CSAE Working Paper WPS/2026-04

Promoting Women to Managerial Roles in the Bangladeshi Garment Sector

Rocco Macchiavello (LSE)

Andreas Menzel (CERGE-EI Prague)

Atonu Rabbani (University of Dhaka)

Christopher Woodruff* (Oxford)

December 2025

Abstract

Women remain disadvantaged in promotion to managerial positions. We conduct a field experiment with 24 large garment factories in Bangladesh to test for inefficient representation of women among line supervisors. We identify the marginal female and male candidates for supervisory positions and randomly assign them to manage production lines. We document four findings: (1) In contrast to widespread negative beliefs about women's ability as supervisors at baseline, female candidates selected by the factories had similar skills to males; (2) during the trial, females performed worse than males, which we show is related to negative bias against them; (3) after the trial, however, many female candidates were retained as supervisors and, conditional on that, performed similarly to males; and (4) after the end of our intervention, factories permanently increased the share of women among newly appointed supervisors. A conceptual framework of experimentation over discrimination rationalizes all these facts and cautions against the standard logic to test for discrimination: when there is uncertainty about the performance of the discriminated group, equal – or even worse – performance of the marginal candidates of that group is no longer sufficient to rule out inefficient discrimination.

Keywords: Gender Discrimination, Productivity, Export Manufacturing

JEL Code: J16, J71, M51, M54, O14, O15

*Corresponding author: christopher.woodruff@economics.ox.ac.uk. We are grateful to the Editor and to two anonymous referees for many comments and suggestions. We also thank seminars and conferences participants at UC San Diego, the University of Washington, Notre Dame, Duke, Leuven, Ecole Polytechnique, MIT / Harvard, LSE, PUC-Chile, the CEPR-IMO and CEPR-IZA workshops, AEA-ASSA 2015, BREAD/CEPR/STICERD/TCD Conference London 2018, and DIW/CRC TRR 190 Workshop Berlin 2019. Remaining failings are the responsibility of the authors. We are grateful for the cooperation and financial support of Deutsche Gesellschaft fuer Internationale Zusammenarbeit (GIZ), who developed the training program that we implement in the project. We are also grateful for financial and logistical support from the International Growth Centre, and financial support from the IPA SME initiative, the ERSC – DFID Growth Research Programme and IFC-Bangladesh, and for the cooperation of the large number of participating workers and factories in Bangladesh. Madhav Malhotra and Oliver Seager provided excellent research assistance. Woodruff and Macchiavello recognize support from the ERC Advanced Grant RMGPP. The project received human subjects approval under the University of Warwick IRB (Approvals 01/11-12 and 86/13-14).

Centre for the Study of African Economies
Department of Economics · University of Oxford · Manor Road Building · Oxford OX1 3UQ
T: +44 (0)1865 271084 · F: +44 (0)1865 281447 · E: csae.enquiries@economics.ox.ac.uk · W: www.csae.ox.ac.uk

1 Introduction

Women remain disadvantaged around the world, notably so in promotion to leadership and managerial positions (Blau and Khan, 2017; Bertrand, 2017; Goldin, 2014). The underlying sources of this disadvantage, however, are not well understood. Alongside cultural gender norms (Ashraf et al., 2023) and policies (Olivetti and Petrongolo, 2016) that hinder female labor force participation, barriers inside the organization may also impede women's access to managerial roles. These barriers might be particularly relevant in newly industrializing countries as they transition from an organization of production where most workers are self-employed to one in which they work under the direction of mid-level managers employed by large firms (Gollin, 2008).

Assessing whether organizational barriers hinder women's access to managerial roles is challenging. On average, men and women might differ in the talent required for, or preferences for, managerial positions in a given organization. Men and women might also be allocated to managerial positions that differ along unobserved dimensions. If so, differences in performance between male and female managers will not reveal whether the under-representation of women is due to a form of discrimination. Whether taste-based or statistical, discrimination implies that the talent requirement to be considered for a promotion is higher for women than men. Testing for discrimination thus requires comparing marginal candidates for promotion – rather than average differences across genders within a hierarchical layer – and finding a higher performance for women relative to men in that group, once differences in task allocation are accounted for.

This paper shows that incorrect beliefs about women's managerial ability – an organizational barrier – hinder the promotion of women to managerial roles and result in non-negligible efficiency losses in the Bangladeshi garment sector. The garment sector has been an engine of economic and social transformation in many countries (Gereffi, 1999; Atkin, 2016; Boudreau et al., 2023) – including Bangladesh. Women account for the majority of workers but only for a negligible share of production line supervisors – the lowest, often internally promoted, echelon of the management ladder and a critical driver of firms' performance (McKinsey, 2011).

We implemented an intervention that nudged factories to identify an equal number of male and female workers as potential candidates for promotion and trial them as supervisors on randomly allocated lines. In contrast to widespread negative beliefs about

women's ability as supervisors at baseline, we found that (1) female candidates selected by the factories had similar skills to males; (2) during the trial, females performed worse than males; (3) after the trial, however, many female candidates were retained as supervisors and, conditional on that, performed similarly to males; and (4) after the end of the intervention, factories permanently increased the share of women among newly appointed supervisors. A conceptual framework of *experimentation* over *discrimination* rationalizes all the experimental results and cautions against the standard logic to test for discrimination: when there is uncertainty about the performance of the discriminated group, equal – or even worse – performance of the marginal candidates of that group is no longer sufficient to rule out inefficient discrimination.

Section 2 describes the industry, our experimental design and data, and motivating facts. We focus on production lines in the sewing departments of large garment export factories. The vast majority of line supervisors are male: women account for only 6% of line supervisors but for around 80% of all workers in the sewing departments. We worked with 24 suppliers of a large UK-based buyer to test whether inefficiently few women were being promoted to line supervisor. To overcome the empirical challenges described above, we asked factories to select sewing operators – equally split between men and women – to attend a supervisor training program. Ultimately, 72 male and 73 female operators completed the training and then worked for two months as co-supervisors on a randomly selected line from a set of lines nominated by the factory. After the trial, factories could return trainees to non-supervisory positions, or keep them as supervisors on the randomly allocated line or on other lines.

Section 3 presents a conceptual framework to guide the interpretation of the experimental design. Our approach is motivated by the observation that employees at all rungs of the organizational hierarchy believe that women are worse supervisors than men. Notably, respondents rate women the worst on "understanding machines", an important skill that can be accurately assessed through a test. The framework embeds the standard discrimination logic into a stylized model of experimentation, and rationalizes both the initial underperformance of female candidates as well as the long-run results. In the model, a decision-maker (DM) considers the best available male and female candidates to fill an internal promotion. There is uncertainty over the performance of the female candidate, which depends on the – potentially also uncertain – extent of bias among subor-

dinates, co-workers, and direct bosses. The DM might have ex-ante biased beliefs about the expected performance of the female candidate, but otherwise has no gender-based preference over promotion. Experimenting with a female candidate generates valuable information that can be used to make better appointments in the future. When beliefs are sufficiently pessimistic, however, the DM doesn't experiment. An intervention – like ours – that nudges experimentation might find that female candidates underperform on average, yet the DM learns that they are better than originally thought. This change in beliefs then leads the DM to promote more women during and after the end of the intervention. Furthermore, an equal, or even worse, initial performance of the discriminated group's marginal candidates no longer suffices to rule out initial discrimination.

Section 4 presents our main results. First, we show that female candidates selected by the factories had similar skills to males. This is most notably the case with their understanding of machines, which we assessed with a comprehensive diagnostic tool. Second, we compare the efficiency of lines randomly allocated a female or a male trainee as line co-supervisor during the trial. During the trial, females performed worse than males. Further analysis reveals that the lower performance of females during the trial is driven, to a significant extent, by co-workers' negative baseline beliefs about women's ability as supervisors on the randomly allocated lines. Third, after the trial, a significant share of female candidates were retained as supervisors and, conditional on that, performed similarly to retained males. A conservative benchmarking exercise reveals that the share of female candidates among newly appointed supervisors is significantly higher than what could have been expected in the absence of our intervention. Finally, within a difference-in-differences framework, we compare the long-run share of female supervisors between factories that participated in an extension of our intervention and similar factories that did not. After the end of the intervention, participating factories permanently increased the share of women among newly appointed supervisors.

Section 5 performs a back-of-the-envelope calculation that, leveraging the experimental estimates, suggests that the initially low share of female supervisors at baseline reduces profits among our partner factories by approximately eight percent. We also find that factories with a higher share of female supervisors sell to buyers that pay higher unit values (a form of upgrading (Verhoogen, 2023)) and have better occupational health and safety conditions. While these correlations are only suggestive and do not prove that fe-

male supervisors cause these outcomes, the evidence is also consistent with a higher share of female supervisors having benefits not captured by our intervention. Finally, Section 6 discusses why factories might have been stuck with wrong beliefs, other organizational barriers to promoting women, and potential spillovers from our intervention.

We contribute to the literature on beliefs and attitudes as barriers to the promotion of women. We demonstrate how, in the presence of biased beliefs over the performance of women, a temporary intervention can lead to a permanent increase in their promotion to managerial roles.[1] Closely related, Benson et al. (2025) find that women in a large company outperform male co-workers with the same predicted potential when appointed as supervisors, a pattern not self-correcting over time. We compare the marginal male and female candidates, controlling for average differences in skills (Goldin et al., 2006) and preferences for managerial positions (Haegele, 2024), and derive the implications of experimentation for discrimination tests. We thus relate to models of discrimination where learning is endogenous and learning traps may arise (see Onuchic (2024) for a review) and to recent studies identifying the sources of discrimination through dynamics and/or ex-ante belief elicitation (Bohren et al., 2019, 2025).[2] We complement this literature by implementing an RCT on promotion decisions in large factories.[3,4]

2 Background and Experimental Design

Garments account for about 80 percent of Bangladesh's exports and an estimated 12 percent of Bangladesh's GDP. Factories are mostly locally owned and managed and are much larger than the typical firm in the country. All factories in the project are located around Dhaka, the largest cluster of the industry in the country.[5]

[1]Exposure to female managers can correct beliefs in the organization (Beaman et al., 2009).

[2]Bohren et al. (2025) finds that less than one in ten published papers measure ex-ante beliefs.

[3]Bardhi et al. (2024); Komiyama and Noda (2024); and Li et al. (2024) also explore experimentation when there is greater uncertainty over the discriminated group, but without focusing on its implications for the validity of the marginal outcome test. Conversely, Canay et al. (2024) describe scenarios in which the marginal outcome test might fail but without focusing on experimentation as a reason for that.

[4]Reforms that mandate minimal representation of women among board members have had mixed effects (see, e.g., Ahern and Dittmar (2012); Matsa and Miller (2013); Bertrand et al. (2019)). Female executives do relatively better in firms with a higher share of female employees (Flabbi et al., 2019).

[5]The Bangladesh garment sector has been widely studied, see, e.g., Heath and Mobarak (2015) on the effects of the sector on gender inequality; Cajal-Grossi et al. (2023) on exporters' margins; and Boudreau (2024); Boudreau et al. (2024) on workplace safety.

2.1 Production Lines and Line Supervisors

Factories are typically organized into a cutting, a sewing, and a finishing department. Most workers are employed on the production lines in the sewing department. The typical factory has between a dozen and 100 lines, with between 20 and 60 sewing operators per line. Within factories, production lines are homogeneous in size, human and physical capital employed, and usually do not specialize by type of garment.

Production lines are supervised by 1 to 4 supervisors, often with a line chief above them. They report to a floor-level production manager, with a typical floor comprising 5-20 lines. Production lines are split into sections of 10-30 workers, each assigned to a line supervisor. Production is organized sequentially from the start to the end of the line, with each sewing operator performing a single step in the process. Therefore, the least productive worker or segment may be a bottleneck for the productivity of the whole line. Line supervisors thus play a crucial role in garment factories. For example, Adhvaryu et al. (2023) documents how better line supervisors increase line efficiency and enable faster learning-by-doing by workers in Indian garment factories.

We conducted extensive interviews with workers at all levels of the factory to elicit the most important tasks performed by line supervisors and identified eight key areas: (i) organize resources on the line; (ii) communicate targets and other upper-level managerial decisions to operators; (iii) ensure quality (correct mistakes); (iv) instruct operators, (v) transmit workers' requests and complaints to upper-level management; (vi) motivate and discipline operators; (vii) teach workers new sewing operations when lines switch garment styles (which they do on average around every two weeks); and (viii) understand machines and fix glitches when they happen. Sewing operators consider this last as the most important in our interviews, while line supervisors and higher-up managers place teaching new techniques and motivating operators at the top.

The vast majority of line supervisors and chiefs are male. Women account for only 6% of line supervisors, and less than 1% of line chiefs or higher-level managers (Menzel and Woodruff, 2021). These figures stand in stark contrast with women accounting for over 80% of all workers in the sewing sections, and with the fact that most line supervisors are internally promoted into their positions from among sewing operators. Line supervisors are thus almost exclusively recruited from among the few male sewing operators. At the time of our intervention, the industry suffered from a shortage of qualified supervisors as

the country's fast growth made it harder to lure the most qualified men into the sector (McKinsey, 2011).Expanding the pool from which line supervisors are recruited to include women could potentially relax these constraints.

2.2 Experimental Design

The under-representation of women among line supervisors may have many causes. In this project, we worked with 24 suppliers of a large UK-based buyer to test whether too few women were being promoted to line supervisor relative to an efficient allocation. Comparing the performance of existing male and female supervisors does not provide a suitable test, since the underlying distribution of talent might differ across the two groups. The test requires comparing the characteristics and performance of the *marginal* candidates, i.e., the best male and female operators available for an internal promotion. This poses two challenges: (1) the marginal male and female operators are not representative of the average male and female operators in the factory and must thus be identified; (2) once identified, we need to compare their skills and performance as line supervisors while they work as operators.

To overcome these challenges, we asked participating factories to identify an equal number of male and female sewing workers and trial them as line supervisors for some time on randomly allocated lines. The design of this intervention required persuading both the factories and the candidates to participate. Given the amount of collaboration involved, the design was informed by the results of an earlier pilot with 58 factories (see Appendix D). The pilot offered a six-week classroom-based supervisor training program to nearly 200 female sewing operators and revealed that (i) factories promoted many of these female trainees as line supervisors – this evidence was crucial to persuading factories in this project to commit to trying candidates as supervisors on production lines; (ii) the number of trainees from each factory should be tailored to the number of anticipated openings for line supervisors; and (iii) the training had not significantly increased the knowledge of trainees but was nevertheless deemed important to ensure that both factories and candidates took the program seriously.

Given point (ii) above, we asked factories how many new line supervisors they anticipated appointing over the next six months. We then asked them to select a commensurate number of sewing workers – equally split between females and males – for the supervisor training program. To incentivize participation, we paid the direct cost of the training

but asked factories to pay workers' salaries while workers were away at the training. We scheduled four training rounds starting between March and May 2014, with each factory randomly allocated to either rounds 1 and 3 or rounds 2 and 4. Within factories, trainees were randomly assigned to an early or a late round, stratified by gender. This process gave us an (aggregate) target of approximately 150 candidates. Based on take-up rates from the pilot, factories approached 121 female and 100 male candidates. Twenty-one females declined the invitation to the training, 11 female and 18 male candidates did not pass the literacy test required to enroll in the training, and, ultimately, 145 operators (henceforth *trainees*) – 72 males and 73 females – completed the training.[6] After the training, the trainees then worked for two months as co-supervisors on a line we randomly selected from a set of lines nominated by the factory, with the number of nominated lines equal to the number of trainees.[7] This randomization allows us to compare the performance of male and female trainees as line supervisors. After the trial, factories were free to return trainees to non-supervisory positions, keep them as supervisors on the randomly allocated line, or move them to other lines as line supervisors.

2.3 Data: Surveys and Administrative Records

Survey Data We surveyed trainees, other workers, and supervisors on four separate occasions. The first survey was a comprehensive diagnostic of trainees conducted on the first day at the training center. This included assessments of technical knowledge of garment production processes, communication, teaching and leadership skills, and numeracy, literacy, and non-verbal reasoning skills. Second, over two days near the end of the training, we surveyed at the factory all of the supervisors and line chiefs and three randomly chosen workers, stratified by gender, from each of the lines nominated by the factory for the trial. We also surveyed line chiefs from the lines where trainees worked as operators before the training, floor managers, and the factory's production and HR managers. We asked trainees to keep a diary during the two-month trial, noting the line on which they worked and how many workers they supervised each day. The third survey was conducted just after the end of the two-month trial. Over a two-day visit, we again surveyed three randomly selected operators and the co-supervisors and line chiefs from

[6] As in the pilot project, a before-after comparison confirms that the training did not increase the knowledge of trainees of either gender (see Appendix C).

[7] Nominated lines are similar to other lines in the factories (Table A.1).

the lines nominated for the trial, and from all lines where any trainee worked even if it was not his or her randomly assigned line. Finally, we conducted a second follow-up survey with trainees, randomly selected operators, and supervisors about 18 weeks after the trial ended for those trained in the early rounds, and 10 weeks after the end of the trial for those trained in the late rounds. We compressed the second follow-up survey into a single day to minimize disruption at the factory. Given time constraints, and the fact that after the trial factories were free to allocate trainees to any line of their choice, we interviewed supervisors and random workers from the lines where trainees were working as supervisors at that time, rather than from the lines to which they were initially assigned.[8]

Administrative Production Data We collected daily data for all production lines in the factories that kept detailed daily production line data, for a period of a year starting from 3 to 5 months before the start of the trial. These data are available for factories that nominated 112 (58 female and 54 male) of the 145 trainees in the study.

We measure daily production line efficiency with a standard engineering measure used in the sector. Efficiency is defined as the ratio of output minutes (given by the number of pieces produced times the garment's "Standard Minute Value" (henceforth SMV)) and input minutes (the number of operators on that line times the minutes the line operated on the day).[9] We also collected daily, line-level data on quality defects and workers' absenteeism. At the end of each production line, quality control inspectors check each piece, with defective pieces not counted in the daily output until the defects have been corrected. Finally, 13 factories provided absenteeism data at the line level. The remaining factories collect these data without line information.

2.4 Randomization Balance and Compliance

Randomization Balance Appendix Table A.2 tests whether the randomization of the nominated lines to receiving male and female trainees is balanced in terms of operator characteristics (from the randomly selected operators surveyed at baseline), line supervisor characteristics (from the same surveys), and administrative production data. Five

[8]We surveyed trainees who left the factories by phone and continued to conduct bi-monthly telephone follow-up surveys with trainees for half a year after the last in-factory survey.

[9]The SMV is calculated by breaking down the sewing process of a garment into individual stitches and assigning a time value to each stitch. The time value of each stitch in the garment is then summed up. Industrial engineers in the factory calculate SMVs taking into account the available machines and sometimes allowing for extra time for handling the garment or cutting thread. SMVs are thus consistent within factories but not across. We include factory fixed-effects in our analysis.

out of 39 variables display an imbalance at the 10 percent significance level. F-tests for the three groups of variables do not indicate imbalance.

Compliance with experimental protocol Eighty-eight percent of both the 72 male and 73 female trainees started their trial after completion of the training program. Fifty-five percent of all trainees were trialed on the randomly assigned line (64 percent of those that were trialed), with again no significant gender difference in the compliance rate. Much of the line-level non-compliance involved the switching of two female, or two male, trainees. With this in mind, we define a third measure of compliance: the share of trainees trialed on a line randomly assigned to a trainee of their gender: 67 and 63 percent of female and male trainees were compliant, or 78 and 71 percent, respectively, conditional on starting the trial. We focus on ITT specifications, comparing lines randomly assigned male or female trainees, to deal with non-compliance.[10]

2.5 Baseline Beliefs and Uncertainty about Men and Women as Supervisors

Employees at all rungs of the organizational hierarchy – including female workers – believe that men are better supervisors than women. During the baseline surveys, we asked managers, supervisors, and workers whether they believe that men are more able (coded as -1), women are more able (coded as $+1$), or men and women are equally able (coded as 0) in each of the eight key aspects of the supervisor role described above.

Figure 1 illustrates the average response on each of the eight skills for workers, supervisors, and managers separately. Respondents at *all* levels report that women are worse supervisors than men. Male operators have the most negative beliefs among all respondents in all dimensions, but female operators also consistently report that women are worse supervisors than their male colleagues, though their beliefs are not as negative as those of male workers. Line supervisors and production managers have beliefs similar to those of female workers. Only HR managers believe there are no gender differences in five of the eight skills. Notably, all groups of respondents give women the lowest relative rating on "understanding machines", a skill that, as mentioned above, is considered most important by operators and that can be accurately assessed through a test.

[10]Compliance rates as low as 50 percent are not uncommon in randomized field experiments (see, for example, the overview by Banerjee et al. (2015)). Half of both the male and female trainees trialed on a line not randomly allocated to a trainee of their gender were allocated to a line not originally nominated by the factory. Conversations with managers revealed that non-compliance was due to idiosyncratic mistakes in communication or shocks to lines, rather than deliberate reassignment of trainees.

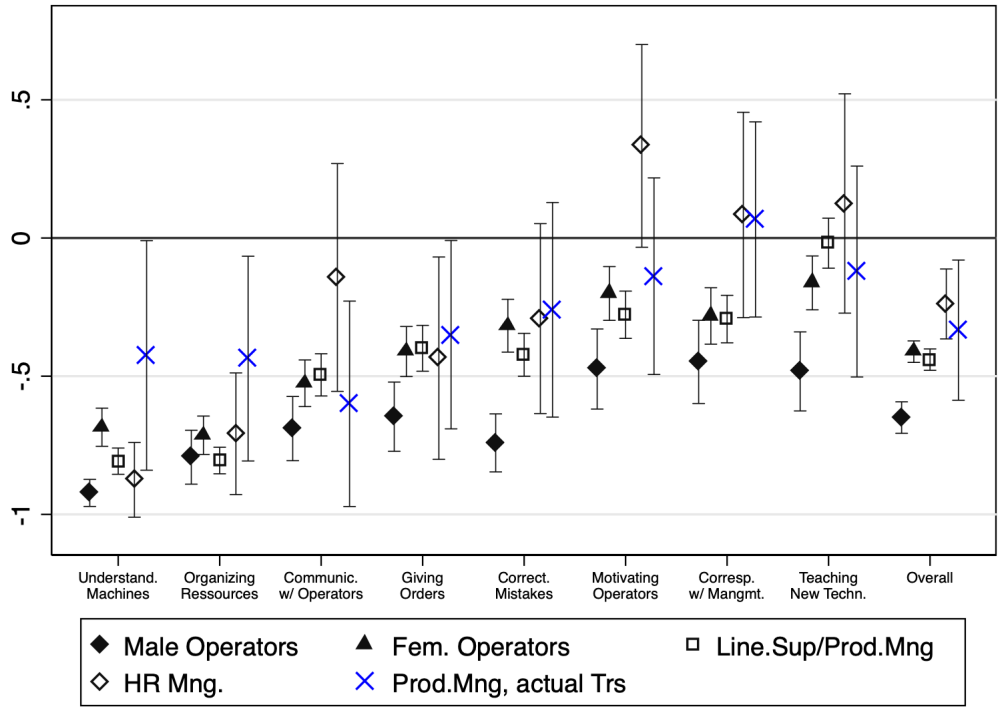

Figure 1: Baseline Beliefs on Female Supervisor Abilities. Figures show mean responses from workers in the factories of the types shown in the legend to the following question: "Do you think that male or female workers are better at following task/have more of the following skill?" Responses coded as -1 for "Men are better", 0 for "No differences", and 1 for "Women are better". "Overall" represents averages over all eight skills/tasks. Capped bars represent 95 percent confidence intervals.

These questions elicited beliefs about a comparison between generic male and female workers, rather than the male and female trainees. The trainees were not directly known to most respondents. However, we also asked the line chiefs and production managers involved in selecting trainees to compare the female and male trainees they selected. Figure 1 shows that beliefs about female trainees are only slightly less negative than those elicited in the generic comparison. Finally, the managers selecting the trainees also reported lower confidence in their assessment of female candidates: the average reported confidence for male (female) trainees was 84 (76) out of 100 (p-value= 0.011, N=114).

3 Conceptual Framework

This section presents a conceptual framework to guide the interpretation of the experimental intervention. The evidence from the previous section motivates a simple model that combines *experimentation* and *discrimination* in which there is (relatively more) ex-ante uncertainty over women's performance as supervisors. Alongside the standard discrimination logic that the marginal female candidate should perform better than the

marginal male, experimentation introduces a countervailing force: trying a female candidate generates valuable information that can be used to make better future appointments. A higher performance of the marginal female candidates is no longer necessary to establish that the baseline share of female supervisors is below the efficient level. Positive updates in beliefs about the performance of female supervisors following experimentation, however, increase the share of women among supervisors in the long run.[11]

3.1 Set-Up

Consider a firm in which a decision-maker (henceforth, DM) needs to fill one position for a line supervisor. There are two periods, $t \in \{0, 1\}$. Period $t = 0$ is the initial learning period. Period $t = 1$ captures the potentially much longer period in which decisions are based on information learned during $t = 0$. For this reason, at $t = 0$, the DM discounts payoffs at $t = 1$ by a factor $\delta \lesseqgtr 1$.

There are two candidates $i \in \{F, M\}$ for the job: a male candidate, M, and a female candidate, F. In each period $t \in \{0, 1\}$ the output of the production line supervised by candidate i is given by

$$y_{it} = \mu_i + \Delta_{it}, \tag{1}$$

where μ_i is assumed to be observed by the decision-maker before deciding which candidate to appoint while Δ_{it} captures drivers of performance (e.g., shocks) not observed by the decision-maker when taking the appointment decision. We normalize $\Delta_{Mt} = 0$ for $t \in \{0, 1\}$, that is, we assume that there is no uncertainty over the performance of the male candidate. In contrast, there is uncertainty over the performance of the female candidate – a natural assumption in contexts like ours in which one group is under-represented.

We model uncertainty over Δ_{Ft} as follows. We assume that Δ_{Ft} is independently and identically distributed over time, and can take one of two values: $\Delta_{Ft} \in \{\Delta - \epsilon, \Delta\}$. $\Delta_{Ft} = \Delta - \epsilon < 0$ occurs with probability ρ_F and its realization corresponds to the case in which the female candidate underperforms due to potentially many factors. $\Delta_{Ft} = \Delta > 0$, instead, occurs with probability $(1 - \rho_F)$ and its realization corresponds to the case in which the female candidate performs better than expected.[12]

[11] The simple model with binary structure in the main text derives testable implications during and after the trial and is particularly convenient for binary promotion decisions. Appendix F introduces experimentation in the canonical model of statistical discrimination (Phelps, 1972; Aigner and Cain, 1977) and shows that: (1) all implications carry through in that framework; (2) experimentation is akin to a more precise signal for the disadvantaged group.

[12] The qualitative insights of the model extend to the case in which Δ_{Ft} is drawn from a distribution

Crucially, at time $t = 0$, the DM is uncertain about the distribution of Δ_{Ft}. The DM can learn about the distribution of Δ_{Ft} by appointing the female candidate and observing her performance at time $t = 0$. That is, there is scope for experimentation. Specifically, we assume that at time $t = 0$ the DM believes that $\rho_F = \underline{\rho}$ with probability ρ_0 and $\rho_F = \bar{\rho}$ with probability $(1 - \rho_0)$, with $1 \geq \bar{\rho} > \underline{\rho} \geq 0$. At $t = 0$, therefore, the DM expects $\mathbf{E}_0[\Delta_F] = \Delta - \beta_0\epsilon$, with $\beta_0 = \rho_0(\bar{\rho} - \underline{\rho}) - \bar{\rho}$ being the probability with which the DM expects to observe $\Delta_{F0} = \Delta - \epsilon$. Prior beliefs, ρ_0, could be the result of past experimentation with other female candidates. Prior beliefs ρ_0, therefore, need not be unbiased, but will be updated after the DM observes the realization of Δ_{F0}.[13]

The distinction between μ_i and Δ_{Ft} captures the idea that the DM has some information about the candidate's performance, μ_i, but also faces uncertainty about the relative performance of female candidates for the job. The DM might be uncertain about the skills of the female candidate: DMs in our context have relatively little experience with female candidates and might have a harder time assessing their skills. Furthermore, the DM might not know how co-workers, including subordinates, line co-supervisors, and line chiefs, behave when working with or under a female line supervisor. Finally, female candidates themselves might initially lack self-confidence. In all these cases, and holding constant her talent, the performance of the line managed by the female candidate will suffer. For simplicity, since μ_i captures the expectation derived from individual-specific skills, we refer to it as the candidate's talent. In contrast, since Δ_{Ft} captures the adjustment to this expectation based on gender expectations and bias, we refer to it as bias. In the reminder of the text, we will use these labels with the understanding that – in practice – the distinction is not as clear-cut and that the insights of the model extend to the case in which the DM must learn about both.

We make three assumptions here to set up the DM's problem, and one more below to interpret our experiment through the model. We will provide evidence in support of each of these four assumptions in Section 4.5.

Assumption 1 *There is an expected negative bias against female supervisors among co-workers and/or subordinates, which lowers their productivity* ($\epsilon > max\{\Delta/\underline{\rho}, 0\}$).[14]

with continuous support. The binary case provides the same intuition with simpler notation.

[13] ρ_F thus captures attributes common across female candidates. Although, for simplicity, we focus on a single appointment decision, this formulation captures how observing the performance of a specific female candidate changes beliefs about the expected performance of other, e.g. future, female candidates.

[14] $\epsilon > max\{\Delta/\underline{\rho}, 0\}$ implies $\mathbf{E}_0[\Delta_{F0}] < 0$ regardless of the DM's prior beliefs ρ_0.

Assumption 2: *The DM is a) non-discriminatory, and b) maximizes the expected discounted output on the line.*

Assumption 2(a) rules out taste-based discrimination on the part of the *DM*. The model distinguishes discrimination by co-workers and subordinates (captured by Δ_{Ft}) from DM's discrimination. The DM can have incorrect prior beliefs, ρ_0, but otherwise has no intrinsic preferences over the gender of the appointed supervisor other than through its effect on performance.[15] Assumption 2(b) abstracts from wage differences between male and female supervisors.

Assumption 3: *At $t = 1$, there is no cost of demoting the trainee trialed at time $t = 0$ and appointing a new supervisor.*

Assumption 3 simplifies the algebra without altering qualitative insights.

3.2 Experimentation and Learning

We now derive the condition under which the DM appoints the female candidate. If the DM appoints the female candidate in $t = 0$, he observes the realization of Δ_{F0} and learns. Let us denote with $\bar{\mathbf{E}}_1[\Delta_F]$ and with $\underline{\mathbf{E}}_1[\Delta_F]$ the updated expected bias at $t = 1$ after a positive ($\Delta_F = \Delta$) and negative ($\Delta_F = \Delta - \epsilon$) update respectively.

Proposition 1 *The optimal decision is as follows:*

1. *If $\mu_F + \bar{\mathbf{E}}_1[\Delta_F] < \mu_M$ the male candidate is appointed in both $t = 0$ and $t = 1$.*

2. *If $\mu_F + \underline{\mathbf{E}}_1[\Delta_F] > \mu_M$ the female candidate is appointed in both $t = 0$ and $t = 1$.*

3. *If $\mu_M \in (\mu_F + \underline{\mathbf{E}}_1[\Delta_F], \mu_F + \bar{\mathbf{E}}_1[\Delta_F])$ the female candidate is appointed in $t = 0$ if*

$$\mu_F - \mu_M > -\frac{\mathbf{E}_0[\Delta_F] + \delta\bar{\mathbf{E}}_1[\Delta_F](1-\beta_0)}{1+\delta(1-\beta_0)}. \tag{2}$$

The female candidate is then retained as supervisor in $t = 1$ if $\Delta_F = \Delta$ is observed. Otherwise, the male candidate is appointed in $t = 1$.

If the DM appoints the male candidate at $t = 0$, he learns nothing and will therefore re-appoint the male again at $t = 1$. If, instead, he appoints the female candidate at $t = 0$, he observes the realization of Δ_{F0} and this may lead to a different decision at $t = 1$.

[15] As discussed in subsection 4.5, this assumption is made for simplicity and ease of interpretation, but the model's insights do not depend on the assumption. Note that we take a reduced-form approach to the bias Δ_i and do not take a stand on whether it arises from taste-based or from statistical discrimination (e.g., it could arise from subordinates' and co-workers' potentially biased beliefs about μ_F). In other words, subordinates and co-workers may or may not share the same source of bias as the DM.

Upon observing the realization of Δ_{F0}, the DM updates his beliefs to $\rho_1 \in \{\underline{\rho}_1, \bar{\rho}_1\}$, where $\underline{\rho}_1$ is the posterior belief after observing $\Delta_{F0} = \Delta - \epsilon$ (the DM becomes more pessimistic about the bias), $\bar{\rho}_1$ after observing $\Delta_{F0} = \Delta$ (the DM becomes more optimistic about the bias), and $\underline{\rho}_1 < \bar{\rho}_1$. Experimentation is, of course, only useful if it potentially changes the decision at $t = 1$. If $\mu_F + \underline{\mathbf{E}}_1[\Delta_F] > \mu_M$ then the DM retains the female candidate at $t = 1$ even when he becomes more pessimistic about her. *A fortiori*, he will appoint her at $t = 0$. If, instead, $\mu_F + \bar{\mathbf{E}}_1[\Delta_F] < \mu_M$, the DM appoints the male candidate at $t = 1$ even after updating positively about the female candidate. If that is the case, the DM should appoint the male candidate also at $t = 0$.

For intermediate values $\mu_M \in (\mu_F + \underline{\mathbf{E}}_1[\Delta_F], \mu_F + \bar{\mathbf{E}}_1[\Delta_F])$, the DM appoints at $t = 0$ the female candidate if condition (2) is satisfied. Holding constant the talent of the male and female candidates, the condition is more likely to be satisfied the more optimistic the DM and the higher the discount factor. A non-discriminatory, but pessimistic, DM will not experiment and fail to learn the share of female candidates that are good supervisors.

Condition (2) characterizes the hurdle, $\mathcal{H} = \mu_F - \mu_M$, that the female candidate must meet to be appointed and captures the main insight from the model. The first term on the numerator of the right-hand side of the condition (2) captures the standard *discrimination* logic. Recall that, by Assumption 1, $\mathbf{E}_0[\Delta_F] < 0$. Absent experimentation ($\delta \to 0$), the female candidate must be strictly better than the male candidate to be appointed: the hurdle is positive, $\mathcal{H} > 0$. *Experimentation* introduces a countervailing force. The second term on the numerator, $\delta \bar{\mathbf{E}}_1[\Delta_F](1-\beta_0)$, captures the value of better future appointment decisions as a result of learning and is positive. If $\delta \bar{\mathbf{E}}_1[\Delta_F](1-\beta_0) > -\mathbf{E}_0[\Delta_F]$, *in expectation* the marginal female candidate might have lower talent – and, *a fortiori*, lower performance – than the marginal male candidate and yet too few female supervisors are appointed if the DM is sufficiently pessimistic to experiment.

3.3 The Intervention through the Lens of the Model

We now discuss how the model relates to our intervention. Given the low share of women among supervisors and the baseline beliefs documented in Section 2, it is natural to assume that, before our intervention, condition (2) was violated and the DM did not experiment with female candidates. Our intervention, in partnership with the buyer, nudged factories to experiment with female candidates they would have otherwise not tried. The two-month trial period then corresponds to $t = 0$ in the model, while $t = 1$

reflects the permanent promotion decision of trainees after the end of the trial. A potential extension with an additional period (not modeled here for simplicity), $t = 2$, in which factories make appointment decisions after the end of our intervention based on updated beliefs maps the model to our long-run evidence.

Assumption 4: *Factories participating in the experiment selected as trainees the marginal male and female candidates.*

The factory might already employ N_M and N_F male and female supervisors which might differ in characteristics and performance. The factory might have more candidates available than positions to fill for the training. We assume factories selected the best – i.e., the *marginal* – candidates of each gender.

Under Assumptions 1-4, the following facts are consistent with too few female supervisors at baseline, relative to a benchmark with correct beliefs:

Fact 1. *At baseline, female trainees have similar (or lower), skills than males ($\mu_F \leq \mu_M$);*

Fact 2. *During the trial ($t = 0$), female trainees perform on average worse than males ($y_{F0} < y_{M0}$);*

Fact 3. *After the trial ($t = 1$), a share ($1-\rho_F$) of female trainees are retained as supervisors and, conditional on that, perform similarly to retained males;*

Fact 4. *After the intervention ($t = 2$), factories increase the share of women among newly appointed supervisors.*

Fact 1 compares male and female trainees. We conduct extensive diagnostics to test the skills of the candidates. The model clarifies that there can be inefficient under-representation of female supervisors at baseline even if the diagnostic reveals that marginal female candidates are similar to, or even worse than, male candidates. Since $\mathbf{E}_0[\Delta_F] < 0$, female candidates have worse performance than male candidates during the trial even when they have similar skills (**Fact 2**). Despite the, on average, worse performance during the trial, the factory learns that some female candidates (a share equal to $(1 - \rho_F)$) perform similarly to male candidates and retains them as line supervisors (**Fact 3**). This share can be significantly higher than the share of female supervisors at baseline, $N_F/(N_M + N_F)$. Finally, after the end of the experiment, some factories might appoint

new female supervisors who were not part of the intervention (**Fact 4**). This happens because the intervention induces factories to update beliefs. Consider a factory that trained E female candidates and observed e instances with $\Delta_{F0} = \Delta$. Posterior beliefs are

$$\rho_1^{E,e} = \frac{\rho_0(1-\underline{\rho})^e(\underline{\rho})^{N-e}}{\rho_0(1-\underline{\rho})^e(\underline{\rho})^{N-e} + (1-\rho_0)(1-\bar{\rho})^e(\bar{\rho})^{N-e}}. \tag{3}$$

Factories that observe many positive realizations of Δ_{F0} during (and after) the trial become (sufficiently) more optimistic to experiment with new female supervisors in the future. Formally, $\rho_1^{E,e} > \rho_0$ if $e/N > \frac{ln(\bar{\rho}/\underline{\rho})}{ln(\bar{\rho}/\underline{\rho})+ln(1-\underline{\rho}/(1-\bar{\rho})}$. Condition (2) might be violated with prior beliefs ρ_0 but satisfied with updated beliefs $\rho_1^{E,e}$. Other factories, however, might observe realizations of Δ_{F0} that do not make them update as positively. These factories might retain successful female trainees but subsequently revert to not appointing female supervisors. Because, in practice, each factory observes a small number of draws N during the experiment, there is no guarantee that any of them learns the true value of ρ_F. Dispersion in the long-run share of newly appointed female supervisors across factories occurs even if prior beliefs ρ_0 are unbiased. Furthermore, initial biases stifle subsequent experimentation: holding constant the number of trials, E, and successes, e, posterior beliefs are lower the more negatively biased the beliefs at $t = 0$.

4 Main Results

This section presents the main empirical results. Subsections 4.1 through 4.4 provide evidence for Facts 1-4. Subsection 4.5 provides empirical evidence on Assumptions 1-4.

4.1 *At baseline, female trainees have similar skills to males.*

Fact 1 concerns the comparison of male and female trainees along drivers of performance that the factories observe before the trial. We use a comprehensive diagnostic of trainees conducted before the trial to proxy for those. Recall that, for simplicity, the model referred to the components of performance the factories observe before the trial – μ_i – as the candidate's talent and to those the factory has to learn about – Δ_{Ft} – as the bias against the (female) candidate. As we discussed in the previous section, however, the distinction is not as clear-cut, and some of the skills compared in this section may not have been observed by the factories. We discuss this issue at the end of the section.

Based on the eight key skills identified as important for line supervisors in Section 2, we conducted a comprehensive diagnostic of trainees that assessed numeracy, literacy,

non-verbal reasoning skills, soft skills including communication, teaching, and leadership as well as technical knowledge of garment production processes.[16]

Table 1 shows that the skills of female and male trainees are broadly similar – establishing Fact 1. If anything, female trainees have slightly lower skills on some dimensions. Starting with demographics, female and male trainees have similar age, education, factory tenure and sector experience. Moving beyond demographics, female trainees perform worse than male trainees on the numeracy test but have similar literacy and reasoning skills. Female trainees perform similarly to male trainees in soft skills in communication, teaching, and leadership, but expressed lower confidence in their ability to perform as supervisors. However, when we repeated the confidence questions on the last day of training, the gap in self-confidence was smaller and no longer statistically significant (bottom rows of Table 1). We also find no differences between male and female trainees in willingness to accept a promotion, measured both before and after the training.

Remarkably, male and female trainees scored equally on a comprehensive battery of 86 questions eliciting technical knowledge of garments' production processes and machines (see Appendix E for the test). This result is particularly significant for several reasons. First, technical knowledge can arguably be assessed with less measurement error than soft-skills. Second, technical knowledge is an important driver of performance and – as mentioned in Section 2.5 – is the most important trait of a good line supervisor according to sewing operators. Technical knowledge assessed on a sample of existing line supervisors positively correlates with the efficiency of their production line (see Figure A.1). Finally, as shown in Figure 1, it is the dimension about which employees at *all* levels – from managers to operators – expressed the most negative beliefs about women at baseline.

These results are consistent with observed skills, μ in the model, of male and female trainees being equal – establishing Fact 1. Note, though, that Fact 1 does not require that *all* skills covered in the diagnostic are observed by factories at baseline. If some were not observed, the strong negative beliefs about female supervisors documented in section 2.5 suggests that factories could have learned about these dimensions – Δ_{Ft} in the model. This may be especially the case for technical knowledge, where the discrepancy between

[16]To measure communication skills, the trainee had to explain several abstract figures verbally while other trainees had to draw them. We use the number of figures the trainees could draw correctly as a measure of communication skills. We measured leadership skills through a game in which trainees had to produce different "products" using Legos. Two enumerators per group scored how often and actively each trainee participated in the group discussions, assigning a leadership "soft score".

Table 1: Fact 1 - At baseline, female trainees have similar skills than males

	Mean Males	Female	SE	N
Demographic:				
Age	24.73	-1.102	(0.758)	145
Married	0.611	0.152*	(0.078)	145
Years Schooling	8.486	-0.334	(0.273)	143
Years in Garment Sector	6.441	-0.520	(0.537)	145
Years in Factory	3.655	-0.714*	(0.392)	145
Nbr Factories	2.069	-0.165	(0.330)	145
Skill Diagnostic:				
Literacy	8.971	-1.107	(0.867)	142
Numeracy	4.826	-1.345***	(0.383)	142
Non-verbal Reasoning	3.159	0.028	(0.376)	142
Technical Knowledge	53.5	-0.348	(1.522)	145
Drawing	0.375	-0.125*	(0.069)	126
Drawing - Soft	-0.561	1.015	(0.661)	127
Communication - Soft	0.136	-0.274	(0.814)	127
Leadership - Soft	0.334	-0.611	(0.667)	124
Beliefs & Attitudes:				
Confidence	-0.142	-0.849***	(0.282)	143
Belief Best	0.652	-0.198**	(0.084)	145
Accept Promotion	0.971	-0.043	(0.040)	143
Beliefs & Attitudes after Training:				
Confidence, after Training	0.197	-0.333	(0.278)	144
Belief Best, after Training	0.791	-0.136*	(0.073)	145
Accept Promotion, after Training	0.957	-0.001	(0.032)	144
Confidence, after Trial	0.238	-0.248	(0.246)	127

Notes: Table compares the 72 Male and 73 Female Trainees on measured ability and other observed characteristics before start of the training. See Section 2 for definition of variables. All comparisons control for factory fixed effects. *** denotes statistical significance at 1%, ** at 5%, and * at 10%.

baseline beliefs and the diagnostic is starkest.

4.2 *During the trial, female trainees perform worse than males.*

Fact 2 compares the performance of female and male trainees during the trial. Female and male trainees were randomly allocated to production lines to work as supervisors during the trial. We use the daily administrative production data from the factories and production-line level outcomes for the periods before and during the trial.[17]

[17]As discussed in Section 2, lines in the factories did not start the trial on the same day, leading to differing pre-trial periods across lines. However, trainees were randomized to lines, and Table A.2 shows that pre-trial outcomes are uncorrelated with the gender of the allocated trainee. We control for pre-trial outcomes in Ancova specifications. Results are similar if we do not control for pre-trial outcomes. We always control for factory fixed-effects.

Table 2: Fact 2 - Lower Productivity of Marginal Female Supervisors

VARIABLES	(1) ITT Efficiency	(2) Efficiency	(3) Efficiency Pre-Trial
Female Tr.	-4.402**	-4.925***	0.905
	(1.835)	(1.719)	(1.401)
Ancova	Yes	Yes	-
PDS Controls	No	Yes	No
Factory FE	Yes	Yes	Yes
Mean Males	60.51	62.86	61.86
N	93	87	87

Notes: Table show differences in average line productivity of lines receiving male versus female trainees. Column 1 show comparisons between lines randomized into receiving male and female trainees (ITT estimates), while Column 2 compares lines on which trainees actually worked, i.e. not correcting for non-compliance with the experimental protocol. Column 3 shows the comparisons for the time before the trial, for lines on which trainees actually worked. Ancova indicates controlling for pre-trial (baseline) average productivity of lines. Column 2 controls further for variables selected by PDS Lasso from line and trainee controls, plus squares of all controls and indicator variables for missing values of each control. Robust standard errors in parentheses; *** denotes statistical significance at 1%, ** at 5%, and * at 10%

Column 1 in Table 2 shows that during the trial lines randomly allocated to female trainees have lower efficiency – establishing Fact 2. The intention-to-treat (ITT) estimate shows that efficiency is four points lower on lines assigned to female trainees, corresponding to a seven percent lower efficiency compared to lines randomly allocated to male trainees. Given non-compliance with the randomization protocol, Column 2 also reports OLS estimates that compare production lines that actually received a female trainee with those that received a male trainee. This specification yields a gap of five efficiency-points, corresponding to a nine percent lower efficiency. Given the selected placement of trainees, and the fact that female and male trainees differed on some characteristics, Column 2 also includes a set of trainee and line-level characteristics selected by PDS Lasso (Belloni et al., 2016). Results are qualitatively the same without these controls. Finally, Column 3 compares the pre-trial efficiency levels of the lines on which female and male trainees actually worked, finding no significant difference: non-compliance with the random allocation of male and female trainees was uncorrelated with baseline line efficiency.[18]

In sum, *on average*, lines receiving female trainees performed worse than those re-

[18]Appendix Table A.3 finds that production lines allocated to female trainees also experienced higher quality defect rates during the trial. Worker absenteeism, however, was not different between production lines with male and female trainees.

ceiving male trainees during the trial. According to the model, multiple channels could account for the under-performance of female trainees during the trial. First, there is bias – $\mathbf{E}_0[\Delta_{F0}] < 0$ – which lowers the performance of female trainees, holding constant their talent μ_i. We provide below direct evidence in support of this channel. Second, Fact 1 implies that female trainees might also have lower skills than male trainees, on average. Regardless of the channel at play, the experimentation logic suggests that the under-performance of female trainees during the trial does not imply that factories should not appoint more females as supervisors after the trial and beyond.

4.3 *After the trial many female trainees are retained as supervisors and, conditional on that, perform similarly to retained male ones.*

Even though female trainees perform worse than male trainees during the trial, factories might learn that some – and perhaps many – female trainees perform as well as males and decide to permanently promote them as line supervisors.

Table 3 confirms both implications, establishing Fact 3. Column 1 shows that 67 percent of the male and 53 percent of the female trainees were retained as supervisors after the end of the trial. While the difference in promotion rates between male and female trainees is statistically significant, the share of female trainees permanently promoted to line supervisors is substantial. Conditional on having been trialed, 60 percent of female trainees were permanently promoted. Column 2 compares the performance of promoted male and female trainees after the end of the trial. Recall that, at this point, factories were free to allocate the newly promoted trainees to any line. Around half of both male and female trainees who were promoted were moved to a different line. We therefore compare performance across these lines (rather than on the ITT), including the line's pre-trial efficiency and controls selected by PDS Lasso. Column 2 shows that the performance gap observed during the trial almost completely vanished: the point estimate is one-seventh of the OLS estimate during the trial and is no longer statistically significant.

The erosion of the performance gap between female and male trainees after the trial could arise from three distinct channels: i) conditional on promotion, female trainees are reallocated to more efficient lines (*reallocation*), ii) relatively better female trainees are retained as line supervisors (*selection*), and/or iii) female trainees improve their performance relatively more (*catch-up*). The reduced-form representation of bias in the model in Section 3 explicitly captures the selection mechanism. However, a less reduced-form

Table 3: Fact 3 – Increased Share of Female Supervisors

VARIABLES	(1) Retention as Superv.	(2) Post-Trial Efficiency
Female Trainee	-0.134*	-0.734
	(0.072)	(1.730)
Ancova	-	Yes
PDS Controls	No	Yes
Factory FE	Yes	Yes
Mean Males	0.667	60.33
N	145	81

Notes: Column 1 regresses on the trainee level a dummy on whether the trainee continues to work as supervisor after the trial period, on an indicator for the trainee being female. Column 2 compares the efficiency of lines on which retained male versus female trainees work as supervisors after the trial period, based on administrative line efficiency data. Both columns control for factory fixed effects. Column 1 further controls for variables selected by PDS Lasso from line and trainee controls, plus squares of all controls and indicator variables for missing values of each control. Robust standard errors in parentheses; *** denotes statistical significance at 1%, ** at 5%, and * at 10%

specification would also be consistent with the catch-up mechanism – factories learn how long it takes for female trainees to catch-up, or which characteristics of female trainees make successful catch-up more likely. We provide evidence below that exposure to female supervisors quickly reduces negative bias.

Column 1 in Table A.4 shows that, at baseline, there was no difference between the efficiency of lines on which retained male and female trainees were appointed as supervisors after the trial. This suggests that reallocation to more efficient lines is unlikely to explain the closing of the performance gap. Column 2 of Table A.4 compares the efficiency *during the trial* of lines with female and male trainees who were subsequently retained as line supervisors. This shows a performance gap similar to the one estimated among *all* trainees in Column 2 of Table 2, speaking against the selection mechanism. In Column 3, we use data during the trial to compare the performance of male and female trainees who were promoted and stayed on the same line after the trial. The gap is smaller and not statistically significant. Among this subset of trainees, the gap remains insignificant and shrinks further in size after the trial (Column 4). Overall, this table thus suggests that the erosion of the performance gap is driven by female trainees becoming more productive over time (catch-up). While many of the female trainees move to different lines when they are promoted, the new lines were not, on average, more

productive before their arrival.[19]

Given that many female trainees were promoted to line supervisor, it is worth asking whether their promotion rate was higher than what would have happened without our intervention. By definition, we do not know how many women would have been promoted without our intervention and can only attempt a conservative comparison. Recall that the number of trainees was selected to match the number of openings for line supervisors that factories expected for the months after the trial. The ratio of female trainees promoted relative to the total number of trainees, therefore, provides a conservative estimate of the proportion of women among the new appointments during that time period.[20] As counterfactual promotion rate for women, we take the share of women among internally promoted supervisors using data from Menzel and Woodruff (2021) (see Appendix C for detail). This share is 14 percent, which is higher – and thus more conservative – than the six percent share of female supervisors at baseline in our sample.[21]

At the time of our endline survey, a few months after the end of the trial, 38 female trainees were working as supervisors. The share of 38 retained female trainees among the 145 trainees, or 0.26, is statistically significantly higher than the 0.14 benchmark share of women among promotions (p-value= 0.018). The evidence thus suggests that the promotion rate in the experiment was higher than what might have happened without our intervention. This evidence – however – doesn't prove that the updating of beliefs at the heart of our experimentation model induced factories to promote more female supervisors. Factories might have done so because our intervention subsidized the costs of identifying suitable female candidates and/or because they faced high costs to demote trainees, once they had been trialed. Fact 4 addresses these concerns.

[19] As we discuss below, in 2019, Better Work Bangladesh scaled up the training program as the GEAR Program. An ex-post matched sample analysis of that program shows that, as in the results presented here, trained female supervisors underperform the mostly male comparison supervisors during the first six months and then close the gap after that (IFC and ILO, 2025). The GEAR program analysis then shows that the female trainees significantly *outperform* the comparison supervisors after two years.

[20] Unfortunately, we lack complete data on all promotions – of both trainees and non-trainees – in the months after the trial. Taking the ratio of retained female trainees over all trainees implicitly assumes that any promoted non-trainees were men, yielding a lower bound for the share of women among promotions.

[21] The share of women among promotions can be higher than that of women as supervisors if either female supervisors exit supervisor positions faster or if there is an upward trend, with the lower stock reflecting lower promotion rates in the past.

4.4 *After our intervention, factories appoint more new female supervisors.*

Our intervention nudged, but also helped, factories to experiment with female supervisors. If such experimentation led to a substantial update in beliefs, then the model suggests that factories should promote more, and new, female supervisors after the end of our intervention. There are two main challenges to testing this hypothesis. First, we need a set of "control" factories that were not nudged to experiment with female supervisors. Second, the hypothesis is about the long run: so we need to gather data for both samples of factories (long) after the end of our trial.

We address these challenges by obtaining data on the share of female supervisors in factories audited for the ILO-IFC Better Work Program (see Cajal Grossi et al. (2022)). The dataset covers a panel of 290 factories from 2015 to 2021, which includes factories that participated in this experiment, its pilot, and an extension trial that followed this project. In particular, 27 factories participated in the extension trial. The experiment described in this paper was conducted in 2014, before the start of the panel. The extension trial, however, was conducted *after* the start of the ILO-IFC Better Work Program data collection. This allows us to explore a difference-in-differences (DID) specification in which we compare changes in the long-run share of female supervisors between factories in the extension trial and a group of control factories that also participated in the ILO-IFC Better Work Program but did not participate in any of the three trials.[22,23]

Table 4 shows that factories in the extension trial appointed more female supervisors after the end of the trial than the control group – establishing Fact 4. Column 1 reports a simple pre-post comparison. The share of female supervisors in the factories participating in the extension trial was 4.6 percentage points higher in the three years after the trial (2017-19) than in the years before (2015-16), when the share was 3.4 percent. Column 2 adds as a control group in a DID specification factories that did not participate in any of the three trials and have at least one pre- and one post- observation. The estimated effect remains similar, at 5.2 percentage point increase. Figure 2 provides suggestive evidence that factories in the extension trial were not on differential trends relative to

[22]Based on the results from the pilot and from this trial, the ILO-IFC added the training to promote women to supervisors in some batches of factories participating in the Better Work Program. Factories in the extension trial were the first batch. If some factories in later cohorts of the Better Work Program also received the female supervisor training program this would arguably work against us.

[23]The pilot randomized invitation to the training but not line assignment. The extension trial did not train men (see Uckat and Woodruff (2020) for details). Both designs preclude an unbiased comparison of male and female trainees' performance, which is the focus of this paper.

Table 4: Fact 4 – Long-Run Increase in Female Supervisor Share

	(1)	(2)
	Pre-Post	Diff-in-Diff
		At least one
VARIABLES		pre- & post obs.
Extension Trial	0.046***	0.052***
	(0.014)	(0.016)
Observations	60	178
Factory FE	Yes	Yes
Year FE	-	Yes
Mean Pre-Trial	0.0338	

Notes: The Table shows results using data from ILO surveys of 290 garment factories in Bangladesh between 2015-19. The outcome variable reported is the share of female supervisors. "Extension Trial" is a dummy variable indicating that factory participated in 2016 in a follow-on project of the main project presented in this paper. Column 1 is a simple pre-post comparison of the share of female supervisors before and after the extension project, with years 2015-16 treated as pre-trial, and 2017-19 as post-trial (we exclude year 2020 and later due to the Covid-19 pandemic). Column 2 adds all factories from the ILO survey that did not participate in the extension project and have at least one observation in both the pre- and the post-treatment period, as a control group. It also includes year fixed-effects. We exclude factories that participated in the main trial or in the pilot. Standard errors clustered at the factory level, *** denotes statistical significance at 1%, ** at 5%, and * at 10%

control factories. While we acknowledge the short length of the pre-period, this assuages concerns that the results in the DID specification may be due to confounding factors.

The estimated increase in the long-run share of female supervisors is too large to be solely due to the female trainees promoted as part of the extension trial. On average, factories in the extension trial promoted 2.14 female trainees to supervisors (see Table A.7), which would imply an 1.6 percentage point increase in the share of female supervisors. The 5.2 percentage point increase is statistically different from such a 1.6 percentage point increase. Furthermore, based on the success of this extension trial, Better Work launched the GEAR program in 2019 (IFC and ILO, 2025). While this program extends beyond the period for which we have data from the full set of Better Work factories, data from the GEAR program shows that by October of 2022, the 40 participating factories for which we have complete data had promoted to supervisory roles 182 GEAR-trained women and an additional 144 women not trained in the GEAR program. There is considerable heterogeneity across factories: 21 factories promoted at least one female in addition to those trained, including 10 factories that promoted at least five in addition to those trained, while for 19 factories the data are consistent with no additional female promotions.

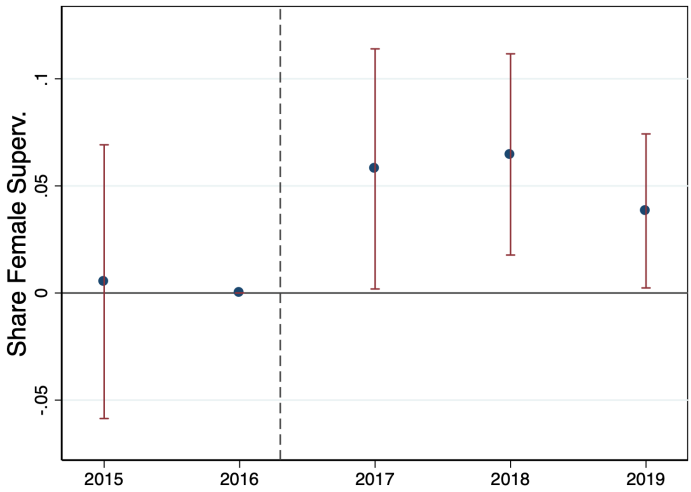

Figure 2: Long-Run Share of Female Supervisors The Figure shows differences in changes in the average shares of female supervisors over time (and associated 95 percent confidence intervals) in factories that participated in the extension trial (but not the main trial or pilot), and factories that participated in none of these trials, for the period 2015-19, based on data from the ILO-IFC Better Work Program. The vertical line indicates the implementation period of the extension trial.

4.5 Evidence in Support of Assumptions 1-4

Having presented the key empirical facts that our model was set up to rationalize, this section provides empirical evidence for the main assumptions in our model.

Assumption 1: There is an average negative bias against female supervisors among co-workers and/or subordinates, which lowers their productivity

The Assumption states that bias against female supervisors lowers their performance. For simplicity, the framework referred to the bias as Δ_{Ft} – the component of female trainees' performance the factory must learn about. The bias could originate from subordinates, co-workers, or direct bosses and could be due to preferences, beliefs, or both. For example, Figure 1 in Section 2 showed that workers across the different levels of the factory hierarchy perceive women as being less able supervisors than men. Exploring Fact 1, we noted that these beliefs are not in line with the evidence from the skill diagnostic for the male and female trainees shown in Table 1.

Table 5 provides direct support for Assumption 1. The Table explores the relationship between the beliefs held by different types of workers, elicited before the trial, and the performance of trainees randomly allocated to the line during the trial. Column 1

investigates the role of beliefs held by line co-supervisors, where beliefs are measured in the same way as in Figure 1. The estimates reveal that lines where co-supervisors have more positive beliefs about female supervisors have higher efficiency when they are allocated a female trainee, relative to their efficiency when allocated a male trainee. Figure 3 shows the same result graphically: the negative beliefs towards potential female supervisors elicited at baseline correlate negatively with the performance of female trainees – while not with that of male trainees – during the trial. Column 2 includes a measure, also elicited at baseline, for the *preference* to co-supervise a production line with a female supervisor. Results show that it is beliefs, rather than preferences, that matter. This provides suggestive evidence that misperceptions about women's abilities might be a source of bias that hinders their performance.

Finally, Column 3 replicates the analysis in Column 2, investigating the role of beliefs and preferences held by three randomly sampled subordinate workers from the line. We do not find the same type of relationship between these and the performance of female trainees during the trial. Note, however, that while we could elicit bias from essentially all co-supervisors on the line, we could do so for only three out of as many as 40 to 60 subordinate workers per line, giving us a noisier measure of beliefs and biases for the subordinates.

Table A.5 in the Appendix provides evidence that random exposure to female supervisors during the trial might have reduced negative attitudes towards female supervisors, potentially contributing to closing the performance gap documented in Fact 3. We surveyed subordinates and co-supervisors after the end of the trial and asked respondents to evaluate the trainees working on their line during the trial on a scale of 1-10, as well as whether they prefer to work with a female (coded as 1), or a male supervisor (coded as -1), or whether they are indifferent (coded as 0). Column 1 shows that both female and male workers evaluate the randomly allocated female trainees worse than the male trainees. Column 2, however, finds that male subordinates who were randomly allocated to a female trainee report better attitudes toward female supervisors than male subordinates randomly allocated to a male trainee. While male workers from lines allocated male trainees score a highly negative preference of -0.716 for female supervisors, those from lines allocated female trainees score a significantly less negative -0.264. The positive update in attitudes despite the worse average performance and evaluations of female

Table 5: Assumption 1 - Bias and Efficiency of Female Trainees during Trial

VARIABLES	(1) ITT-Trial Co-Superv. Beliefs	(2) ITT-Trial Co-Superv. Beliefs	(3) ITT-Trial Worker's Beliefs
Female Trainee	0.098	0.720	-7.788
	(2.565)	(2.979)	(4.782)
Basel. Beliefs	-6.321*	-7.474*	12.485*
	(3.502)	(4.197)	(7.024)
Basel. Beliefs x Fem. Trainee	10.793*	12.669**	-6.225
	(5.442)	(6.319)	(9.539)
Basel. Preference		3.255	1.458
		(3.532)	(3.397)
Basel. Preference x Fem. Trainee		0.545	-2.178
		(3.756)	(4.217)
Factory FE	Yes	Yes	Yes
Ancova	Yes	Yes	Yes
Missing Attitudes Ind.	Yes	Yes	Yes
N	93	93	93

Notes: Column 1 replicates Column 1 from Table 2, adding an interaction between the female trainee dummy with co-supervisors' average beliefs about female supervisors from Figure 1. Column 2 further adds an interaction between the female trainee dummy and co-supervisors' stated preferences for male vs female supervisors. Column 3 replicates Column 2 with beliefs and preferences elicited from three randomly sampled subordinates from the line. Regressions control for factory fixed effects, baseline line efficiency, and missing data indicators. Robust Standard Errors: *** denotes stat. sign. at 1%, ** at 5%, and * at 10%.

trainees is consistent with beliefs being too negative at baseline.[24]

Assumption 2: The decision-maker is (a) non-discriminatory and (b) maximizes the expected discounted output on the line.

To focus on the implications of the experimentation logic and the pivotal role played by belief updating, the model assumes that the decision-maker does not have an intrinsic preference against appointing female supervisors. Note that the decision-maker in the model can still have incorrect prior beliefs about female supervisors – perhaps due to previous attempted experimentation or forms of statistical discrimination – which are consistent with the evidence above.

Table A.6 provides direct support for the first part of Assumption 2. We conducted

[24]Columns 3-4 of Table A.5 detect no statistically significant differences in the preferences between male and female trainees among co-supervisors. Given the small number of female supervisors in the sample, we do not separate male and female respondents.

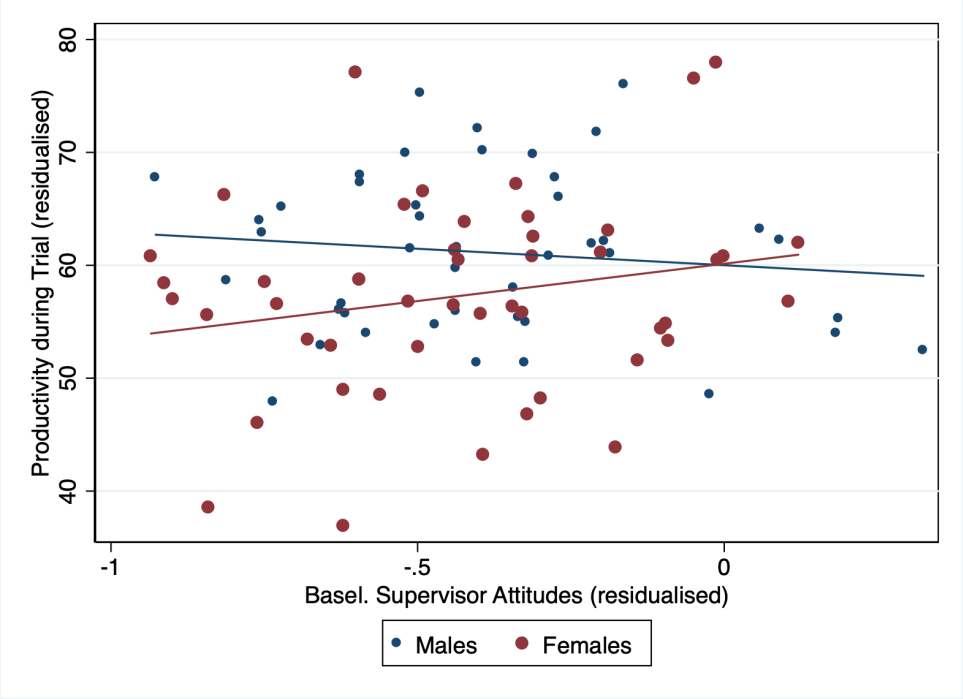

Figure 3: Assumption 1 - Beliefs among co-supervisors and Productivity of Trainees during Trial. Figure plots the productivity during the trial period of lines randomly selected to receive male trainees (blue dots) or female trainees (red dots) against the average beliefs among the baseline team of line supervisors on the line before the arrival of the trainee. Beliefs are an average over the eight skills coded as in Figure 1. Averages were first taken for each supervisor, with the mean then taken over these averages for each supervisor on the line at baseline.

implicit association tests (IATs) with the HR managers in the factories that participated in the study. HR managers are formally responsible for promotion decisions, often based on recommendations provided by production managers and/or line chiefs, and we thus treat them as representing the decision-maker in the model.[25] Column 1 finds no evidence of gender bias among these decision-makers. While the evidence should be interpreted cautiously due to the limited number of observations, we note that Figure 1 found that the HR managers also view females as less suitable supervisors than men, albeit to a somewhat lesser extent than other workers in the factory. This is consistent with our assumption that the decision-maker has no intrinsic preference against female supervisors, but might have relatively pessimistic prior beliefs ρ_0 about them.[26]

While the evidence supports the Assumption, it is worth noting that DM's taste-based discrimination would strengthen our findings. To see why, consider a version of our model

[25]HR managers are the only non-production managers we surveyed. For logistical reasons we refrained from surveying CEOs, financial or marketing managers from our partner factories.

[26]Factories in which managers had an intrinsic distaste towards female supervisors might not participate in the project. To the extent that they participated, they could potentially sabotage the intervention. As we discuss below, this would likely work against our results.

in which the DM has a distaste $\mathcal{D} > 0$ against promoting women. An experimentation threshold can still be derived, along the lines of equation (2). It is then straightforward to show that Facts 1 and 2 go qualitatively unchanged. If anything, $\mathcal{D} > 0$ suggests that marginal women have to be better, which is against the evidence. Facts 3 and 4 are also unchanged, but fewer women would be retained/appointed both in the experiment and in the long-run, which implies our results underestimate the extent of misallocation.

We also conducted implicit association tests (IATs) with the direct bosses (production managers and line chiefs) and with line supervisors. Columns 2 and 3 find evidence of a large, and statistically significant, bias against women among these middle-level managers. This evidence further reinforces Assumption 1 that bias against female supervisors potentially lowers their performance.

Table A.6 also provides direct support for the second part of Assumption 2. Using data from Menzel and Woodruff (2021), which covers more than 1,500 supervisors from 33 factories, including some from this project, Columns 4 and 5 find no significant differences in the base wage and overall pay to male and female supervisors. This is consistent with Menzel and Woodruff (2021)'s finding of generally small gender gaps in wages for workers with the same job title in this setting. This reinforces the importance of studying gender gaps in promotion – as we do in this project – as well as the framework's focus on productivity differences between male and female supervisors, without considering potentially different wage costs between them.

Assumption 3: There is no cost of demoting the trainee after the trial and appointing a new supervisor.

Assumption 3 concerns the absence of costs of demoting trainees, particularly female trainees, after they have been trialed. While the assumption could be relaxed without altering the qualitative insights of the model – in the presence of significant demotion costs, the DM would be even more cautious in experimenting – the assumption is important to interpret the *short-run* results of the trial. In particular, the concern is that the high retention rates of female trainees after the trial (Fact 3) could be due to the cost of demoting a worker once she has been trialed as a line supervisor. Unfortunately, we do not have direct information on the costs incurred by factories to demote trialed trainees. Note, however, that demotion costs can not explain Fact 4 – the higher share of women

among supervisors newly appointed after the end of our intervention.

Table A.7 reports the number of trialed and retained female trainees across the three experiments – the large pilot (see Appendix D for more details), the main trial discussed in this paper, and the extension project discussed above.[27] Across the three projects, factories promoted a consistent share of female trainees from among those they trialed: 53% in the pilot, and 59% in both the main trial and the extension project. In all three experiments, a significant share of female trainees was thus first tried on the line as supervisor, and then subsequently demoted, regardless of whether they had been told, or promised, that they would be trialed after the training. The fact that almost half of the trainees were not retained also suggests that factories can demote trialed trainees if they need to. While this evidence doesn't prove the absence of (differential) demotion costs, it suggests that demotion costs are not so high as to significantly distort factories' promotion choices.

Assumption 4: Factories in the experiment selected as trainees the marginal male and female candidates.

Assumption 4 concerns the selection of trainees for the trial and states that the factories considered the best male and female candidates available. The assumption could be relaxed without altering the qualitative insights from the experimentation model. The assumption, however, facilitates the interpretation of the results and the quantification of losses explored in the next section.

In practice, it was not feasible to conduct the diagnostic on all workers that the factories could have potentially selected and demonstrate that male and female trainees were indeed the best available candidates among their gender for promotion. Two pieces of evidence, however, suggest that the assumption might be apt for our study.

First, trainees are positively selected relative to the typical workers in the factories. Table A.8 confirms this to be the case by comparing trainees to randomly sampled workers from the baseline surveys. Like existing supervisors, both female and male trainees have significantly higher educational attainment than workers despite being no older. Compared with the average male worker, male trainees have longer tenure in the factory

[27]Unlike the main trial, the pilot did not require factories to trial trainees as supervisors; while the extension project required a shorter trial during the training.

but not the sector; while no such difference is observed for female trainees.[28] Trainees are also much more likely to report being willing to accept an offer of promotion to supervisor. Meanwhile, compared to existing supervisors at baseline, both male and female trainees have fewer years of schooling, in line with them representing the marginal supervisors. However, we do not detect differences between trainees and existing supervisors in technical knowledge, numeracy, literacy, or non-verbal reasoning (Raven tests), based on data we collected from existing supervisors assigned as mentors for the trainees, who were surveyed when they came to the training center.

Second, the post-trial performance of retained trainees is comparable to that of supervisors who were promoted just before our intervention. If factories selected the best available candidates, retained trainees should perform similarly to recent non-trainee promoted supervisors. While we cannot identify the most recently appointed supervisors in the factories that participated in the main experiment, we can do so in the factories that participated in the large pilot. Given that trainees in the pilot experiment were not trialed on randomly allocated lines, we exploit the panel nature of the data and estimate the change in line efficiency associated with the arrival of new supervisors within a DID framework that controls for production line and factory-month fixed effects. Table A.9 shows that the most recently promoted supervisor is associated with a mild decrease in the efficiency of the production line he is assigned to. None of the estimates associated with the arrival of a trainee are statistically significant, suggesting that retained trainees perform similarly to supervisors recently promoted before our intervention.

The similar performance of male trainees to that of recently promoted candidates outside the experiment suggests that, consistent with Assumption 4, the male trainees were likely to be the marginal candidates. This, of course, leaves open the possibility that the female trainees weren't the marginal ones, i.e., that *better* female candidates existed. If that were the case, however, it would be even harder to find evidence consistent with the under-promotion of female supervisors at baseline: either Assumption 4 is satisfied, or, if it is violated, then the share of female trainees promoted following the trial would be a lower bound for the share promoted under optimal selection. Maintaining Assumption 4 as a conservative benchmark, we next attempt to quantify the resulting efficiency losses.

[28] Among trainees, the strongest predictors of promotion to supervisor after the trial are the education level for female trainees and the tenure in the factory for male trainees. Thus trainees are positively selected particularly along those dimensions that matter for the subsequent promotion decision.

5 Female Supervisors and Factory Performance

Our intervention nudged factories to experiment with female supervisors, leading to the promotion of many women who participated in the program and to a sustained increase in the share of women among newly appointed supervisors after its end. Given the misalignment between the initial perceptions about women's skills as supervisors and the reality of their skills and relative performance during and after the trial, our framework suggests there may have been an inefficiently low share of female supervisors at baseline, relative to a benchmark with (more) accurate beliefs. This section provides a tentative, back-of-the-envelope, quantification of the losses incurred by factories due to the wrong beliefs. Our calculation focuses on the efficiency of production lines, leveraging estimates from the experimental intervention. Further below, we show that the share of female supervisors correlates with other desirable factory-level outcomes. While such evidence isn't sufficient to prove a causal link, it is consistent with a higher share of female supervisors potentially benefiting factories through other channels not captured by our intervention.

5.1 Quantification of output loss

Figure 4 reports the estimated change in line efficiency associated with the arrival of each trainee as a co-supervisor on the line, as well as illustrating our approach. In particular, the figure reports the estimated difference between the pre-trial average efficiency and the efficiency during the trial. The estimated changes for male trainees (in blue) are sorted from the smallest to the largest, while those of the female trainees (in red) are sorted in the opposite direction. Based on these estimates, the allocation that maximizes efficiency equalizes the estimated effect of the marginal male and female supervisors, given by the point at which the two curves cross. Efficiency in the optimal allocation is then given by the area underneath the envelope of the two curves: the sum of the estimated changes for female candidates to the left of the crossing point and the sum of the estimated changes for males to the right.[29]

Figure 4 shows that the optimal allocation is achieved with a share of female candidates of around 40 percent, remarkably close to the share of female trainees among all

[29]We focus on the trial period to leverage the random allocation of trainees to lines. However, recall that female trainees under-performed during the trial (Fact 2) but then closed the gap after the trial, conditional on retention (Fact 3). Focusing on the trial period thus provides a lower bound for the losses.

retained trainees after the trial, which is 42 percent.[30] At baseline, however, only around six percent of supervisors are female. Under the conservative assumption that – conditional on that share – factories select male and female supervisors to maximize efficiency, efficiency at baseline is represented by the estimated changes for female candidates to the left, and for male candidates to the right, of the corresponding vertical line.

The two horizontal lines in Figure 4 indicate the average production line efficiency resulting from the optimal share of female supervisors and the baseline share. This back-of-the-envelope calculation suggests that the efficiency of the baseline allocation is 2.44 units lower than the one resulting from the optimal allocation. Given an average efficiency of 60 units in the sample, this corresponds to a ca. four percent lower labor efficiency due to the initial misallocation. Based on Cajal-Grossi et al. (2023) estimates of the labor share and profit margins in the industry, this corresponds to eight percent lower profits, comparable to losses incurred by firms that do not offer the soft skills training studied by Adhvaryu et al. (2023) or adopt the new technology introduced by Atkin et al. (2017).

5.2 Further potential benefits of female supervisors

The appointment of qualified women as line supervisors can yield factory-level benefits beyond the efficiency gains on the production lines identified by our experiment. We provide suggestive, factory-level, evidence that a higher share of female supervisors positively correlates with other performance indicators. As in Section 4.4, we use survey data from Cajal Grossi et al. (2022) to measure the share of female supervisors in a sample of 290 factories in the industry that participated in the IFC-ILO Better Work program. Given the similar recruitment strategy between our intervention and the Better Work program, this sample of factories is comparable to – and, in fact, overlaps with – the sample of factories in our interventions.

Figure 5 illustrates the correlation between the share of female supervisors on the horizontal axis and three factory-level outcomes on the vertical axis, after controlling for factory size and year fixed-effects. Panel (a) shows that factories with a higher share of female supervisors supply "better" buyers. Merging detailed production and custom records from the industry, Cajal-Grossi et al. (2023) show that the mix of buyers a factory supplies to is an important dimension of upgrading. In particular, international buyers differ in the unit values and margins paid to suppliers for identical garments. Some buyers,

[30] This evidence is also consistent with an unbiased DM (Assumption 2a).

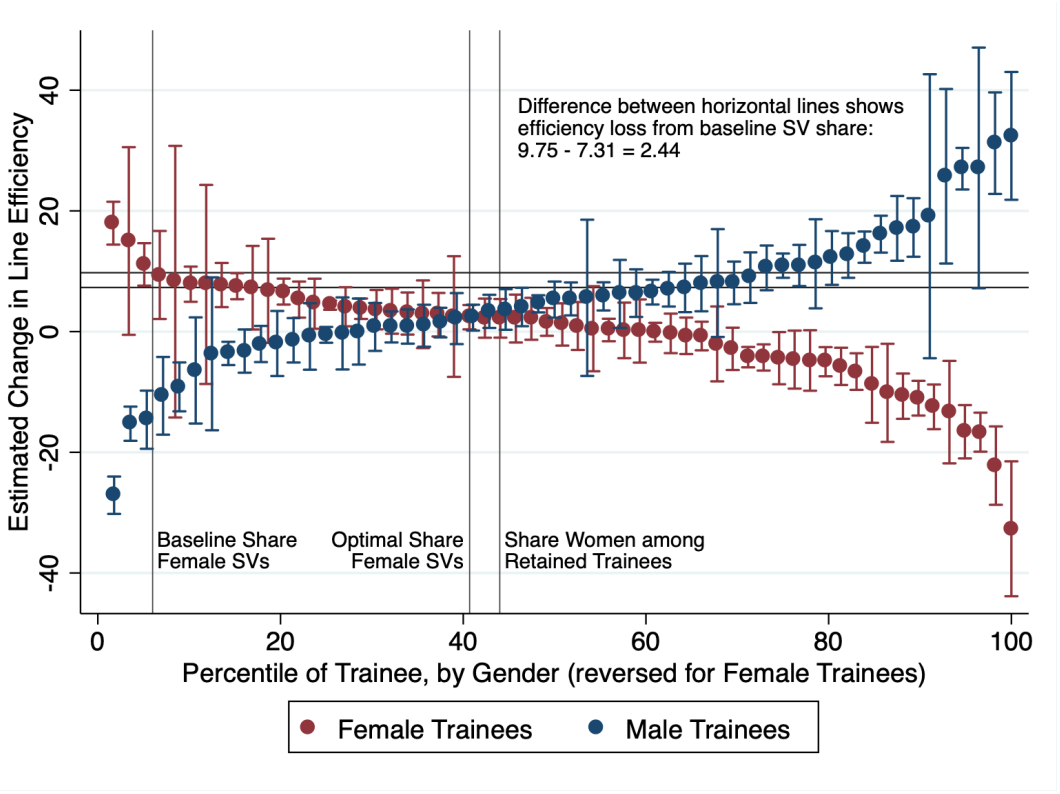

Figure 4: Quantification of Efficiency Loss. The Figure shows the effects of the arrival of individual trainees as supervisors on a line on the line's productivity, based on day-line panel data regressions with line and factory-month fixed effects. 95 percent confidence intervals for all trainee effects are shown as well. Effects sorted in ascending order from left to right for male trainees, and in descending order for female trainees. X-axis shows percentiles of effects for trainees of a given gender.

particularly those that adopt relational sourcing practices in their global supply chains, systematically pay higher unit values to suppliers. Following Cajal-Grossi et al. (2023), we estimate buyers' fixed-effects on unit values and compute for each factory an index of the quality of its buyers taking the weighted average of the estimated fixed effects for the buyers the factory supplies to. There is a positive and statistically significant ($p-value < 0.001$) correlation between this index of buyer quality and the share of female supervisors in the factory. Different mechanisms could account for this correlation. One possibility is that a higher share of female supervisors correlates with factories' capabilities to deliver on dimensions better buyers are willing to pay for, e.g., quality or reliability. Better buyers might also push factories to increase the share of female supervisors. The evidence, however, is also consistent with the possibility that factories that aim to upgrade their buyer mix may find themselves wanting to increase the share of female supervisors.

Panel (b) in Figure 5 documents a positive correlation between the share of female su-

pervisors and an index of occupational health and safety (OHS) in the factory, constructed by aggregating workers' answers across several questions (see Appendix in Boudreau et al. (2024) for details). Line supervisors can influence OHS, for example, by ensuring that workplaces are kept orderly and safe, by explaining basic safety measures to workers, and by communicating workers' concerns – including on OHS – to management. The correlation is thus suggestive that female supervisors, who are more representative of the pool of workers, might facilitate that process. Like the pattern in Panel (a), this correlation could also simply indicate that factories that are better managed in general also have a higher share of female supervisors. Panel (c), however, finds no correlation between the share of female supervisors and processes in place to reliably pay salaries to workers – a proxy for the quality of management practices in the factory. Unlike OHS conditions, this proxy is determined by managers higher up the hierarchy, with limited involvement of line supervisors. If the share of female supervisors was driven by omitted factors – e.g., better management practices or attitudes toward more positive industrial relations – we may expect it to positively correlate with both OHS and the pay score, rather than only with the dimension influenced by line supervisors.

Taken together, these correlations suggest that a higher share of female supervisors might benefit factories through channels that are not captured by our experimental design. If so, the quantification in the previous sub-section may provide a lower bound to the losses factories incur due to their wrong baseline beliefs about female supervisors.

6 Discussion and Concluding Remarks

Women account for the vast majority of workers in the Bangladeshi garment sector but only for a negligible share of line supervisors – the lowest rung of the management ladder in the industry. A combination of experimental evidence, factories' internal records, workers' surveys, and a conceptual framework blending *discrimination* and *experimentation*, reveals that incorrect beliefs about women's ability as line supervisors hinder their promotion to managerial roles.

We implemented an intervention that induced factories to experiment with trialing women (and men) as supervisors. In contrast to widespread negative beliefs about women's ability as supervisors at baseline, we found that (1) female candidates selected by the factories had similar skills to males, (2) during the trial, females performed worse

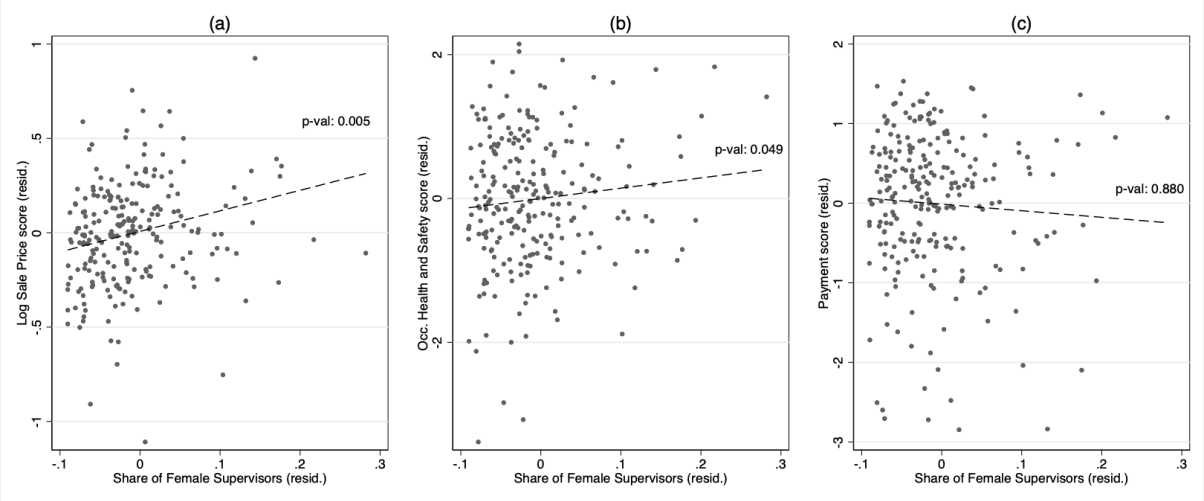

Figure 5: Female Supervisors and Factory Characteristics. Panel (a) plots the average price per kilogram paid by the average buyer of a factory (from custom export records) against the share of female supervisors in factory (from ILO survey data). Panel (b) plots the values of an Occupational Health and Safety (OSH) score of factories against their share of female supervisors, while Panel (c) plots the values of a score summarising the reliability of wage payment practices against the share. Both the OSH and the pay score are z-scores based on variables collected in the same ILO survey, with the variables listed in Appendix of Boudreau et al. (2024). All variables residualised against number of workers in the factory and survey year. If a factory was surveyed in multiple years, the first year is used to construct the graphs.

than males due, to a large extent, to bias among co-workers against them, (3) after the trial, however, a significant share of female candidates were retained as supervisors and, conditional on that, performed similarly to retained males; and (4) after the end of our intervention, factories permanently increased the share of women among newly appointed supervisors. Our framework rationalizes both the initial underperformance and the long-run results. Uncertainty over female candidates' abilities implies that experimenting with a female candidate generates valuable information that can be used to make better future appointments. An equal or worse initial performance of the marginal candidates of the discriminated group is not sufficient to rule out discrimination.

Based on the experimental estimates, a back-of-the-envelope calculation suggests that the resulting misallocation of managerial talent leads to non-negligible losses – approximately eight percent of profits. We also provide suggestive evidence that a higher share of female supervisors may be associated with other benefits not captured by our experiment. This begs the question of why didn't factories run this experiment themselves. We are, of course, not alone in finding positive impacts from an intervention that firms had not adopted on their own (see, e.g., Bloom et al. (2012), Sandvik et al. (2020), Adhvaryu et al.

(2023)). The literature discusses barriers to *information*, *incentives*, and *implementation* (Gibbons and Henderson (2012)), and, accordingly, we offer a few considerations. First, even a rational decision-maker will not experiment if prior beliefs are sufficiently pessimistic: incorrect beliefs – a type of *information* barrier – played a role. Some factories, however, had experimented with female supervisors before our intervention. Even then, a factory replicating our trial in isolation would find it difficult to learn due to a small sample size. The average factory in our trial experimented with three female trainees and at most with seven. In contrast, pooling observations across 24 factories yields higher statistical power and confidence in the results: experimentation is an individually costly public good that tends to be under-provided. Second, intense competitive pressures in the industry might lower *incentives* to experiment. Concomitantly with our intervention, Macchiavello et al. (2015) offered a training program for existing line supervisors – an intervention in which factories expressed a keen interest. Despite the initial interest, only 6% of 135 factories took up a free slot to train existing supervisors. Follow-up phone interviews with firms that had expressed interest but ultimately turned down the offer reveal that production pressures and the inability to spare workers were the main reasons behind the lack of participation. Such pressures likely hamper the more complex experimentation with female supervisors.[31] Finally, appointing female supervisors requires changing factory practices and norms around promotion and might pose significant *implementation* challenges. For example, qualified men who work in the factories expecting that they will not compete with women for promotion might resist the change, or quit. Even an unbiased DM might find it difficult to establish an unbiased workplace, and might be uncertain about the transition costs involved.

This study identifies incorrect beliefs as a particular organizational barrier to women's promotion. There might be other reasons for the under-promotion of women even without differences in talent, discrimination, or prejudiced beliefs. For example, there could be gender-specific preferences over job characteristics (Haegele, 2024). Focusing on marginal candidates we found no gender differences in preferences for a promotion. However, many women approached by factories turned down the offer to participate, suggesting that such

[31]Buyers' involvement might foster factories' incentives to trial female supervisors. For the pilot project, we contacted approximately 200 factories to enroll 90 (of which almost half subsequently dropped out) without the help of a buyer. The main trial and the extension trial – introduced by two different buyers in their supplier base – had 100% participation.

differences in preferences are also a barrier to promotion.[32] While outside the scope of this paper, understanding how differences in preferences hinder the advancement of women is extremely important. If factories increase the share of female supervisors substantially, gender-specific preferences might become a constraint. As more women become line supervisors and demonstrate that they can perform well and be respected in this role, however, preferences might also evolve, relaxing that constraint. Being a supervisor is one step in a longer career and there could thus also be dynamic considerations. Factories may use the supervisor position to test for promotion to the next level where there could be gender differences in performance, preferences, and so on. There were no women among higher-level managers in our factories, so we can't investigate this source of bias. However, the long-run evidence suggests that such dynamic considerations could not be the only reason why equally talented women were not promoted at baseline.

Finally, our results are also relevant from a gender equity perspective. A wide literature has shown that empowering women in decision-making carries positive effects, for example, on child welfare and health (Duflo, 2012, 2003; Miller, 2008). Uckat (2020) shows that women selected for the trainee role in the above-mentioned extension project had a larger bargaining power within their household, resulting, among other aspects, in a larger share of household budgets spent on goods for women (clothing, accessories). Female workers working under the newly appointed female supervisors featured similar effects. If increased responsibilities for women on the job are not compensated by reduced household duties, complementary policies may be needed to increase the effectiveness of policies targeted at promoting career advancement for women (McKelway, 2021). Factories with more female supervisors also have better health and safety conditions, consistent with the idea that empowering women might change factories' human-resources practices and priorities in ways beneficial to workers (Boudreau et al. (2023)).

References

Adhvaryu, A., N. Kala, and A. Nyshadham (2023). Returns to on-the-job soft skills training. *Journal of Political Economy 131*(8), 2165–2208.

Adhvaryu, A., A. Nyshadham, and J. Tamayo (2023). Managerial quality and productivity dynamics. *The Review of Economic Studies 90*(4), 1569–1607.

[32] In our framework, a very capable female trainee who doesn't want to become a line supervisor would not be considered a misallocation of talent.

Ahern, K. R. and A. K. Dittmar (2012). The Changing of the Boards: The Impact on Firm Valuation of mandated Female Board Representation. *Quarterly Journal of Economics 127*(1), 137–197.

Aigner, D. J. and G. G. Cain (1977). Statistical theories of discrimination in labor markets. *Ilr Review 30*(2), 175–187.

Ashraf, N., O. Bandiera, V. Minni, and V. Quintas-Martínez (2023). Gender and the misallocation of labor across countries. *Work in progress, June*.

Atkin, D. (2016). Endogenous Skill Acquisition and Export Manufacturing in Mexico. *American Economic Review 106*(8), 2046–85.

Atkin, D., A. Chaudhry, S. Chaudry, A. Khandelwal, and E. Verhoogen (2017). Organizational Barriers to Technology Adoption: Evidence from Soccer-Ball Producers in Pakistan. *Quarterly Journal of Economics 132*(3), 1101–1164.

Banerjee, A., D. Karlan, and J. Zinman (2015, January). Six randomized evaluations of microcredit: Introduction and further steps. *American Economic Journal: Applied Economics 7*(1), 1–21.

Bardhi, A., Y. Guo, and B. Strulovici (2024). Early-career discrimination: Spiraling or self-correcting? *mimeo*.

Beaman, L., R. Chattopadhyay, E. Duflo, R. Pande, and P. Topalova (2009). Powerful Women: Does Exposure Reduce Bias? *Quarterly Journal of Economics 124*(4), 1497–1540.

Belloni, A., V. Chernozhukov, and Y. Wei (2016). Post-Selection Inference for Generalized Linear Models with Many Controls. *Journal of Business & Economic Statistics 34*(4), 606–619.

Benson, A., D. Li, and K. Shue (2025). Potential" and the gender promotion gap. *American Economic Review*.

Bertrand, M. (2017). The Glass Ceiling. Becker Friedman Institute for Research in Economics Working Paper No. 2018-38.

Bertrand, M., S. E. Black, S. Jensen, and A. Lleras-Muney (2019). Breaking the Glass Ceiling? The Effect of Board Quotas on Female Labour Market Outcomes in Norway. *Review of Economic Studies 86*(1), 191–239.

Blau, F. and L. Khan (2017). The Gender Wage Gap: Extent, Trends and Explanations. *Journal of Economic Literature 55*(3), 789–865.

Bloom, N., B. Eifert, A. Mahajan, D. McKenzie, and J. Roberts (2012). Does Management Matter? Evidence from India. *Quarterly Journal of Economics 128*(1), 1–51.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2025). Inaccurate statistical discrimination: An identification problem. *Review of Economics and Statistics 107*(3), 605–620.

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

Boudreau, L. (2024). Multinational Enforcement of Labor Law: Experimental Evidence from Bangladesh's apparel sector.

Boudreau, L., J. Cajal-Grossi, C. Can, and R. Macchiavello (2024). Relationships and responsibility. *mimeo*.

Boudreau, L., J. Cajal-Grossi, and R. Macchiavello (2023). Global value chains in developing countries: a relational perspective from coffee and garments. *Journal of Economic Perspectives 37*(3), 59–86.

Cajal Grossi, J., R. Macchiavello, and C. K. Can (2022). Is better work better?: evidence from the garment sector in bangladesh.

Cajal-Grossi, J., R. Macchiavello, and G. Noguera (2023). Buyers' sourcing strategies and suppliers' markups in bangladeshi garments. *The Quarterly Journal of Economics*, 2391–2450.

Canay, I. A., M. Mogstad, and J. Mountjoy (2024). On the use of outcome tests for detecting bias in decision making. *Review of Economic Studies 91*(4), 2135–2167.

Duflo, E. (2003). Grandmothers and granddaughters: Old-age pensions and intrahousehold allocation in south africa. *The World Bank Economic Review 17*(1), 1–25.

Duflo, E. (2012). Women Empowerment and Economics Development. *Journal of Economic Literature 50*(4), 1051–1079.

Flabbi, L., M. Macis, A. Moro, and F. Schivardi (2019). Do Female Executives Make a Difference? The Impact of Female Leadership on Firm Performance and Gender Gaps. *Economic Journal 129*(622), 2390–2423.

Gereffi, G. (1999). International Trade and Industrial Upgrading in the Apparel Commodity Chain. *Journal of International Economics 48*(1), 37–70.

Gibbons, R. and R. Henderson (2012). Relational Contracts and Organizational Capabilities. *Organization Science 23*(5), 1350–1364.

Goldin, C. (2014). A Grand Gender Convergence: Its Last Chapter. *American Economic Review 104*(4), 1091–1119.

Goldin, C., L. F. Katz, and I. Kuziemko (2006). The homecoming of american college women: The reversal of the college gender gap. *Journal of Economic perspectives 20*(4), 133–156.

Gollin, D. (2008). Nobody's business but my own: Self-employment and small enterprise in economic development. *Journal of Monetary Economics 55*(2), 219–233.

Haegele, I. (2024). The broken rung: Gender and the leadership gap. *arXiv preprint arXiv:2404.07750*.

Heath, R. and M. Mobarak (2015). Manufacturing growth and the lives of Bangladeshi women. *Journal of Development Economics 115*, 1–15.

IFC and ILO (2025). Creating better jobs for women and boosting productivity in bangladesh's garment factories: An assessment of the gender equality and returns program. Technical report, IFC / ILO.

Komiyama, J. and S. Noda (2024). On statistical discrimination as a failure of social learning: A multiarmed bandit approach. *Management Science*.

Li, D., L. R. Raymond, and P. Bergman (2024). Hiring as exploration. *Review of Economic Studies*.

Macchiavello, R., A. Rabani, and C. Woodruff (2015, May). The Market for Training Services: A Demand Experiment with Bangladeshi Garment Factories. *American Economic Review, Papers & Proceedings*.

Matsa, D. A. and A. R. Miller (2013). A Female Style in Corporate Leadership? Evidence from Quotas. *AEJ: Applied Economics 5*(3), 136–69.

McKelway, M. (2021). How Does Women's Employment Affect Household Decision-Making? Experimental Evidence from India. Working Paper, Dartmouth College.

McKinsey (2011). Bangladesh's ready made garments landscape: The challenge of growth. McKinsey&Company, Apparel, Fashion & Luxury Practice.

Menzel, A. and C. Woodruff (2021). Gender wage gaps and worker mobility: Evidence from the garment sector in bangladesh. *Labour Economics 71*, 102000.

Miller, G. (2008, 08). Women's Suffrage, Political Responsiveness and Child Survival in American History. *The Quarterly Journal of Economics 123*(3), 1287–1327.

Olivetti, C. and B. Petrongolo (2016). The Evolution of Gender Gaps in Industrialized Countries. *Annual Review of Economics 8*, 405–434.

Onuchic, P. (2024). Recent contributions to theories of discrimination. *mimeo*.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The american economic review 62*(4), 659–661.

Sandvik, J. J., R. E. Saouma, N. T. Seegert, and C. T. Stanton (2020, 04). Workplace knowledge flows*. *The Quarterly Journal of Economics 135*(3), 1635–1680.

Uckat, H. (2020). Womens promotion and intra-household bargaining: Evidence from bangladesh. Technical report, Working paper, Oxford University.

Uckat, H. and C. Woodruff (2020). Learning What to Look For: Hard Measures on Soft Skills in Promotion. Working paper, Oxford University.

Verhoogen, E. (2023, December). Firm-level upgrading in developing countries. *Journal of Economic Literature 61*(4), 1410–64.

Online Appendices:

A Further Results

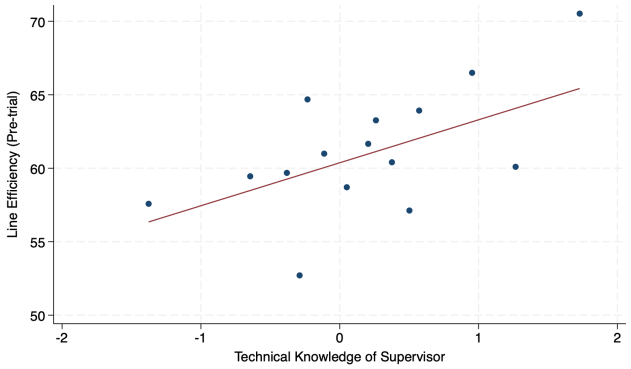

Figure A.1: Technical Knowledge of Baseline Supervisors and Efficiency of their Line. Binscatter based on data from 70 mentor supervisors who were invited to the training center, where they conducted the same technical knowledge test that also trainees took at the beginning of training phase (see Table 1). Efficiency is averaged over 60 days before arrival of trainees (50 days before test taken until 10 days after). Highest and lowest outlier observations from outcome variable omitted from sample. Robust p-value of fitted line 0.058.

Table A.1: Comparing Nominated to other Lines

	Non-Nominated	Nominated Lines	SE	N
SMV	11.31	0.285	(0.376)	314
Order Quantity	57477	-326	(6394)	320
Running Days	6.877	-0.203	(0.449)	477
Hourly Target	152.1	-1.613	(3.891)	439
Total Target	1479	-18.15	(34.85)	459
Daily Hours	9.570	0.059	(0.048)	477
Nbr Operators	33.47	0.318	(0.801)	464
Nbr Helpers	13.78	-0.036	(0.521)	458
Buyer Size	0.181	0.017*	(0.009)	399
Order Size	0.249	0.000	(0.009)	528
Sh.Fem.Worker (HR Data,7 Fact)	0.781	-0.012	(0.016)	244
Fem. SV (HR Data, 3 Fact)	0.049	0.040	(0.071)	75
Fem. SV (Survey Data)	0.041	0.008	(0.037)	184

Notes: Table shows average characteristics of production lines not nominated by the factories for the trial (Column 1), and the difference in average characteristics to the nominated lines (Column 2), including standard errors of the difference (Column 3). The last Column shows the number of lines for which the characteristic was available. * indicates a p-value of the difference of below 0.1.

Table A.2: Balance, Lines allocated Male vs Female Trainees

	Female	S E	Mean Males	N
Line Operators:				
Gender	-0.019	(0.042)	0.732	142
Age	-0.264	(0.398)	24.05	142
Married	0.026	(0.043)	0.760	142
Years Education	-0.375	(0.262)	5.967	142
Grade	0.189	(0.135)	3.748	115
Months Factory	4.344	(3.183)	30.91	142
Months Industry	2.932	(4.031)	71.27	142
Months Design	1.397	(3.762)	57.95	142
Months Line	0.080	(2.354)	16.37	142
Nbr Factories	-0.120	(0.152)	1.880	142
Previous Fem SV	-0.009	(0.055)	0.441	142
Prefer Fem SV	0.147*	(0.074)	-0.417	142
Accept Promotion	-0.084	(0.052)	0.5	139
F-Test:		0.134		
Line Supervisors:				
Gender	0.025	(0.026)	0.022	136
Age	0.595	(0.761)	28.72	136
Married	0.050	(0.049)	0.843	136
Years Education	0.196	(0.278)	9.276	135
Months Factory	5.655	(4.958)	40.48	136
Months Industry	5.725	(6.167)	102.4	136
Months Design	2.640	(4.761)	51.02	136
Months Line	2.383	(3.314)	21.21	136
Nbr Factories	0.340	(0.330)	2.701	136
Previous Fem SV	-0.100	(0.066)	0.327	136
Prefer Fem SV	0.045	(0.086)	-0.722	136
Spouse Works	0.097	(0.076)	0.327	128
F-Test:		0.105		
Production Data:				
SMV	0.430	(0.696)	12.08	75
Order Quantity	18896*	(9779)	50415	76
Running Days	1.187*	(0.642)	6.129	111
Total Target	-10.19	(71.77)	1519	104
Daily Hours	0.052	(0.080)	9.607	107
Nbr Operators	0.501	(1.470)	36.43	109
Nbr Helpers	0.544	(0.959)	16.59	109
Efficiency	0.020	(0.017)	0.590	98
Absenteeism	0.000	(0.004)	0.043	76
Defects Rate	0.001	(0.006)	0.061	118
Spot Rate	0.018	(0.012)	0.032	75
Reject Rate	0.000	(0.000)	0.004	90
Buyer Size	0.037*	(0.020)	0.283	99
Style Size	0.012*	(0.006)	0.053	111
F-Test:		0.77		

Notes: Table compares lines that were randomized into receiving male vs. female trainees on a) average baseline observables from three randomly selected operators per line (upper panel), b) average baseline observables from two supervisors per line (middle panel), and c) on variables from the administrative production data (lower panel). Tests control for factory fixed effects. *** denotes statistical significance at 1%, ** at 5%, and * at 10%. "F-Test" show p-values for joint significance of all variables in each of the three panels when regressing indicator variable of receiving female trainee on them, with factory fixed effects. For production data, this regression also controls for indicator variables for each these variables indicating missing values in them (with missing value set to 0 in the main variables).

Table A.3: Main Trial Effects on Other Administrative Outcome Variables

	(1) Defects	(2) Defects	(3) Defects Basel.	(4) Absence R.	(5) Absence R.	(6) Absence R. Basel.
VARIABLES	ITT			ITT		
Female Trainee	0.181	1.204**	-0.366	-0.410	-0.487	-0.062
	(0.570)	(0.577)	(1.209)	(0.673)	(0.540)	(0.586)
Factory FE	Yes	Yes	Yes	Yes	Yes	Yes
Ancova	-	Yes	-	-	Yes	-
PDS Controls	No	Yes	No	No	Yes	No
Mean Males	6.220	6.220	7.912	6.331	6.331	5.776
N	105	98	83	81	74	63

Notes: Table replicate Columns 1-3 of Table 2 for the two other outcome variables based on administrative production data: Defect rates among produced output (Columns 1-3), and Absenteeism rates of workers on lines (Columns 4-6). Robust standard errors in parentheses: *** denotes statistical significance at 1%, ** at 5%, and * at 10%

Table A.4: Closure of Gender Efficiency Gap

	(1) Baseline: Lines w/ Trainees	(2) Trial: Lines w/ Retained	(3) Trial: Lines w/ Tr. Retained	(4) Post-Trial: Tr. Retained
VARIABLES	post-trial	Trainees	on Same Line	on Same Line
Female Trainee	-0.619	-6.645***	-4.619	-2.938
	(1.614)	(2.193)	(4.071)	(3.193)
Ancova	-	Yes	Yes	Yes
PDS Controls	No	Yes	as col 2	as col 2
Factory FE	Yes	Yes	Yes	Yes
Mean Males	63.48	63.24	65.35	65.81
N	81	78	35	35

Notes: Column 1 shows comparisons of pre-trial average efficiency of lines on which retained trainees work as supervisors after the trial period. Column 2 shows efficiency comparisons of those lines on which *retained* trainees worked during the trial period, while Column 3 replicates Column 2 but further restricts the sample to trainees that continued to work as SVs on the same line. Column 4 compares the trainees from Column 3 during the post-trial period. All columns control for controls selected by PDS Lasso and factory fixed effects. For consistency, Column 4 controls only for those controls and factory fixed effects selected by PDS Lasso for Column 3. Robust standard errors in parantheses; *** denotes statistical significance at 1%, ** at 5%, and * at 10%

Table A.5: Trainee Exposure and Update of Preference for Female SVs

VARIABLES	(1) Evalaluations by Workers ITT	(2) Preferences by Workers ITT	(3) Evalaluations by Co-SVs ITT	(4) Preferences by Co-SVs ITT
Female Tr. x Fem. Resp.	-0.622*	0.004		
	(0.323)	(0.106)		
Female Tr. x Male Resp.	-0.924*	0.446***		
	(0.509)	(0.118)		
Female Trainee			-0.360	0.068
			(0.265)	(0.079)
Fem. Respondent	-0.037	0.498***		
	(0.520)	(0.094)		
Observations	266	396	150	184
PDS Controls	Yes	Yes	Yes	Yes
Factory FE	Yes	Yes	Yes	Yes
Mean Omitted Cat.	6.979	-0.716	6.275	-0.763
Nbr Male Tr.s	46	66	43	43
Nbr Female Tr.s	55	65	51	51

Notes: Table compares evaluations of male versus female trainees, as well as preferences for male versus female supervisors in the future, by subordinate workers (Columns 1-2), and by co-supervisors (Columns 3-4), from the lines on which the trainees did their trial, collected during the follow-up 1 surveys after the end of the trial. All columns show ITT specifications, showing comparisons between lines that were randomly allocated male or female trainees. Preference for female supervisor coded as 1 "Prefer Female SV", 0 "Indifferent", and -1 "Prefer Male SV", while evaluations were asked on a Likert scale from 1-10, with 10 the most positive possible evaluation. Controls selected by PDS Lasso from among respondent and trainee controls. *** denotes statistical significance at 1%, ** at 5%, and * at 10%

Table A.6: Assumption 2 - Non-discriminatory decision-maker.

VARIABLES	(1) IAT HR Managers	(2) IAT Prod Mngs & Line Ch.	(3) IAT Line Superv.	(4) Superv. Gender Wage Gap Base Wage	(5) Superv. Gender Wage Gap Paid Wage
i	-0.042	-0.213***	-0.269***		
	(0.106)	(0.051)	(0.044)		
Female Superv.				-0.016	0.003
				(0.013)	(0.019)
Observations	29	145	223	1,552	1,358
Factory FE	-	-	-	Yes	Yes

Notes: Columns 1-3 show results from Implicit Association Tests (IAT) on bias against female supervisors among employees on different levels of the factory hierarchy, as indicated in the column headers. Column 4 regresses the log of the base wage of line supervisors from administrative data from 33 factories in the sector from Menzel and Woodruff (2021) on an indicator variable for a female line supervisor, while Column 5 does the same for the paid-out wage, which includes overtime pay and subtracts pay for missing days. Columns 4-5 control for factory fixed effects. Standard errors clustered at the factory level: *** denotes stat. significance at 1%, ** at 5%, and * at 10%

Table A.7: Assumption 3 - Demotion of Trainees across three Trials

	N Factories	N Fem. Trainees	Fem. Tr. trialed	% Fem. Tr. promoted
Pilot	58	182	124	54%
Main Trial	24	73	64	59%
Extension Tr.	27	145	98	59%

Notes: Table shows number of female trainees trained, trialed, and the share trialed trainees retained after the end of the trial (i.e. permanently promoted to supervisors, not demoted back to non-supervisor workers) during three trials conducted by the authors in the Bangladeshi garment sector: the pilot trial to the one discussed in this paper (see Appendix D for more details), the main trial discussed in this paper, and the extension project discussed in more detail in Section 4.4.

Table A.8: Comparison of Trainees to other Workers and Supervisors

N	Worker Male 115	Worker Female 313	Trainee Male 72	Trainee Female 73	Superv. Male 223	Superv. Female 10	Tr. vs. Worker (Male)	Tr. vs. Worker (Female)	Tr. vs. Superv. (Male)	Tr. vs. Superv. (Female)
Age	24.3	23.9	24.7	23.7	29.2	27.7	0.39	-0.23	-4.53***	-4.07**
Years Educ.	6.09	5.69	8.48	8.05	9.23	9.39	2.39***	2.35***	-0.75***	-1.34**
Years in Garment Sec.	6.63	5.90	6.44	6.20	8.93	11.7	-0.19	0.29	-2.49***	-5.52***
Years in Factory	2.38	2.92	3.65	3.20	3.48	6.25	1.27***	0.28	0.17	-3.04***
Nbr.Other Factories	2.41	1.65	2.06	2	2.91	1.20	-0.34	0.34	-0.84***	0.80
Accept Prom.someday	0.57	0.41	0.89	0.83			0.32***	0.42***		
Techn. Knowledge			53.5	53.3	55.5	52			-2.03	1.38
Numeracy			4.82	3.38	5.20	2.66			-0.37	0.71
Non-verbal Reason.			3.15	3.01	2.84	2.66			0.31	0.34
Literacy			8.97	7.84	9.78	7			-0.81	0.84

Notes: Table shows comparisons of baseline survey observables of trainees to randomly selected workers of the same sex from lines nominated by factories for the trial, and to line supervisors of the same sex from these lines, including tests for statistical significance of these differences. See Section 4.1 for more details on the variables. *** denotes statistical significance at 1%, ** at 5%, and * at 10%

Table A.9: Assumption 4 - Marginality of Trainees, Pilot Data

VARIABLES	(1) Line Prod.
Most recent promoted LSV	-1.245
	(1.161)
Male Trainee	2.203
	(1.749)
Female Trainee	1.023
	(1.368)
Observations	104,130
Factory-Month FE	Yes
Prod. Line FE	Yes
Avg. Productivity	45.74

Notes: Table regresses daily line productivity from the pilot phase on an indicator variable for a newly promoted line supervisor (LSV) supervising the line on that day, including any trainee from the pilot phase, and for a male or female trainee specifically supervising the line on that day. Regressions control for production line and factory-month fixed effects. Standard errors clustered at the factory level: *** denotes stat. significance at 1%, ** at 5%, and * at 10%

B Training, and its Effects

This appendix describes the training that the trainees attended before the start of the trial in more detail, and shows how some skill measures changed from before to after the training. The training program was designed by the German bilateral aid agency (GIZ) with the aim to provide sewing machine operators the necessary skills to be sewing line supervisors. GIZ's goal in developing the program was to increase the number of women working as supervisors in the sector. The training was viewed by GIZ as important to build skills of female operators, and to encourage factories to experiment to learn whether women were equipped to be supervisors. The training was implemented through a number of private training centres contracted by GIZ with many years of experience in training staff at different levels from Bangladeshi garment factories. All trainees from this project were trained by the same training centre. The training lasts six weeks, with eight-hour sessions held at the classrooms at the training provider's offices on six days per week. The curriculum was divided more or less equally into modules on production planning and technical knowledge, quality control, and leadership and social compliance, and included both class-room sessions as well as instructions directly with sewing machines available in the training centres. The trainees received an allowance to travel daily to the training centre by bus or other public transport, with the distance between the different participating factories and the training centre varying between less than 1 and around 20km.

To understand to which extent the training affected the skills of the trainees, Table B.1 below shows a simple pre-post comparison on a number of supervisor skills of the trainees, for which we already show baseline gender comparisons in Table 1. Given that we neither have a randomly selected, nor any other type of control group to which we could compare time trends for the skills, and given the relative short time-period of the training of six weeks, we show simple pre-post comparisons in these skills. We start in Column 1 with actual technical skills as measured in our 86-question diagnostic test, on which we do not find a significant effect of the training, neither for male nor female trainees. We see the strongest pre-post differences for confidence in own ability: after

the training participants rate their own skills higher (Column 2), and are more likely to state that they consider themselves the best candidate from the workers from their line for promotion to supervisors (Column 3). In both cases the pre-post difference is considerably larger for female trainees, which was already reflected by the smaller post-training differences in confidence shown in Table 1, but the difference to the effect for male trainees is only statistically significant for the first of the two confidence measures. We also see some effects on our measure of communication skills, the number of drawings that trainees can explain within a limited time period to other trainees such that they can successfully draw them (Column 4). The effect is again larger for women, eliminating the small baseline difference that could be observed for this measure, though the difference to the effect on male trainees is again not statistically significant. On the remaining three soft-skill measures, we do not find any effects (Columns 5-7). A tentative conclusion is that the training may have provided a lot of information to the trainees that allowed them to update their beliefs about their supervisor skills relative to existing supervisors, but otherwise did not provide additional skills to the trainees. Even the positive effect on the drawing exercise may be due to the trainees being more familiar and comfortable with the test procedure when going through it a second time after training.

We therefore do not believe that the presence of the training distorts in a first order way the lessons we can draw from the trial for the effects that factories can expect when they promote more women to supervisor positions (without sending them through a comparable training program first). The effects that we see on confidence in particular may imply a "head-start" for our trainees compared to other newly promoted supervisors, which may need a few more days or weeks to reach the same productivity as the trainees achieve in the first days after promotion to supervisors. Meanwhile, the mild evidence for the effect of the training being larger for female trainees implies that our male-female comparisons may be somewhat biased in favour of women. In particular, this would mean that the initial negative effects on productivity and evaluations we estimate for female trainees may be lower bounds for the effect we would find had there been no training, and that the catch-up we observe between the trial and post-trial period may take somewhat

Table B.1: Pre-Post Training Differences in Trainee skills

VARIABLES	(1) Technical knowledge	(2) Confidence	(3) Belief best	(4) Drawings correct	(5) Drawing -soft	(6) Communic. -soft	(7) Leadership -soft
Post Training	-0.040	0.340*	0.139*	0.131*	0.484	0.330	0.091
	(0.670)	(0.199)	(0.078)	(0.068)	(0.596)	(0.857)	(0.794)
Post Training × Female Tr.	0.684	0.564*	0.094	0.069	-0.944	-0.653	-0.205
	(0.995)	(0.331)	(0.100)	(0.091)	(0.842)	(1.207)	(1.015)
Female Trainee	-1.596	-0.816***	-0.228***	-0.115*	1.097*	-0.266	-0.668
	(1.215)	(0.286)	(0.081)	(0.069)	(0.639)	(0.880)	(0.665)
Constant	55.829***	-0.143	0.653***	0.375***	-0.561	0.136	0.334
	(0.835)	(0.193)	(0.057)	(0.051)	(0.426)	(0.617)	(0.519)
Observations	287	287	290	268	271	271	196

Notes: Table regresses measures of supervisor skills of the female and male trainees, measured during the first and the last day at the training center, on an indicator for post-training measurement, interacted with an indicator variable for female trainees. The skill measures are explained in more detail in Section 4.1. Sample restricted to those trainees for which both pre- and post measures of skills are available. Robust standard errors in brackets: *** denotes statistical significance at 1%, ** at 5%, and * at 10%

longer if there were no training.

Appendix C: Estimation of Baseline Female Promotion Share in sector, with data from Menzel and Woodruff (2019)

Using administrative wage data from 36 garment factories from Menzel and Woodruff (2021) that include information on supervisors, we estimate that six percent of the stock of supervisors in these factories are female. However, this number may differ from the share of women among *promotions* to supervisors, if, for example, women quit from supervisor positions more or less quickly on average.

We observe 99 internal promotions to supervisors in 28 of the 36 factories in the data from Menzel and Woodruff (2021) over the course of one year, with 20 of them being of women. This share, however, is partly driven by one factory with 18 promotions, six of which are women. This suggests that highly factory-specific promotion rounds in the year from which we have data from a given factory (e.g. due to opening of new floors or lines) could distort this ratio. Thus we reweight the factory-specific promotion gender ratios by the total number of workers in the factories, and obtain a new average gender promotion ratio of 13.7 percent across the 28 factories.

Promotions to supervisors need not all be internal, at times workers move between factories, entering the new factory on a higher position than the previous one, a process which Menzel and Woodruff (2021) refer to as "external promotions". However, data on external promotion rates in the sector are difficult to come by, because factory records do not record whether a worker that has joined a factory as supervisor has already worked on that position in the previous factory. However, Menzel and Woodruff (2021) estimate external promotions to be more common for male workers, and to be generally low for promotions to supervisor and higher positions (as opposed to entry level positions, i.e. between helper and machine operator). The estimated share of women among internal promotions can therefore be considered a lower bound for their share among all (internal + external) positions.

Online Appendix D: Description Pilot Phase

Design

We started a first pilot run of the project in November 2011, in cooperation with the German Development Corporation GIZ, who had designed a supervisor training program with the goal of increasing the number of female supervisors in factories. GIZ initially expressed a preference that we train only female operators as part of the project. Recognizing the value of having some comparison sample of male operators, we agreed with GIZ to train four female and one male worker from each of the participating factories.

We began contacting potential factories, with a letter of introduction from a large UK-based buyer, in August 2011. Our aim was to work with a sample of factories capable of selling directly to large international buyers. Using transaction-level import- and export-data obtained from the Bangladeshi National Bureau of Revenue, we calculated the average unit value of shipments (USD per kilogram) on the exporter- and exporter-product-year-level. Using these two measures, we selected a sample of 230 firms with annual shipment volumes large enough to sell directly to large foreign buyers. We started to contact these factories per telephone, offering participation in this evaluation scheme of the training course. By November 2011, we had received an initial commitment to participate in the project from 85 factories from the list, with 58 completing the pilot, including all worker survey rounds.

Selection of Trainees

Our aim was to select from each factory four female and one male operator for training, and a valid control group of workers not attending the training. In all rounds the selection process started with factories selecting a pool of potential trainees to which we administered a literacy and simple production knowledge test. Potential trainees were excluded if they did not pass the literacy test or said their families would not allow them to participate in the training.

Initially, we asked the factories to identify 16 female and 4 male operators who were good candidates for the training. We ranked the nominees according to their score on

the knowledge and literacy test and then selected the two females with top marks on the diagnostic test as trainees. We then assigned a random number to the female trainees ranked 3rd to 6th on the test, and assigned the two with the highest random numbers to training, and the two with the lowest random numbers to control. Among the males, we followed a similar procedure by taking the males with the top two marks and randomly assigning one to treatment and one to control. Halfway through the pilot, we modified the selection process to allow the factory to choose two females they wanted to send to training, conditional only on them demonstrating a basic level of literacy, and, even later, by reducing the number of operators the factory nominated to eight females and four males. Overall, 271 operators (213 females and 58 males) were selected this way and received the training.

Description of Trainees

Around half a year after the return from the training, 90 percent of the male and 77 percent of the female trainees self-reported that they have been tried out as supervisors, and 77 percent of the male and 53 percent of the female trainees report to be still working as supervisors in the factory. Meanwhile 20 percent of the male and 23 percent of the female trainees had left the factories. These numbers are close to those we see in the main trial, as discussed in Sections 2.4 and 4.5 of the paper.

Table D.1 shows basic demographic characteristics for male and female trainees from the pilot, and comparisons to the trainees of the same sex from the main trial. Overall, the two sets of trainees look very similar. Among ten comparisons, only one shows significant differences; women from the main trial were more likely to be married (77 vs 63 percent).

Table D.1: Pilot Phase, Trainee Demographic Characteristics

Variable	Mean Pilot Men	Difference Pilot Men-Women	Difference Pilot-Main Tr. Men	Difference Pilot-Main Tr. Women
Age	24.171	-0.955	0.565	0.511
	(3.162)	(0.621)	(0.714)	(0.548)
Married	0.537	0.084	0.075	0.133**
	(0.505)	(0.085)	(0.097)	(0.065)
Years Schooling	9.098	-0.653	-0.209	-0.135
	(2.427)	(0.401)	(0.448)	(0.313)
Years in Sector	6.488	-0.742	-0.047	0.455
	(3.362)	(0.577)	(0.670)	(0.457)
Years in Factory	3.064	0.264	0.591	-0.123
	(2.409)	(0.490)	(0.514)	(0.393)
Observations	41	233	113	255

Notes: Table compares male and female trainees from the pilot phase against the male and female trainees from the main trial. The first column shows averages for male trainees from the pilot phase, while the second column the difference to female trainees from the pilot phase. The third column shows the difference of male trainees from main trial to those from the pilot, while the fourth column the same for female trainees from the main trial and the pilot. ** denotes statistical significance of differences at 5%

Online Appendix E: Knowledge Diagnostic Test

Figure E.1

Section 2: APTITUDE EVALUATION

Section 2.1: Garment Processes

2.1	**READ:** Now I am going to ask you some questions about a type of garment that you are most accustomed to working with, so please tell me the type of garment you produce most often. Is it 001=knit tops such as t-shirts and polo-shirts, 002=bottoms, or 003=woven shirts?	

READ: Now I am going to hand you a garment. Please look at the 5 parts identified with a tag, name the process required to cre name the best machine to use for this process, and tell me how many pieces could be completed in an hour. What is the name of this part? [tag 1] What is the best machine to use to create this part?

A 003=single needle/plain machine, 004=over lock, 005=flat lock, 006=button hole, 007=Feed Of Arm, or 008=Double Needle? [Show Picture booklet section 2.2] How many of these pieces could be completed in an hour?

Instruction: Please give the appropriate garment which was selected in question 2.1 to the respondent and let them identify the p machines. Point out the location on the garment indicated by the tags, start with tag 1, and finish with tag 5. **Put in 001=yes in t "process" if the process is identified, and 002=no if the process is incorrect.**
Show section 2.1.1 of the picture booklet when asking about the machine. Record the machine identified in the column titled "m the number of pieces per hour in the column titled "Pieces / hr." Use -88= Refuse to answer, -99=Don't know if needed

If 2.1 was 001= knit tops such as t-shirts and polo-shirts

Instruction: Please give the t-shirt to the respondent and let them identify the processes and machines.

		Process	Machine	
2.1.1	Attach rib at neck position	___	___	
2.1.2	Attach main label at back	___	___	
2.1.3	Shoulder top stitch	___	___	
2.1.4	Hem sleeves	___	___	
2.1.5	Tack at sleeve ends	___	___	

If 2.1 was 002=bottoms

Instruction: Please give the bottoms to the respondent and let them identify the processes and machines.

		Process	Machine	
2.1.6	Back Rise	___	___	
2.1.7	Top Stitch at Fly/ J Stitch	___	___	
2.1.8	Attach Zipper	___	___	
2.1.9	Hem bottom	___	___	
2.1.10	Attach Waist Band to body	___	___	

If 2.1 was 003=woven shirts

Instruction: Please give the shirt to the respondent and let them identify the processes and machines.

		Process	Machine	
2.1.11	Join Back Yoke	___	___	
2.1.12	Attach front pocket	___	___	
2.1.13	Attach Placket to Front side	___	___	
2.1.14	Join Cuff	___	___	
2.1.15	Armhole Top Stitch	___	___	

Section 2.2:Machine Parts and Functions

READ: Now I am going to show you a photo of a single needle machine, and I would like you to name the parts indicated by th numbers.

Instruction: Show section 2.2.1 of the picture booklet and ask the following questions.
[Codes: codebook serial number 9]

2.2.1.1	**READ:** What is the name of (1)? *Instruction:* Was the name "Arm/ Take-Up-Arm"?	___
2.2.1.2	**READ:** What is the name of (2)? *Instruction:* Was the name "Pin/Spool pin in"?	___
2.2.1.3	**READ:** What is the name of (5)? *Instruction:* Was the name "Eye clamp"?	___
2.2.1.4	**READ:** Now, please tell me the order in which you would thread a single needle machine by pointing at the numbers shown in the picture? *Instruction:* Indicate the order given in the space below. Indicate here: ___>___>___>___>___>___ Correct order: _3_>_2_>_5_>_4_>_1_>_6_ Was the order correct?	___

Figure E.2

Section 2.4:Cause of Quality Issue

READ: Now I am going to show you some fabric with stitches in it. Please choose the cause of the problem from the options I g
than one cause can be selected.

Instruction: Show the sample fabric. Point out the location on the fabric indicated in writing when asking each question. Start w
with 6. If the respondent thinks a numbered sentence is a correct answer to the question above it, then indicate 001=yes. If the r
a numbered sentence is not a correct answer to the question above it, then indicate 002=no.
[Codes: codebook serial number 9]

READ: What is wrong with the single needle machine that made this puckered / Sheared stitch?		
2.4.1.1	Low tension	_____
2.4.1.2	High tension	_____
2.4.1.3	Roughness of thread	_____
2.4.1.4	Oversized needle	_____
2.4.1.5	Improper oiling	_____
READ: What is wrong with the single needle machine that made this stitch with skipped/drop/false stitches, also known as drop Stitch?		
2.4.2.1	Lopper / Bobbin is not adjusted with Needle	_____
2.4.2.2	Upper tension is too tight	_____
2.4.2.3	Needle is too short for machine	_____
2.4.2.4	Needle is not set correctly in needle clamp	_____
2.4.2.5	Ball point needle	_____
READ: What is the cause of this loose stitch or tension loose made by a single needle machine that leads to thread breaking?		
2.4.3.1	Improper-sized needle	_____
2.4.3.2	tension between upper thread and lower thread or between Bobbin and Lopper is not adjusted	_____
2.4.3.3	Incorrect threading sequence	_____
2.4.3.4	Needle head is broken	_____
2.4.3.5	Improper oiling of the machine	_____
READ: What is the cause of this needle-cut made by a Single Needle machine?		
2.4.4.1	needle chosen according to fabric thickness	_____
2.4.4.2	Improper size of needle	_____
2.4.4.3	Blunt needle	_____
2.4.4.4	Upper tension too tight	_____
2.4.4.5	the tension is not adjusted between lower and upper thread	_____
READ: What is the possible cause of this stain / spot on the garment?		
2.4.5.1	Machine cleaned once in last 24 hours	_____
2.4.5.2	No hand gloves on operator	_____
2.4.5.3	Tension is not adjusted between bobbin and lopper	_____
2.4.5.4	Garment is sewn just after oiling the machine	_____
2.4.5.5	Food is not taken in sewing floor	_____
READ: What is the cause of this uneven stitch made by a Single Needle machine?		
2.4.6.1	Tension too tight	_____
2.4.6.2	Improper machine guide	_____
2.4.6.3	Oversized feed dogs	_____
2.4.6.4	Improper machine handling	_____
2.4.6.5	Improper oiling of machine	_____

Section 2.5:Quality Issues (Picture)

READ: Now I am going to show you a couple of pictures of activities and things inside a factory. In each of these pictures there
or more things that are wrong or missing and need to be corrected. Suppose you are given the task of maintaining good work pra
situations depicted by the pictures. Try to look into each picture very minutely and identify as many problems as necessary in it,
exam so do not worry if you are unable to do so.
I will show you a picture that shows a finished garment awaiting final inspection. You will have 1 minute to look at this picture
many problems as necessary in the situation depicted.
Instruction: Show section 2.5 of the picture booklet and take it back after one minute. Put in 1=yes in the rightmost column agai
correctly mentioned and 2=no for the one not mentioned by the respondent. For problems mentioned other than the ones in the
keep a count of such problems and state the total number in 2.5.8

2.5.1	Chalk marks around the buttonholes	
2.5.2	Broken needle stuck inside a seam	
2.5.3	No label at all	

Figure E.3

Section 2.7: Operation Breakdown		
READ: I will now ask you some questions about how production is best arranged on a sewing line in a garments factory. Please take a look at the table shown here. This is a simplified operation breakdown. An operation breakdown separates each step of the garment production process, and this information is used to make decisions about how a production line is arranged. This example is not realistic, but we would like you to answer some questions about it, given the information that is provided. *Instruction: Show section 2.7 of the picture booklet and explain what each part of the table means.* [Use -88= Refuse to answer, -99=Don't know if needed]		
2.7.1	What is the hourly target? You have 1 minute to answer the question. *Instructions: Start the stop-watch when you finish asking the question, and if no answer is given after 1 minute, use -99=Don't know, and move to the next question.*	|___|___|___|
2.7.2	How many operators would you require for operation 7? You have 1 minute to answer the question. *Instructions: Start the stop-watch when you finish asking the question, and if no answer is given after 1 minute, use -99=Don't know, and move to the next question.*	|___|
2.7.3	How many operators would you require for operations 5? You have 1 minute to answer the question. *Instructions: Start the stop-watch when you finish asking the question, and if no answer is given after 1 minute, use -99=Don't know, and move to the next question. When the respondent answers, check if the respondent means X persons working on each operation, or X persons working on both operations 6 and 7.The number of persons working on both operations should be written in the space provided (E.g. If the respondent answers 2 people on operation 6 and 2 people on operation 7, then write 4 in the space provided).*	|___|
2.7.4	On which operation you put your fastest-working single needle operator?	|___|
2.7.5	On which operation you put your slowest-working single needle operator?	|___|

Appendix F: Experimentation in the Canonical Model

We model experimentation in the canonical Phelps–Aigner–Cain (PAC) model of statistical discrimination (Phelps, 1972; Aigner and Cain, 1977). This extension shows that all our main implications (Facts 1-4) can also be derived in the canonical model.

In the canonical PAC model, the decision-maker (DM) learns about *individual talent* from a noisy signal and sets wages using Bayesian predictions. In our setting, to sharpen our focus on experimentation, we let the DM be certain about the performance of men and instead learn about the expected performance of women. To align with our empirical context and evidence, we label the known component of a woman's performance μ "talent", and the uncertain component Δ about which DM learns "bias". This distinction is purely semantic: ultimately, the DM is learning about drivers of performance in both models. So, we assume that talent is *known* and instead let the DM learn about a *group-specific bias* against women. We retain the linear-normal structure familiar from PAC. The key implication is that even with correct Bayesian inference, *marginal-candidate comparisons fail* as tests for discrimination once optimal experimentation about the bias is valuable.

Environment and Learning

There are two periods, $t = 0, 1$, with discount factor β. The female and male candidates have known talent μ_F and μ_M. The output when the candidate is appointed as a line supervisor is given by

$$y_M^t = \mu_M + \varepsilon^t, \qquad y_F^t = \mu_F + \varepsilon^t + \Delta^t,$$

where $\varepsilon^t \sim \mathcal{N}(0, 1/h^\varepsilon)$ and, for female assignments, $\Delta^t \mid \theta \sim \mathcal{N}(\theta, 1/h^\Delta)$. The *expected* bias θ is the parameter that the DM is trying to learn about, and for which she has prior $\theta \sim \mathcal{N}(\theta_0, 1/h^\theta)$ with $\theta_0 < 0$.

As in our model, if DM appoints the male candidate in $t = 0$, she learns nothing and thus appoints a woman in period $t = 1$ only if $\mu_F + \theta_0 > \mu_M$. Of course, if this condition is satisfied, the DM would have appointed a woman in $t = 0$ to begin with. In other words, as in our model, the only interesting case to consider is the one in which in period

$t = 0$ the DM appoints a woman, and learns information about θ.

A single female appointment at $t = 0$ reveals the residual

$$r \;\equiv\; y_F^0 - \mu_F \;=\; \varepsilon^0 + \Delta^0 \;\sim\; \mathcal{N}\Big(\theta,\, \frac{1}{h^r}\Big), \qquad h^r \;\equiv\; \Big(\frac{1}{h^\varepsilon} + \frac{1}{h^\Delta}\Big)^{-1}.$$

Bayesian updating after having tried the female candidate yields posterior precision and mean

$$h_1^\theta = h^\theta + h^r, \qquad \theta^1 = \frac{h^\theta \theta_0 + h^r r}{h^\theta + h^r} \;=\; w\,\theta_0 + (1-w)\,r, \quad w = \frac{h^\theta}{h^\theta + h^r}.$$

If the prior is very precise (i.e., DM already knows θ, $h^\theta \to \infty$) or if the performance metric is very noisy (i.e., $h^\epsilon \to 0$ and thus $h^r \to 0$), the DM learns little from the trial, and $\theta^1 \to \theta^0$ regardless of realized output y_F^0.

Observation: *Experimentation is akin to obtaining a signal with higher precision.*

Define the *information content* (precision) of one female observation about θ as

$$h^r \;\equiv\; \Big(\frac{1}{h^\varepsilon} + \frac{1}{h^\Delta}\Big)^{-1}.$$

More generally, after k female trials, the posterior precision on the bias mean is

$$h_k^\theta \;=\; h^\theta + k\,h^r \qquad (k = 0, 1, 2, \ldots).$$

The male predictive precision is constant at h^ε.

For women, before any female experiment ($k = 0$),

$$Var(y_F - \mu_F) \;=\; \underbrace{\tfrac{1}{h^\varepsilon}}_{\text{measurement}} \;+\; \underbrace{\tfrac{1}{h^\Delta}}_{\text{idiosyncratic bias}} \;+\; \underbrace{\tfrac{1}{h^\theta}}_{\text{mean-bias uncertainty}} \qquad \Rightarrow \quad h_{F,0}^{\text{pred}} \;=\; \Big(\tfrac{1}{h^\varepsilon} + \tfrac{1}{h^\Delta} + \tfrac{1}{h^\theta}\Big)^{-1}.$$

After k female experiments (so the posterior on θ has precision h_k^θ),

$$Var(y_F - \mu_F) \;=\; \tfrac{1}{h^{\varepsilon}} + \tfrac{1}{h^{\Delta}} + \tfrac{1}{h^{\theta}_k} \quad \Rightarrow \quad h^{\text{pred}}_{F,k} \;=\; \left(\tfrac{1}{h^{\varepsilon}} + \tfrac{1}{h^{\Delta}} + \tfrac{1}{h^{\theta}+k\,h^{r}}\right)^{-1}.$$

Because $h^{\theta}_k = h^{\theta} + kh^{r}$ is strictly increasing in k and enters the denominator as $1/h^{\theta}_k$,

$$h^{\text{pred}}_{F,k+1} \;>\; h^{\text{pred}}_{F,k}.$$

Hence, the *predictive precision* for women strictly increases with experimentation (k). Moreover,

$$\lim_{k\to\infty} h^{\text{pred}}_{F,k} \;=\; \left(\tfrac{1}{h^{\varepsilon}} + \tfrac{1}{h^{\Delta}}\right)^{-1},$$

which is the upper bound set by irreducible noise ε and idiosyncratic bias Δ.

Testable Implications

We now derive the four key implications (Facts 1-4) from our framework.

Fact 1 (Exploration threshold and selection at $t = 0$). Let the DM choose in $t = 0$ between assigning a man or a woman. Expected output in $t = 0$ are $E[y^0_M] = \mu_M$ and $E[y^0_F] = \mu_F + \theta_0$. Let the $t = 1$ continuation values be

$$V_1^{\text{noF}} = \max\{\mu_M,\ \mu_F + \theta_0\}, \qquad V_1^{\text{tryF}} = E_r\left[\max\{\mu_M,\ \mu_F + \theta^1(r)\}\right].$$

The DM assigns a woman in $t = 0$ iff

$$\underbrace{(\mu_F - \mu_M)}_{\text{talent gap}} + \underbrace{\theta_0}_{\text{prior bias}} + \underbrace{\beta\Big(V_1^{\text{tryF}} - V_1^{\text{noF}}\Big)}_{\text{value of information (VOI)}} \;\geq\; 0. \tag{4}$$

The VOI term is strictly positive whenever learning can flip the $t = 1$ choice with nonzero probability (i.e., when $\mu_M \approx \mu_F + \theta_0$ and h^{θ}, h^{r} are finite).

Implications. Assigning a woman can be optimal even when $\mu_F < \mu_M$ provided the VOI is large (higher δ, weaker prior precision h^{θ}, or more informative female residual h^{r}) relative to the expected bias.

Fact 2 (Failure of marginal-candidate tests). Even with a fully-rational Bayesian DM, the initial *marginal* performance gap need not diagnose efficiency or ability.

(a) Initial marginal comparison (static rule).

Without experimentation (or with $\beta \to 0$), the $t = 0$ choice is male iff

$$\mu_M > \mu_F + \theta_0 \quad \Longleftrightarrow \quad \mu_F - \mu_M < -\theta_0. \tag{5}$$

Hence if $\mu_F - \mu_M < 0$ and $\theta_0 < 0$, the static margin selects men.

(b) Initial expected performance gap.

Regardless of the assignment decision, the *expected* period-0 outcome difference is

$$E\big[y_F^0 - y_M^0\big] = \ (\mu_F + \theta_0) - \mu_M = \ (\mu_F - \mu_M) + \theta_0. \tag{6}$$

This can be negative, for two reasons: (1) the expected negative bias, and (2) due to a positive VOI term, μ_F can be lower than μ_M. Thus, observing a worse female margin at $t = 0$ is not a sufficient test for (the absence of) discrimination.

Fact 3 (Dynamic retention and convergence). At $t = 1$, given belief θ', the DM assigns a woman iff $\mu_F + \theta' \geq \mu_M$. After a female trial, the posterior mean is $\theta^1(r) = \frac{h^\theta \theta_0 + h^r r}{h^\theta + h^r}$, so retention (female again in $t = 1$) occurs iff

$$\mu_F + \theta^1(r) > \mu_M \quad \Longleftrightarrow \quad \theta^1(r) > \Delta\mu, \quad \Delta\mu \equiv \mu_M - \mu_F.$$

Solving for the residual threshold $r^\star$:

$$\frac{h^\theta \theta_0 + h^r r^\star}{h^\theta + h^r} > \Delta\mu \quad \Longleftrightarrow \quad r^\star = \frac{(h^\theta + h^r)\,\Delta\mu - h^\theta \theta_0}{h^r} = \ \Delta\mu + \frac{h^\theta}{h^r}\big(\Delta\mu - \theta_0\big). \tag{7}$$

Since $r \sim \mathcal{N}(\theta, 1/h^r)$, the *retention share* (probability a trial female is assigned again in $t = 1$) is

$$\text{Retain}_F = \Pr\big(r > r^\star\big) = 1 - \Phi\Big(\big(r^\star - \theta\big)\sqrt{h^r}\Big), \tag{8}$$

with $\Phi(\cdot)$ the standard normal cdf. Conditional on retention, posterior beliefs are

higher (less negative), implying convergence of expected female performance toward male levels in $t = 1$; the $t = 0$ dispersion of female outcomes is larger than in $t = 1$ due to selection on r.

Fact 4 (Policy and identification comparative statics). Female trials and retention increase when (i) the horizon is longer (higher β), (ii) prior precision h^θ is lower (beliefs more uncertain), or (iii) the female residual is more informative (higher h^r: better measurement or less idiosyncratic bias).

When θ^0 is lower than the true value, experimentation will lead to positive beliefs updating and to a higher share of appointed women in the future.

Experimentation thus acts like a PAC-style *comparative static* that *raises the effective precision for the discriminated group over time*. Concretely, define an *effective female noise* for prediction

$$\tau^2_{F,k} \equiv \frac{1}{h^{\text{pred}}_{F,k}} = \underbrace{\tfrac{1}{h^\varepsilon}}_{\text{measurement}} + \underbrace{\tfrac{1}{h^\Delta}}_{\text{idiosyncratic bias}} + \underbrace{\tfrac{1}{h^\theta + k\,h^r}}_{\text{mean-bias uncertainty (falls with } k)},$$

so $\tau^2_{F,k}$ *decreases* with k, i.e., $h^{\text{pred}}_{F,k}$ *increases*. Although experimentation is dynamic (learning about θ), its reduced-form manifestation in a PAC model is that *the group-F signal becomes more precise with experimentation*. The canonical Phelps–Aigner–Cain (PAC) model is naturally suited to *wage setting*: with linear–Gaussian assumptions, the employer sets $w = E[\theta \mid \text{signal}, g]$. Since, by contrast, we focus on *promotion/assignment* decisions, which are inherently *threshold-based*, a binary choice formulation provides simpler algebra.

References

Adhvaryu, A., N. Kala, and A. Nyshadham (2023). Returns to on-the-job soft skills training. *Journal of Political Economy 131*(8), 2165–2208.

Adhvaryu, A., A. Nyshadham, and J. Tamayo (2023). Managerial quality and productivity dynamics. *The Review of Economic Studies 90*(4), 1569–1607.

Ahern, K. R. and A. K. Dittmar (2012). The Changing of the Boards: The Impact on Firm Valuation of mandated Female Board Representation. *Quarterly Journal of Economics 127*(1), 137–197.

Aigner, D. J. and G. G. Cain (1977). Statistical theories of discrimination in labor markets. *Ilr Review 30*(2), 175–187.

Ashraf, N., O. Bandiera, V. Minni, and V. Quintas-Martínez (2023). Gender and the misallocation of labor across countries. *Work in progress, June*.

Atkin, D. (2016). Endogenous Skill Acquisition and Export Manufacturing in Mexico. *American Economic Review 106*(8), 2046–85.

Atkin, D., A. Chaudhry, S. Chaudry, A. Khandelwal, and E. Verhoogen (2017). Organizational Barriers to Technology Adoption: Evidence from Soccer-Ball Producers in Pakistan. *Quarterly Journal of Economics 132*(3), 1101–1164.

Banerjee, A., D. Karlan, and J. Zinman (2015, January). Six randomized evaluations of microcredit: Introduction and further steps. *American Economic Journal: Applied Economics 7*(1), 1–21.

Bardhi, A., Y. Guo, and B. Strulovici (2024). Early-career discrimination: Spiraling or self-correcting? *mimeo*.

Beaman, L., R. Chattopadhyay, E. Duflo, R. Pande, and P. Topalova (2009). Powerful Women: Does Exposure Reduce Bias? *Quarterly Journal of Economics 124*(4), 1497–1540.

Belloni, A., V. Chernozhukov, and Y. Wei (2016). Post-Selection Inference for Generalized Linear Models with Many Controls. *Journal of Business & Economic Statistics 34*(4), 606–619.

Benson, A., D. Li, and K. Shue (2025). Potential" and the gender promotion gap. *American Economic Review*.

Bertrand, M. (2017). The Glass Ceiling. Becker Friedman Institute for Research in Economics Working Paper No. 2018-38.

Bertrand, M., S. E. Black, S. Jensen, and A. Lleras-Muney (2019). Breaking the Glass Ceiling? The Effect of Board Quotas on Female Labour Market Outcomes in Norway. *Review of Economic Studies 86*(1), 191–239.

Blau, F. and L. Khan (2017). The Gender Wage Gap: Extent, Trends and Explanations. *Journal of Economic Literature 55*(3), 789–865.

Bloom, N., B. Eifert, A. Mahajan, D. McKenzie, and J. Roberts (2012). Does Management Matter? Evidence from India. *Quarterly Journal of Economics 128*(1), 1–51.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2025). Inaccurate statistical discrimination: An identification problem. *Review of Economics and Statistics 107*(3), 605–620.

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

Boudreau, L. (2024). Multinational Enforcement of Labor Law: Experimental Evidence from Bangladesh's apparel sector.

Boudreau, L., J. Cajal-Grossi, C. Can, and R. Macchiavello (2024). Relationships and responsibility. *mimeo*.

Boudreau, L., J. Cajal-Grossi, and R. Macchiavello (2023). Global value chains in developing countries: a relational perspective from coffee and garments. *Journal of Economic Perspectives 37*(3), 59–86.

Cajal Grossi, J., R. Macchiavello, and C. K. Can (2022). Is better work better?: evidence from the garment sector in bangladesh.

Cajal-Grossi, J., R. Macchiavello, and G. Noguera (2023). Buyers' sourcing strategies and suppliers' markups in bangladeshi garments. *The Quarterly Journal of Economics*, 2391–2450.

Canay, I. A., M. Mogstad, and J. Mountjoy (2024). On the use of outcome tests for detecting bias in decision making. *Review of Economic Studies 91*(4), 2135–2167.

Duflo, E. (2003). Grandmothers and granddaughters: Old-age pensions and intrahousehold allocation in south africa. *The World Bank Economic Review 17*(1), 1–25.

Duflo, E. (2012). Women Empowerment and Economics Development. *Journal of Economic Literature 50*(4), 1051–1079.

Flabbi, L., M. Macis, A. Moro, and F. Schivardi (2019). Do Female Executives Make a Difference? The Impact of Female Leadership on Firm Performance and Gender Gaps. *Economic Journal 129*(622), 2390–2423.

Gereffi, G. (1999). International Trade and Industrial Upgrading in the Apparel Commodity Chain. *Journal of International Economics 48*(1), 37–70.

Gibbons, R. and R. Henderson (2012). Relational Contracts and Organizational Capabilities. *Organization Science 23*(5), 1350–1364.

Goldin, C. (2014). A Grand Gender Convergence: Its Last Chapter. *American Economic Review 104*(4), 1091–1119.

Goldin, C., L. F. Katz, and I. Kuziemko (2006). The homecoming of american college women: The reversal of the college gender gap. *Journal of Economic perspectives 20*(4), 133–156.

Gollin, D. (2008). Nobody's business but my own: Self-employment and small enterprise in economic development. *Journal of Monetary Economics 55*(2), 219–233.

Haegele, I. (2024). The broken rung: Gender and the leadership gap. *arXiv preprint arXiv:2404.07750*.

Heath, R. and M. Mobarak (2015). Manufacturing growth and the lives of Bangladeshi women. *Journal of Development Economics 115*, 1–15.

IFC and ILO (2025). Creating better jobs for women and boosting productivity in bangladesh's garment factories: An assessment of the gender equality and returns program. Technical report, IFC / ILO.

Komiyama, J. and S. Noda (2024). On statistical discrimination as a failure of social learning: A multiarmed bandit approach. *Management Science*.

Li, D., L. R. Raymond, and P. Bergman (2024). Hiring as exploration. *Review of Economic Studies*.

Macchiavello, R., A. Rabani, and C. Woodruff (2015, May). The Market for Training Services: A Demand Experiment with Bangladeshi Garment Factories. *American Economic Review, Papers & Proceedings*.

Matsa, D. A. and A. R. Miller (2013). A Female Style in Corporate Leadership? Evidence from Quotas. *AEJ: Applied Economics 5*(3), 136–69.

McKelway, M. (2021). How Does Women's Employment Affect Household Decision-Making? Experimental Evidence from India. Working Paper, Dartmouth College.

McKinsey (2011). Bangladesh's ready made garments landscape: The challenge of growth. McKinsey&Company, Apparel, Fashion & Luxury Practice.

Menzel, A. and C. Woodruff (2021). Gender wage gaps and worker mobility: Evidence from the garment sector in bangladesh. *Labour Economics 71*, 102000.

Miller, G. (2008, 08). Women's Suffrage, Political Responsiveness and Child Survival in American History. *The Quarterly Journal of Economics 123*(3), 1287–1327.

Olivetti, C. and B. Petrongolo (2016). The Evolution of Gender Gaps in Industrialized Countries. *Annual Review of Economics 8*, 405–434.

Onuchic, P. (2024). Recent contributions to theories of discrimination. *mimeo*.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The american economic review 62*(4), 659–661.

Sandvik, J. J., R. E. Saouma, N. T. Seegert, and C. T. Stanton (2020, 04). Workplace knowledge flows*. *The Quarterly Journal of Economics 135*(3), 1635–1680.

Uckat, H. (2020). Womens promotion and intra-household bargaining: Evidence from bangladesh. Technical report, Working paper, Oxford University.

Uckat, H. and C. Woodruff (2020). Learning What to Look For: Hard Measures on Soft Skills in Promotion. Working paper, Oxford University.

Verhoogen, E. (2023, December). Firm-level upgrading in developing countries. *Journal of Economic Literature 61*(4), 1410–64.