Challenges of Change: An Experiment 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)

July 2020

Abstract

Women remain disadvantaged in access to management positions around the world. 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. Three sets of results emerge: (i) extensive diagnostic testing at baseline reveal few skill differences between marginal female and male supervisor candidates; (ii) initially, marginal female candidates have lower productivity and evaluations from sub-ordinate workers, though after four to six months, these gaps disappear; and (iii) the share of the female candidates retained as line supervisor after the trial is significantly higher than the share of female supervisors in the factories at baseline. This suggests that factories previously promoted fewer women than would have been optimal. Additional surveys and a lab-in-the-field experiment suggest that the initially worse performance stems from negative beliefs of workers about the abilities of female supervisors.

Keywords: Gender Discrimination, Productivity, Export Manufacturing
JEL Code: J16, J71, M51, M54, O14, O15

---

*Corresponding author: christopher.woodruff@qeh.oc.ac.uk.

The project has benefited from comments from seminars and conferences 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. 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).

1 Introduction

Women remain disadvantaged around the world, notably so in access to management positions (Blau and Khan (2017); Bertrand (2017); Olivetti and Petrongolo (2016); Goldin (2014)). There is ample evidence that women are rewarded less for the same performance on the job, and that they are less likely to be promoted.[1] The underlying source of women's disadvantage in managerial promotions, however, is not well understood. One possibility is taste-based discrimination (Becker (1957)): those in control of a promotion decision may have a preference for promoting males even when they know that the marginal female candidate will be a better manager than the marginal male candidate. Alternatively, suboptimal promotion may result from statistical discrimination originating from inaccurate beliefs that decision-makers may have about the future performance of candidates for managerial positions (Bohren et al. (2019)).[2]

Causal evidence of the effects of promoting women to management positions on productivity and worker beliefs is key to understanding the nature of bias in promotion decisions. The scarcity of evidence on this question owes to three distinct challenges. First, research on female managers is largely observational (e.g., Flabbi et al. (2019)) and thus struggles to causally attribute differences in performance to managers' gender. Second, measuring the efficiency cost of bias from any source requires a comparison of the marginal female and male managers, and marginal candidates are especially difficult to identify in observational data. And third, a manager's performance depends not just on her intrinsic skills, but also on typically unobservable beliefs held by her subordinates, peers, and superiors about her skills.

We address these challenges with a field experiment conducted with 24 large garment factories in Bangladesh. The garment sector is of interest both for its role in drawing women into full-time wage work and because it is a key driver of export-led industrialization globally (Gereffi (1999)). The Bangladeshi garment sector also offers advantages to overcome the three challenges. Production in the typical garment factory in Bangladesh is organized in a large number of independent, separate production lines, each of which is managed by one or more line supervisors. Although supervisors are

[1] See, e.g., Sarsons (2019) Mengel et al. (2018), Hengel (2018), Egan et al. (2017),, MacNell et al. (2015), Bowles et al. (2007), Goldin and Rouse (2000).

[2] Beliefs may be based on performance in nonmanagerial positions and thus be inaccurate about performance in managerial positions (Benson et al. (2019)).

almost always promoted from the pool of production workers, more than 90 percent of supervisors are male, while around 80 percent of sewing floor workers are female. For our experiment, participating factories identified 73 female and 72 male candidates for supervisory positions (*trainees* henceforth). This addresses the challenge of identifying the marginal candidates considered for promotion. After participating in a short training program, trainees were randomly allocated to an equal number of production lines to work as assistant supervisors for a period of two months. The randomization of supervisors to workers allows us to make causal inferences about the effect of supervisor gender on productivity.[3] We use a combination of surveys with subordinate workers and co-supervisors, and lab-in-the-field experiments, to directly measure beliefs and attitudes toward female supervisors before, during, and after the trial to address the challenge of measuring skills, beliefs, and outcomes.

The literature distinguishes tastes and beliefs by their malleability. Tastes are "stable over time" (Stigler and Becker (1977), p. 76), while beliefs calibrate rapidly to new information, especially when there are objective measures of performance (Bohren et al. (2019)). However, initial bias may lead to under-experimentation. As a result, beliefs may not be self-correcting even when objective measures are available (Beaman et al. (2009)). Learning is particularly challenging in the context of leadership in firms, where leadership depends on a willingness by subordinates to be led. Even if firms decide to increase experimentation with the goal of learning, if subordinate workers doubt the wisdom of their supervisors, they will not pay attention to their advice.[4] This suggests that even objective output measures will not be sufficient to calibrate beliefs to reality. Indeed, if teams underperform because subordinates believe female leaders are less effective, beliefs will be reinforced. We write down a simple framework that translates these insights into three potential forms of misallocation of supervisor talent along gender lines, with differing implications for firm productivity. We use the framework to guide the experimental design and the interpretation of the results.

[3]Mixed results surface from analysis of policy changes in Norway and Italy that mandate minimal representation of women among board members of publicly listed companies (see Bertrand et al. (2019), Matsa and Miller (2013), Ahern and Dittmar (2012) for Norway, and Maida and Weber (2019) for Italy). By contrast, our work focuses on lower-level, direct supervisors whose effectiveness may be determined by different factors, such as the (informal) authority invested in them by their sub-ordinates.

[4]Recent lab experiments show that workers pay less attention to advice provided by supervisors who are framed as being female, resulting in lower output (Ayalew et al. (2018); Grossman et al. (2016)). Similarly, Abel (2019) shows that workers lose more motivation on, and satisfaction with, a job after negative feedback from supervisors who are framed as female.

We present three core results. First, we show that beliefs about female supervisor skills are misaligned at baseline. We conducted extensive skills tests of the trainees before the start of the training and before their promotion to assistant supervisors on the randomly assigned lines. We find few differences between the female and male trainees. They are the same in terms of their age, level of schooling, experience in the garment sector, and tenure at the current factory.[5] However, employees at all levels of the factory hierarchy believe that males are better prepared to be supervisors. Most strikingly, employees are particularly likely to say that males have more technical skills, though the female and male candidates perform no differently on an 86-question technical skills diagnostic.

Second, the new female trainees had, on average, seven percent lower productivity during the first two months during which they were assigned to work in supervisory roles on random lines (henceforth the *trial* period). Female trainees also received lower evaluations from the workers on those lines at the end of this trial period, though co-supervisors rated the female and male trainees as equally effective. A lab-in-the-field exercise conducted with a comparable set of trainees from a different set of factories suggests that male workers, in particular, respond negatively to female leadership. However, workers randomly exposed to female supervisors during our trial soften their preference for male supervisors, and after the end of the trial they report being more indifferent to the sex of their supervisor. This effect is again mostly driven by male workers.

Finally, we find that, when factories were free to move trainees to other lines or demote them at the end of the two-month trial, the share of females among the trainees that the factories retained as supervisors (44 percent) was significantly higher than the share of women working as supervisors in the sector (7 percent), or the share of women among recently promoted supervisors (16 percent), before the trial. We also find evidence that the retained female trainees no longer underperform their male counterparts, measured either by production data or surveys of subordinate workers. Taken together, these results suggest that factories in our sample were under-promoting women to managerial positions before our intervention, with the evidence suggesting that the under-promotion of women was due to inaccurate beliefs about their managerial skills.

[5] Among the few differences we find, the most notable is lower self-reported confidence in own supervisor abilities of female trainees. The gap in self-reported self-confidence largely disappears by the end of the short training, before candidates are trialed as supervisors.

Our trial contributes to a number of different literatures on the causes and consequences of gender gaps in leadership positions. Chattopadhyay and Duflo (2004) and Beaman et al. (2009) analyze a policy experiment generating random exposure to female political leaders in Indian municipalities. Like our setting, theirs is characterized by strong negative baseline beliefs about (political) leadership abilities of women. They show that female village heads provide more public goods and take fewer bribes. Still, evaluations of female politicians by their constituents are initially more negative, and only catch up after repeated election cycles with female reservations, though beliefs about the general abilities of women as leaders adjust more rapidly. We extend this analysis into the realm of large, export-oriented firms, a setting in which one might have expected that competitive forces would have eliminated the bias.[6] Our results are in line with observed productivity differences across firms (Syverson (2011)) being at least in part driven by differences in beliefs and attitudes of mid-managers' subordinates, peers, and bosses in the organizational hierarchy (Gibbons and Henderson (2012)).

Our results also connect to Atkin et al. (2017), who show that misaligned incentives between employees on different firm hierarchy levels can hamper the adoption of new technologies. We argue that poorly aligned beliefs can similarly hamper the adoption of better HR practices such as promoting from a wider pool of talent. More broadly, we contribute to an emerging literature on firm upgrading in developing countries (see Verhoogen (2020) for an up-to-date review) focusing on an export-oriented industry of great socioeconomic relevance (Atkin (2016)).

Finally, we contribute to the literature on discrimination and stereotypes. A recent literature has shown that group stereotypes may induce discrimination by another route: negative beliefs may directly alter group members' performance. For example, Glover et al. (2017) show that managers' racial bias negatively affects minority cashier's job performance in a French grocery store chain, and Carlana (2019) finds that a teacher's bias affects gender gaps in math performance. We complement this literature by emphasizing the importance of subordinate beliefs about gender roles in affecting performance of female managers on the job.[7]

[6] Cajal (2016) and Cajal et al. (2019) study competitive pressures, sourcing arrangements and exporters' performance in the Bangladeshi garment sector.

[7] We also contribute to the broader literature on the effects of exposure to members from discriminated against groups on beliefs. Dahl et al. (2018), show that Norwegian soldiers with randomly allocated female fellow squad members significantly improve their beliefs about women's abilities as soldiers and beyond.

This paper proceeds as follows: Section 2 provides background on the Bangladeshi garment sector. Section 3 presents a simple conceptual framework. Section 4 summarizes our experimental design and data. Section 5 presents baseline beliefs about the abilities of male and female supervisors, comparing them to the supervisor skills we measure among our trainees at baseline, and Section 6 presents the main experimental outcomes on the lines supervised by the male and female trainees. Section 7 provides further evidence on mechanisms. Section 8 concludes.

2 Background

2.1 Context

Bangladesh is the second largest exporter of garments worldwide, after China. From a handful of factories in the early 1980s, the sector has grown to more than 4,000 factories employing around 4 million workers. Garments account for about 80 percent of Bangladesh's exports and an estimated 12 percent of Bangladesh's GDP (McKinsey (2011))[8]. The factories are very large by developing countries' standards. While data spanning the whole sector are hard to come by, membership data from Alliance and Accord, two organizations with the goal to improve factory safety in around 2,000 member factories, show that these factories employ on average almost 1,600 workers, with a median of 1,140.[9] While the Accord/Alliance members are clearly selected from among the larger and more modern factories, they are representative for the type of factories that we recruited for this project. International clothing brands (henceforth "buyers") design the garments and contract with factories in Bangladesh for their mass production. The factories are mostly locally owned and managed, and the majority are located in and around the two largest cities of the country, Dhaka and Chittagong. For logistical reasons,

Similar results have been shown for exposure to racial minorities (Carrell et al. (ming); Corno et al. (2019); Boisjoly et al. (2006)), immigrants (Finseraas and Kotsadam (2017)), or members from other castes (Lowe (2019)). Ben-Yishay et al. (2019) exogenously vary the gender of agricultural extension workers in Malawi. They and find that female and male extension agents perform equally well on knowledge tests and that farmers learn equally from female and male extension agents. Yet, farmers believe that female extension workers are less effective than males in that role.

[8] The Bangladesh garment sector has been widely studied, see, e.g., Heath and Mobarak (2015) on the effects of the sector on gender norms and gender inequality in the country; Boudreau (2019) for a RCT aimed at improving work safety, Heath (2018) on hiring practices in garment factories, and Boudreau et al. (2020) on worker migration patterns between factories.

[9] Own calculations, based on the membership records.

all factories participating in the project are located around Dhaka.

2.2 Production Lines

A general challenge to identifying the relative productivity of female and male supervisors is the availability of accurate and consistent productivity data spanning the workers managed by female and male supervisors. This makes the Bangladeshi garment an excellent setting for our study. Most factories have at least a cutting, sewing, and finishing department, with the sewing departments employing the majority of the workers. The sewing departments of most factories are organized into autonomous production lines, with between a dozen and 100 lines in the factories where we work, and 20 to 60 workers per line. Within factories, production lines are typically homogeneous in size, human and physical capital employed, and rarely specialize by types of garment. Furthermore, factories collect consistent daily line-wise productivity data based on physical output and employ a common labor-intensive technology, facilitating the comparison of performance.

2.3 Line Supervisors

Sewing line supervisors are promoted into their positions from among the ordinary sewing workers in the sections. Supervisors maintain worker discipline and function as first point of contact if any worker on their line encounters a problem during work. They also teach workers new sewing operations when lines switch garment styles, which they do on average around every two weeks. Lines can have one to four supervisors, depending on their size, and a supervisor typically manages 5-20 workers. When lines have more than one supervisor, there is often a line chief directly above them. Alternatively, the most senior line supervisor on a line may be designated as line chief.

Traditionally, while the large majority of non-supervisory workers in the sewing departments are women, supervisors are almost exclusively recruited from among the male sewing workers. Menzel and Woodruff (2019) study administrative staff records from 70 large garment factories in Bangladesh, and find that women account for 80 percent of non-supervisory sewing workers, but only seven percent of line supervisors, and one percent of line chiefs or higher-level supervisors. Put differently, in sewing sections more than 17 percent of male workers, but only 0.3 percent of female workers, are in supervisory positions. Almost all female workers leave the sector by the age of 35, while men in supervisor positions often have much longer careers in the sector.

A shortage of qualified supervisors is perceived to be an important barrier to better performance of the factories (e.g., McKinsey (2011)). As the economy and the sector continue to grow at more than five and ten percent annually, respectively, the sector appears less and less able to attract the most qualified men. As shown in Figure 1, the level of education of female and male operators entering in the sector has converged over time across successive cohorts, with a notable decline in the average education of male operators even during a period in which education levels in Bangladesh more generally were increasing. Accessing the vast pool of female workers for supervisor talent is thus an increasingly important need for the sector.

3 Conceptual Framework

This section provides a framework to guide the experimental design and determine the conditions under which promotion practices are efficient. We begin with a standard scenario in which the output of a production line depends on the talent of its supervisor and characteristics of the production line. In this case comparing the output of marginal male and female supervisors on any set of production lines is sufficient to identify misallocation. We discuss how practical details of the experimental implementation potentially bias the test against female supervisors and propose an alternative test that accounts for this bias.

We then extend the framework to consider the case in which output also depends on the gender of the supervisor (e.g., because of biases among co-workers and subordinates). In this case, the set of production lines on which the experiment is run matters and comparing performance no longer identifies misallocation. This highlights a general issue in experiments in which those implementing the experiment have unobservable knowledge about optimal allocation decisions. We sum up the discussion introducing three distinct notions of misallocation and how they relate to our evidence.

3.1 Experimental Design in a Standard Production Framework

Consider a standard framework in which the output of a worker depends on his or her talent for the job and other characteristics of the job s/he is assigned to. Specifically, the factory employs a mass of female workers and a mass of male workers on a continuum of production lines indexed by $j \in [0,1]$. Each worker in the factory is endowed with a potential level of supervisor talent μ, which is distributed among female and male workers according to continuous cumulative distribution functions $F^F(\mu)$ and $F^M(\mu)$, respectively. We allow these two distributions to differ, for example because of gender-specific endowments of managerial talent or differential selection into the sector. Output y_{ij} of a production line j depends on a set of line characteristics (such as the skills of workers on the line) aggregated into a scalar score x_j and on the managerial talent μ_i of its supervisor i, or $y_{ij} = Y(\mu_i, x_j)$, with $\frac{dY}{d\mu_i} > 0$.

Assume factories need to fill a mass B of supervisor positions. In line with Becker (1957), we consider a scenario in which the factory might have a preference over the number of appointed male and female supervisors but, conditional on those preferences, appoints the best available male and female candidates first. Denote with μ_B^g the talent of the marginal supervisor of gender $g = \{F, M\}$, that is, the supervisor with the lowest level of talent among those of gender g. We have $B = M\left(1 - F^M(\mu_B^M)\right) + F\left(1 - F^F(\mu_B^F)\right)$.

Let $R_B = \left[F\left(1 - F^F(\mu_B^F)\right)\right] / \left[M\left(1 - F^M(\mu_B^M)\right)\right]$ denote the ratio of female to male supervisors. This ratio is around seven percent in the Bangladeshi garment sector at the time of the baseline of our trial. Since factories employ many more female then male workers ($F > M$) it must be that $F^M(\mu_B^M) < F^F(\mu_B^F)$. The low baseline ratio can thus arise if either the upper tail of the distribution for males is thicker than for females and/or if the threshold μ is lower for males than for females, $\mu_B^M < \mu_B^F$. Figure 2 is drawn illustratively to depict these possibilities. In the absence of preferences over the gender composition of supervisors, the factory maximizes output by equating the talent of the marginal male and female supervisors, $\mu_B^M = \mu_B^F$. If, instead, the factory is biased against female supervisors, the marginal female supervisor will be better than the marginal male, i.e., $\mu_B^M < \mu_B^F$. In the simple standard framework, this would result in a production loss for the factories.

To test for misallocation in the basic scenario we therefore need to compare the performance of the lines managed by the *marginal* female and male supervisors. The first step in the experimental design asks factories to identify male and female operators for a training plus trial program that yields a pool from which to select the next internally promoted supervisors in the factory. Under the assumption that the factory selects for the program the next best available male and female candidates, the selection procedure identifies the marginal male and female supervisors with talent μ_B^M and μ_B^F. In the second step, the experimental design compares the performance of these candidates by *randomly* allocating them to manage production lines during the trial period. This test for misallocation compares output $Y(\mu_i, x_j)$ on a set of experimental lines ϵ to which the newly promoted marginal male and female supervisors have been randomly assigned, i.e.,

$$\Delta_\epsilon = \int_{e \in \epsilon \subset [0,1]} \left(Y(\mu_B^F, x_e) - Y(\mu_B^M, x_e) \right) de$$

Note that, since $Y(\mu, .)$ is increasing in μ on all lines, in this baseline scenario (i) we do not need to measure underlying managerial talent μ for the trainees, and (ii) the test is valid regardless of which production lines the experiment is conducted upon. In practice, we asked factories to identify lines where they expected to need supervisors in the near future to maximize compliance with the experimental protocol.

3.2 Experimental Design in Practice: Selecting Trainees

Although straightforward from a conceptual point of view, the implementation of the experiment presents certain practical difficulties. To increase the statistical power of the experiment we asked factories to select several female and male candidates, in line with their projected need for new supervisors over the half year following the baseline, and to select male and female trainees in equal proportion. Suppose, for example, the factory wanted to promote 10 workers to supervisory positions. We ask them to select five women and five men. In the absence of the intervention, however, the factory would have likely selected a pool of 10 workers with a very different gender composition. For example, if we use the prevailing ratio at baseline, the relevant counterfactual is one in which the factory would select nine males and a single female. Relative to this counterfactual, our experiment will overstate the talent of the marginal male supervisor (who would have a μ lower than any male in our experiment pool) and understate the talent of the marginal female supervisor (who would have the highest μ in the experimental pool). Comparing the average male and female trainee in our sample is thus a more stringent test than if factories had promoted only a single female trainee. We note that this bias is not likely to be large given that our trainees represent a tiny fraction of the female and male workers in the factories.[10]

3.3 Experimental Design Under Gender-Dependent Production

So far we have considered a standard scenario in which production only depends on supervisor talent and line characteristics. We now extend the model to allow the output of female and male supervisors on the same line to differ either because of differences in supervisory talent μ, or as a result of other factors related to the gender of the supervisor. For example, the level of authority bestowed by workers on their supervisor could vary with the gender of the supervisor: (some) workers might be more inclined to follow the orders of a male supervisor than of a female supervisor of equal talent. Alternatively, line operators might have less confidence in female supervisors' ability to help them sort out problems with machines, and so request less help. Formally, we capture this by extending the production function to $y_{ij} = Y(\mu_i, x_j, \phi_j)$, where $\phi_j = \{\phi_j^M, \phi_j^F\}$ depends on the sex of the line's supervisor. We refer to this mechanism as the "authority" channel as a short-hand, where we allow the amenability to female leadership to vary across lines.

Comparing output differences across lines randomly assigned to male and female trainees no longer identifies misallocation. Moreover, the set of experimental lines, ϵ, and the process through which the factory selects those lines, now matters. It is useful to define, for all lines supervised by males, the gap in productivity when the line is instead supervised by the marginal female supervisor, $\Delta_j = Y(\mu_B^F, x_j, \phi_j^F) - Y(\mu_j^M, x_j, \phi_j^M)$, with μ_j^M the skill level of the current male supervisor of the line.[11] Assuming the factory knows the bias on each line and allocates existing supervisors such that output is maximised, we can assume the factory would assign the marginal female supervisor to the line that maximizes Δ_j. We refer to this line as the *marginal* line. A necessary and sufficient

[10] Note that the source of bias described in this subsection is in addition to another potential source of bias against the female candidates. Factories have a lot of experience promoting male candidates and can thus be expected to be able to correctly identify the marginal male candidate. Factories have a lot less experience identifying female candidates for promotion, and might thus make mistakes (Uckat and Woodruff (2020)). Indeed, line chiefs involved in the selection of trainees report to be less confident in their assessment of the female trainees relative to the male trainees (see section 8 for more details).

[11] Given that more than 90 percent of lines are managed by males, we assume that the variation in ϕ_j is driven by variance in amenability to female leadership ϕ_j^F.

condition for misallocation in favor of men is a positive $\Delta_j > 0$ on the marginal line.

Based on an extensive pilot phase (see Appendix D), we asked factories to identify lines where they expected to need supervisors in the near future to maximize compliance with the experimental protocol. We then randomly allocated the female and male trainees for work during the trial period to one of the lines selected by the factory. This choice, inconsequential in the standard framework, potentially biases the test in the scenario with gender-dependent production *against* detecting misallocation in favor of men if the factories chose experimental lines with Δ_j smaller than the marginal line.

Two considerations suggest that, possibly because factories have limited information on Δ_j (and/or might have based their selection of experimental lines on factors only weekly correlated with Δ_j), the experimental set of lines is likely to be more representative of the whole set of lines in the factory, and less of the marginal lines. First, on observable characteristics, experimental lines ϵ do not look different from other lines in the factory (see Appendix A.1). Second, if it was indeed the case that trainees were allocated on lines particularly favorable to women relative to other lines in the factory, we should observe that factories disproportionately re-allocate male trainees away from the experimental set of lines after the trial. We don't see such differential reallocation, consistent with the experimental lines being fairly representative.

Due to the presence of bias ϕ, the estimated productivity differences between female and male trainees no longer identify differences in managerial talent among the marginal candidates. We address this challenge in four ways. First, we proxy for managerial talent through an extensive set of diagnostic tests on the trainees before the beginning of the program. Second, we complement the diagnostic with a set of lab-in-the-field-experiments in which we simulate managerial exercises with a comparable sample of trainees. Third, we use evaluations by peers, a popular tool in businesses around the world to measure staff skills (Sol (2016); Edwards and Ewen (1996); Bohl (1996)). Fourth, we rely on factories' revealed preferences on which of the trainees they keep as supervisors after the initial trial.

3.4 Experimental Design and Experiment Effects

A final challenge is the possibility that the experiment itself differentially changes the performance of the marginal male and female supervisor. In the framework above, this could arise either from a differential increase in managerial talent μ, or from a reduction in the bias against female supervisors, ϕ. To account for such possibility, we would need to extend the model to a dynamic framework. We interpret our short trial period (two months) as a relatively minor cost to find out whether there are suitable female supervisors. To rationalize the low rate of female supervisors at baseline with the high retention rate following the trial, it must be that management has a very high discount factor, and hence is unwilling to pay even the modest cost implied by the trial period, and/or that initial beliefs about μ and/or post-trial ϕ were widely inaccurate at baseline.

We will provide evidence that the training itself is unlikely to have (differentially) changed skills μ. At baseline, beliefs could be biased about both μ and (post-trial) ϕ. We will document widely held beliefs biased against the efficacy of female supervisors at baseline, i.e., incorrect beliefs about μ. The experimental design thus identifies misallocation at baseline relative to (easily attainable) production possibilities after the short trial period.

3.5 Summing Up: Three Forms of Misallocation

The framework allows us to define three types of misallocation of managerial talent along gender lines. Recalling that Δ_j represents the productivity gap of a line j when managed by the current male or marginal female supervisor, we denote:

1. *Weak Misallocation*: $\mu_B^M < \mu_B^F$
2. *Strong Misallocation*: $\Delta_S = \int_{j \in [0,1]} Y(\mu_B^F, x_j, \phi_j^F) - Y(\mu_B^M, x_j, \phi_j^M)\ dj$
3. *Misallocation at the Margin*: $\max(\Delta_j) > 0$ (among lines j supervised by men)

While weak misallocation denotes a form of discrimination against women that can be consistent with profit maximization, strong misallocation and misallocation at the margin are invariably costly to the firm. The absence of strong misallocation doesn't imply the absence of misallocation at the margin and is thus also not sufficient for establishing maximizing behavior by firms. The combination of our experimental design and data collection strategy allow us to examine these forms of misallocation. Data from the extensive baseline diagnostic exercise allow us to potentially identify *weak misallocation*.

The experimental assignment to lines identifies strong misallocation under the additional assumption that the experimental lines are representative of the whole set of lines in the factories, and misallocation at the margin under the assumption that the factories choose the lines with the highest Δ_j. As it was practically not possible to run the experiment on a random set of lines, and we are unable to precisely measure Δ_j, we cannot be sure which the experiment identifies. However, note that strong misallocation implies misallocation at the margin, and hence the experiment will identify a sufficient condition for misallocation that is costly to factories. Because managers may have information about ϕ_j (or Δ_j) that is unobservable to the researchers, their decisions with regard to retention of trainees after the trial may be informative above and beyond measures of productivity on the lines managed by the trainees. We therefore also test for misallocation at the margin using the retention decisions of the factories after the initial two-month trial. If the gender ratio among the trainees that the factories retained as supervisors is larger than the gender ratio of supervisors in the factories at baseline, this suggests that the share of lines better served by female supervisors is larger than the baseline share of women among supervisors. We discuss further assumptions underlying this test when we implement it in section 6.

4 Design and Data

This section describes the implementation of the experiment and the data we collected. We highlight the three design challenges identified in the previous section - identification of the marginal male and female supervisors, allocation of trainees to experimental lines and measurement of performance.

4.1 Experimental Design

For the main trial of the project we worked with 24 factories that are suppliers of a large UK-based buyer.[12] To identify the marginal candidates for promotion to line supervisors, we asked participating factories to select an equal number of male and female sewing workers to be sent to a six-week classroom-based supervisor training program, developed by the German Development Corporation GIZ. The researchers paid the direct cost of training, though factories continued to pay worker salaries.

The training program was offered for free mainly as an incentive for factories to

[12] As is typical for factories in Bangladesh, these factories also produce for other buyers.

participate in the experiment and identify the marginal candidate supervisor. As shown and discussed in more detail in Appendix B, we do not find many effects of training on measured skills. Given that our analysis is based on comparisons of female and male trainees, both of whom participated equally in the training, the internal validity of the analysis should not be affected. The training might, however, affect the external validity of our results to situations in which firms consider promoting more women without a prior training course. With this in mind, we measured a rich set of trainee skills both before and after the training course, to understand the effects of the training.

The selected trainees were all current non-supervisory workers. To determine the total number of workers nominated for the program, we asked each factory to consider the expected demand for new supervisors in the factory in the six months following training. Given that much of the training material was written, we asked that nominated workers have at least basic literacy skills. We scheduled four training rounds starting between March and May 2014, with half the nominees from each factory randomly allocated to one of two training rounds, either rounds 1 and 3 or rounds 2 and 4. Within factories, trainees were randomly assigned to early or late training rounds, stratified at the gender level, to stagger their return to the factories.

The factories initially selected 121 female and 100 male candidates. Twenty-one female candidates declined to start the training course, either because they decided they did not want to be supervisors or because of resistance from their families. Given the use of written material, we screened on a test of basic literacy. Eleven female and 18 male candidates did not pass the literacy test we conducted at the start of the training. After the initial assessment, 13 female and four male candidates dropped out of the program, and three further female and six male candidates were excluded for other reasons. The final remaining sample, all of whom completed the training course, consisted of 73 female and 72 male trainees.[13]

We then agreed with the factory management that the trainees would work for the first two months after the training as co-supervisors on a line randomly selected from a set of lines nominated by the factories. After the trial, factories were then free to return trainees to non-supervisory positions, keep them as supervisors on the trial line, or move

[13]Five of the factories sent operators to the first training session but stopped sending the agreed upon trainees in the second round. However, given the within-factory randomization of trainees into training rounds, stratified for male and female trainees, this does not bias the estimates from the sample of firms finalizing the trial.

them to other lines as a supervisor.

4.2 Data

We collected three types of data: (i) surveys of the trainees, other workers, and other supervisors before and after the training, and after the trial period (ii) productivity data for a period of 12 months covering the whole trial from all lines in a subset of factories that provided data of sufficient quality, and (iii) a managerial lab-in-the field experiment at 48 separate factories that participated in a pilot run of the project a year before the main trial. We describe the survey data here in more detail, and the productivity and managerial experiment data later when we bring them into the analysis.

We carried out surveys of trainees, other workers and current supervisors on four separate occasions. A prime focus of all surveys was to collect information on beliefs about and preferences for female supervisors held by the respondents from various different positions in the factories. The first survey collected detailed information from trainees themselves on the first day at the training center, including attitudes, assessments on knowledge of machine and production processes, communication, teaching and leadership skills, and numeracy, literacy and non-verbal reasoning skills. In the second survey, a few days before the arrival of the trainees on their designated lines, we surveyed both the trainees at the training center and workers in the factory. In the factory, we surveyed 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. Finally, we surveyed the floor managers and the factory-level production- and HR managers.

The third survey was conducted in the factory just after the two-month trial ended. We again surveyed three randomly selected operators and the co-supervisors and line chiefs from the lines that were nominated for the trial, and on 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 four and a half months after the trial ended for those trained in the early rounds, and ten weeks after the end of the trial for those trained in the late rounds. Because each factory participated in two training rounds, the first follow-up survey was conducted on two separate days, roughly two months apart. However, in an effort to minimize disruption of work at the factory, the second follow-up survey was conducted on a single day for trainees and lines from both of the training rounds in which the given factory participated. This meant that we were unable to survey both the originally assigned lines and the lines where trainees actually worked. But as factories were free to move workers after the two-month trial period, results on the original trial lines would be difficult to interpret. We therefore chose to survey supervisors and random workers from the lines where trainees were actually working as supervisor at the time of the second follow-up. We surveyed trainees who had left the factories by that time by phone, and we continued to conduct bi-monthly telephone follow-up surveys with trainees for around half a year after this fourth survey round in the factories. Finally, trainees kept a daily diary during the two months of the trial period, noting the line they worked on each day, and if, and how many, workers they were supervising on these lines on that day. We use these data to identify where the trainees actually worked during the trial.

5 Comparison of Trainees and Beliefs at Baseline

This Section describes evidence at baseline. We focus on three aspects: we document beliefs about the relative supervisory skills of females and males, proxies for μ; we compare actual measures of skills of the female and male trainees selected by the factories; and we compare the trainees against representative sets of line operators and supervisors. We show that the beliefs are inconsistent with the evidence from the basic comparison of female and male trainees. As we argued in the conceptual framework, this gap between beliefs and measured skills could translate into lower performance of female supervisors on the line – an issue we investigate in Section 6.

5.1 Attitudes Towards Female Leadership

We begin by documenting negative beliefs about the suitability of females as supervisors. We examine attitudes towards female leadership in the 24 participating factories using baseline surveys with workers, supervisors and managers. In pilots, managers and supervisors reported eight different skills as being important for supervisory roles (see Figure 3). At baseline, we asked respondents whether they believe that men are more able, women are more able, or men and women are equally able in each of these eight skills. We code the answers as -1 if the respondent thinks that men are more able, 0 if they think women and men are equally able, and 1 if they think women are more able.

The upper part of Figure 3 shows the results on the eight skills and the average response across the eight skills. The responses are shown separately for sewing workers, supervisors, line chiefs, HR managers, and Production managers. Sewing workers have the most negative beliefs about women as supervisors, reporting that they are less able than men on each of the eight skills. The gender gaps are largest for "hard" skills, particularly "understanding machines", and "organizing resources". Fully 77 and 92 percent of female and male workers, respectively, think that men understand machines better, while nine and zero percent, respectively, say that women understand machines better. The gaps are smallest for "soft skills" such as "motivating workers", "corresponding with managers", and "teaching new techniques".

Line supervisors report beliefs that are indistinguishable from operators, except for "teaching new techniques" where they see no gender difference. As we move up the organizational hierarchy, the beliefs favor women more, particularly for the "soft skills". Production and HR managers believe that there are no gender differences in "correcting mistakes", "motivating workers", "corresponding with management", and "teaching new techniques". However, on the "hard" skills of "understanding machines" and "organizing resources", the beliefs of higher-level managers are no different of those of workers and supervisors, with 82 percent of factory level production and HR managers stating that men understand machines better.

The lower part of Figure 3 disaggregates the evaluations from sewing workers by the gender of the respondent. Both female and male workers rate women lower on each of the eight skills, though the gaps are larger for male respondents on all eight skills. We also asked the female and male trainees themselves to rate the supervisory skills of females and males, with the results shown in the same figure. The responses of male trainees lie between those of male and female workers, while female trainees report lower gender gaps than female workers on six out of eight dimensions. But even the female trainees rate men better in five of the eight skills, while reporting no gender gap on the three "soft skills" of "motivating workers", "corresponding with management", and "teaching new techniques".

Table 1, Columns 1-2 shows that the differences in the beliefs held by female and male sewing workers, aggregated over all eight skills, are highly statistically significant even when controlling for a large number of demographic and socio-economic respondent controls. These beliefs translate into both female and male workers saying they prefer to work under male supervisors. We asked each worker surveyed at baseline whether s/he prefers to work under a male supervisor, a female supervisor, or is indifferent between the two. Among respondents, 70 percent of males and 50 percent of females say they prefer to work for men, while 7 and 27 percent, respectively, report they prefer to work for female supervisors. The remaining 23 percent of both male and female production workers report being indifferent. Columns 3 and 4 of Table 1 show that the preference for male supervisors is significantly stronger among men, though even for women there is a statistically significant preference for male supervisors. These gaps remain when adding respondent controls and factory fixed effects. Meanwhile, the 34 percent of the respondents reporting that they had previously worked under female supervisors do not report different beliefs about skills (Column 2 of Table 1), but do report stronger preferences for female supervisors (Column 4). The effect of exposure to female leadership does not differ significantly between male and female respondents (results not shown).

While the survey questions underlying Figure 3 and Table 1 do not target the marginal supervisor directly, we also asked the line chiefs and production managers to rate the trainees they were involved in selecting for training. We asked the line chiefs to compare the female and male trainees on all eight skills; because of time constraints, we asked production managers only for one aggregate ranking. The relative skill patterns are similar to the general comparisons of men and women as supervisors shown in Figures 3, though these beliefs are measured less precisely due to the much smaller number of respondents. We show in the next subsection that these widely held beliefs are, prima facie, inconsistent with the results of the extensive diagnostic conducted before the training in which we found little difference in skills of female and male trainees.

5.2 Comparing the Female and Male Trainees

Using the data collected during the first day at the training center we see few differences in characteristics between the female and male trainees in our final sample (Table 2). Note that the design was not intended to yield a set of female and male trainees that are balanced on observable characteristics, but rather to identify the marginal female and male supervisors, the workers that the factories would chose to promote if they would need to find additional (female and male) supervisors. Among the diverse set of characteristics that we measure, only two differ significantly; female trainees scored on average lower on the numeracy test (but not on the literacy nor on the Raven test), and they expressed lower confidence into their ability as supervisors. However, when we repeated the confidence questions on the last day of training, when the female and male trainees knew they would soon be trialed as supervisors, the gender gap in confidence levels had notably shrunk, as shown in the bottom two rows of Table 2. The majority of the measures are similar for female and male trainees. From a demographic perspective, they are of equal age (25 years), educational attainment (8 years of schooling), seniority in the factory (around 3 years) or the garment sector (around 6 years).[14]

We also measured more direct proxies for managerial talent. During the baseline diagnostic, we measured proxies for each of the eight skills for which we elicited beliefs as described above. Male and female trainees showed no significant difference in group game exercises designed to test communication and leadership skills.[15] Crucially, they scored no differently on a battery of 86 technical questions on the garment production process (see Appendix E for a copy of the test). This last dimension is of particular importance for two reasons. First, technical knowledge was ranked as the most important trait of a good line supervisor by ordinary workers.[16] Second, as we have documented, employees at all levels expressed a belief at baseline that women are particularly weak in this dimension.

The comparison of means is arguably not the relevant comparison. Figure 4 shows the full distribution of technical knowledge scores for female and male trainees. Females were only 16 percent of promotions in the period prior to the baseline. As we elaborate in more detail further below, a conservative measure indicates that the factories retained 86 of the 145 trainees as supervisors after the trial period. If factories had followed their

[14]We note that female trainees are slightly more likely to be married, in line with female workers in the sector overall (Menzel and Woodruff 2019). This could indicate that female trainees have more household responsibilities besides their work at the factories. We control for these features in our analysis.

[15]For the communications game, the trainee had to explain a number of abstract figures verbally while other trainees attempted to draw them. We use the number of figures the other trainees could draw as a measure of communication skills. We also conducted a production game in mixed groups of female and male trainees, in which the 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", and we do not see a difference on this score either.

[16]Ordinary workers found technical knowledge of machines as the most important skill for a line supervisor, while current supervisors and higher up managers considered "teaching new techniques" and "motivating operators" as most important.

baseline promotion ratio of 16 percent women, they would have promoted 14 women and 72 men. The relevant comparison is thus the 14th best female and the 72nd best male. Given that there were only 72 male trainees in our sample, this would be the worst male trainee. So we instead conservatively highlight in the upper part of Figure 4 the male trainee at the 20th percentile of the knowledge test score distribution. As the vertical lines indicate, the score of the 14th best female trainee far exceeds that of the 20th percentile male, indicating that baseline promotion practices did not equate technical knowledge at the margin, as measured in our knowledge diagnostic.

5.3 Trainees as Marginal Candidates: Comparison to Other Workers and Supervisors

It is useful to compare trainees selected by the factory with average sewing workers and line supervisors in the factory. If the factory selected marginal candidates, we expect the trainees to be relatively more similar to existing supervisors rather than to sewing workers. Table 3 provides supporting evidence. First, we compare the trainees with random workers of the same sex we surveyed from the lines the factories nominated for the trial. Both female and male trainees have significantly higher educational attainment than the operators. On the other hand, we see no difference in age. Compared with the average male worker, male trainees have longer tenure in the factory but not the sector; there are no differences between female trainees and workers in either factory or sector tenure.

Second, we compare trainees to existing supervisors. Compared to current male supervisors from the nominated lines, male trainees are younger, have spent less time in the sector, and worked at fewer garment factories, which is not surprising given that they are just about to be promoted to supervisor. Indeed, the average age of the male trainees is almost exactly the same as the age of promotion reported by the average existing supervisor. The trainees also have somewhat lower years of schooling. The patterns are similar when we compare female trainees to the small number (10) of female supervisors we surveyed, except that the female trainees are younger than the age at which the existing female supervisors reported to having been promoted.

In sum, the diagnostic exercise and baseline surveys provide evidence consistent with a bias in beliefs widely held inside the factory: women are rated to be worse supervisors in general, and particularly so on knowledge of machines – a dimension we can accurately measure and on which we saw no meaningful difference between male and female trainees. The bias in beliefs against women is widely shared across layers of the hierarchy and particularly pronounced among potential male subordinate workers.

6 Female vs Male Trainees: Experimental Results

This section first shows that, when randomly allocated to supervise production lines, female trainees initially have lower productivity on their lines than male trainees, and receive worse reviews from subordinates. We then show that this gap is short-lived: a few months after the end of the trial, female trainees have caught up with men in both dimensions. Furthermore, after the end of the trial, when factories were free to demote trainees again to non-supervisory positions, they retained a significantly higher share of female trainees as supervisors than what could have been expected given baseline levels. This indicates that factories were under-promoting female workers to supervisors before our intervention. The section concludes relating the evidence to the notions of misallocation introduced in the conceptual framework in Section 3.

6.1 Randomization Balance and Compliance

Before presenting the experimental results, we describe randomization balance and compliance with the experimental protocol. In Appendix A, Table A.2, we test whether the randomization of the nominated lines to receiving male and female trainees is balanced on three types of observables: average operator characteristics (from the randomly selected operators surveyed before the trainees return from the training), average line supervisor observables from the same surveys, and different variables in the production data (e.g. productivity, daily work time, number of workers on line, etc.). Among 39 variables, we detect imbalances at the 10 percent statistical significance level for 4 variables, in line with expectations. F-tests indicate joint imbalance of worker characteristics at the five percent level, and for production data at ten percent. We show that our results are robust to controlling for these variables (individually or jointly in a PDS Lasso framework).

Table 4 shows that compliance with the agreement that trainees would work as assistant supervisors was fairly high: 88 percent of the 145 trainees started their trial after completion of the training program. This compliance rate did not differ significantly by gender (column 1). However, compliance with the random allocation of workers to specific production lines was lower. Among the 88 percent of trainees who began their trial, 64 percent did so on their assigned line (Panel 3, column 2). This implies that 56 percent of all trainees started their trial on the randomly allocated line. Again, there is no significant gender difference in this compliance measure (Panel 1, column 2). Factory managers reported that non-compliance often resulted from a lack of internal communication of line assignments to floor-level managers. Consistent with this, we see that much of the line-level non-compliance involved the switching of two female, or two male, trainees. With this in mind, we can define a third measure of compliance by asking what share of trainees worked on a line randomly assigned to a trainee of the same sex. The compliance rate is higher by this measure, with 69 and 63 percent of female and male trainees working on a line assigned to their gender (Panel 1, column 3), or 78 and 71 percent, respectively, conditional on starting the trial (Panel 3, column 3).[17] We address this non-compliance by reporting ITT specifications.

6.2 Line-level productivity during the trial

We first compare lines allocated female and male trainees using daily line-wise production data, which we collected for all production lines in the 14 factories, where available.[18] The production data cover a period of twelve months starting from around 3 to 5 months prior to the start of the trial (depending on the training round of the trainee). We measure line productivity with a standard engineering measure used in the sector. Daily piecewise output of a production line of a given garment is multiplied by the garment's "Standard Minute Value" (henceforth SMV). The SMV is calculated by industrial engineers by breaking down the sewing process of a garment into its individual stitches and assigning a time value to each stitch under ideal production conditions. The time value of each stitch in the garment is then summed up.[19] The SMV typically

[17]Note that 86 percent of male and 89 percent of female trainees were trialed on any line nominated by the factory, conditional on doing the trial. This implies that half of both the male and female trainees that did the trial on a line not allocated to a trainee of their sex did the trial on lines outside the set of lines nominated by the factory.

[18]These 14 factories nominated 112 of the 145 trainees, 58 females and 54 males.

[19]The SMV is sometimes referred to as the SAM (Standard Allowed Minutes).

depends on the machinery the factory uses, and engineers sometimes allow for extra time for moving or folding the garment, or cutting thread. Thus, while the SMVs are consistent within factories, they are not necessarily consistent across factories. Therefore, we include factory fixed effects for all of our analysis, yielding within-factory comparisons of lines with male and female trainees. Once standardized by multiplication with the SMV, daily output is divided by the number of minutes of labor available on that day and line, the number of workers on the line times the minutes of operation on the day.

At the end of each production line there is a quality control table where quality inspectors check each piece. Defective pieces are not counted in the daily output used for productivity until the defects have been corrected. Instead, the ratio of defective pieces over the total output yields the defect rate, our second outcome variable, available from 17 factories. Finally, most factories track the daily number of absent workers, with 13 factories in our sample having this information available at the individual line level. We collected this data as our third outcome variable.

Table 5 shows that lines randomly allocated female trainees had significantly lower productivity during the two-month trial period when the trainees worked as assistant supervisors on these lines. The first two columns show ITT effects, comparing average productivity between lines randomly allocated female and male trainees, regardless of whether the allocated trainee worked on the line. Column 1 shows the ITT comparison for the period prior to the start of the trial. As predicted by random allocation of the trainees, there is no statistically significant or economically meaningful difference. However, during the trial period, lines allocated female trainees have on average more than four percentage points lower productivity, controlling for baseline productivity of the line (column 2). This amounts to an around seven percent lower productivity compared with lines allocated male trainees.

Given that non-compliance with the randomization protocol was not trivial, in columns 3 to 5 of Table 5 we compare the productivity of lines on which the female and male trainees actually worked, using data from the worker diaries. In column 3, we compare baseline productivity levels on those lines, finding no significant difference prior to the start of the trial between the lines on which female and male trainees actually did their trial. Controlling for baseline productivity, we find that lines actually receiving female trainees have more than six percentage points lower productivity (column 4), broadly in line with the fact that only around two thirds of the lines received a trainee of the randomly allocated sex. Given the selected placement of trainees, and the fact that female and male trainees differed, on average, in confidence and numeracy skills, column 5 repeats column 4, but adds a battery of observable line and trainee characteristics controls. These include the average SMV of the garments produced, the average order size, the number of workers on the line, or the average daily runtime of the line. On the trainee side we control for age, schooling, and the measures of numeracy or confidence in own abilities. We apply PDS Lasso to select the relevant controls from those listed above, including squares of, and indicator variables for missing values in, each control (Duflo (2018); Belloni et al. (2016)). These controls have only a very small effect on the results.

While we see clear patterns in the effect of female and male trainees on line productivity, there are no similar patterns for defect rates or worker absenteeism during the trial period. As shown in Columns 1 to 5 of Table A.3 in Appendix A, the only effect that we find is a significant increase in defect rates during the trial period on lines on which female supervisors actually worked. This effect is significant only when we add line and trainee controls selected by PDS Lasso, and never in the ITT specification. Thus, while the defect rate results are in line with lower performance of lines receiving female trainees during the trial, the effect is not robust.

6.3 Performance during the trial: subordinate and co-worker evaluations

We next look at evaluations of the trainees by co-supervisors and subordinate workers. At the end of the two-month trial period, we surveyed two co-supervisors and three randomly selected workers on each line that was allocated a female or male trainee. We asked them to evaluate on a scale from zero to 10 both a “typical supervisor” and the trainee supervisor on their line along the eight dimensions shown in Figure 3. Columns 6 to 8 in Table 5 report the results of regressions with each respondent's average evaluation across the eight dimensions as the dependent variable. We find no statistically significant difference in the average evaluation of female and male trainees by co-supervisors on the same lines, either without (Panel 1, column 6) or with (Panel 1, column 7) demographic controls. Columns 6 and 7 are labelled “ITT”, but note that co-supervisor had no way to evaluate trainees that never worked on their line, and hence data are missing for this form of non-compliance.[20] Meanwhile, column 8 compares evaluations from any lines on which female and male trainees actually worked, similar to the specifications using production data in columns 3 to 5 of Table 5. The estimated difference is slightly larger but far from significant at conventional levels. All of the regressions in columns 6 through 8 control for the respondent's average ranking of a typical supervisor. The additional controls in columns 7 and 8 are selected again using PDS Lasso, from among respondent and trainee controls, similar to column 5.

Panel 2 of Table 5 replicates the analysis of survey responses with those from subordinate workers. Given the larger number of female respondents among workers, the differences between evaluations of female and male trainees are shown separately for female and male worker respondents. Controlling only for the respondent's rating of a typical supervisor (column 6), female workers evaluate female trainees on average around 0.6 points lower than male trainees on the 10-point scale, while male workers around 1.2 points lower. Both gaps are statistically significant at the five and one percent level, respectively. While the difference between male and female respondents is large, it is not statistically significant at conventional levels. On the other hand, the gap estimated from the combined responses by male and female workers is significantly different from both zero and from the (small and insignificant) gap elicited from co-supervisors. Controlling for respondent and trainee characteristics selected by PDS Lasso (column 7) does not affect these results in a meaningful way, nor does expanding the sample to all lines on which male of female trainees actually worked, regardless of their random allocation status (column 8).

The more negative evaluation of female supervisors by their sub-ordinate workers after their first two months on this position mirrors the more negative evaluation of female village heads in close-by West Bengal (India) at the end of the first election cycle in which the village head position was reserved for women (Beaman et al. (2009)). However, in contrast to our results, the female village politicians outperformed their male

[20]We managed to survey workers from 131 of the 145 lines nominated for the trial in the first follow-up survey, resulting in an attrition rate of 9.7 percent. Out of the 14 trainees that attrited, 13 are due to whole factories attriting from the trial. Given that randomization was stratified at the factory level, potential bias from this attrition should be limited. A trainee was trialed on 101 of the 131 lines, resulting in an available rating of the trainee. Evaluations are also available for 20 further trainees that were not trialed on a line originally nominated by the factories for the trial.

counterparts on some objective metric (public good provision, corruption). This should caution us, however, to necessarily interpret the lower evaluation of female supervisors as being caused by the lower productivity on their lines. We return to this discussion in section 7.

6.4 Post-Trial Outcomes

Columns 4-6 of Table 4 describe outcomes for trainees after the trial, when our agreement with the factories on the random allocation of trainees had ended. We find that 53 percent of female and 67 percent of male trainees continued to work as supervisors on some line, with a p-value of this gender difference of 0.07 (column 4). Those continuing to work as supervisors represent 62 and 75 percent, respectively, of those who started the trial (Panel 3, column 4). However, the majority of those continuing as supervisors were moved to another line. Only 36 percent of male, and 34 percent of female, trainees that did their trial worked for at least another three months on the same line as during the trial, and only 21 and 15 percent, respectively on their originally allocated line (Panel 3, column 5).[21] Finally, 22 percent of the trainees had left the factory within half a year of the trial's end, with no significant gender difference (Panel 1, column 7).[22]

Table 6 explores the productivity differences between the lines of the female and male trainees who continued to work as supervisors after the end of the trial, on the lines where they actually worked. Given that the majority of trainees were relocated after the trial, we do not report ITT effects. As shown in column 1, retained female trainees worked after the trial on lines that were insignificantly less productive before the start of the trial. Controlling for this baseline productivity, we find only a very small and highly insignificant gender gap in line productivity in the post-trial period of around 1 percentage point (column 2). Adding again a rich set of respondent and trainee controls selected through PDS Lasso does not change the result (column 3).

[21] Note that generally line supervisors work on fixed assigned lines and are reallocated only rarely. Existing supervisors at baseline in our sample factories report to have worked on average 1.9 years on their current line as supervisors (median: 1.3 years). Thus, we interpret the observed line-moves of retained trainees after the trial as one-off movements from the lines we randomly allocated them to initially, to those lines which the factory deemed as optimal fit for them.

[22] Among men, 50 percent of those not trialed and 17 percent of those trialed had left, while for women the numbers are 55 and 19 percent, respectively. In further phone surveys we conducted after the second follow-up survey in the factories, two male trainees and one female trainee reported they were working as supervisors at other factories, while four more trainees reported they were non-supervisor workers in other factories.

If we replicate the basic comparison of female and male trainees during the trial with the sample limited to those trainees who continued to work as supervisors post-trial, the gap during the trial period remains very similar to that in the overall sample during the trial, and highly significant (Table 6, column 4). This suggests that the observed closure of the productivity gap of female trainees between the trial and post-trial period is not due to factories retaining as supervisors only those female trainees with the highest productivity during the trial. Furthermore, even looking at the small sample of 30 trainees whose modal line in the post-trial period coincides with the line on which they did their trial, we see a similar, though insignificant, productivity gap during the trial (column 5). This suggests that the closure of the gap is not driven by reallocation of female trainees post-trial to lines with, for example, a lower bias ϕ_j.

As with the trial period, we also elicited evaluations of supervisor performance from co-supervisors and subordinate workers of the retained trainees at the second follow-up survey, two to five months after the end of the trial period. For co-supervisors, there is again no gender gap in the evaluation of the trainees (Table 6, Panel 1, column 6). The initial gap in the evaluations by workers, however, disappeared. Two to four months later, neither male nor female workers evaluate the retained female trainees lower than male trainees (Panel 2, column 6). Again, we can ask whether the closure of this gap relative to the trial period is due to those (female) trainees with the lowest evaluations after the trial not continuing as supervisor. As with the production data, the results suggest that this is not the case. Using data from the trial period for those who are later retained as a supervisor, we find the gap in evaluations remains almost unchanged and is still highly significant for male workers, though it is somewhat smaller than in the full sample for female workers (column 7). Thus, it does not appear that the closure of the evaluations gap from trial to post-trial period is driven by factories retaining only those (female) trainees that had the highest evaluations by workers.

Finally, given that at baseline, supervisors and workers expressed beliefs that women have lower supervisory abilities, we might ask whether the strength of these beliefs measured on a given line is associated with the lower productivity or evaluations of female trainees assigned to those lines. We do this in Appendix A.4, finding that controlling for the best measures we have of baseline beliefs about relative female/male supervisory skills do not meaningfully change the estimates. However, as we discuss in the appendix, the available measures are based on responses of a small percentage of the operators on each line, and so are likely to be noisy.

6.5 Female Share among Retained Supervisors: Misallocation at the Margin

As discussed in the conceptual framework section, comparing the average performance of male and female trainees likely biases the comparison against the marginal female supervisor for several reasons. Assuming the managers understand something about the bias in beliefs of subordinate workers, an assumption supported by the co-supervisor performance ratings, they will take this into account when they make decisions about which trainees to retain as supervisors. Given this, a more informative test of misallocation at the margin compares the ratio of women among the trainees retained by the factories as supervisors after the end of the trial against the share of women among supervisors in the factories at baseline.

Out of the pool of 143 trainees, 86 still work as supervisors by the time of the second follow-up survey.[23] Of those, 38 were women, representing a 44 percent share. Data from Menzel and Woodruff (2019) reveal that of 101 promotions to supervisors in 36 garment factories in Bangladesh around the time of the trial the share of women among *promotions* to supervisors is around 16 percent (see Appendix C for more details).[24] A t-test that takes into account the different sample sizes from which the shares of 44 and 16 percent were estimated reveals that the 44 percent ratio in the experiment is statistically different from the 16 percent counterfactual marginal ratio (p-value < 0.001).

Factories may retain a larger share of female trainees than they find optimal due to perceived experimenter demand, or perceived costs of demoting trainees. In fact, many trainees were demoted, so these do not appear to be particularly strong assumptions. Furthermore, based on our framework, another way to check for the efficiency of the share of women among the retained supervisors is to test again whether it equalizes trainee skills

[23]We define promoted here as working as a supervisor at the time of the second follow-up survey. Alternative definitions of working as a supervisor for at least 30 days or at least 60 days after the end of the two-month trial produce higher promotion rates of both female and male trainees, but the gender ratio remains unchanged (increases minimally in favour of female trainees).

[24]The 16 percent female promotion percentage is higher than the seven percent female supervisor rate that we estimate among the stock of supervisors, suggesting either that promotions of women are increasing over time or that women remain in supervisory positions for a shorter amount of time.

at the margin. For that, in the lower part of Figure 4, we repeat the exercise from the upper part of the same Figure, but now comparing the technical knowledge scores of the 38th best female trainees against that of the 48th male trainee, given the numbers of retained female and male trainees. We find that these two have precisely the same test score in our data, suggesting that factories moved much closer to an efficient gender ratio in promotion decisions. Overall our findings indicate that a larger share of lines in the factories, including the marginal lines from our conceptual framework, were better served by female supervisors than the share that had female supervisors at baseline, pointing towards the presence of misallocation at the margin in these factories.

7 Mechanisms

Do lines managed by the female trainees initially underperform because the female trainees are less skilled than the male trainees (lower μ), or because workers bestow less "authority" on the female supervisors, as captured by ϕ in our framework (Abel (2019); Ayalew et al. (2018); Grossman et al. (2016))? The combination of few objective skill differences and strong negative baseline beliefs about the supervisory skills of women provides prima facie evidence that beliefs are misaligned. The few differences we do find, such as the significant differences in confidence into own supervisor skills, could in principle drive the initial productivity differences. Note, however, that controlling for trainee's confidence levels did not affect our results. However, the misaligned beliefs may be a reason why workers bestow less "authority" on female supervisors, for example paying less attention to their instructions. In this section, we present two additional results that are consistent with an explanation based on misaligned beliefs, but more difficult to reconcile with a pure skill-based explanation.

7.1 Management Simulation

The first additional piece of evidence favoring the "authority" hypothesis comes from a lab-in-the-field management simulation exercise conducted with 169 trainees at a set of 48 similar factories that participated in an extensive pilot of this project around a year before the implementation of the main trial. These trainees were similarly selected by the factories and went to the same training program as the trainees from the main trial.

The simulation was conducted around two months after the trainees in this pilot returned from the training, with most of them having worked as supervisors in the meantime, as per our agreement with the factories. The trainees are thus similarly skilled and experienced as the trainees from the main experiment after the completion of their two-month trial. However, as opposed to the main experiment, the pilot factories could freely select the lines on which they put the trainee supervisors during the trial period. We thus implemented the management simulation to measure the trainee's productivity as supervisors when matched with randomly selected workers. Due to the logistical complexity of implementing the management simulation in the factories, we did not conduct them during the main experiment, where we could exploit the random allocation of trainees to lines during the first two months after the training.

The simulation allows us to study, for example, whether the effectiveness of female trainees differs when they supervise female or male workers. On the lines in the factories, this share fluctuates and factories typically do not keep day-to-day records of which workers work on which lines.[25] Moreover, line productivity will be affected by bottlenecks on specific processes, and factories do not record the gender of workers on these key processes. We provide more detail on the trial phase in Appendix D. In particular, Table D.3 compares demographic characteristics of male and female trainees in both samples, finding few significant differences.

We find that these female trainees performed better than their male counterparts in these simulations, in which each trainee was matched with a team of two randomly selected workers from their factory. The trainees were tasked with explaining to the two workers one of two types of stylized production exercises, either sorting buttons according to their color and size, or building figures with Lego pieces. The first exercise was designed to be familiar to people working in the factories, while the second exercise to test the ability to supervise unfamiliar tasks. The trainees received cash pay-offs proportional to the output of their team.[26] Each simulation consisted of four rounds, with the pay-off of the trainee determined in slightly different ways in each round (twice based on the total

[25] We were able to collect administrative data from eight factories on the share of men on production lines, but did not find that variation in this share affected our results based on productivity data. However, this administrative data offered only monthly snapshots of allocation of workers to lines, but not daily changes.

[26] The simulations were done during the work-time of the workers, and so the workers in the teams received their standard pay for the time participating in the simulations, but no bonus for performance in the game.

output of team, and once each on output of the team-member with the largest or smallest output, respectively). We analyze the results of the simulation in Table 7, which regresses the payoffs on the trainee-game round level on gender of trainee and fixed effects for the two types of games.

Female trainees generate almost 0.3 standard deviations higher payoffs, a difference that is statistically significant. The difference persists controlling for trainee characteristics (Table 7, column 2), for fixed effects for the team of two operators (column 3), and for dummies indicating whether the trainee was actually tried out, or retained as line supervisor (columns 4 and 5, respectively). Column 6 allows the results to be heterogeneous according to the composition of the production team. We see that female trainees outperformed male trainees only when they supervised a team of two women. When at least one of the two workers was male, productivity was significantly lower. This result is consistent with men having more negative baseline beliefs about female supervisors, and with attitudes based on such beliefs affecting female supervisors, instead of some unobserved skill differences.[27]

7.2 Preference Update for Male and Female Supervisors

The second piece of evidence more in line with the "authority" mechanisms stems from a positive update of preferences for female supervisors that we observe among workers randomly allocated a female trainee supervisor, as shown in Table 8. We elicit preferences for male versus female supervisors from workers at the end of the two-month trial in the same way as we did at baseline, coding a preference for male supervisors as -1, for female supervisors as 1, and indifference as 0. Under the assumption that workers prefer to work under supervisors they believe to be more skilled, this result indicates that the random exposure corrected initial beliefs about female supervisors that were too negative. This conclusion is furthermore supported by the positive effect of exposure on preference being driven by male workers, which held particularly negative, and therefore potentially misaligned, beliefs at baseline. While a simultaneously lower evaluation of female trainees and positive updating on preferences may appear puzzling,

[27] Female trainees may have outperformed males by providing more effort in the management simulations to "prove" that they are equal to their male counterparts. However, it is not clear why female trainees would not similarly provide more effort as supervisors on their lines. Furthermore, even if that was the case, note that this additional effort would have lower returns when managing teams with male workers.

it is in line with beliefs having been excessively negative at baseline, especially among men. It is also consistent with Beaman et al. (2009), who find that male constituents initially evaluated female village politicians lower, while they positively updated their perceptions of female leader's abilities at the same time. They also argue that due to behavioral biases, evaluations could still reflect earlier negative beliefs (Mullainathan and Washington (2009)), and be therefore biased themselves.

Finally, the results on updated preferences also speak to the question whether the lower authority bestowed on female supervisors would arise from taste-based discrimination against female supervisors or negative beliefs. A long theoretical and empirical literature has argued that while beliefs can be quickly updated, tastes are typically assumed to be stable over time and much less sensitive to new information (Bohren et al. (2019); Beaman et al. (2009); Bisin and Verdier (2001); Stigler and Becker (1977)). The rapid update in preferences to female supervisors suggests that excessively negative beliefs rather than more deeply rooted tastes cause the lower authority exhibited towards female supervisors.

8 Discussion and Conclusion

We show results from a factory initiative to promote more female workers to supervisor positions. Female supervisors initially have lower productivity on their assigned lines, but results suggest that they catch up with other lines after four to six months. These and other results are best rationalized by a model in which otherwise identical workers bestow different levels of authority on female and male supervisors, affecting the way their supervisory skills translate into team output.

We do not find clear evidence for weak misallocation, in that the marginal female trainees are not clearly more qualified than the marginal male trainees based on measured baseline skill levels. But efficiency depends on misallocation at the margin, and the results do suggest there was such misallocation. The significantly higher share of female trainees among the retained supervisors, compared to the baseline share in the sector, suggests that the baseline share of female supervisors was lower than the share of lines that had supervisor openings and would have fared better with a female supervisor at the time of the RCT. This suggests that indeed, at baseline, some female candidates for promotion to supervisors were overlooked in favor of less suitable male candidates.

The experiment illustrates the difficulty of isolating misallocation at the relevant margin in real-world settings. It is not clear either how well managers understood the amenability of lines to female leadership or the attributes to take into account when selecting the marginal female supervisors. On the first, our framework assumes that the amenability of lines to female leadership (ϕ_j) varies across lines, but is fixed across time. The results suggest that the beliefs of subordinate workers adjusted rapidly, increasing the amenability of lines exposed to female leadership. Moreover, there is anecdotal evidence of learning by factory management with regard to ϕ_j. Two of the participating factories established all-female production lines after our final survey, but before we shared any information about the experimental results. That they did this well after they were free to move the trainees to other lines indicates that they learned that amenability matters, and that it can be manipulated.

Evidence from baseline surveys indicates that managers found it more difficult to select the marginal female supervisor. Line chiefs, who are typically involved in promotion decisions from workers to supervisors, reported lower confidence into their ability to identify the marginal female supervisors, compared to the marginal male supervisors. We asked 94 line chiefs to evaluate the trainees from their lines, and to state how confident they were in their evaluations, on scale from zero to 100 percent. While they reported an average confidence level of 84 percent for male trainees, their confidence in selecting female trainees was almost 10 percentage points lower on average, with a p-value for the difference of less than 0.015.[28]

Learning in this regard is hampered by excessively negative beliefs about women as supervisors. Because subordinates' beliefs affect the performance of the female supervisors who are promoted, these beliefs may be self-reinforcing. Factories had promoted some women to supervisory roles before. However, if it is more difficult for the factory management to identify the "true" marginal female trainee, a lower learning rate from experimentation may result (Uckat and Woodruff (2020)). Risk aversion and negative tastes towards female supervisors may further reduce the rate of experimentation (Beaman et al. (2009)). Due to their much higher ex-ante promotion likelihood, men may also

[28]Variations in these confidence differences across factories or line chiefs do not, however, explain the gender gaps in productivity or evaluations we found in the paper. However, the variation in these confidence measures across factories may be noisy, and our sample may be too underpowered to detect such effects. Results available on request.

more actively signal to factory managements their skills for, and interest in promotions.

Our results, particularly those on updates to preferences for the gender of supervisors, mirror the findings from Beaman et al. (2009) and Chattopadhyay and Duflo (2004) that exposure to women in political leadership positions positively affects beliefs about their skills and abilities. Our paper shows that these results also hold for leadership positions in firms. This is an environment in which the effectiveness of leaders likely even more depends on the attitudes displayed towards them by subordinates, peers, and superiors. It is also one in which competitive pressures should attenuate bias. In terms of policy implications, given the positive spillovers into households, families and the broader society that have been shown to come with increased power and income of women (Uckat (2020); Duflo (2012)), our results suggests that targeted subsidies for employers to overcome initial, temporary frictions from promoting more women in the workplace can be promising tools for policy makers.

References

Abel, M. (2019). Do Workers Discriminate against Female Bosses? Working Paper, Middlesbury College.

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.

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.

Ayalew, S., S. Manian, and K. Sheth (2018). Discrimination from below: Experimental evidence on female leadership in Ethiopia. WPS 079, Center for Effective Global Action.

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.

Becker, G. S. (1957). The Economics of Discrimination. University of Chicago Press: Economic Research Studies.

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.

Ben-Yishay, A., M. Jones, F. Kondylis, and A. M. Mobarak (2019). Gender gaps in technology diffusion. *Journal of Development Economics 143*.

Benson, A., D. Li, and K. Shue (2019). Promotions and the Peter Principle. *Quarterly Journal of Economics 134*(4), 2085–2134.

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.

Bisin, A. and T. Verdier (2001). The Economics of Cultural Transmission and the Dynamics of Preferences. *Journal of Economic Theory 97*(2), 298–319.

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

Bohl, D. L. (1996). Minisurvey: 360-Degree Appraisals Yield Superior Results, Survey Shows. *Compensation and Benefits Review 28*, 16–19.

Bohren, J. A., A. Imas, and M. Rosenberg (2019). The Dynamics of Discrimination: Theory and Evidence. *American Economic Review 109*(10), 3395–3436.

Boisjoly, J., G. Duncan, M. Kremer, D. Levy, and J. Eccles (2006). Empathy or Antipathy? The Consequences of Racially and Socially Diverse Peers on Attitudes and Behaviors. *American Economic Review 96*.

Boudreau, L. (2019). Multinational Enforcement of Labor Law: Experimental Evidence from Bangladeshs apparel sector. Working Paper, Columbia University.

Boudreau, L., R. Heath, and T. McCormick (2020). Migrants, Information, and Working Conditions in Bangladeshi Garment Factories. Working Paper, Columbia University.

Bowles, H. R., L. Babcock, and L. Lai (2007). Social Incentives for Sex Differences in the Propensity to Initiate Negotiation: Sometimes It Does Hurt to Ask. *Organizational Behavior and Human Decision Processes 103*(1), 84–103.

Cajal, J. (2016). Searching for Trade Partners in Developing Countries: Testing Firms in the Fast Fashion Industry. Working Paper, Graduate Institute (IHEID), Geneva.

Cajal, J., R. Macchiavello, and G. Noguera (2019). International Buyers Sourcing and Suppliers Markups in Bangladeshi Garments. CEPR Discussion Paper 13482.

Carlana, M. (2019). Implicit Stereotypes: Evidence from Teachers Gender Bias. *Quarterly Journal of Economics 134*(3), 1163–1224.

Carrell, S., M. Hoekstra, and J. E. West (Forthcoming). The Impact of College Diversity on Behavior toward Minorities. *AEJ: Economic Policy*.

Chattopadhyay, R. and E. Duflo (2004). Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. *Econometrica 72*(5), 1409–1443.

Corno, L., E. L. Ferrara, and J. Burns (2019). Interaction, stereotypes and performance. Evidence from South Africa. IFS Working PapersW19/03.

Dahl, G. B., A. Kotsadam, and D.-O. Rooth (2018). Does Integration Change Gender Attitudes? The Effect of Randomly Assigning Women to Traditionally Male Teams. NBER Working Paper 24351.

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

Duflo, E. (2018). Machinistas meet Randomistas: Some useful ML tools for RCT Researchers. NBER Summer Institute 2018, Master Lecture in Development.

Edwards, M. R. and A. J. Ewen (1996). How to Manage Performance and Pay with 360Degree Feedback. *Compensation and Benefits Review 28*, 41–46.

Egan, M. L., G. Matvos, and A. Seru (2017). When Harry Fired Sally: The Double Standard in Punishing Misconduct. NBER Working Paper 23242.

Finseraas, H. and A. Kotsadam (2017). Does Personal Contact with Ethnic Minorities affect anti-immigrant sentiments? Evidence from a Field Experiment. *European Journal of Political Research 56*(3), 703–722.

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.

Glover, D., A. Pallais, and W. Pariente (2017). Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores. *Quarterly Journal of Economics 132*(3), 1219–1260.

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

Goldin, C. and C. Rouse (2000). Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review 90*(4), 715–741.

Grossman, P. J., C. Eckel, M. Komai, and W. Zhan (2016). It pays to be a man: Rewards for leaders in a coordination game. *Journal of Economic Behavior & Organizations 161*, 197–215.

Heath, R. (2018). Why do Firms Hire using Referrals? Evidence from Bangladeshi Garment Factories. *Journal of Political Economy 126*(4), 1691–1746.

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

Hengel, E. (2018). Publishing while Female. Working Paper, University of Liverpool.

Hsieh, C.-T. and B. A. Olken (2014). The Missing "Missing Middle". *Journal of Economic Perspectives 28*(3), 89–108.

Lowe, M. (2019). Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration. Working Paper, University of British Columbia.

MacNell, L., A. Driscoll, and A. N. Hunt (2015). Whats in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innovative Higher Education 40.*

Maida, A. and A. Weber (2019). Female Leadership and Gender Gap within Firms: Evidence from an Italian Board Reform. IZA Discussion Paper No. 12099.

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

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

Mengel, F., J. Sauermann, and U. Zlitz (2018). Gender Bias in Teaching Evaluations. *Journal of the European Economic Association 17*(2), 535–566.

Menzel, A. and C. Woodruff (2019). Gender Wage Gaps and Worker Mobility: Evidence from the Garment Sector in Bangladesh. NBER Working Paper 25982.

Mullainathan, S. and E. Washington (2009). Sticking with Your Vote: Cognitive Dissonance and Political Attitudes. *AEJ: Applied Economics 1*(1), 86–111.

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

Sarsons, H. (2019). Interpreting Signals in the Labor Market: Evidence from Medical Referrals. Working Paper, University of Toronto.

Sol, J. (2016). Peer Evaluations: Incentives and Coworker Relations. *Journal of Economics and Management Strategy 25*(1), 56–76.

Stigler, G. and G. Becker (1977). De Gustibus Non Est Disputandum. *American Economic Review 67*(2), 76–90.

Syverson, C. (2011). What Determines Productivity? *Journal of Economic Literature 49*(2), 326–365.

Uckat, H. (2020). Womens Promotion and Intra-Household Bargaining: Evidence from Bangladesh. 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. (2020). Firm-Level Upgrading in Developing Countries. CDEP-CGEG Working Paper No. 83.

Figures:

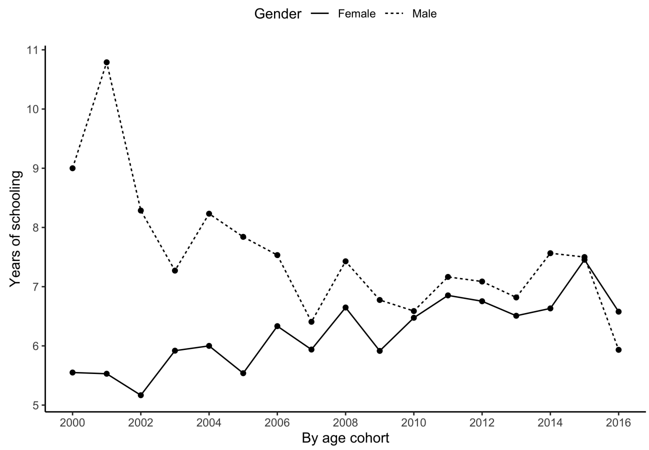


Figure 1: Education levels of Entering Sewing Workers over Time. Calculations by authors using survey data from 1,435 female and 1,049 male sewing workers (line operators) in 30 factories. The surveys were carried out between 2012 and 2017. The data show the reported years of schooling for males and females who turned 18 in the year shown, roughly the age in which workers enter the sector.

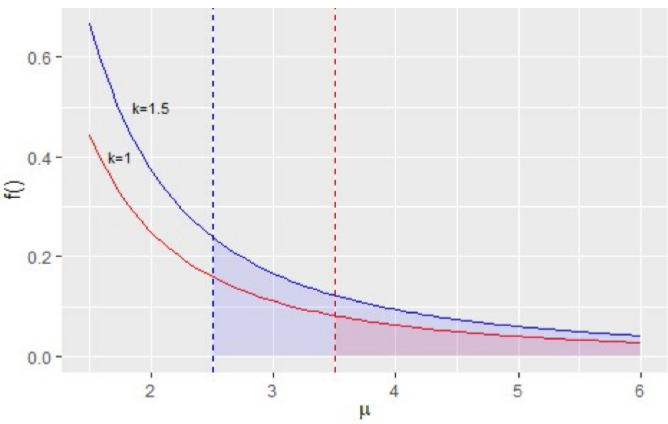


Figure 2: Hypothetical Allocation of Male and Female Supervisors. Figure illustrates a hypothetical scenario that rationalizes the low share of female supervisors. It plots a hypothetical right tail of the distributions of managerial talent among male (blue) and female (red) workers. The blue (red) vertical bar denotes the talent of a hypothetical marginal male (female) supervisor. The ratio of the red area to the blue area is the ratio of female supervisors. A low share of female supervisors can, in principle, arise from either a larger mass in the right tail of the male distribution relative to the female distribution, and/or from $\mu_B^M < \mu_B^F$, i.e., a bias against female supervisors.

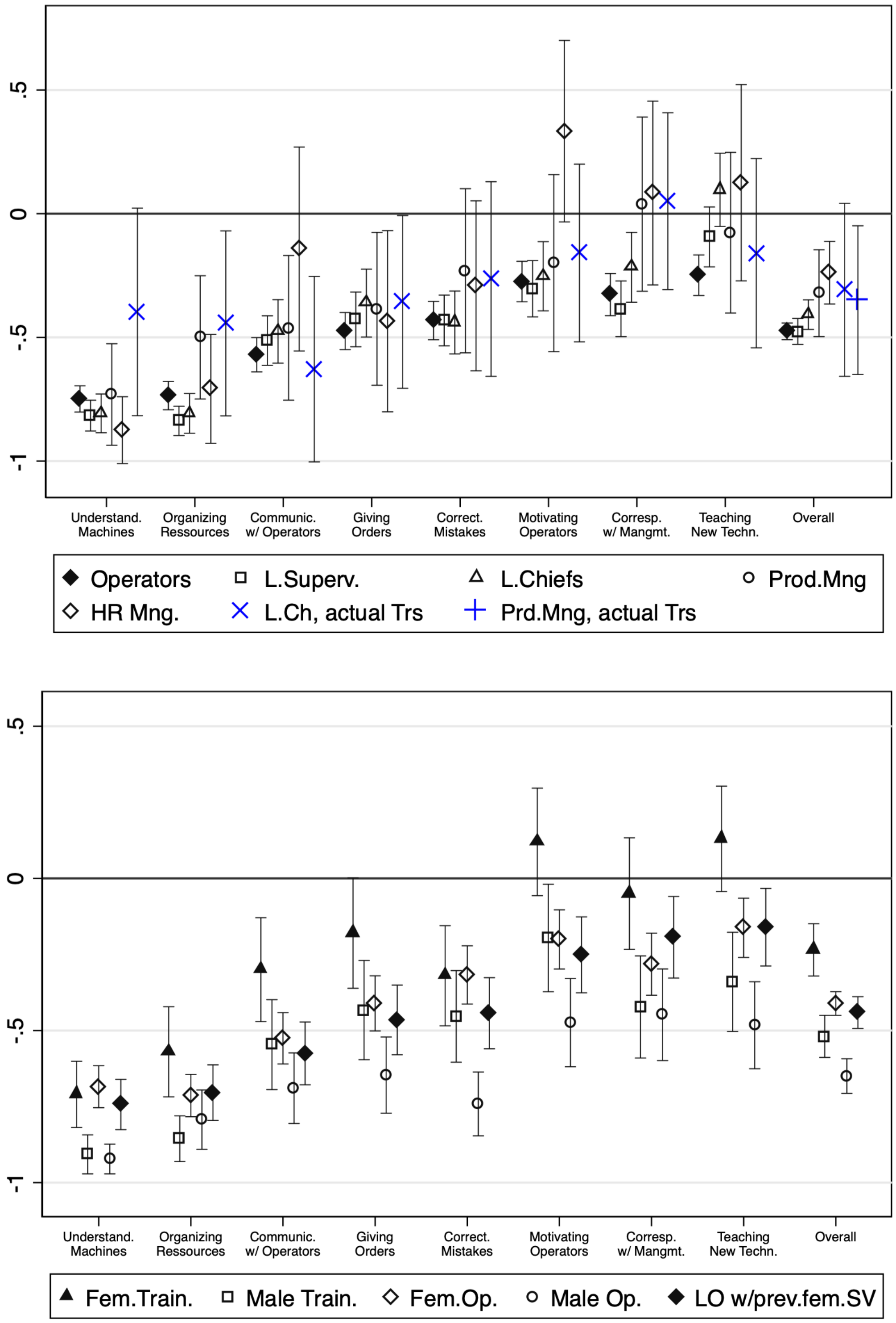
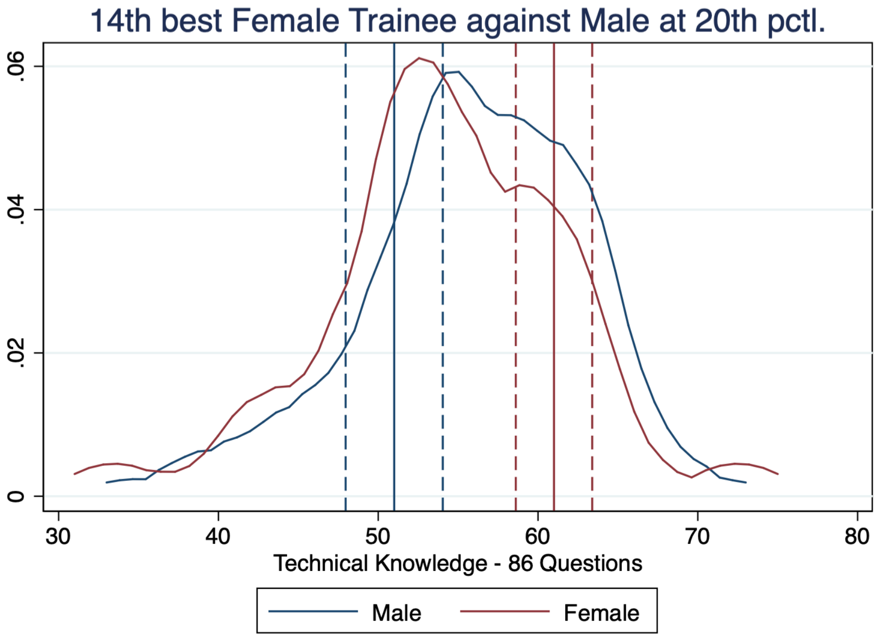


Figure 3: Baseline Beliefs on Female Supervisor Abilities. Figures show mean responses from workers in the garment factories of the types shown in the legend to the following question: "Do you think that male or female workers are better at following task/have more of the following skill?" Responses 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. "LO w/prev.fem.SV" stands for workers who already worked under female line supervisors. Capped bars represent 95 percent confidence intervals.



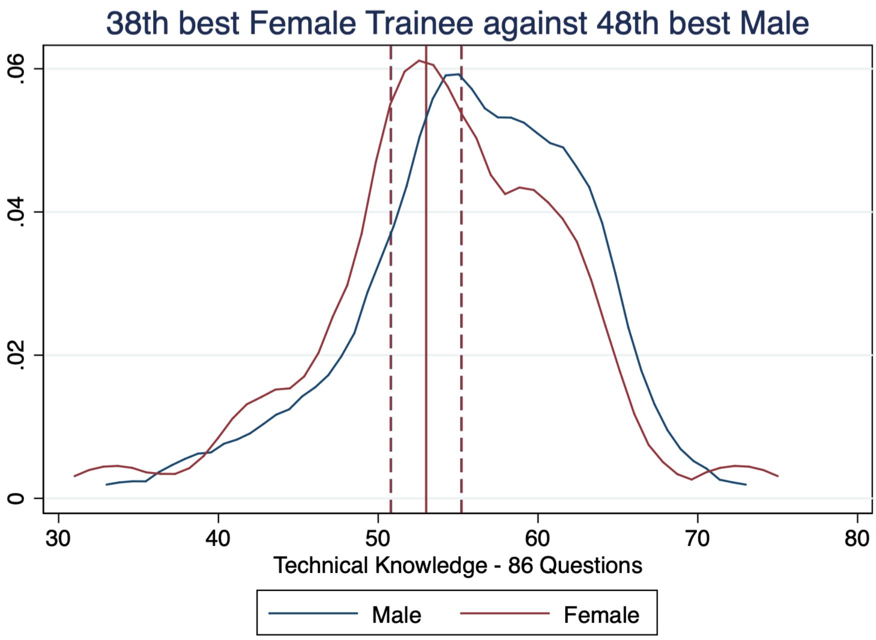


Figure 4: Distribution of Technical Knowledge. Figures show the distribution of scores in the 86-item technical knowledge diagnostic conducted at baseline with the female and male trainees. In the upper figure, the 20th percentile in the distribution of male trainee's test scores, and the score of the 14th best female trainee is marked illustratively: had the factories recruited the 86 trainees it retained as supervisors after the trial according to the baseline ratio of women among supervisors (16%), then only 14 female and all of the male trainees would have been promoted. The percentiles indicate down to which trainee in the score distribution the factories would have chosen trainees, had they strictly promoted following the test scores (we indicate conservatively the male at the 20th percentile instead of the one at the very bottom of the distribution). The lower figure shows the quantiles of the 38th best female and 48th best male trainees, due to the factories retaining after the trial 38 female and 48 male trainees as supervisors. Dashed lines indicate 95 percent confidence intervals of estimated percentiles. Note that in lower figure, the shown percentiles of male and female trainees coincide.

Tables:

Table 1: Beliefs and Preferences on Female Supervisor, Male vs. Female Workers

	(1)	(2)	(3)	(4)
	Overall evaluation		Female SV Preference	
Male Respondent	-0.245***	-0.266***	-0.359***	-0.355***
	(0.039)	(0.043)	(0.088)	(0.095)
Previous Female SV		0.007		0.274***
		(0.038)		(0.083)
Mean Female Respond.	-0.411		-0.253	
Observations	430	418	427	416
Factory FE	YES	YES	YES	YES
Demogr.Controls	NO	YES	NO	YES

Notes: Columns 1-2 shows results from regressing the average relative comparisons across eight supervisor skills of male and female supervisors (see Figure 3) by male and female workers at baseline, on a dummy that respondent is male. Columns 3-4 show equivalent regressions with preference for male/female supervisors, coded as 1 "Prefer female SV", 0 "Indifference" and -1 "Prefer Male Trainee". "Previous Female SV" is an indicator for respondent having worked under female supervisors before. *** denotes stat. significance at 1%, ** at 5%, and * at 10%

Table 2: Comparison of Male and Female Trainees

	Mean Males	Female	SE	N
Age	24.73	-1.010	(0.70)	145
Married	0.611	0.142*	(0.08)	145
Years Educ.	8.486	-0.429	(0.29)	143
Years in Garment Sec.	6.441	-0.240	(0.55)	145
Years in Factory	3.655	-0.449	(0.45)	145
Nbr Factories	2.069	-0.069	(0.33)	145
Literacy	8.842	-0.993	(0.86)	143
Numeracy	4.757	-1.370***	(0.39)	143
Non-verbal Reason.	3.114	-0.100	(0.37)	143
Technical Knowledge	55.82	-1.590	(1.21)	143
Drawing	0.375	-0.114*	(0.07)	126
Drawing - Soft	-0.561	1.096*	(0.64)	127
Communic. - Soft	0.136	-0.266	(0.88)	127
Leadersh. - Soft	0.334	-0.668	(0.66)	124
Confidence, Baseline	-0.142	-0.816***	(0.29)	143
Belief Best, Baseline	0.653	-0.228***	(0.08)	145
Confidence, after Training	0.197	-0.251	(0.28)	144
Belief Best, after Training	0.792	-0.134*	(0.07)	145
Confidence, after Trial	0.238	-0.238	(0.24)	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 5 for definition of variables. *** denotes statistical significance at 1%, ** at 5%, and * at 10%.

Table 3: Comparison of Trainees to other Workers and Supervisors

	Op.Male	Op.Fem	Tr Male	Tr Fem	SV Male	SV Fem	Tr vs.Op	Tr vs.Op	Tr.vs. SV	Tr.vs.SV
N	115	313	72	73	223	10	(Male)	(Female)	(Male)	(Female)
Age	24.33	23.96	24.73	23.72	29.26	27.79	0.40	-0.24	-4.53***	-4.07**
Married	0.655	0.809	0.611	0.753	0.875	0.800	-0.04	-0.06	-0.26***	-0.05
Years Educ.	6.094	5.697	8.486	8.056	9.238	9.399	2.39***	2.36***	-0.75***	-1.34**
Years in Garm.Sec.	6.632	5.906	6.441	6.200	8.939	11.72	-0.19	0.29	-2.49***	-5.52***
Years in Factory	2.380	2.924	3.655	3.205	3.484	6.250	1.27***	0.28	0.17	-3.04***
Nbr. Factories	2.413	1.652	2.069	2.000	2.92	1.20	-0.34	0.35	-0.85***	0.80
Age Promoted SV			24.73	23.72	24.85	21.09			-0.12	2.63*

Notes: Table shows comparisons of baseline survey observables of Trainees (Tr), randomly selected workers (Operators – Op) and line supervisors (SV) from lines nominated by factories for trial, including tests for statistical differences between male and female trainees, operators and supervisors of the same sex. *** denotes statistical significance at 1%, ** at 5%, and * at 10%

Table 4: Individual Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Trial Period:			**Post-Trial Period:**			
	Any Line	Correct Line	Correct Gender	Any Line	Correct Line	Correct Gender	Left
Panel 1: Without Controls							
Female Trainee	-0.004	-0.004	0.061	-0.134*	-0.047	0.024	-0.007
	(0.04)	(0.07)	(0.07)	(0.07)	(0.06)	(0.07)	(0.06)
Mean Males	0.875	0.556	0.625	0.667	0.181	0.222	0.222
N	145	145	145	145	145	145	145
Panel 2: With Trainee Controls							
Female Trainee	-0.030	0.133	0.081	-0.095	0.041	0.024	0.037
	(0.05)	(0.10)	(0.10)	(0.11)	(0.09)	(0.10)	(0.09)
Mean Males	0.875	0.556	0.625	0.667	0.181	0.222	0.222
N	141	141	141	141	141	141	141
Panel 3: Cond. on doing Trial, no contr.							
Female Trainee		-0.016	0.062	-0.129	-0.054	0.028	-0.004
		(0.08)	(0.07)	(0.08)	(0.06)	(0.08)	(0.07)
Mean Males		0.635	0.714	0.746	0.206	0.254	0.159
N		127	127	127	127	127	127
Factory FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Regressions of individual trainee outcomes (e.g. whether trainee was trialed on assigned line, on any line, was being promoted to supervisor, left factory,...) on dummy for being a female trainee, without (Panel 1) and with worker-level controls from baseline surveys (age, confidence, technical knowledge, months in factory, months in industry, number of factories previously worked – Panel 2). Panel 3 restricts sample to trainees that did trial (without controls). *** denotes statistical significance at 1%, ** at 5%, and * at 10%

Table 5: Main Outcomes during Trial

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Administrative Productivity Data					**Survey Based Evaluations**		
Period of Comparison:	Pre-Trial	Trial	Pre-Trial	Trial	Trial	Trial	Trial	Trial
Specification:	ITT	ITT Ancova	Lines *With* Trainee d.Trial	Ancova	Ancova			
						Panel 1: Co-SV Evaluations		
Female Trainee	1.362	-4.402**	0.905	-6.404***	-6.919***	-0.002	-0.163	-0.278
	(1.261)	(1.835)	(1.400)	(1.960)	(2.118)	(0.282)	(0.274)	(0.230)
Baseline value, same line		0.492**		0.461***	0.468***			
		(0.195)		(0.169)	(0.165)			
Observations	93	93	87	87	87	146	146	167
Mean Males	61.79	60.51	61.86	62.86	62.86	6.275	6.275	6.561
						Panel 2: Worker Evaluations		
Fem. Tr.ee x Fem. Resp.						-0.597**	-0.543*	-0.785***
						(0.284)	(0.288)	(0.248)
Fem. Tr.ee x Male Resp.						-1.251***	-1.316***	-1.182***
						(0.473)	(0.488)	(0.409)
Female Respondent						-0.058	-0.403	-0.220
						(0.479)	(0.457)	(0.356)
Observations						266	266	324
Mean Males						6.960	6.960	7.315
p. Men-Women						0.276	0.175	0.412
p SVs-OPs						0.0273	0.115	0.0507
Factory FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls					Yes		Yes	Yes

Notes: Columns 1-5 show differences in average productivity of lines, and columns 6-8 in survey-based evaluations, between female and male trainees, before and during the trial. For survey-based evaluations, Panel 1 shows differences in evaluations from co-supervisors from lines of trainees, and Panel 2 from sub-ordinate workers working on the lines of the trainee supervisors. Productivity data collapsed on the line-level. Controls selected by PDS Lasso from line and trainee controls, plus squares of all controls and indicator variables for missing values of each control. Columns 1-2 and 6-7 show comparisons between lines randomized into receiving male and female trainees, while columns 3-5 and 8 compare lines on which trainees actually worked during the trial. Column 1 shows comparison during pre-trial period between those lines randomly allocated female and male trainees, while column 3 between lines on which male and female trainees actually work as trainee during trial. Columns 6-8 control for evaluation by respondent of the "average" supervisor in the factory. Robust standard errors, clustered at line level for columns 6-8: *** denotes statistical significance at 1%, ** at 5%, and * at 10%

Table 6: Main Outcomes in Post-Trial Period

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Administrative Productivity Data					**Survey Based Evaluations**	
Period of Comparison:	Pre-Trial	Post-Tr.	Post-Tr.	Trial	Trial	Post-Trial	Trial
Sample restriction:	Lines *With* Train.Post-Tr.			Lines *Of* Train.Post-Tr.	Lines *Of* Tr. P-Tr.,Same L.		Lines *Of* Train.Post-Tr.
Specification:		Ancova	Ancova				
						Panel 1: Co-SV Evaluations	
Female Trainee	-0.619	-1.073	-1.021	-7.928***	-6.039	-0.037	0.035
	(1.614)	(1.647)	(1.752)	(2.477)	(4.903)	(0.348)	(0.382)
Baseline value, same Line		0.251*	0.169	0.484***	0.460		
		(0.134)	(0.148)	(0.151)	(0.442)		
Observations	81	81	81	78	34	87	105
Mean Males	63.48	60.32	60.32	60.32	60.32	7.283	6.561
						Panel 2: Worker Evaluations	
Fem. Tr.ee x Fem. Resp.						-0.381	-0.396
						(0.274)	(0.294)
Fem. Tr.ee x Male Resp.						0.117	-1.222**
						(0.505)	(0.614)
Female Respondent						0.193	-0.344
						(0.312)	(0.472)
Observations						250	221
Mean Males						7.951	7.315
p. Men-Women						0.324	0.208
p SVs-OPs						0.957	0.174
Factory FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls			Yes	Yes	Yes	Yes	Yes

Notes: Columns 1-5 show differences in the post-trial period in average productivity, and columns 6-7 in survey-based evaluations, between female and male trainees who worked at least for 30 days as supervisors after the end of the trial period. For survey-based evaluations, Panel 1 shows differences in evaluations from co-supervisors from same lines, and Panel 2 from subordinate workers working on the lines of the trainee supervisors. Column 1 shows comparison during pre-trial period of those lines on which male and female trainees will work as supervisor in post-trial period. Columns 4 and 7 show comparisons during trial period between female and male trainees that will continue to work as supervisors after end of trial period, while column 5 shows the comparison between those that will continue to work as supervisors on the same line as during the trial (for at least 60 days). Productivity data collapsed on the line-level. Controls selected by PDS Lasso from line and trainee controls, plus squares of all controls and indicator variables for missing values of each control. Columns 6-7 control for evaluation by respondent of the "average" supervisor in the factory. Robust standard errors, clustered at line level for columns 6-7: *** denotes statistical significance at 1%, ** at 5%, and * at 10%

Table 7: Management Simulation

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Variable:	Pay-off in Games (Standardized)					
Female Team Leader	0.290***	0.259**	0.466***	0.305**	0.332***	0.412***
	(0.109)	(0.122)	(0.156)	(0.131)	(0.124)	(0.152)
Tried as Line Supervisor				0.329		
				(0.212)		
Promoted to Line Supervisor					0.508**	
					(0.206)	
Mixed Gender / Male Team						-0.092
						(0.725)
Mixed Gender / Male T. x Fem. TL						-0.433*
						(0.240)
Observations	676	612	612	612	608	612
Factory FE	Yes	Yes	Yes	Yes	Yes	Yes
Trainee Controls		Yes	Yes	Yes	Yes	Yes
Team FE			Yes			
Game Type FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Table compares pay-out of male and female trainees in a lab-in-the-field management simulation during pilot of the project, in which the trainees had to explain tasks to a team of two randomly selected workers from the factory, and supervise them doing the task, being paid-out according to the output the two workers produced. Each trainee played four rounds of the game with slightly different incentive structures (see text for more details). Unit of observation is on the trainee-game round level. Standard errors clustered at trainee level. *** denotes significance at 1%, ** at 5%, and * at 10%

Table 8: Trainee Exposure and Preference for Female SVs Update

	(1)	(2)	(3)	(4)	(5)	(6)
	ITT	ITT				
	Trial	Trial	Trial	Trial	Post-Trial	Post-Trial
Fem. Tr.ee x Fem. Operator	-0.007	0.000	0.076	0.078	0.294***	0.201*
	(0.105)	(0.103)	(0.093)	(0.092)	(0.104)	(0.115)
Fem. Tr.ee x Male Operator	0.434***	0.429***	0.318***	0.305***	0.316	0.314*
	(0.113)	(0.130)	(0.114)	(0.111)	(0.193)	(0.175)
Female Respondent	0.497***	0.512***	0.352***	0.344***	0.309**	0.399***
	(0.096)	(0.126)	(0.094)	(0.092)	(0.135)	(0.147)
Mean Male Workers	-0.424	-0.424	-0.420	-0.420	-0.431	-0.431
Observations	396	343	448	448	239	239
Nbr Male Tr.s	46	46	60	60	49	49
Nbr Female Tr.s	55	55	61	61	38	38
Factory FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls		Yes		Yes		Yes
BL Preference, same line		Yes				

Notes: Table compares preference for female supervisor between operators working on lines with male and female trainee supervisors, surveyed right after the end of the trial period (Columns 1-4) or 2-4 month after the end (columns 5-6). Preference for female supervisor coded as 1 "Prefer Female SV", 0 "Indifferent", and -1 "Prefer Male SV". Controls in columns 2, 4, and 6 selected by PDS Lasso from among respondent and trainee controls. *** denotes statistical significance at 1%, ** at 5%, and * at 10%

Appendix A: Further Results

Table A.1: Nominated vs Other Lines

	Non-Nominated	Nominated Lines	SE	N
Style Complexity	12.11	0.286	(0.376)	314
Order Quantity	72322	-3267	(6394)	320
Running Days	7.865	-0.203	(0.449)	477
Hourly Target	148.1	-1.610	(3.890)	439
Daily Target	1424	-18.10	(34.80)	459
Daily Hours	9.517	0.060	(0.048)	477
Nbr. Operators	34.56	0.319	(0.801)	464
Nbr. Helpers	13.20	-0.036	(0.521)	458
Buyer Size	0.181	0.018 *	(0.009)	399
Style Size	0.293	0.001	(0.009)	528

Notes: Table compares lines nominated by the factories for the trial against all other lines in the factories, based on baseline observables from the administrative production data. With factory fixed effects. *** denotes significance at 1%, ** at 5%, and * at 10%

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

	Female	SE	Mean Males	N
Line Operators:				
Gender	-0.019	(0.043)	0.732	142
Age	-0.264	(0.399)	24.05	142
Married	0.026	(0.044)	0.761	142
Years Education	-0.375	(0.262)	5.967	142
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	0.081	(2.354)	16.37	142
Nbr Factories	-0.120	(0.152)	1.880	142
Prev Fem SV	-0.009	(0.055)	0.441	142
Pref Fem SV	0.147 *	(0.075)	-0.417	142
AcceptProm	-0.084	(0.053)	0.500	139
F-Test:		0.024		
Line Supervisors:				
Gender	0.025	(0.026)	0.022	136
Age	0.595	(0.761)	28.72	136
Married	0.050	(0.050)	0.843	136
Years Education	0.196	(0.279)	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
Prev Fem SV	-0.100	(0.066)	0.327	136
Pref Fem SV	0.045	(0.087)	-0.722	136
Spouse Works	0.097	(0.077)	0.328	128
F-Test:		0.105		
Production Data:				
SMV	0.430	(0.696)	12.18	75
Order Quantity	18896 *	(9779)	34070	76
Running Days	1.187 *	(0.642)	5.468	111
Hourly Target	-6.480	(8.098)	161.1	98
Total Target	-10.10	(71.77)	1491	104
Daily Hours	0.053	(0.081)	9.709	107
Nbr Operators	0.501	(1.470)	34.37	109
Nbr Helpers	0.544	(0.959)	14.87	109
Efficiency	0.020	(0.017)	0.566	98
Absenteeism	0.000	(0.005)	0.047	76
Defects Rate	0.002	(0.006)	0.061	118
Spot Rate	0.0185	(0.013)	0.017	75
Reject Rate	0.000	(0.000)	0.003	90
Buyer Size	0.030 *	(0.017)	0.182	99
Style Size	0.002	(0.003)	0.018	111
F-Test:		0.085		

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 Effects on Other Administrative Outcome Variables, Trial and Post-Trial

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pre-Trial	Trial	Pre-Trial	Trial	Trial	Pre-Trial	Post-Tr.	Post-Tr.
	ITT	ITT Ancova		Ancova	Ancova	Lines *With* Train.Post-Tr.	Ancova	Ancova
Panel 1: Defects Rates								
Female Trainee	0.636	0.181	0.065	0.611	1.274**	1.187	0.286	0.551
	(0.646)	(0.570)	(0.930)	(0.566)	(0.622)	(0.852)	(0.427)	(0.478)
Baseline Value, same line		0.839***		0.427***	0.424***		0.591***	0.618***
		(0.182)		(0.134)	(0.120)		(0.118)	(0.106)
Mean Males	6.113	6.220	6.734	5.997	5.997	6.696	6.741	6.741
Observations	105	105	98	98	98	86	86	86
Panel 2: Absenteeism Rate								
Female Trainee	-0.452	-0.410	-0.190	-0.653	-0.568	-0.759	-0.790	-0.865
	(0.449)	(0.673)	(0.508)	(0.500)	(0.516)	(0.556)	(0.699)	(0.691)
Baseline Value, same line		0.302		0.319	0.327		0.387	0.311
		(0.327)		(0.215)	(0.202)		(0.240)	(0.284)
Mean Males	5.727	6.331	5.619	6.363	6.363	6.030	7.438	7.438
Observations	81	81	74	74	74	65	65	65
Factory FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls					Yes			Yes

Notes: Columns 1-5 replicate columns 1-5 of Table 5, while columns 6-8 replicate columns 1-3 of Table 6, for the two other outcome variables based on administrative production data: Defect rates among produced output (Panel 1), and Absenteeism rates of workers on lines (Panel 2). *** denotes statistical significance at 1%, ** at 5%, and * at 10%

Table A.4: Co-Supervisor and Subordinate Beliefs & Outcomes during Trial

Given that at baseline, supervisors and workers expressed beliefs that women have lower supervisory abilities, the strength of these beliefs measured on a given line could be associated with the lower productivity or evaluations of female trainees on those lines. As shown in Table A.4, controlling for average baseline beliefs about relative female/ male supervisory skills by the supervisors or randomly sampled workers on the lines does not cause meaningful changes in the estimates. However, the worker belief measures on the line level in particular are likely very noisy, as we only surveyed three workers on lines which typically have 20-60 workers. Furthermore, especially for workers, turn-over between the baseline survey and the end of the trial period may have been high. Worker turnover in comparable factories is around 6 percent per month (Menzel and Woodruff (2019)), and workers may also be assigned to other lines by the factory management between the surveys.

	(1)	(2)	(3)	(4)
Dep. Variable:	**Productivity**		**Survey Evaluations by Workers**	
Basel. beliefs by:	Co-SV	Worker	Co-SV	Worker
Specification:	Ancova	Ancova		
Female Trainee	-4.307**	-3.976**		
	(1.808)	(1.811)		
Fem. Tr.ee x Fem. Respondent			-0.587*	-0.598**
			(0.324)	(0.283)
Fem. Tr.ee x Male Respondent			-1.192**	-1.201**
			(0.501)	(0.463)
Female Respondent			-0.008	-0.027
			(0.498)	(0.479)
Baseline Beliefs	-6.321*	12.565*	-0.100	-0.663
	(3.502)	(7.159)	(0.612)	(0.902)
Baseline Beliefs x Fem. Trainee	10.793*	-6.646	-0.249	-0.911
	(5.442)	(8.749)	(0.677)	(1.094)
Observations	93	93	266	266
Factory FE	Yes	Yes	Yes	Yes

Notes: Column 1 replicates column 2 from Table 5, using productivity data, but controlling for average baseline beliefs about female vs male supervisor skills across the eight skill dimensions shown in Figure 3, as expressed by the *supervisors* on the lines. The average belief is interacted with whether the supervisors are from lines assigned female trainees, allowing it to affect male and female trainees differentially. Column 2 repeats the regression but controlling in the same way for the baseline beliefs of the surveyed sub-ordinate *workers* from the lines. Columns 3 and 4 repeat this analysis, but with evaluations of the trainees by sub-ordinate workers as outcome. Column 3 thus replicates column 6 of Panel 2, Table 5, controlling for co-supervisor baseline attitudes, while column 4 controls for worker baseline attitudes. *** denotes statistical significance at 1%, ** at 5%, and * at 10%

Appendix 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 2]. 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 see the strongest pre-post differences for confidence in own ability: after the training participants rate their own skills higher (column 1), and are more likely to state that they consider themselves the best candidate from the workers from their line for promotion to supervisors (column 2). 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 2, 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 3). 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 other four skill measures, notably the technical knowledge test with 86 questions, we do not find any effects. 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 longer if there were no training.

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

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Confidence	Belief best	Drawings correct	Technical knowledge	Drawing -soft	Communic. -soft	Leadership -soft
Post-Training	0.333*	0.139*	0.122*	-0.040	0.382	0.973	0.310
	(0.201)	(0.078)	(0.071)	(0.670)	(0.643)	(0.869)	(0.973)
Post-Training x Fem. Trainee	0.571*	0.094	0.128	0.684	-0.683	-1.469	-1.039
	(0.332)	(0.100)	(0.095)	(0.995)	(0.882)	(1.235)	(1.213)
Female Trainee	-0.814***	-0.228***	-0.118*	-1.596	1.034	-0.213	0.144
	(0.288)	(0.081)	(0.071)	(1.215)	(0.641)	(0.887)	(0.977)
Baseline Mean Male Trainees	-0.145	0.653***	0.381***	55.829***	-0.499	0.083	0.137
Post-Training, non-interacted	0.627***	0.186***	0.187***	0.308	0.030	0.215	-0.210
	(0.169)	(0.050)	(0.048)	(0.498)	(0.440)	(0.620)	(0.607)
Observations	284	290	246	287	252	252	140

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 trainee. The skill measures are explained in more detail in section 4. 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%

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

Using administrative wage data from 36 garment factories from Menzel and Woodruff (2019) that include information on supervisors, we estimate that seven 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. However, data on promotion rates in the sector are difficult to come by, because promotions can occur either "internally" at a factory, or "externally", when a worker moves from one to another factory, leaving the previous factory from a worker position and starting work as supervisor at the new factory. 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. Menzel and Woodruff (2019) estimate substantial shares of external promotions between paygrades among non-supervisor workers in the sector, with higher shares among men. But we lack rich enough data to repeat this exercise for promotions to supervisors.

What we do observe in the data from Menzel and Woodruff (2019) are 101 internal promotions to supervisors in 29 of the 36 factories over the course of one year, with 22 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 (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 16.4 percent across the 29 factories. Given that external promotions in the sector below supervisor ranks are more common for men, and given the prevailing rarity of female supervisors in the sector, we assume that among external promotions to supervisors, the share of men is higher than for internal promotions. Furthermore, some of the factories in our sample participated in projects which encouraged the promotion of female supervisors. Excluding these factories reduces our estimate to 13 percent. For these reasons, we interpret the 16 percent share of women among internal promotions to supervisors as an upper bound for the overall (external + internal) promotion share of women in the sector.

Appendix D: Description Pilot Phase

D.1 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. We obtained an initial sample frame of factories from 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, with unit values in the range of mid-level 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. Table D.1 shows the characteristics of these 85 factories, and of the 145 other factories on the list. We find that that those factories agreeing to participate sell to more buyers, and sell to higher-end buyers (in terms of average unit price paid by buyer). We also find some evidence that the participating factories had higher rates of recent growth and export products to a larger number of countries. Participating factories were randomly placed into one of six treatment rounds of 12 factories each. In practice several factories decided in the end not to participate. Furthermore, the management simulation, which were conducted during the follow-up surveys, were not implemented with factories from the first two training rounds, reducing the number of factories with whose trained trainees the simulations were implemented to 48.

Table D.2 shows additional characteristics of the participating factories, collected during management surveys at baseline. The factories are large, averaging 19 production lines and 2,100 workers. Somewhat more than half of the employees in a typical factory work in the sewing section. The distributions are slightly right-skewed, with the median factory having 15 production lines, with 2,000 workers in total, of which 59% are in the sewing section. A typical factory had been operating for 12 years.

Table D.1: Take-up of Pilot

	Signed-Up N = 85	Not Signed-Up N = 145	p-value	p-value (Probit)
Size (Export, 1000 Kgs)	830.4	683.8	0.11	0.44
Avg. Unit Value (per Kg)	925.9	883.8	0.15	0.01
Growth (Sales 2009-10)	1.89	1.46	0.08	–
Number of Destinations	10.1	8.3	0.09	0.18
Number of Buyers	9.75	8.3	0.06	0.02
Number of Products	3.01	2.91	0.32	0.31
Main Product in Woven	0.59	0.54	0.26	–
Year of first export	2006	2006.2	0.2	–

Notes: Table compares factory-level observables between the pool of factories contacted for offering participation in the pilot round and declining participation, against those agreeing to participate. 'Growth (Sales 2009-10)' on a sample of 80 and 135 exporters respectively. *** denotes statistical significance at 1%, ** at 5%, and * at 10%

D.2 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

Table D.2: Description, Factories Pilot Phase

	Mean	Median
Number of sewing lines	19	14
Number of employees, total	2116	2000
Number of employees, Sewing	1171	1000
Operators per sewing line	48	47
Number of sewing supervisors (total)	48	36
Percentage female supervisors	10.80%	5.60%
Percent factories conducting training	68.10%	NA
Percent training outside factory	8.90%	NA
Year factory established	1999	2001

Notes: Table shows descriptive statistics of the 85 factories that agreed to participate in the pilot, collected using surveys of factory managers at baseline.

excluded if they did not pass the literacy test or said their families would not allow them to participate in the training.

For training rounds 1 to 3, 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 diagnostic score 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 diagnostic 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. In round 4, 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. In round 5, we modified the process further by reducing the number of operators the factory nominated to eight females and four males. Over the first six training rounds, 271 operators (213 females and 58 males) received training.

D.3 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 from the factories in which the management simulation was implemented self-reported that they have been tried out as supervisors, and 77 percent of the male and 54 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 have left the factories. These numbers are very close to those we see in the main trial, as reported in Table 4, except for the number of male trainees still working as supervisors being around 10 percentage points lower.

Table D.3 shows basic demographic characteristics, separately for male and female trainees from the pilot, and comparisons for each to the trainees of the same sex from the main trial. Overall, the two sets of trainees look very similar; none of the ten comparison for the five characteristics, for each male and female trainees, shows statistically significant differences. Furthermore, the marriage gap that we observed between male and female trainees from the main sample is visible also among the trainees from the pilot sample. There seems to be a mild pattern that male trainees in the main trial are somewhat older and more experienced in the sector, and with more tenure in their current factory, and somewhat fewer years of schooling on the other hand. But as already mentioned, none of these differences is statistically significant.

Table D.3: Pilot Phase, Trainee Demographic Characteristics

	Trainees Pilot Male N = 42	Diff. Pilot Female N = 127	Diff. Main-Tr. Male N = 72	Diff. Main Tr. Female N = 73
Age	24.05	-0.43	0.68	0.10
Married	0.60	0.16**	0.01	-0.01
Years in School	8.90	-0.44	-0.42	-0.40
Experience in Garments	6.03	0.06	0.41	0.11
Tenure in Factory	3.02	0.45	0.64	-0.26

Notes: Table compares male and female trainees from the pilot phase, who participated in the management simulation, 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%

Appendix E: Knowledge Diagnostic Test

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?	__ __ __

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 l 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	

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?	|__|