that if women are not actually more trustworthy, as our trainee data suggest, the trust premium itself reflects biased beliefs; in this case, biased in favor of women (or against men).

[35] This comes from a proportionality assumption: if harassment accounts for the same share of the gap under Monitoring-Behavior as it does under business-as-usual, then the 16.7 percentage point gap under Monitoring-Behavior is 48% bias (8.1 percentage points) and 52% harassment (8.6 percentage points), and the 6.4 percentage points hidden by trust splits in the same proportion.

More generally, our experiment shows that residual estimates of bias depend on which attributes are measured and treated as productive. In a multidimensional framework, omitting relevant attributes can both overstate and understate bias, complicating the standard presumption that omitting productive attributes leads to overestimating discrimination (Lang and Spitzer, 2020). In our setting, residual bias ranges from 1.7 percentage points (when both trustworthiness and harassment are accounted for) to 16.7 percentage points (when only trustworthiness is accounted for). Moreover, even once an attribute is measured, the researcher must take a stance on whether it reflects productivity or discrimination. What should be considered a determinant of productivity is itself a conceptual choice, with first-order consequences for the estimate of bias.

5.4 Discussion

This section establishes three facts about the hidden gender gap. First, the penalty does not reflect statistical discrimination on technical performance: high-performance signals do not narrow the gap. Second, as far as we can measure, male and female workers have comparable skills, and employers' beliefs are inaccurate. Third, the safety benchmark suggests that approximately half of the hidden gap operates through harassment and safety concerns (Appendix A.1). This evidence points to hidden discrimination against women.

A limitation of our experiment is that we cannot observe all attributes that employers value, nor establish whether safety concerns and trustworthiness operate only through costs borne by the employer or also through worker productivity. Safety concerns could reflect only employer costs — paternalism, reputational risk — or also genuine productivity differences, if harassment reduces women's output. As we show above, this distinction has first-order consequences for the estimate of bias. Similarly, women may be genuinely more trustworthy in ways our behavioral task does not capture, in which case the trust premium partly reflects productivity differences rather than biased beliefs alone. These limitations are not specific to our setting: it is difficult for any discrimination study to observe every attribute that employers value. Our contribution is to show that residual bias estimates depend on which attributes are measured and whether they are assumed to reflect productivity or discrimination, that the friction environment determines which attributes matter, and that failing to account for some dimensions can lead to underestimating bias.

Given the evidence that the gender gap does not respond to performance signals and that employers' beliefs about both technical performance and trustworthiness are inaccurate, we conclude that there is bias against women and that the trust premium reflects biased beliefs rather than real productivity differences. The trust advantage that offsets the gender penalty under business-as-usual is therefore itself a distortion. Under these two conclusions, our results provide experimental evidence of a second-best problem (Lipsey and Lancaster, 1956; Amodio et al., 2025): the monitoring friction generates compensating demand for women through inaccurate beliefs, partially offsetting bias. Solving it pushing the allocation further against women than efficiency would require.

Even without these assumptions— for example, if, as noted above, one questions the adequacy of our trustworthiness measure or the strength of the evidence on bias — the opposite-signed effects of the two interventions are themselves sufficient: the two cannot both be efficiency-improving, regardless of assumptions about productivity differences between men and women.[36]

6 Conclusions

We study hiring preferences for women in three male-dominated manufacturing sectors in Uganda and their relationship to workplace frictions. We document substantial unmet demand for female workers, despite employers perceiving women as less technically skilled. Employers also hold gendered beliefs about attributes such as cooperation and trustworthiness, whose value depends on the frictions employers face. We show that technologies addressing different frictions affect the hiring gender gap in opposite directions, predicted by employers' gendered beliefs about attributes. Reducing monitoring constraints over new worker behavior widens the hiring gender gap in the experiment; addressing harassment concerns by monitoring new workers' interactions closes it. Monitoring frictions hide discrimination, as the hiring gender penalty we uncover is not explained by statistical discrimination on performance. The extent of hidden discrimination varies significantly depending on which attributes we measure and consider as productive, and is concentrated among employers that have the strongest stated diversity preferences. These preferences predict the hiring gender gap—the larger the preferred share of women, the smaller the gap—and back-of-the-envelope calculations suggest that 56% of this predictive power is

[36]For these results, we focus on efficiency and we do not formally analyze welfare. However, in our setting women earn less than men and the sectors we study are among the higher-paying vocational occupations, suggesting that a more gender-balanced workforce may reduce earnings inequality.

attributable to the trustworthiness premium sustained by monitoring frictions (the share of the preference–hiring gradient that disappears when monitoring frictions are relaxed; Table A11).

Our experiment measures employers' hiring preferences, not realized labor market outcomes. How these preferences translate into retention and female labor force participation depends on many factors we do not observe. With this caveat, apparently gender-neutral technologies can have gendered effects on hiring. This need not be inefficient: if beliefs are accurate and workers' productivity is genuinely changing, resolving frictions improves allocations. However, when multiple wedges coexist, solving one friction can reduce efficiency in a second-best sense. In our setting, beliefs about productive attributes by gender are inaccurate on various dimensions and there is bias. Our experiment provides direct evidence of this logic, as the two interventions push the gender gap in opposite directions. Moreover, the fact that gender bundles multiple potentially productive attributes also complicates estimation of bias: positive perceived selection along friction-rewarded traits can lead to underestimating discrimination, and the researcher must take a stance on what employers treat as productive. More speculatively, since the production function defines which attributes are productive but is itself endogenous, discriminatory employers may choose production technologies that favor the attributes of preferred groups, blurring the line between taste-based and statistical discrimination.

The channel we document operates whenever three conditions hold: workers differ along multiple attributes, those attributes carry identity connotations, and frictions determine their relative weight in hiring. These conditions are common. Soft skills matter increasingly for labor market outcomes (**?**Deming, 2017), and are harder to signal credibly, particularly in developing countries (Carranza et al., 2022). Gendered perceptions of social preferences are widespread and often inaccurate (Exley et al., 2025). The same logic extends to any identity dimension—including race, ethnicity or caste—where group-correlated beliefs about non-cognitive attributes interact with workplace frictions.

Turning to policy implications, relatively simple safety interventions can meaningfully reduce the perceived cost of hiring women. Our treatment, however, operates on employers' beliefs about risk and may not improve actual workplace safety, unlike interventions that directly address harassment reporting (Boudreau et al., 2023) or safety through policing (Amaral et al., 2023). Our results also speak to the effectiveness of supply-side interventions: the same trained female worker may face a different hiring likelihood depending on which frictions bind.

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.

Agte, Patrick and Arielle Bernhardt, "The economics of caste norms: Purity, status, and women's work in India," *Job Market Paper*, 2023.

Ajayi, Kehinde F, Aziz Dao, and Estelle Koussoubé, "The effects of childcare on women and children: Evidence from a randomized evaluation in Burkina Faso," Technical Report, World Bank Policy Research Working Paper 2023.

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.

Amaral, Sofia, Girija Borker, Nathan Fiala, Anjani Kumar, Nishith Prakash, and Maria Micaela Sviatschi, "Sexual harassment in public spaces and police patrols: Experimental evidence from urban India," Technical Report w31734, National Bureau of Economic Research 2023.

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, 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*.

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," Technical Report, National Bureau of Economic Research 2023.

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.

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

Bertrand, Marianne and 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.

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

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

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*.

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

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

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

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.

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

Daher, Mohamad Abou, Hala Kobeissi, Holger Sieg, and Chamna Yoon Wang, "Drivers of change: Employment responses to the lifting of the Saudi female driving ban," *American Economic Review*, 2025, *115* (9), 3248–3271.

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.

Fernández, Raquel, Alessandra Fogli, and Claudia Olivetti, "Mothers and sons: Preference formation and female labor force dynamics," *The Quarterly Journal of Economics*, 2004, *119* (4), 1249–1299.

__ and __, "Culture: An empirical investigation of beliefs, work, and fertility," *American economic journal: Macroeconomics*, 2009, *1* (1), 146–177.

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

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.

Halim, Daniel, Elizaveta Perova, and Sarah Reynolds, "Childcare and mothers' labor market outcomes in lower-and middle-income countries," *The World Bank Research Observer*, 2023, *38* (1), 73–114.

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

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

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

ILO, International Labour Organization, "Violence and Harassment Convention, 2019 (No. 190)," https://www.ilo.org/dyn/normlex/en/f?p=NORMLEXPUB:12100:0::NO:12100:P12100_INSTRUMENT_ID:3999810 2019. Adopted at the 108th International Labour Conference, Geneva, 21 June 2019.

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

Jayachandran, Seema, "The roots of gender inequality in developing countries," *Annual review of economics*, 2015, *7* (1), 63–88.

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

Kala, Namrata and Madeline McKelway, "The Power of Persuasion: Causal Effects of Household Communication on Women's Employment," Technical Report, National Bureau of Economic Research 2025.

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.

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.

Macchiavello, Rocco, Andreas Menzel, Atonu Rabbani, and Christopher Woodruff, "Promoting Women to Managerial Roles in the Bangladeshi Garment Sector," Technical Report, Working Paper 2026.

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.

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

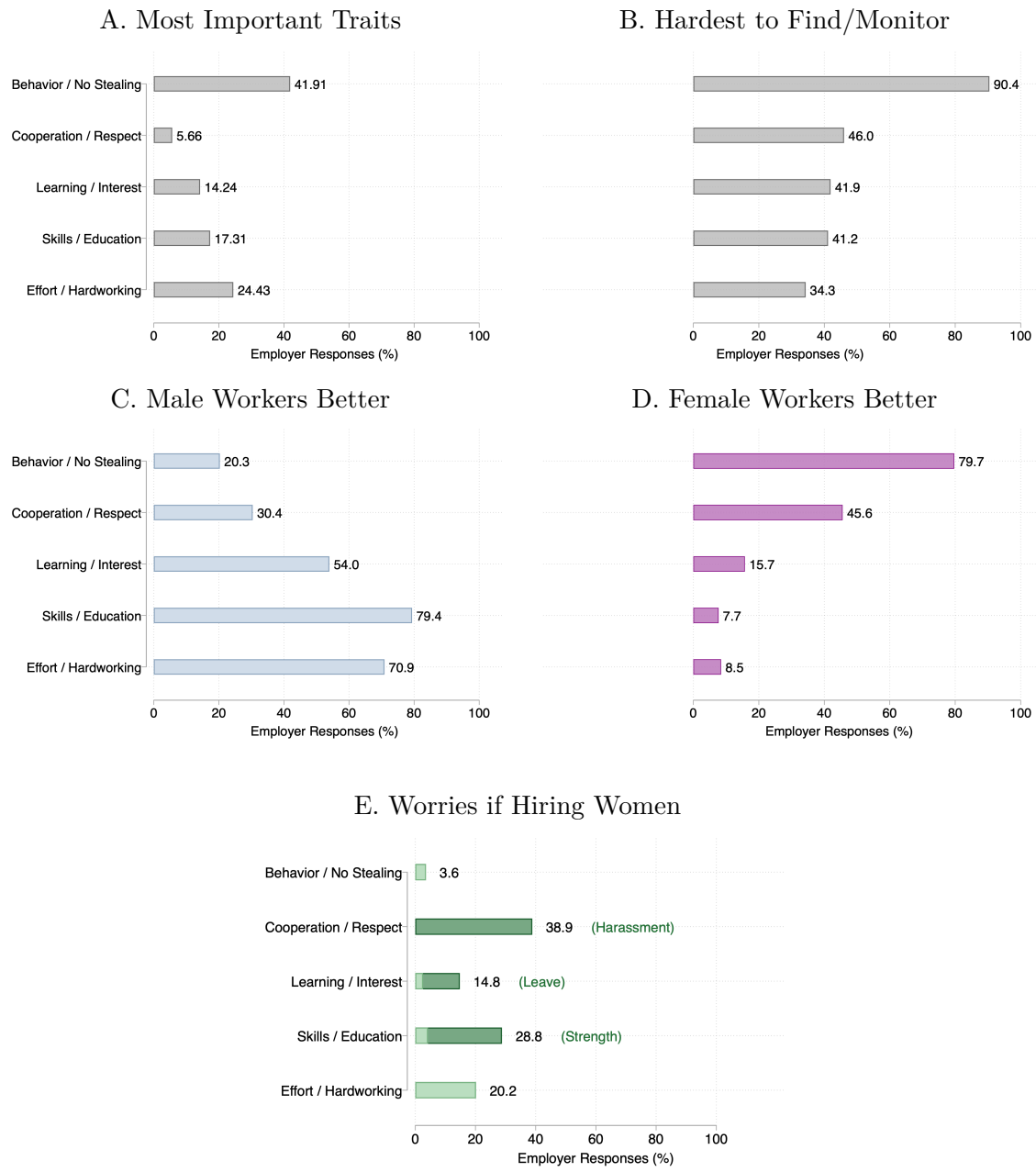
Quidt, Jonathan De, Johannes Haushofer, and Christopher Roth, "Measuring and bounding experimenter demand," *American Economic Review*, 2018, *108* (11), 3266–3302.

Ronchi, Maddalena and Viola Salvestrini, "Gender Diversity and Decision-Making in Teams," 2025. Working paper.

Woessmann, Ludger, "Skills and earnings: A multidimensional perspective on human capital," *Annual Review of Economics*, 2024, *17*.

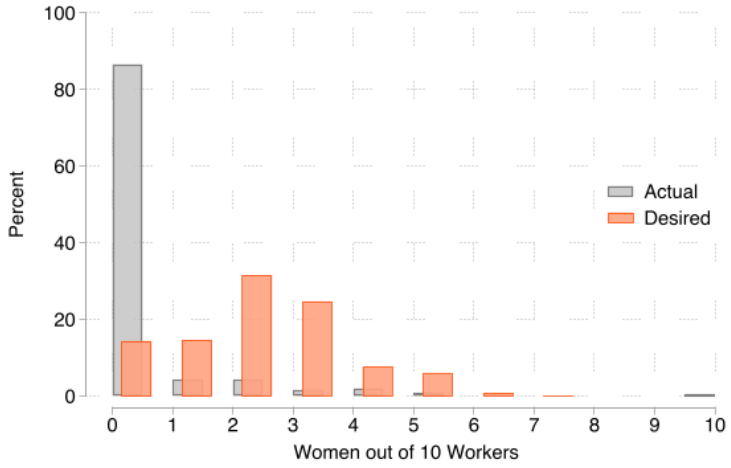
Figures

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



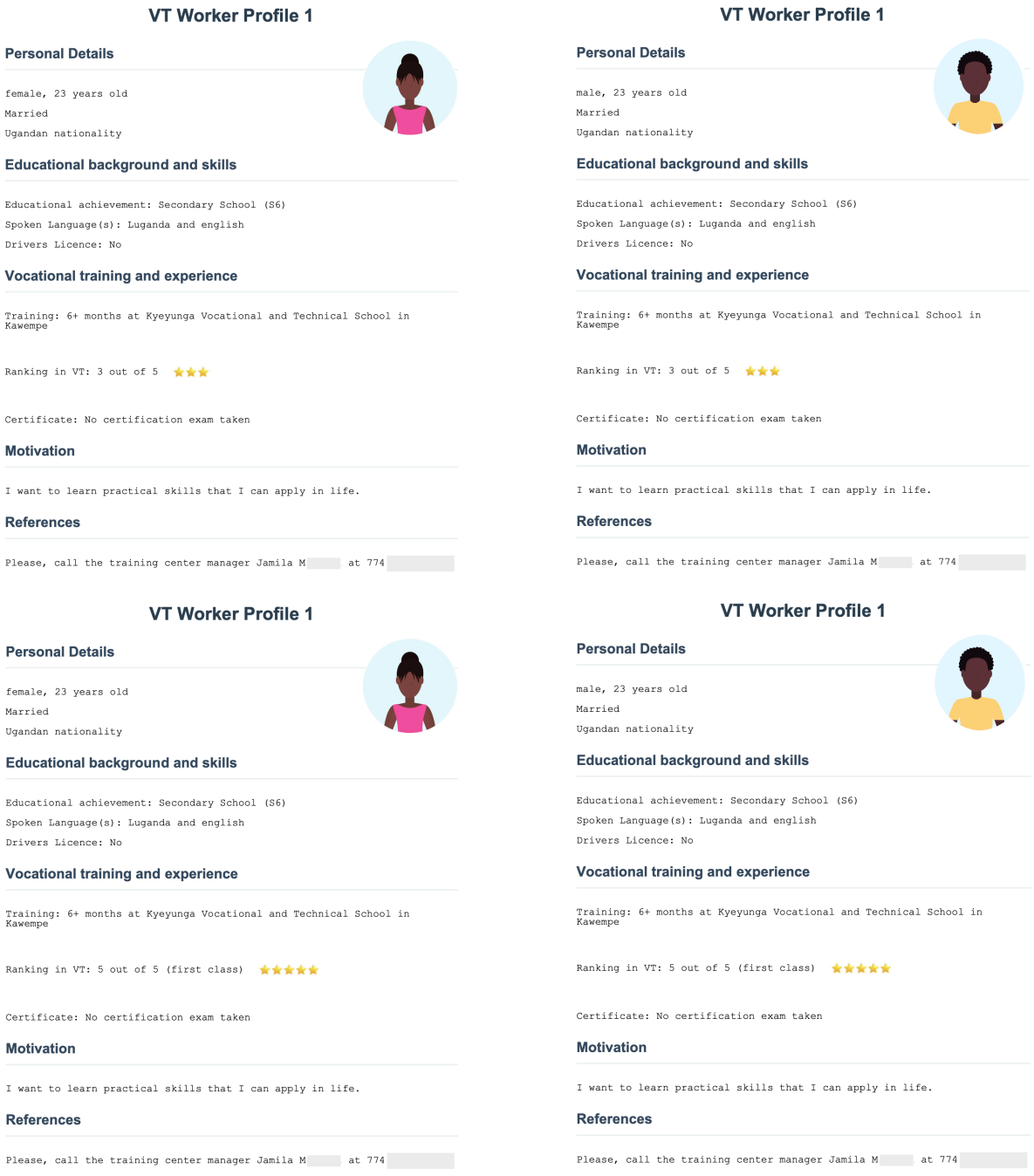
Notes: Descriptive statistics from the full sample of employers. Panel A: share citing each trait as most important in a worker (open-ended responses, manually recoded). Panel B: share reporting each trait as hardest to find or monitor. Panels C–D: share answering "Male" or "Female" to "Do you think male or female workers are better at [trait]?", including indifference. Panel E: open-ended responses to "If you think of hiring a woman, what do you worry about?", recategorized into the same trait categories. Darker bar portions isolate gender-specific sub-worries: "Strength" within Skills/Education, "Leave" within Learning/Interest, "Safety/Harassment" within Cooperation/Respect. Of respondents, 10.73% answered "Nothing" and 1.79% "Other".

Figure 2: Employers' Desired and Actual Gender Mix Among Workforce



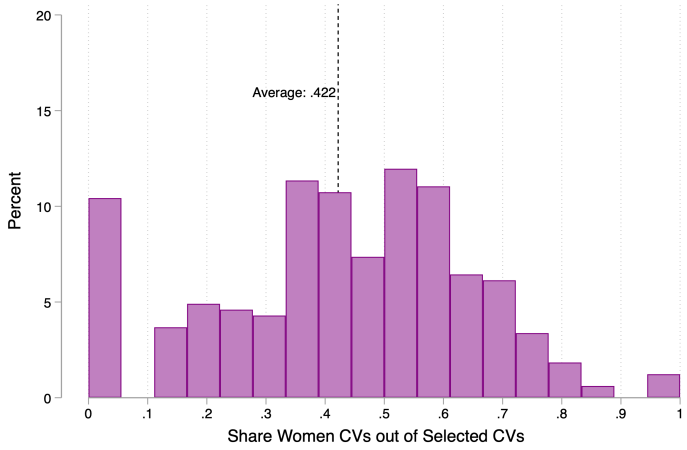
Notes: The figure is based on the full sample of employers. It presents the distributions of the desired gender composition (on a scale from 0 women out of 10 (all men) to 10 women out of 10 (all women)) versus the actual share of women among current workers, rescaled to 10.

Figure 3: CV Examples



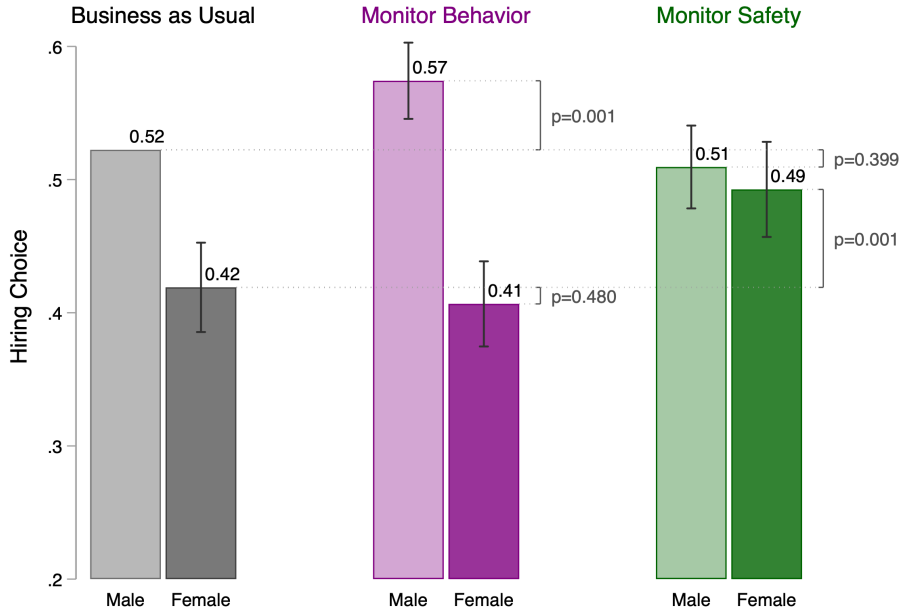
Notes: Examples of hypothetical profile 1 in both female and male versions and for high and low performance levels. In total, there are 24 profiles, with four variations for each profile that reflect the 2×2 design based on gender and performance.

Figure 4: Demand for Female Workers in the Experiment Under Business as Usual



Notes: Data from the main experiment, restricted to the 326 employers in the Control group (business-as-usual), comprising 7,365 hiring decisions. The figure shows the distribution of the share of female candidates among those selected for a meeting, with an average female share of 42.2%.

Figure 5: Hiring Choice by Gender and Monitoring Arm

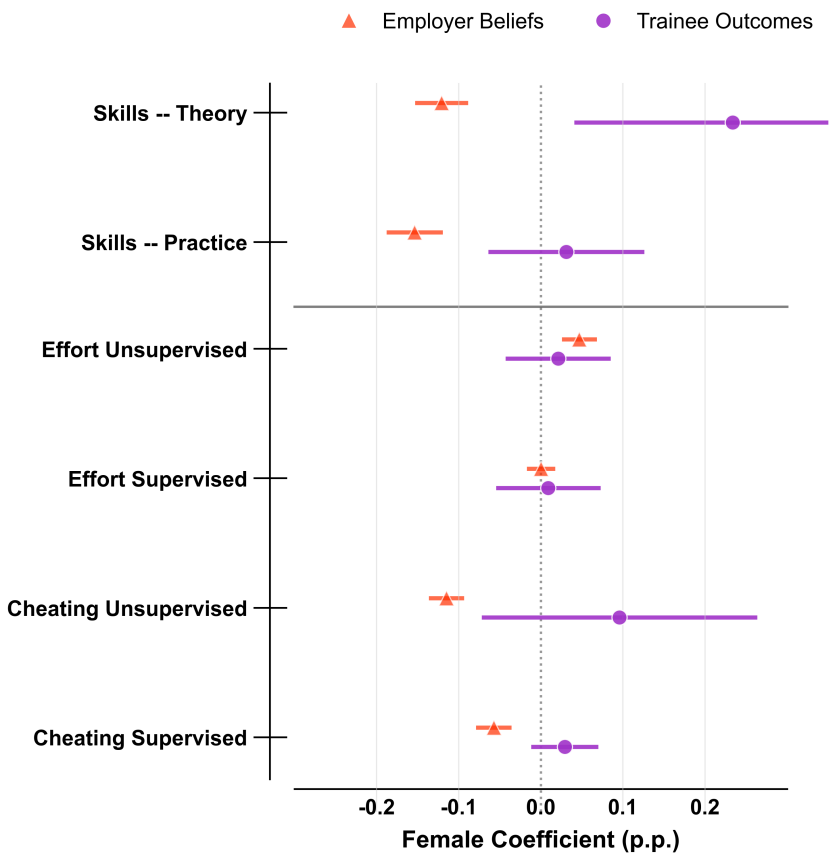


Notes: Bars report the likelihood of the employer to meet the candidate for hiring on probation for male and female candidates across the three monitoring arms: Business as Usual (Control), Monitor Behavior, and Monitor Safety. Levels are estimated from a fully interacted specification of candidate gender 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 ($p = 0.480$), while Monitor Safety significantly increases hiring of female candidates ($p < 0.001$) with no effect on male candidates ($p = 0.399$). $N = 20{,}726$ evaluations from 921 employers.

Figure 6: 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 ($N = 20{,}726$). Squares denote estimates for employers with any stated preferences for workforce gender mix ($N = 17{,}748$); these are obtained from a regression that includes an indicator for having any stated preferences, fully interacted with candidate gender and the randomly assigned monitoring condition. Triangles denote estimates from the same specification using an indicator for employers in the top quintile of stated preferences (i.e., 40–60% female workers; $N = 3{,}031$).

Figure 7: Gender Gap in Skills and Trustworthiness



Notes: Data from the baseline employer survey and the trainee survey. The figure plots the coefficient on gender for employers' beliefs and for actual trainee behavior. 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)	8.34	15.16	[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	5.40	[2.0 ; 6.0]	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)	10.73	12.11	[6.0 ; 16.0]	10.01	[5.0 ; 14.0]	9.99	[5.0 ; 13.0]
Firm capital (USD)	1,215	1,594	[0.0 ; 1,852]	933	[0.0 ; 1,037]	1,079	[0.0 ; 1,481]
Firm revenue (USD)	718	589	[0.0 ; 926]	705	[0.0 ; 1,111]	874	[0.0 ; 1,407]
Firm profits (USD)	286	246	[0.0 ; 370]	277	[0.0 ; 370]	339	[0.0 ; 556]
Customers (N/day)	11.76	6.44	[4.0 ; 8.0]	14.37	[6.0 ; 20.0]	14.76	[5.0 ; 20.0]
Any stealing	0.83	0.90	[1.0 ; 1.0]	0.81	[1.0 ; 1.0]	0.77	[1.0 ; 1.0]
Monitor constrained	0.90	0.89	[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)	37.08	40.08	[32.0 ; 47.0]	35.78	[28.0 ; 42.0]	35.24	[29.0 ; 40.0]
Experience (years)	15.21	18.66	[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)	10.43	10.47	[10.0 ; 12.0]	10.43	[9.0 ; 12.0]	10.38	[10.0 ; 11.0]
Hours monitor (per day)	2.06	2.03	[1.0 ; 2.0]	2.10	[1.0 ; 2.0]	2.07	[1.0 ; 2.0]
Observations	921	318		300		303	

Notes: The table reports summary statistics for the full 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)	8.24	(8.96)	8.67	(8.66)	7.95	(9.14)	0.085	0.502	903
Female workers (N)	0.19	(0.55)	0.21	(0.56)	0.17	(0.48)	0.464	0.831	906
Trainees (N)	2.52	(3.43)	2.57	(3.25)	2.81	(5.48)	0.381	0.097	906
Family workers (share)	0.15	(0.25)	0.15	(0.23)	0.13	(0.22)	0.527	0.115	898
Age (years)	10.62	(7.82)	11.27	(8.11)	10.35	(7.23)	0.244	0.671	906
Firm capital (USD)	1245	(2776.4)	1437	(3547.9)	974.7	(1421.8)	0.220	0.136	866
Firm revenue (USD)	649.2	(861.9)	699.9	(885.8)	805.6	(1055.4)	0.556	0.122	861
Firm profits (USD)	259.8	(396.4)	279.8	(405.2)	317.1	(468.7)	0.634	0.187	865
Customers (N/day)	11.9	(17.65)	12.03	(13.2)	11.09	(11.87)	0.810	0.109	906
Any stealing	0.83	(0.38)	0.85	(0.36)	0.81	(0.39)	0.667	0.528	906
Monitor constrained	0.9	(0.31)	0.92	(0.28)	0.9	(0.3)	0.395	0.930	906
Panel B: Employers									
Male	0.97	(0.17)	0.95	(0.23)	0.98	(0.14)	0.115	0.387	906
Age (years)	37.12	(9.84)	36.47	(9.25)	37.26	(10.26)	0.472	0.620	903
Experience (years)	15.12	(8.96)	15.38	(9.11)	14.81	(9.01)	0.580	0.719	904
Higher education	0.62	(0.49)	0.62	(0.49)	0.66	(0.47)	0.966	0.491	906
VTI trained	0.27	(0.45)	0.34	(0.47)	0.32	(0.47)	0.058	0.159	906
Hours worked (per day)	10.55	(1.54)	10.45	(1.7)	10.31	(1.64)	0.592	0.102	905
Hours monitor (per day)	2.08	(1.58)	2	(1.29)	2.08	(1.52)	0.459	0.970	902

Notes: The table presents baseline balance for the main experimental 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: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Hiring Gender Gap in Control Arm and by Stated Preferences

			Stated Preferences for Workforce Gender Mix				
Meet (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	-0.103***	-0.077***	-0.233***	-0.153***	-0.098***	-0.083***	0.026
	[0.016]	[0.015]	[0.050]	[0.038]	[0.026]	[0.027]	[0.032]
Observations	7,365	6,206	1,159	870	2,340	1,764	1,232
R-squared	0.123	0.123	0.200	0.190	0.149	0.129	0.137
Control mean	0.522	0.517	0.552	0.497	0.523	0.535	0.491

Notes: The table reports treatment effects from regression model 1 for employers in the Control (business-as-usual) group, overall and by quintiles of employers' preferred workforce gender composition. Column (1) presents results for the full business-as-usual sample; column (2) excludes employers who prefer an all-male workforce. Columns (3)–(7) restrict the sample to each quintile of the preferred gender composition distribution. The dependent variable *Meet* is a binary indicator equal to 1 if the employer wants to meet the candidate to hire them on probation. *Female* is an indicator for female candidates. Standard errors, reported in brackets, are clustered at the respondent and profile levels. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Hiring Gender Gap Across Treatment Arms and by Stated Preferences

			Stated Preferences for Workforce Gender Mix				
Meet (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	-0.103***	-0.078***	-0.235***	-0.158***	-0.099***	-0.087***	0.029
	[0.016]	[0.016]	[0.052]	[0.039]	[0.026]	[0.028]	[0.034]
Monitor Behavior	0.052***	0.061***	-0.014	0.038	0.052*	0.060**	0.104***
	[0.014]	[0.015]	[0.040]	[0.037]	[0.027]	[0.026]	[0.034]
Monitor Safety	-0.013	-0.017	-0.000	-0.061*	-0.016	-0.020	-0.005
	[0.015]	[0.017]	[0.047]	[0.035]	[0.026]	[0.027]	[0.044]
Female x Monitor Behavior	-0.064***	-0.077***	0.003	-0.050	-0.061	-0.040	-0.178***
	[0.022]	[0.023]	[0.071]	[0.061]	[0.038]	[0.044]	[0.047]
Female x Monitor Safety	0.086***	0.091***	0.026	0.137**	0.104**	0.091*	0.070
	[0.028]	[0.028]	[0.089]	[0.054]	[0.044]	[0.045]	[0.073]
Observations	20,725	17,748	2,977	3,002	6,471	5,244	3,031
R-squared	0.127	0.126	0.177	0.175	0.147	0.119	0.124
Control Mean	0.522	0.517	0.552	0.497	0.523	0.535	0.491

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. The dependent variable *Meet* is a binary indicator equal to 1 if the employer wants to meet the candidate to hire them 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 by respondent and profile levels. Standard errors in brackets * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A Appendix

A.1 Residual Estimates of Bias: Formal Framework and Supporting Algebra

We decompose the observed hiring gender gap under business-as-usual conditions into three components: residual bias (B), a trust premium (T) that offsets the gap, and a harassment penalty (H) that widens it.

Under each experimental condition, the gender gap reflects a different combination of these components:

$$\text{C gap} = B - T + H = 10.30 \text{ pp} \tag{3}$$
$$\text{MB gap} = B + H = 16.70 \text{ pp} \tag{4}$$
$$\text{MS gap} = B - T = 1.70 \text{ pp} \tag{5}$$

The identifying assumption is additivity: the trust premium does not depend on whether harassment is present, and vice versa.

Solving for the components. From the three equations:

$$T = \text{MB gap} - \text{C gap} = 16.70 - 10.30 = 6.40 \text{ pp} \tag{6}$$
$$H = \text{C gap} - \text{MS gap} = 10.30 - 1.70 = 8.60 \text{ pp} \tag{7}$$
$$B = \text{C gap} + T - H = 10.30 + 6.40 - 8.60 = 8.10 \text{ pp} \tag{8}$$

These satisfy all three conditions: $B - T + H = 8.10 - 6.40 + 8.60 = 10.30$ (C); $B + H = 8.10 + 8.60 = 16.70$ (MB); $B - T = 8.10 - 6.40 = 1.70$ (MS).

The *trust premium* ($T = 6.40$ pp) captures how women's perceived trustworthiness reduces the hiring penalty under business-as-usual conditions. Monitoring-Behavior removes this premium by reducing monitoring costs.

The *harassment penalty* ($H = 8.60$ pp) captures how concerns about harassment and safety increase the hiring penalty against women. Monitoring-Safety removes this penalty by reducing integration costs.

The *residual bias* ($B = 8.10$ pp) is the hiring penalty that remains when both frictions are accounted for. This residual is inconsistent with statistical discrimination on technical performance: the gender gap does not respond to performance signals in any experimental arm (Section 5.1).

Residual estimates under different measurement assumptions. The residual estimate depends on which attributes the researcher measures and considers hiring determinants, and whether they are treated as productive (reflecting efficient statistical discrimination) or as distortions. Attributes that are separated from the residual are treated as legitimate hiring determinants. Attributes that are not measured remain in the residual alongside bias. The estimate of bias is accurate as long as these attributes are non-productive or the beliefs about them are inaccurate.

Dimensions	Trust	Harassment	Residual
1D: performance only	in residual	in residual	10.30 pp
2D: + trust	separated	in residual	16.70 pp
2D: + harassment	in residual	separated	1.70 pp
3D: both	separated	separated	8.10 pp

"Separated" means the researcher measures the attribute and treats it as a productive hiring determinant; its contribution is removed from the residual. "In residual" means the attribute is either not measured or not treated as productive; its contribution remains in the residual alongside bias. When trust is separated, the residual increases by $T = 6.40$ pp, because the trust premium was masking discrimination. When harassment is separated, the residual decreases by $H = 8.60$ pp, because the harassment penalty was inflating the residual beyond pure bias.

The residual ranges from 1.70 to 16.70 percentage points.

- *Lower bound* (1.70 pp): Harassment concerns reflect genuine productivity differences (e.g., harassment reduces women's output or disrupts teamwork) and are separated from the residual. Trust beliefs are fully inaccurate: women are not actually more trustworthy than men, so the trust premium is a distortion that remains in the residual. Under these assumptions, the trust premium partially offsets bias in the observed gap, and the residual equals the MS gap. This is a lower bound because it requires both that safety is entirely productive and that trust beliefs are entirely wrong.
- *Upper bound* (16.70 pp): Trust reflects genuine productivity differences (women are more trustworthy, and therefore more productive) and is separated. Harassment is treated as bias (paternalism or reputational costs), so it remains in the residual. Under these assumptions, the trust premium was masking a penalty of nearly 7 percentage points, which would be fully bias. The residual equals the MB gap and includes both pure bias and the unseparated harassment penalty.
- *Intermediate case* (8.10 pp): Both trust and harassment reflect only productivity differences and are separated. The residual reflects only the component of the gender gap that cannot be attributed to either friction.

Estimates of hidden discrimination under measurement assumptions. Separating trust reveals 6.40 pp of hidden discrimination (the difference between the MB and C gaps). The amount of this hidden gap that reflects bias depends on whether safety is also accounted for.

One-dimensional framework (technical performance only): The researcher accounts only for technical performance. The residual is 10.30 pp. Separating trust increases the residual to 16.70 pp, revealing 6.40 pp of hidden discrimination. Since harassment is not separated, the entire 6.40 pp is attributed to hidden bias.

Two-dimensional framework (technical performance and safety): The researcher also accounts for safety, reducing the residual to 1.70 pp. Separating trust then increases the residual to 8.10 pp, again revealing 6.40 pp of hidden gap. However, the researcher now knows that the MB residual ($B + H = 16.70$) contains two components: bias ($B = 8.10$, 48%) and harassment ($H = 8.60$, 52%). The trust premium was offsetting the total gap, not bias alone. Since the gap it was offsetting is 48% bias and 52% harassment,

the 6.40 pp hidden by trust is also 48% bias and 52% harassment:

$$\text{Hidden bias} = T \times \frac{B}{B+H} = 6.40 \times \frac{8.10}{16.70} = 3.10 \text{ pp} \tag{9}$$

$$\text{Hidden harassment concerns} = T \times \frac{H}{B+H} = 6.40 \times \frac{8.60}{16.70} = 3.30 \text{ pp} \tag{10}$$

Framework	Trust in residual	Trust separated	Hidden bias
Performance only	10.30 pp	16.70 pp	6.40 pp
Performance + safety	1.70 pp	8.10 pp	3.10 pp

The amount of hidden discrimination thus depends on how many dimensions the researcher accounts for. From a one-dimensional framework, separating trust reveals 6.40 pp of hidden bias. From a two-dimensional framework that already accounts for safety, separating trust reveals 3.10 pp of hidden bias.

Caveats. First, the decomposition assumes that the trust premium and harassment penalty are independent (additivity). The experimental design varies the two frictions across employers (not within), so we cannot directly test for interactions. If the trust premium changes when harassment is addressed (or vice versa), the decomposition is only approximate. As a partial test, we examine whether the Monitor Behavior effect on the gender gap varies with employers' baseline harassment concerns. It does not: the interaction between the MB treatment effect and baseline harassment risk is small and statistically not significant ($p = 0.139$), consistent with the trust and harassment channels operating independently.

Second, the treatments may not fully eliminate each friction. If Monitoring-Behavior only partially removes the trust premium, $T \geq 6.40$ and the true residual bias exceeds the estimates above. Similarly, if Monitoring-Safety only partially addresses harassment, $H \geq 8.60$. The estimates should be interpreted as bounds.

Third, there may be productive attributes beyond technical performance, trustworthiness, and safety that we do not observe. The residual includes any unobserved attribute effects that have not been separated.

Fourth, the table treats each attribute as either fully productive or fully non-productive. In practice, each attribute most likely has a combination of productivity and non-productivity effects. For example, the trust premium may partly reflect genuine productivity differences and partly reflect inaccurate beliefs; harassment concerns may partly reduce women's output and partly reflect paternalism. The bounds above correspond to the polar cases; the true bias estimate lies within them.

A.2 Differences From 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 *Offer*—elicited using the question, "How likely would you be to offer this worker a 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 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.

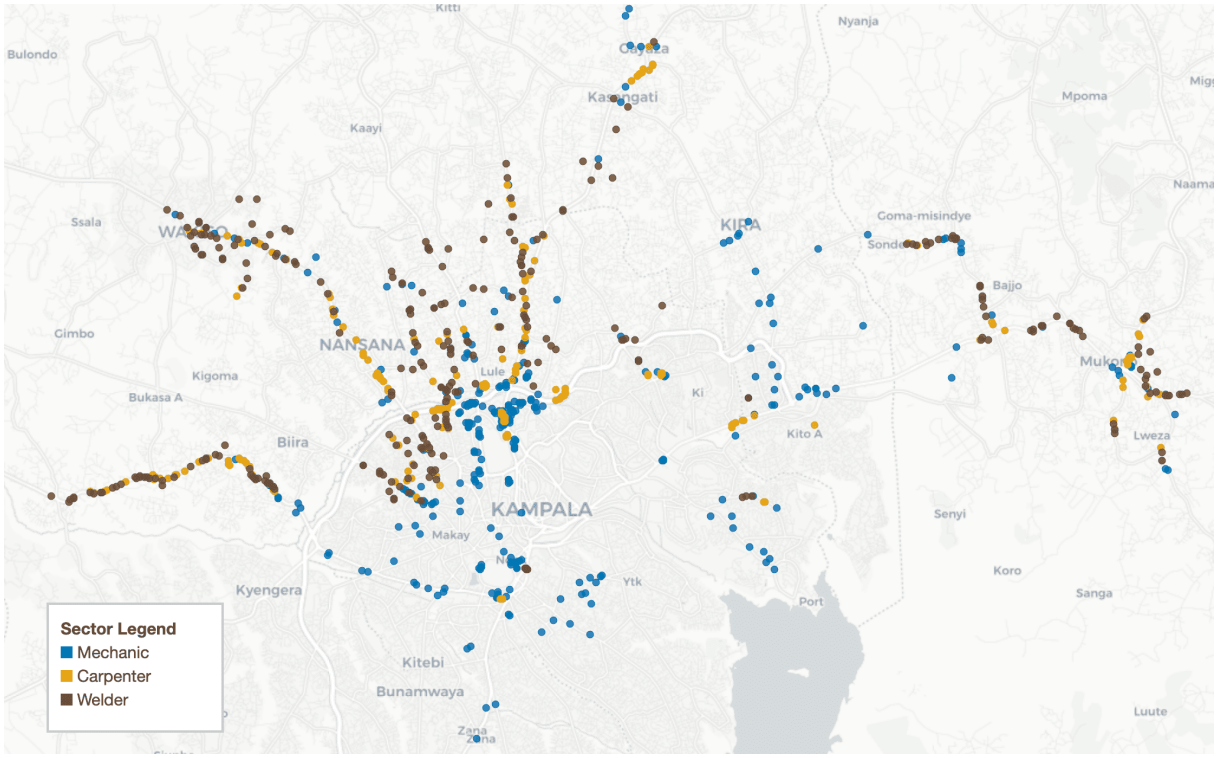
We amended the preregistration before expanding the sample to reflect these changes (except the last one) and to document the motivations.

A.3 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 A12, 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 A12, 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.4 Appendix Figures

Figure A1: Firm Location



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 A2: Firm and Vocational Training Environment

(A) Firm Layout

(B) Practical VTI Class

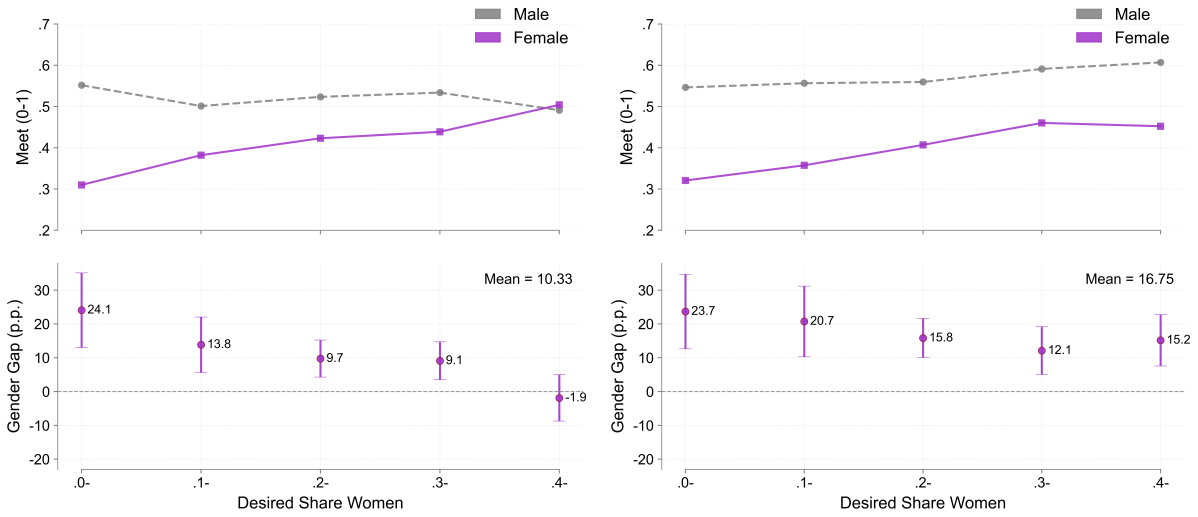
(C) Internship

Photo credit: © Mariajose Silva-Vargas.

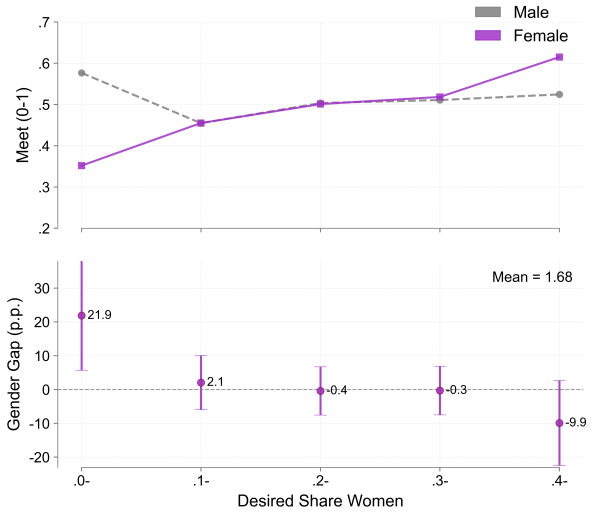
Figure A3: Hiring Gender Gap by Treatment Arm and by Stated Preferences for Hiring Women (Raw Data)

(A) Control (C)

(B) Monitoring Behavior (MB)



(C) Monitoring Safety (MS)

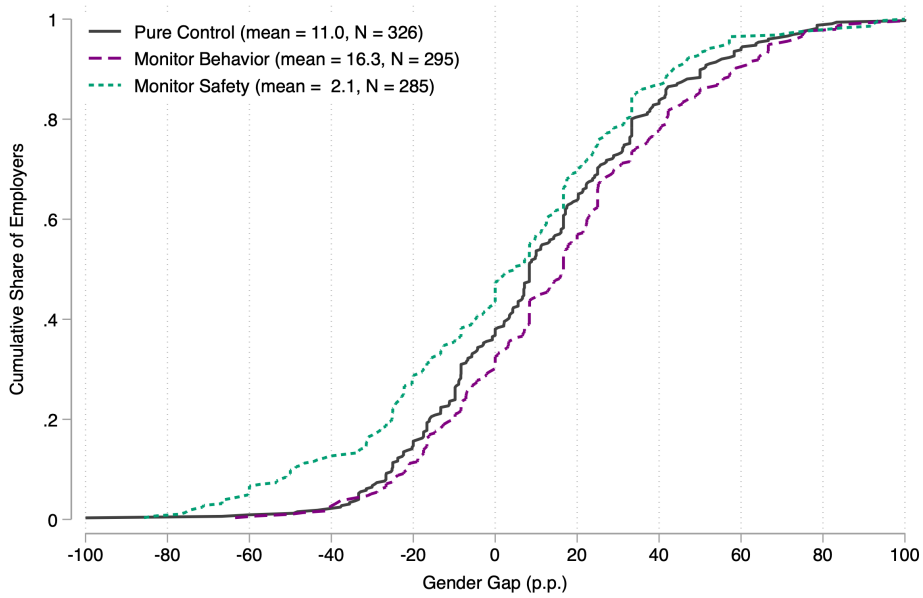


Notes: The figures show the heterogeneity in the gender gap by stated preferences for workforce gender mix, by treatment arm. Panel A presents descriptive statistics for the Control group, while panels B and C present the corresponding statistics for the Monitoring Behavior and Monitoring Safety groups, respectively. The top graph of each panel presents the raw data average of the employer's interest in meeting the candidate by profile gender 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 A4: Hiring Rates by Gender and Trainee Performance, by Treatment Arm

Notes: Each panel reports regression-adjusted hiring rates for male and female candidates carrying either an average-performance or a top-performance CV signal, separately by treatment arm. Lighter bars correspond to male candidates; darker bars to female candidates. The bracket spanning each panel reports the p-value on the main effect of trainee performance (pooled across gender), from a regression of the hiring indicator on gender and performance indicators with the standard fixed effects and clustering. "Female × Performance" reports the p-value on the gender × performance interaction term from the fully interacted specification. Confidence intervals are shown for all groups except the reference category (Male, Avg Performance). The null interaction term indicates that top-performance CVs raise hiring rates equally for male and female candidates across all three arms.

Figure A5: 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 A6: Theory Technical Exam — An Example (Motor Mechanics)

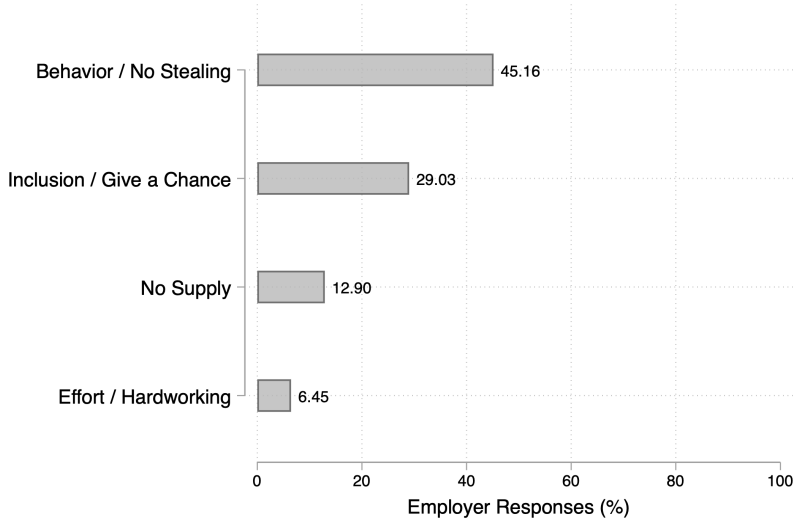
FULL NAME:_____ ☐ MAN

ID:_____ ☐ WOMAN

MOTOR-MECHANICS

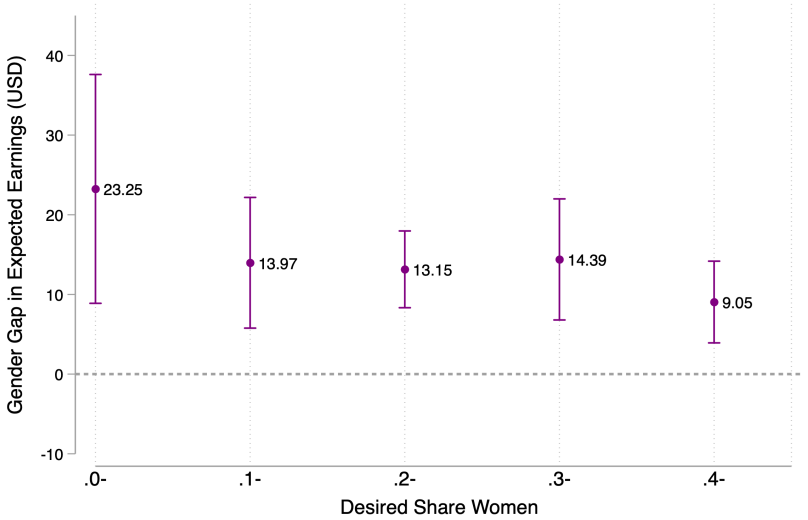
	Question	Answers
1	*multiple-choice* What are you advised to do when servicing the engine by changing oil?	A. Top up lubricating oil B. Replace oil filter C. Over hand engine D. Over hand cylinder head
2	*multiple-choice* What immediate remedy can you give to a vehicle with a problem of excessive tyre 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 A7: 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 A8: 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.5 Appendix Tables

Table A1: Attention Checks and Attrition

	Monitor Treat		Female Treat
	(1) C vs. MB Sample	(2) C vs. MS Sample	(3) C Sample
Main Sample	0.140	-0.092	-0.032
	[0.125]	[0.058]	[0.022]
Observations	14,954	14,916	7,812
R-squared	0.001	0.002	0.000
p-value	0.264	0.113	0.155

Notes: The table tests whether attrition due to attention checks is correlated with treatment assignment. The dependent variable is the treatment indicator: Monitoring Behavior (MB) assignment in column (1), Monitoring Safety (MS) assignment in column (2), and Female profile assignment 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. Samples used to estimate the regressions are restricted to Control (C) and MB, C and MS, and C 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 A2: Hiring Gender Gap by Performance Signals and Monitoring

	(1)	(2)
Monitor Behavior	0.021	0.055***
	[0.014]	[0.018]
Monitor Safety	0.046***	0.006
	[0.013]	[0.019]
High Performance	0.101***	0.118***
	[0.013]	[0.019]
Monitor Behavior × High Performance	-0.004	-0.007
	[0.017]	[0.021]
Monitor Safety × High Performance	-0.032*	-0.037*
	[0.018]	[0.022]
Female		-0.087***
		[0.023]
Female × Monitor Behavior		-0.067**
		[0.027]
Female × Monitor Safety		0.082**
		[0.030]
Female × High Performance		-0.033
		[0.032]
Female × Monitor Behavior × High Performance		0.005
		[0.030]
Female × Monitor Safety × High Performance		0.009
		[0.029]
Observations	20,725	20,725
Control Mean (median-performance, male)	0.419	0.464

Notes: The table reports treatment effects for the main sample. The dependent variable *Meet* is a binary indicator equal to 1 if the employer wants to meet the candidate to hire them on probation. *Female* is a binary variable indicating whether the profile corresponds to a woman. *Top Performance* is a dummy indicating whether the profile has a top GPA score (5/5) within their vocational training program. *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. Column (1) tests whether employers respond to performance signals independently of gender. Column (2) adds gender interactions and tests whether treatment effects on the gender gap differ by candidate performance. The control mean is that of male candidates with median performance (i.e. median GPA score within their vocational training program) in the *Control* arm. Profile and strata fixed effects are included (not reported). Standard errors are clustered by respondent and profile levels and reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Hiring Gender Gap by Performance Signal and by Treatment Arm

			Stated Preferences for Workforce Gender Mix				
	(1) Main Sample	(2) Gender-Mix Pref. > 0	(3) Bottom QT	(4) 2nd QT	(5) 3rd QT	(6) 4th QT	(7) Top QT
Panel A: **Control**							
Female	-0.087***	-0.058**	-0.233***	-0.148***	-0.048	-0.083**	0.034
	[0.022]	[0.021]	[0.052]	[0.051]	[0.031]	[0.037]	[0.039]
Female x Top Performance	-0.033	-0.042	0.004	-0.008	-0.103***	-0.002	-0.016
	[0.032]	[0.029]	[0.062]	[0.061]	[0.033]	[0.053]	[0.052]
Top Performance	0.118***	0.121***	0.105**	0.121**	0.141***	0.117***	0.093*
	[0.018]	[0.019]	[0.044]	[0.053]	[0.025]	[0.040]	[0.047]
Observations	7,365	6,206	1,159	870	2,340	1,764	1,232
R-squared	0.134	0.134	0.212	0.204	0.160	0.142	0.144
Control mean	0.464	0.459	0.491	0.445	0.448	0.481	0.455
Panel B: **Monitor Behavior**							
Female	-0.153***	-0.148***	-0.169***	-0.249***	-0.147***	-0.094**	-0.140***
	[0.024]	[0.025]	[0.053]	[0.064]	[0.036]	[0.038]	[0.050]
Female x Top Performance	-0.028	-0.015	-0.117**	0.077	-0.021	-0.063	-0.021
	[0.023]	[0.023]	[0.048]	[0.057]	[0.037]	[0.048]	[0.054]
Top Performance	0.111***	0.107***	0.147***	0.010	0.099***	0.190***	0.103**
	[0.021]	[0.020]	[0.040]	[0.044]	[0.032]	[0.037]	[0.046]
Observations	6,830	5,847	983	1,070	2,134	1,620	1,023
R-squared	0.148	0.150	0.205	0.190	0.185	0.132	0.182
Control mean	0.516	0.523	0.475	0.555	0.510	0.499	0.556
Panel C: **Monitor Safety**							
Female	-0.005	0.027	-0.228**	-0.029	0.000	0.046	0.109
	[0.025]	[0.025]	[0.091]	[0.042]	[0.040]	[0.041]	[0.071]
Female x Top Performance	-0.024	-0.029	-0.011	0.008	0.005	-0.084**	-0.019
	[0.025]	[0.027]	[0.047]	[0.037]	[0.043]	[0.040]	[0.068]
Top Performance	0.081***	0.090***	0.058	0.017	0.084*	0.131***	0.070
	[0.016]	[0.016]	[0.047]	[0.023]	[0.044]	[0.022]	[0.055]
Observations	6,530	5,695	835	1,062	1,997	1,860	776
R-squared	0.102	0.103	0.219	0.178	0.111	0.116	0.103
Control mean	0.468	0.452	0.556	0.437	0.454	0.446	0.484

Notes: The table reports the treatment effects from estimation augmented regression model 2 with triple interaction with technical performance signal, separately for the sample of respondents in Control (C) arm (Panel A), the Monitoring Behavior (MB) arm (Panel B) and in the Monitoring Safety (MS) arm (Panel C), and by quintiles distribution of 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. The dependent variable *Meet* is a binary indicator equal to 1 if the employer wants to meet the candidate to hire them on probation. *Female* is a binary variable indicating whether the profile corresponds to a female candidate. *Top Performance* is a dummy indicating whether the profile has a top GPA score (5/5) within their vocational training program. Profile and strata fixed effects are included (not reported in the table). Standard errors are clustered by respondent and profile levels. Standard errors in brackets * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Hiring Gender Gap by Treatment Arm and Performance Index

			Stated Preferences for Workforce Gender Mix				
	(1) Main Sample	(2) Gender-Mix Pref. > 0	(3) Bottom QT	(4) 2nd QT	(5) 3rd QT	(6) 4th QT	(7) Top QT
Panel A: **Control**							
Female	-0.104***	-0.078***	-0.231***	-0.138***	-0.100***	-0.090***	0.026
	[0.016]	[0.016]	[0.050]	[0.039]	[0.025]	[0.026]	[0.033]
Female x Performance Index	-0.009	-0.010	-0.010	-0.029	0.021	-0.043*	0.000
	[0.011]	[0.012]	[0.010]	[0.032]	[0.017]	[0.024]	[0.030]
Performance Index	0.133***	0.131***	0.141***	0.170***	0.104***	0.167***	0.107**
	[0.017]	[0.018]	[0.044]	[0.053]	[0.033]	[0.038]	[0.046]
Observations	7,365	6,206	1,159	894	2,340	1,740	1,232
R-squared	0.133	0.133	0.212	0.195	0.157	0.146	0.144
Control mean	0.478	0.480	0.466	0.460	0.503	0.477	0.451
Panel B: **Monitor Behavior**							
Female	-0.168***	-0.155***	-0.228***	-0.210***	-0.158***	-0.126***	-0.150***
	[0.017]	[0.018]	[0.051]	[0.050]	[0.028]	[0.033]	[0.038]
Female x Performance Index	0.001	0.010	-0.038*	0.022	0.002	-0.019	0.045
	[0.013]	[0.014]	[0.022]	[0.038]	[0.024]	[0.019]	[0.029]
Performance Index	0.121***	0.121***	0.128***	0.050	0.110***	0.210***	0.094
	[0.020]	[0.021]	[0.045]	[0.048]	[0.027]	[0.042]	[0.057]
Observations	6,830	5,847	983	1,070	2,134	1,620	1,023
R-squared	0.148	0.150	0.204	0.189	0.185	0.131	0.184
Control mean	0.538	0.545	0.500	0.540	0.502	0.539	0.648
Panel C: **Monitor Safety**							
Female	-0.017	0.013	-0.234***	-0.025	0.002	0.004	0.099
	[0.022]	[0.021]	[0.079]	[0.038]	[0.035]	[0.035]	[0.060]
Female x Performance Index	-0.017	-0.019	-0.042	-0.009	-0.004	-0.027	-0.027
	[0.012]	[0.013]	[0.030]	[0.018]	[0.023]	[0.023]	[0.033]
Performance Index	0.095***	0.104***	0.085	0.030	0.110***	0.125***	0.090
	[0.016]	[0.015]	[0.058]	[0.035]	[0.034]	[0.026]	[0.063]
Observations	6,530	5,695	835	1,062	1,997	1,860	776
R-squared	0.102	0.104	0.221	0.178	0.111	0.115	0.104
Control mean	0.451	0.447	0.480	0.382	0.458	0.458	0.485

Notes: The table reports the treatment effects from estimation of the same regression model as in A3 for the sample of respondents in the Monitoring Behavior (MB) arm (Panel A) and in the Monitoring Safety (MS) arm (Panel B), and by quintiles distribution of 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. The dependent variable *Meet* is a binary indicator equal to 1 if the employer wants to meet the candidate to hire them on probation. *Female* is a binary variable indicating whether the profile corresponds to a female candidate. *Performance Index* is a standardized index that combines all high-quality trainee characteristics in the CVs with equal weights: completion of secondary education, a Directorate of Industrial Training (DIT) certification, a 12- (vs. 6-month) study period, and a high GPA. Profile and strata fixed effects are included (not reported in the table). Standard errors are clustered by respondent and profile levels. Standard errors in brackets * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Hiring Gender Gap by Treatment Arm: Secondary Outcomes

	(1) Quality	(2) Trust-worthiness	(3) Effort	(4) Expected Earnings
Female	-0.169***	0.058**	-0.209***	-12.395***
	[0.026]	[0.026]	[0.028]	[1.578]
Monitor Behavior	-0.013	-0.034	-0.012	2.267
	[0.041]	[0.049]	[0.042]	[3.193]
Monitor Safety	-0.030	-0.035	-0.016	-1.611
	[0.048]	[0.053]	[0.044]	[3.200]
Female x Monitor Behavior	-0.043	-0.012	-0.054	-0.933
	[0.028]	[0.035]	[0.037]	[2.394]
Female x Monitor Safety	0.034	0.059	0.019	-0.556
	[0.047]	[0.042]	[0.044]	[2.414]
Observations	20,724	20,613	20,684	20,725
R-squared	0.226	0.083	0.173	0.196
Control Mean	0.085	-0.028	0.106	114.913

Notes: The table reports the treatment effects for the main sample, by pre-registered secondary outcomes. Columns (1)-(3) reports standardized outcomes relative to C group. Column (1) shows estimates for the perceived quality and skills of candidates: "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) shows estimates for the perceived trustworthiness of candidates: "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) shows estimates for the perceived effort of candidates: "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) shows estimates for the expected monthly earnings of the candidates: "What is your best guess of the monthly earnings of this worker a year from now?". Profile and strata fixed effects are included (not reported in the table). Standard errors are clustered by respondent and profile levels. Standard errors in brackets * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: Effect of Audit Visits on Number of Selected Profiles

	(1) Meet (0-1)
Monitor Behavior	0.019*
	[0.011]
Monitor Safety	0.030**
	[0.011]
Observations	20,725
R-squared	0.114
Control Mean	0.470

Notes: The table reports the treatment effects for the main sample. We estimate the following regression model: $\text{Meet}_{ijs} = \beta_0 + \beta_1 \text{MB}j + \beta_2 \text{MS}j + \delta_i + \sigma_s + u_{ijs}$. *Meet* is a binary indicator equal to 1 if the employer wants to meet the candidate for a probation period at the firm. *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 *Control* arm. Profile and strata fixed effects are included (not reported). Standard errors are clustered at the respondent and profile levels and reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: Hiring Gender Gap by Treatment Arm and by Resume Characteristics

Meet (0-1)	(1) Age ≥ Median	(2) Married	(3) Second. Educ.	(4) English	(5) Trained 1+ year	(6) DIT Certified
CHRS	-0.094***	-0.209***	0.007	0.067***	0.209***	0.145***
	[0.012]	[0.011]	[0.012]	[0.018]	[0.011]	[0.013]
Monitor Behavior	0.012	0.009	0.025*	0.016*	0.029**	0.029***
	[0.014]	[0.012]	[0.015]	[0.009]	[0.012]	[0.010]
Monitor Safety	0.037***	0.022*	0.034**	0.033***	0.040***	0.038***
	[0.014]	[0.012]	[0.015]	[0.009]	[0.012]	[0.010]
CHRS x Monitor Behavior	0.010	0.019	-0.010	0.016	-0.019	-0.035*
	[0.017]	[0.016]	[0.018]	[0.025]	[0.016]	[0.018]
CHRS x Monitor Safety	-0.008	0.019	-0.003	-0.009	-0.019	-0.021
	[0.017]	[0.017]	[0.018]	[0.026]	[0.017]	[0.018]
Observations	20,725	20,725	20,725	20,725	20,725	20,725
Control Mean	0.470	0.574	0.466	0.462	0.365	0.428

Notes: The table reports the treatment effects for the main sample, by resume characteristics (CHRS). Estimates are shown in Columns (1)–(6), respectively, for whether the candidate: is above the median age (22.5 years), is married, has attended secondary education, speaks English, attended VTI one- or two-year training relative to 6-month training, holds a Directorate of Industrial Training (DIT) certification. Strata fixed effects are included (not reported in the table). Standard errors in brackets * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A8: Hiring Gender Gap by Treatment Arm – Heterogeneous Treatment Effects (Preregistered)

	Gender-Mix Pref. > 0		Employer Gender		VTI Search	
	(1) Low	(2) High	(3) Female	(4) Male	(5) Yes	(6) No
Female	-0.146***	-0.041*	0.001	-0.105***	-0.086***	-0.174***
	[0.021]	[0.022]	[0.111]	[0.017]	[0.018]	[0.042]
Monitor Behavior	0.035*	0.080***	0.149	0.049***	0.067***	-0.027
	[0.019]	[0.020]	[0.100]	[0.014]	[0.017]	[0.032]
Monitor Safety	-0.021	-0.006	0.280***	-0.018	-0.014	-0.008
	[0.018]	[0.024]	[0.075]	[0.016]	[0.017]	[0.041]
Female x Monitor Behavior	-0.043	-0.092***	-0.206	-0.061**	-0.086***	0.027
	[0.031]	[0.032]	[0.138]	[0.023]	[0.024]	[0.065]
Female x Monitor Safety	0.098***	0.075*	-0.218*	0.092***	0.081**	0.096
	[0.033]	[0.041]	[0.125]	[0.028]	[0.032]	[0.069]
Observations	12,450	8,275	727	19,998	17,108	3,545
R-squared	0.147	0.112	0.173	0.130	0.127	0.172
Control Mean	0.528	0.520	0.465	0.524	0.513	0.566

Notes: The table reports the treatment effects for the main sample, by pre-registered heterogeneity dimensions: Columns (1)-(2) show estimates for above- (high) and below-median (low) optimal gender composition of the workforce; 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 *Meet* is a binary indicator equal to 1 if the employer wants to meet the candidate to hire 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. Profile and strata fixed effects are included. Standard errors are clustered by respondent and profile levels. Standard errors in brackets * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A9: Robustness to Experimenter Demand

	(1) Baseline	(2) Excl. employers who believe we prefer women
Female	-0.103^{***}	-0.118^{***}
	[0.016]	[0.016]
Monitor Safety	-0.012	-0.014
	[0.016]	[0.017]
Monitor Behavior	0.051^{***}	0.058^{***}
	[0.016]	[0.017]
Female × Monitor Safety	0.088^{***}	0.098^{***}
	[0.026]	[0.026]
Female × Monitor Behavior	-0.064^{***}	-0.070^{***}
	[0.023]	[0.024]
Observations	20,725	19,302
R-squared	0.119	0.127
Control mean	0.522	0.528

Notes: The table replicates the main treatment effect regression under two specifications. Column (1) is the baseline on the full main sample. Column (2) excludes the 6.8% of employers who believe the research team would prefer they hire women. All columns include CV profile fixed effects interacted with trainee performance, sector, survey-wave, and enumerator fixed effects. Standard errors clustered at the employer level in brackets. The gender gap and both treatment effects are stable, ruling out experimenter demand as a driver. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A10: Customer Discrimination Channel Test

	(1) Full sample	(2) Restricted sample
Female	-0.118***	-0.129*
	[0.022]	[0.072]
Female × Monitor Behavior	-0.074**	-0.092
	[0.030]	[0.093]
Female × Monitor Safety	0.040	0.101
	[0.035]	[0.095]
Female × *Customer channel*	0.038	0.050
	[0.032]	[0.075]
Female × MB × *Customer channel*	0.033	0.050
	[0.044]	[0.101]
Female × MS × *Customer channel*	0.113**	0.050
	[0.050]	[0.101]
Observations	20,629	9,166
Control Mean (male)	0.522	0.544
Employer FE × Performance	Yes	Yes
Sector, Year, Enumerator FE	Yes	Yes

Notes: The dependent variable is an indicator for meeting the applicant (= 1). *Customer channel* is a binary variable 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? Can you explain why?" It equals 1 if the employer cited customer attraction, customer care, or customer trust as the reason why profits would increase (86.5% of coded responses among employers who expected profits to rise). Column (1) uses the full analysis sample: employers who did not expect profits to rise and employers who cited non-customer reasons are both assigned *Customer channel* = 0. Column (2) restricts to employers who expected profits to rise from hiring more women. Standard errors clustered at the business and enumerator level in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A11: Gender Gap by Treatment Assignment and Interaction with Gender Preferences

	(1)	(2)	(3)	(4)	(5)
Meet (0-1)	Pure Control	Monitor Behavior	Monitor Safety	Main Sample Linear	Main Sample Quadratic
Female	-0.213***	-0.216***	-0.141***	-0.214***	-0.232***
	[0.034]	[0.033]	[0.044]	[0.034]	[0.046]
Monitor Behavior				-0.002	0.009
				[0.027]	[0.036]
Monitor Safety				-0.026	-0.006
				[0.029]	[0.043]
Diversity	-0.107	0.135	-0.051	-0.115	0.083
	[0.086]	[0.079]	[0.100]	[0.079]	[0.213]
Diversity^2					-0.407
					[0.402]
Female × Monitor Behavior				-0.001	-0.008
				[0.046]	[0.063]
Female × Monitor Safety				0.073	0.067
				[0.054]	[0.078]
Female × Diversity	0.495***	0.221*	0.573***	0.499***	0.736**
	[0.122]	[0.108]	[0.168]	[0.122]	[0.352]
Female × Diversity^2					-0.499
					[0.647]
Monitor Behavior × Diversity				0.246**	0.093
				[0.101]	[0.288]
Monitor Safety × Diversity				0.065	-0.214
				[0.123]	[0.389]
Monitor Behavior × Diversity^2					0.333
					[0.543]
Monitor Safety × Diversity^2					0.616
					[0.757]
Female × Monitor Behavior × Diversity				-0.281*	-0.217
				[0.161]	[0.465]
Female × Monitor Safety × Diversity				0.077	0.152
				[0.208]	[0.647]
Female × Monitor Behavior × Diversity^2					-0.102
					[0.793]
Female × Monitor Safety × Diversity^2					-0.196
					[1.269]
Observations	7,365	6,830	6,530	20,725	20,725
Baseline Mean	0.522	0.572	0.522	0.552	0.552

Notes: The table reports treatment effects on employer interest to meet candidates, 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 *Meet* is a binary indicator equal to 1 if the employer wants to meet the candidate for a probation period. *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 A12: Hiring Gender Gap by Treatment Arm: Within-Subject Design

			Stated Preferences for Workforce Gender Mix				
Meet (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]	[0.055]	[0.049]	[0.024]	[0.026]	[0.024]
Monitor Behavior	0.092***	0.103***	0.037	0.076**	0.121***	0.083***	0.106***
	[0.013]	[0.013]	[0.038]	[0.032]	[0.017]	[0.025]	[0.031]
Monitor Safety	-0.046***	-0.045***	-0.018	-0.092***	-0.041*	-0.050**	-0.027
	[0.013]	[0.014]	[0.043]	[0.027]	[0.024]	[0.021]	[0.025]
Female x Monitor Behavior	-0.122***	-0.125***	-0.119*	-0.149**	-0.129***	-0.096**	-0.123***
	[0.020]	[0.020]	[0.064]	[0.059]	[0.029]	[0.039]	[0.036]
Female x Monitor Safety	0.149***	0.163***	-0.019	0.179***	0.163***	0.174***	0.166***
	[0.024]	[0.024]	[0.085]	[0.057]	[0.042]	[0.040]	[0.046]
Observations	20,800	18,409	2,391	2,447	6,521	5,714	3,727
R-squared	0.221	0.214	0.314	0.249	0.237	0.208	0.191
Control Mean	0.529	0.529	0.536	0.522	0.513	0.550	0.526
p-value MB vs. MS	0.000	0.000	0.240	0.000	0.000	0.000	0.000

Notes: The table reports the treatment effects for the sample for which we implemented the within-subject randomization. The table also reports the treatment effects by quintile distribution of 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 quintile's distribution of preferred gender workforce composition. The dependent variable *Meet* is a binary indicator equal to 1 if the employer wants to meet the candidate to hire them on probation. *Female* is a binary variable indicating whether the profile corresponds to a woman. *Monitoring Behavior* is a dummy indicating whether the firm was randomly assigned to audit visits to monitor trainees' behavior. *Monitoring Safety* is a dummy indicating whether the firm was randomly assigned to audit visits to monitor trainees' safety. We report results for the primary outcome variable: interest of the employer to meet the candidates to hire them on probation (0-1). Respondent and profile-by-ability fixed effects are included (not reported in the table). Standard errors are clustered by respondent and profile-by-ability levels. Standard errors in brackets * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A13: 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 A14: Trainee Outcomes and Employer's Beliefs

Panel A: **Trainees**

	Social Preferences				Skills		Supervision & Honesty Game				Dictator Game	Trust Game		Cooperation Game
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	Patience	Altruism	Pos. Recipr.	Trusting	Theory	Practice	Task Completed Unsupervised	Task Completed Supervised	Lying Unsupervised	Lying Supervised	Share	Trust Pl. 1	Reciprocate Pl. 2	Cooperate
Female	0.089	0.067	0.064	0.041	0.230**	0.031	0.021	0.009	0.096	0.029	0.177*	-0.375***	-0.005	0.131
	[0.073]	[0.071]	[0.045]	[0.059]	[0.097]	[0.047]	[0.031]	[0.031]	[0.084]	[0.020]	[0.106]	[0.093]	[0.099]	[0.099]
Observations	171	171	171	171	182	182	182	182	182	182	182	182	182	182
R-squared	0.028	0.038	0.064	0.034	0.125	0.021	0.097	0.106	0.032	0.083	0.133	0.096	0.088	0.059
Control Male	0.696	0.681	0.768	0.450	0.752	0.920	0.989	0.989	0.071	0.018	0.723	0.679	0.875	0.830

Panel B: **Employers**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Female	-0.132***	-0.129**	0.069	0.481***	-0.121***	-0.154***	0.047***	0.000	-0.115***	-0.057***	-0.089***	-0.011	-0.035***	-0.018*
	[0.048]	[0.051]	[0.045]	[0.042]	[0.015]	[0.016]	[0.010]	[0.007]	[0.010]	[0.010]	[0.012]	[0.010]	[0.010]	[0.010]
Observations	919	919	920	921	1,205	1,206	1,842	1,829	1,831	1,826	1,838	1,835	1,829	1,831
R-squared	0.012	0.010	0.046	0.312	0.093	0.120	0.016	0.007	0.077	0.025	0.032	0.002	0.008	0.020
Control Male	0.435	0.449	0.253	0.130	0.580	0.607	0.646	0.881	0.466	0.256	0.573	0.534	0.602	0.571

Notes: The table reports gender differences in trainees' outcomes (Panel A) and the corresponding employer beliefs (Panel B). Columns (1)–(4) present standardized survey measures of social preferences following Falk et al. (2018): patience, altruism, positive reciprocity, and trust, respectively. In Panel A, the coefficient on *Female* captures the difference between female and male trainees. In Panel B, *Female* captures the difference in employer beliefs about female versus male trainees. Columns (5)–(6) present the share of correct answers to sector-specific theory and practical skill questions, respectively. Theory questions are randomly drawn from the technical exam taken by trainees prior to the survey. Practical skills are measured through a self-reported binary indicator on the performance to perform a specific technical task, designed by VTI instructors. The remaining columns shows results of four behavioral games. Columns (7)–(8) report the share of tasks completed in a novel lab-style task — Supervision & Honesty Game — by trainees (Panel A), and employers' beliefs about the same task outcomes under conditions without and with supervision (Panel B). Columns (9)–(10) report the share of lying behavior in the same game. Column (11) reports sharing behavior in a dictator game. Columns (12)–(13) correspond to player 1 (trusting) and player 2 (reciprocating) behavior in a trust game. Column (14) reports decisions to cooperate in a cooperation game. Standard errors are clustered at VTI level in Panel A and at the sector level in Panel B. Standard errors in brackets * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.