Warwick Economics Research Paper Series

Challenges of Change: An Experiment Training Women to Manage in the Bangladeshi Garment Sector

Rocco Macchiavello, Andreas Menzel, Atonu Rabbani and Christopher Woodruff

December, 2015
Series Number: 1100
ISSN 2059-4283 (online)
ISSN 0083-7350 (print)

This paper also appears as *CAGE Working Paper No: 256*

Challenges of Change: An Experiment Training Women to Manage in the Bangladeshi Garment Sector

November 15, 2015

Rocco Macchiavello, University of Warwick

Andreas Menzel, University of Warwick

Atonu Rabbani, University of Dhaka

Christopher Woodruff, University of Warwick[#]

Abstract:

Large private firms are still relatively rare in low-income countries, and we know little about how entry-level managers in these firms are selected. We examine a context in which nearly 80 percent of production line workers are female, but 95 percent of supervisors are male. We evaluate the effectiveness of female supervisors by implementing a training program for selected production line workers. Prior to the training, we find that workers at all level of the factory believe males are more effective supervisors than females. Careful skills diagnostics indicate that those perceptions do not always match reality. When the trainees are deployed in supervisory roles, production line workers initially judge females to be significantly less effective, and there is some evidence that the lines on which they work underperform. But after around four months of exposure, both perceptions and performance of female supervisors catch up to those of males. We document evidence that the exposure to female supervisors changes the expectations of male production workers with regard to promotion and expected tenure in the factory.

[#] Corresponding author: c.woodruff@warwick.ac.uk. The project has benefitted from comments from seminars at UC San Diego, the University of Washington, Notre Dame, Duke, Leuven, Ecole Polytechnique, MIT / Harvard, PUC-Chile and the CEPR IMO workshop. 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.

Challenges of Change: An Experiment Training Women to Manage in the Bangladeshi Garment Sector

Management of large firms in low-income countries is highly variable and, on average, poor (Bloom et al. [2012]). The recent literature has focused on the implementation of a broad set of management practices pioneered by Bloom and Reenen [2007]. However, effective management, including the adoption of such practices, rests on successfully managing relationships and perceptions in the workplace (Gibbons and Henderson [2012]). This observation shifts our attention from practices to managers. Shortages of qualified managers are perceived to be an important barrier to better management in developing countries (e.g., McKinsey [2011]); yet, we still know little about how companies in these countries develop and select managerial talent.

We study low-level management in the ready-made garment industry in Bangladesh, a sector with more than 4,000 factories, employing around 4 million workers and accounting for an estimated 12 percent of Bangladesh's GDP. Besides its intrinsic relevance, the sector provides an ideal context to study low-level managers. The sewing section in a typical factory is organized along several production lines employing between 20 and 80 workers (operators) directly managed by line supervisors, the lowest level of management. We focus on one distinctive feature of the industry: while women account for about 75 to 80 percent of workers in the sewing operations, men account for around 95 percent of supervisors and higher-level managers. The situation is stark: Figure 1 contrasts employment patterns in Bangladesh with the historical evolution in the United States and shows just how strong the gender imbalance is in Bangladesh.

Why are there so few female supervisors? Does this gender imbalance result in a large misallocation of managerial talent in the sector? To address these questions, we start from a simple observation: in a static sense, managerial capital is misallocated if the marginal female supervisor is more effective than the marginal male supervisor.[1] If

[1] Note that the observation is correct for any distribution of potential supervisor's effectiveness across genders. In particular, it is possible that in the current industry equilibrium men self-select and/or invest in additional skills with the expectation of becoming supervisors. This

this was the case, factories could improve efficiency by promoting additional women and fewer men.[2]

Empirically, we face several challenges in answering these questions. First, given there are so few female supervisors to begin with, it is difficult to identify the marginal female supervisor in observational data. To overcome this problem, we implement a six-week operator-to-supervisor training program in 24 factories.[3] The program induces factories to try out (and possibly promote) more female supervisors than they otherwise would. The pool of female and male trainees for the program are selected by factory management. The initial lack of female supervisors may also pose a challenge to factory management because the managers have little experience selecting females for promotion. We examine the selection environment using uniquely detailed baseline surveys and diagnostics tools implemented with workers and managers at all levels in the factories. Data from these exercises help us understand what supervisors are expected to do, and how – in both perception and in reality – the skills of females compare with the skills of males.

Second, we need to observe the performance of both male and female line supervisors. We implement an experimental design in which returning trainees are tried as assistant supervisors on randomly assigned production lines. This allows us to identify the causal impact of female supervisors on performance. We compare the performance of females and males trained in the program, and the response of operators working for them, using both very detailed production data and in-factory surveys.

We show four sets of results. First, we ask what supervisors do, and what the perceived weaknesses of females as supervisors are. Across eight broadly defined sets of tasks, we find remarkable agreement across hierarchical layers in the factories about

---

could result in the pool of men available for promotion being on average better than the pool of available women for promotion.

[2] Large inefficiencies would be at odds with the fact that all factories in our sample are large exporters operating in highly competitive product markets. A large literature shows that competition increases efficiency (Syverson [2004]; Foster et al. [2008]; Backus [2014]), improves management practices (Bloom and Reenen [2007]; Bloom et al. [2012]) and that export status is associated with higher productivity (Bernard et al. [2007]) and better management (Bloom et al. [2015]). On the other hand the factories in our setting are typically owned by a small group of investors and might face lower pressure on the financial market side.

[3] The training program was designed by the German bilateral aid agency, Deutsche Gesellschaft fuer Internationale Zusammenarbeit (GIZ), together with local training companies.

what supervisors are supposed to be doing. There is also remarkable agreement in the factory that women are weaker than men in essentially all eight dimensions. In particular, women are perceived to be less competent than man in understanding machines and operations - crucially, the most important task for a supervisor from the point of view of operators. These negative perceptions are less strong, but nevertheless present, even among female operators and among those operators with experience working under a female supervisor.

Second, we compare these perceptions to reality. Before the training began, we conducted an extensive skills assessment with the trainees. Three results emerge. First, there is no difference between female and male trainees in technical knowledge of machines and operations - despite the widely held opinion to the contrary. Second, in simple leadership exercises women are less likely to be selected by their team for a leadership position and women perform slightly worse in an exercise in which they instruct other team members to perform a simple task. Third, in essentially all eight broad tasks, females rate themselves as being weaker than existing supervisors while male trainees do not.

Third, we examine the performance of female and male trainees once they return from the training. Two sets of results emerge. First, immediately upon returning from training female trainees underperform relative to male trainees. This initial gap in performance is measured both using surveys of operators supervised by the trainees as well as detailed daily line-level production data. The gap in performance, however, completely closes after few months working on the line as supervisors. In simulated management exercises, female trainees outperform male trainees on average but not when managing small teams that include a male operator.

Finally, we explore attitudes of male operators exposed to the program. These are of particular importance given that the bulk of future line supervisors is currently recruited from this pool. Two results stand out: first, male operators exposed to female trainees improve their view of female as supervisors. Second, male operators exposed to female trainees are more pessimistic about their prospects of being later promoted to supervisor roles and expect to work for a shorter period of time in the factory. In short, the promotion of female supervisor appears to demotivate male workers.

Taken all together, these results portray a nuanced but comprehensive picture of the causes and consequences of gender imbalance in the sector. The evidence is

consistent with an industry equilibrium in which factories have not experimented with female supervisors due to misperceptions about their relative effectiveness. The equilibrium is supported by the fact that misperceptions are widespread across the organization, including among workers and potential female supervisors themselves. Shifting to a new equilibrium requires coordinated changes in beliefs. In a static sense, even a profit maximizing manager with correct beliefs might not promote women if - in our case - he believes other co-workers won't respond adequately due to their beliefs. Dynamically, such a manager might believe workers' perceptions can be aligned to reality, but at the cost of alienating and demotivating male operators - from which the bulk of managerial talent is still likely to be supplied to the factory in the short-run. In the conclusions, we distil some implications of this interpretation for our understanding of organization's failure to adopt adequate management practices, the sources and consequences of gender imbalances in general, and the design of policies that could ameliorate those.

This paper contributes to different strands of literature. It complements a literature examining the causes and consequences of the (lack of) female leadership. Although there are numerous contributions studying the gender gap in labour markets and in the private sector (see, e.g., Bertrand et al. [2014]; Matsa and Miller [2013]; Bertrand and Hallock [2000]; Dezso and Ross [2012]; Glover et al. [2015]), our work is conceptually closer to studies of female politicians in India by Chattopadhyay and Duflo [2004] and Beaman et al. [2009].[4]

As Chattopadhyay and Duflo [2004] we focus on establishing the causal impact of female leaderships on outcomes. As Beaman et al. [2009] we emphasize the importance and evolution of perceptions of female leadership. Our analysis, however, needs to be adapted to reflect the operations and incentives of large firms operating in a competitive export sector. First, the performance - not just the appointment - of female leaders is affected by beliefs and perceptions of co-workers. Second, we investigate the costs associated with appointing female leaders.

In so doing, the paper also contributes to the literature on management and productivity (see, e.g., Bloom and Reenen [2007]; Bloom et al. [2012, 2013]; Bruhn et al.

[4] Some of our results are also related to a large experimental literature documenting gender differences in attitudes and preferences, see, e.g., Gneezy and Rustichini [2004]; Niederle and Vesterlund [2007]; Niederle, Segal, and Vesterlund [2013].

[2012]; McKenzie and Woodruff [2015]).[5] The work by Bloom and various co-authors raises a puzzle: the management practices they study are well-known and seemingly simple to implement. Why do firms fail to implement them? Gibbons and Henderson [2012] argue that changing practices is actually quite complex, both because individual practices are complementary to one another (see also Ichniowski et al. [1997]) and because management involves both formal rules and informal norms. Managers may know what is wrong, know how to fix what is wrong, but yet be unable to implement the required changes because they are unable to shift the equilibrium of the game between managers and workers (or between managers at different levels of the hierarchy). Our research design and emphasis on understanding misalignment of perceptions within the firm borrows from this perspective. The difficulties of implementing change echo recent work by Atkin et al. [2015] in the soccer ball industry in Pakistan. They show that firms may fail to adopt productivity-increasing changes in technology because the pay structure of production workers encourages them to misreport to management the productivity of the technology. We instead highlight how resistance to change is embedded in a set of norms and perceptions we set out to measure.

Finally, the paper contributes to our understanding of the garment sector in Bangladesh and elsewhere. Historically, the sector has represented one of the first opportunities for women to enter the formal labour force. Heath and Mobarak [2015] study the relationship between garments, female labour force participation and schooling in Bangladesh respectively. Line supervisors in the garment industry are also studied by Schoar [2011] and Achyuta et al. [2014] with a different focus and research design.

2. Design and Data

At the core of our study is a training program designed by the German bilateral aid agency (GIZ) and local training companies which aims to provide sewing machine

[5] There are two additional methodological contributions of the paper. With respect to the productivity literature, the paper uses a physical measure of productivity in a multi-product industry with product differentiation. Line-level productivity is measured taking advantage of "standard minute values" which allow to convert units of differentiated garment pieces into standardized measures of time value of output. With respect to the literature on the evaluation of training program, we directly investigate the impact of the training on productivity, not just on the wages paid to trainees. This is important as for a variety of reasons wages might fail to reflect the marginal value of labour.

operators skills necessary 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 as important to build skills of female operators, and to convince factories that women were equipped to be supervisors. 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. We initially contracted with three training providers and then later selected one of them with the capacity to conduct all of the sessions.

Our project was carried out in two phases. Phase 1 began in November 2011 and continued through February 2013, with 56 factories sending five participants each to training. After analysing the data from the first phase, we made several changes to the project design and launched the second phase in February 2014. Lessons from the first phase were incorporated into the design of the second phase. As a result of incorporating these lessons, the quality of the data are generally higher in the second phase. Aside from a management simulation exercise that we conducted only in Phase 1, we rely on the data from the second phase for this paper. We describe the design for the second phase here, and refer the reader to Appendix A for a description of the design of the first phase, and a comparison of results where they overlap.

In the second phase of the project we worked with direct and indirect suppliers of a large UK-based buyer. We started with a pool of 26 suppliers of woven and light-knit products located in the Dhaka area.[6] The buyer invited these suppliers to an information session in February 2014. At the end of the information session, 24 factories expressed interest in the project, all of whom ultimately participated.[7]

We asked each factory to consider the expected demand for new supervisors in the factory in the months following training, and select a number of trainees matching that demand. Because the size of the factories varied and because, for example, some factories were planning to open new production lines, the number of trainees varied from as few as four to as many as 24. Where an even number of trainees was provided,

[6] We limited the sample to the Dhaka area for logistical reasons and to woven and light knit because production in these products is organized by sewing lines in Bangladeshi factories. Direct suppliers are managed by employees working directly for the buyer; indirect suppliers are managed on behalf of the buyer by intermediaries.

[7] Five of the factories sent operators to the first training session, but dropped out in the second half of the program.

we asked factories to select an equal number of male and female trainees. Where an odd number of trainees was selected, we asked them to select one more female than male.

We informed the factories that much of the training material was written, and therefore the trainees needed to have at least basic literacy skills. We gave them no other criteria, but did encourage them to involve in the selection decisions managers down to at least the level of the line chief – the immediate superior of line supervisors. The factories sent 99 males and 100 female trainees to the training centre for the initial diagnostic. Note that this represents a significant movement toward female supervisors, because in the typical factory at baseline only around 4 percent of supervisors were female.

We scheduled four training sessions, the first beginning March 9th, 2014, and the last beginning June 1st, 2014. In order to stagger the return of trainees to the factory, half the nominees from each factory were randomly allocated to on of two training rounds, either rounds 1 and 3 or rounds 2 and 4. Within the factory, the trainees were randomly assigned to receive early or late training. Randomization at the trainee level was stratified on gender so that a nearly equal number of female and male trainees were trained in each session.

Factories agreed to give each trainee a six- to eight-week trial as an assistant line supervisor immediately after the end of the training program. We asked factories to identify the lines which were suitable for the trainee trials and to identify an experienced supervisor working on each of those lines who could act as a mentor for the trainee. On the penultimate day of training, we invited the mentor supervisors to the training centre and matched them randomly with one of the trainees from their factory - thus assigning the trainee randomly to a production line for the trial period. Over two days with the mentors in attendance, we conducted a series of team building exercises between trainees and mentors. After the six- to eight-week trial, factories were free to return the trainee to a position as operator, leave them as an assistant supervisor, or promote them to supervisor.

There was dropout of trainees at various points, detailed in Figure 2. The factories initially selected 121 females and 96 males for training. All were invited to the training centre for the initial assessment. On the allocated day, 100 females and 99 males actually showed up. Twenty-one females declined to come to the training centre, either because they decided they did not want to be supervisors or because of

resistance from their families. Meanwhile, three additional males came as some factories replaced the females who declined to attend. Admission to the full training program depended on passing the literacy and numeracy test administered at the training centre. The literacy exam was developed in conjunction with researchers at BRAC University.[8] Nominees were disqualified if they scored zero on either the literacy or numeracy exam, or if they scored below 25 percent on both parts of the exam. Eleven females and 18 males did not pass the literacy / numeracy threshold. An additional three females and five males were disqualified for other reasons, mainly because the factory sent a male rather than a female.[9] Finally, after the assessment day, 13 females and four males decided they did not want to complete training and dropped out of the program. The remaining sample, all of whom completed the training course, was 73 females and 72 males. Figure 2 also shows the number of trainees working as a supervisor at various points after training, which we discuss in more detail below.

2.1 Data

We conducted surveys on six separate occasions. First, prior to the start of training we conducted a combined survey and skills assessment for the trainees at the training centre. The survey and assessment lasted a full day. In addition to gathering basic information on demographics, work history and attitudes, we assessed knowledge of machine and production processes, conducted communication, teaching and leadership exercises, and tested numeracy, literacy and non-verbal reasoning skills. The assessment is described in more detail below.

Second, near the end of the six-week training program, we asked factories to nominate production lines and mentor supervisors in a number matching the number of trainees. With the list of lines and mentors in hand, we conducted a baseline survey in the factory prior to the end of training and the start of the trail. For the factory survey, we surveyed line operators, line supervisors, line chiefs, floor supervisors assistant production managers (floor managers), production managers and HR managers. Three operators and all of the supervisors and line chiefs were surveyed at the lines where

[8] The literacy/numeracy test was developed by Sameeo Sheesh and Badrul Alam of BRAC University's Institute of Education Development (IED). The content is based on the skills required to benefit from the Operator to Supervisor Training material, and content taught in grades 5 through 8.

[9] In a couple of cases, the literacy exam was mismarked so that a failing score was given when the exam was a marginal pass.

trainees were assigned to have their trial. Line chiefs from the lines where trainees were working at the start of the training were also surveyed. The three operators were randomly selected from the line in a way which ensured that at least two of these operators work directly under the mentor supervisor, and we select both male and female operators wherever possible.

Third, on the penultimate day of the training, the mentors were invited to the training centre and paired with their matched trainee. We conducted team building exercises and also conducted a survey and skills assessment with both the trainees and the mentors. The survey and assessment was designed to capture any effects of the training on the trainees and to measure the skills of experienced mentor supervisors for comparative purposes. Fourth, at the end of the six- to eight-week trial period, we again invited the trainees back to the training centre for refresher sessions and group discussions of their experience during the trial. On the refresher day we also conducted a final skills assessment for trainees to measure the effect of the factory trial.

The fifth survey was conducted in the factory just after the trial period ended. We again surveyed three randomly selected operators, the supervisors and line chiefs of the lines that were nominated for the trial, and the assistant production managers, production managers, and HR managers. In addition, where there was either non-compliance with the assignment of trainees to lines, or where trainees had moved from the assigned line to another line after the trail began, we surveyed operators and supervisors on the lines which were not nominated for the trial, but where trainees were actually working as assistant supervisors.

Finally, we conducted a second follow-up survey in the factory in October (training rounds 1 and 2) and November (training rounds 3 and 4). The last follow-up was thus about four and a half months after the trail ended for those trained in the early rounds, and two and a half moths after the end of the trail for those trained in the late rounds. The survey sample was selected using the same criteria as in the previous factory survey, but because of time constraints, we were able to survey operators and supervisors only from the lines where a trainee was working as either an assistant line supervisor or a line supervisor. In addition, all of the trainees were surveyed in-person if they were still working at the same factory, and over the phone, if they had left.

In addition to the face-to-face surveys, we conducted telephone follow-up surveys with trainees at regular intervals. During the six- to eight-week trial, we

contacted the trainees every week to track the line they were working on, and the level of responsibility given to them. We also asked the trainees to keep a daily diary of their experience working as an assistant supervisor or supervisor. After the trial ended, we contacted the trainees every month until March 2015 (four to nine months after the trial) to track where they were working, and their designation.

In addition to the survey data, we also collected daily line-level production data from each factory. We describe these data in more detail in Appendix B and in Section 6 below.

2.2 Characteristics of Trainees

Table 1 shows basic demographic and skills data for the pool of trainees, compared to existing supervisors and random operators where the comparison data are available. Compared with a sample of random operators, the trainees have two additional years of schooling and just more than half a year more tenure in the factory. Age, marital status and experience in the garment sector are similar to other operators. We split the supervisor sample into mentors and non-mentors for the purposes of comparing trainees with existing supervisors. We see that, while the trainees have much more schooling than typical operators, they have almost a year less schooling than typical supervisors. They are also 4.7 years younger with 2.3 year less experience in the sector. However, the age of the trainees is statistically identical to the age of the random supervisors at the time of their promotion to supervisor.

With regard to the relative skills of female and male trainees (not shown on table), we find that females are just over a year younger ($p=0.05$), but there are no differences in schooling or experience. Whether the trainees have less schooling than existing supervisors because factories face a shortage of workers with higher schooling levels, or whether the factories have not selected the very best supervisory talent for the training program is not clear. But while 62 percent of existing supervisors have at least a lower secondary certificate (that is, they have passed O-level exams), only 14 of 430 random operators (3 percent) have achieved this level of education. This suggests that factories do face a very limited pool of workers with education levels comparable to the pool of existing supervisors. This, combined with the age and experience profiles of the trainees suggests that the factories selected trainees in a manner similar to those selected in the usual promotion routine.

We can also compare the skills of trainees and the mentor supervisor using tests administered at the training centre during the skills assessment, though we lack similar data for other operators and supervisors. The bottom half of Table 1 shows that the literacy and numeracy scores of the trainees are significantly below those of the mentor supervisors. These data provide further evidence that the skills of the trainees are below those of the mentor supervisors.

3. Perceptions of female supervisors

A typical factory in our sample has only one or two female supervisors at baseline. Therefore, operators and managers have little direct experience working with female supervisors. Nevertheless, they have perceptions about the relative ability of females and males as supervisors. As a first step in exploring these perception, we asked employees at all levels of the factories to tell us which tasks are the most important for line supervisors. We constructed a list of eight main tasks from an initial set of open-ended conversations with managers. We then gave each respondent 10 tokens and asked him or her to place the 10 tokens on the list of the eight tasks plus an "other" category in a way which indicated the relative importance of each. Respondents were told they could place all 10 tokens on a single task if they thought that it was the only task that is important, or spread the tokens across the tasks as they wished. Surveys were conducted with HR Managers, Production Managers, Assistant Production Managers, Line Chiefs, Line Supervisors and Operators.

Figure 3 shows the percentage of tokens placed on each of the eight tasks by respondents holding different positions at the factory. The characteristics given the highest weights are shown to the left of the graph. One pattern that emerges is that all levels of managers agree about which characteristics are important. Teaching and motivating operators are given the largest weights by all managers. Operators, on the other hand give somewhat different weights. They appear to prefer problem solvers, giving higher weights to understanding machines and correcting mistakes. There is agreement across the hierarchy that organizing resources, corresponding with management, and giving order are less important tasks of supervisors.

We then asked the same set of respondents whether, based on their own experience, they thought females or males were better at each of the eight tasks of being a supervisor. The allowed responses included the option of "both are equal". We code

these data in a way that indicates the perceived deficit that females face in each of the tasks. A response “males are better" is coded as -1, “females are better" is coded as +1 and “both are equal" is coded as 0. The scores are shown in Figure 4, again by type of respondent.[10] The first takeaway from the table is that males are overwhelmingly seen as having an advantage in every supervisory task. Line operators and line supervisors rate males better in all eight tasks, line chiefs and production managers rate males better in seven of the eight tasks, and HR managers see males as being better in five of them.

We also find a very high level of agreement about the specific tasks where females are most lacking. According to every category of respondent, females have the largest deficits in understanding machines and organizing resources. All respondents also agree that the three areas where females are closest to males are teaching new techniques, motivating operators, and corresponding with management, though there is some disagreement about the ranking of these three. Notice that the two tasks rated as most important by managers are two of those where the gap between females and males is perceived to be the smallest. On the other hand, machine knowledge, rated highest by operators, is the area where females are perceived to be the weakest.

The sample of operators is the largest and most diverse, so in Figure 5, we show the same comparisons for different subgroups of operators. First we split the randomly selected operators by gender. The relative rankings are very similar for female and male operators - the correlation is 0.87 - though female operators uniformly describe a smaller gap. Next we split the operators into those who have and those who have not worked for a female supervisor at some point in their career. Past experience working for a female supervisor has no significant effect on the perceived gap in female skills. Finally, when we asked the trainees the same comparisons between generic male and female supervisors, the responses are very close to those of other operators. As Figure 5 shows, female trainees do rate women somewhat higher than do other operators.

We also asked trainees about their own ability relative to typical supervisors in their factory. We first asked the trainees to rate the typical supervisor on a scale of 1-10 with regard to each of the eight supervisory roles, and then asked the trainee to rate her- or himself on the same scale. Female trainees rate themselves as worse than the

[10] We did not ask the Assistant Production Managers to make this comparison because of time constraints on the survey instrument.

typical supervisor on each of the eight characteristics, while males rate themselves better at motivating workers and giving orders. The average gap for males is only 0.09, while for females it is 0.45. Across skills, the females' self-assessments largely match the pattern of the gender perceptions more generally. The correlation between the gaps the female trainees perceive in themselves and the gaps that operators perceive in female supervisors is 0.68.

We aggregate the ratings of males and females on all eight skills to create a single variable indicating each respondent's beliefs about the relative skills of males and females. For the aggregation, we assign a value of 1 to “females are better", 0 to “males are better" and 0.5 to the indifferent response. The first column of Table 2 shows how the average deficit for females across the eight tasks is affected by the gender of the operator and past experience working with female supervisors. Consistent with the data in Figure 5, we find that female operators have slightly higher opinions of female supervisors, being about 12 percent more likely to choose “female is better" over “male is better". Previous reported experience working for a female supervisor does not change the perceived skill level of females and males. In the second column, we spilt the experience effect by the gender of the operator. There is no effect for female operators, while there is a small effect for male operators (p-value 0.101).

We also asked operators whether they prefer to work for a female or male supervisor. Similar to the coding for skills, we code the responses as 1 for “prefer female", 0 for “prefer male" and 0.5 for indifferent. As a group, the operators say they prefer to work for male supervisors by a margin of about two to one. However, female operators are 17 percent more likely to say they prefer females, and those with previous experience working for female supervisors are 12 percent more likely to say they prefer working for a female supervisor (Table 2, column 3). Again there appears to be, if anything, a somewhat stronger effect for male operators (column 4) - though as with the skills assessment, the gap between female and male operators is not statistically significant. Among the 140 female operators reporting experience working for a female supervisor, 40 percent say they prefer to work for males, 30 percent for females and 30 percent are indifferent. Among males with no experience working for females, the percentages are 81, 16, and 3.

In sum, the skills assessment provides little evidence that perceptions are influenced by experience. However, when asked to express a preference to work for

male or female supervisors, previous experience working for women does appear to matter, especially for male operators.

4. Do measured skills match the perceptions?

The surveys indicate that female supervisors are viewed as less skilled than male supervisors in each of eight supervisory tasks. The female trainees see similar weaknesses in themselves. Do these perceptions match reality? We conducted an extensive skills assessment of the female and male trainees selected by the participating factories during their first day at the training centre. We administered tests of numeracy, literacy, and non-verbal reasoning. We also assessed technical skills and knowledge of machines, and conducted teaching, communication, and leadership exercises. The data from this assessment provide evidence on several dimensions of the actual skills gaps between females and males selected by factories as having supervisory potential. We use these data for two purposes. The first is to assess the extent to which perceptions match reality at the baseline. The second is to measure the effects of training and the trial period working as an assistant supervisor on the trainees' skills. For the latter purpose, we repeat some of the exercises at the end of training and after the factory trial period.

4.1 Baseline measures: Do the skills gaps match the perceptions?

The most direct and extensive comparison we can make between perceptions and reality is on the question of technical and machine knowledge. The assessment asked the trainees to name different parts of sewing machines, and to tell us which type of machine (e.g., at lock, single needle, etc.) would be used for different sewing processes. We showed the trainees garments of the type they typically produce with faults in them, and asked them to identify what machine problem (e.g., loose thread tension) would most likely cause the particular type of fault. We showed the trainees pictures of production lines and asked them to identify issues where worker safety was being compromised. In all, the diagnostic included 86 questions. We conducted a very similar exercise after training and then again after the trainees completed the trial in the factory.

We examine differences between the female and male trainees in Table 3. The first column of the table shows results of factory fixed effect regressions using all three

rounds of the assessment. For now, we focus on the top line of the table, which shows the difference between females and males on the baseline assessment. In raw scores, males outperform females by a single percentage point, scoring 65 per cent compared with 64 per cent for females. The regression shows a similar gap, with females on average score one point lower on the 86-point scale. The female – male difference is highly insignificant. In other words, even though close to 90 percent of survey respondents say that male supervisors have more technical knowledge than female supervisors, we find no statistical difference between the female and male trainees selected by the factories.

We also conducted exercises to measure teaching, communication and leadership. In the teaching exercise, we divided the trainees into groups of four to six. We assigned each trainee the role of teacher in one round of the exercise, with the others being students. The teacher was given an abstract figure, which might be for example several triangles and circles with some coloured in. The teacher's task was to instruct the students to reproduce the figure using only verbal instructions. She could not show the figure to the students or use her hands. We examine two outcome measures. The simplest is the number of drawing that were correct. The first row of column 2 on Table 3 shows that males obtain a slightly higher percentage of correct drawings, with the gap being marginally insignificant with a p-value of 0.10.

The second outcome from the teaching assessment comes from observations recorded by two enumerators observing the exercise. For example, the enumerators recorded whether the instruction was given at an appropriate pace, and the number of times the teacher explained the task in more than one way. We take six such observations and construct standardized measures for each assessment round. We then sum the six standardized indicator variables to create an index of “soft teaching skills". Column 3 on Table 3 shows a factory fixed-effect regression with this index as the dependent variable. We see no significant difference between females and males in baseline teaching techniques, though the standard errors are larger than we might like. We also note that the soft skills measure is not significantly associated with the harder outcome - the percentage of correct drawings - though the measured effect is positive ($p=0.22$).

We create similar ‘soft' measures as our main outcome in the communication exercise and the leadership exercise. In the communications exercise, the trainees were

asked to give a short speech on a topic related to rules in the factory, such as: “Describe to a new operator, all the things that you need to do when your machine breaks". During the speech, the trainee was interrupted with questions on two occasions. (For example, “What should I do if I think I can fix the machine myself?"). Two enumerators recorded judgements on whether the trainees spoke clearly, at a reasonable pace, whether she had confidence, etc. The top row of column 4 in Table 3 shows that female trainees perform insignificantly worse by these measures. Finally, in the leadership exercise we asked the group to create a production hierarchy, and then asked them to produce some ‘products' using Legos. The precise hierarchy depended on the size of the group, but we measure whether there are differences across the genders in the probability of being appointed a management role, and in soft measures reflecting the extent to which the individual participated actively in the discussion. The top row of column 5 shows that females are scored insignificantly lower on the soft skills measure. But we do find that males are significantly more likely to be appointed to management (75 percent vs. 32 percent, $p<0.001$).

The teaching, communication and leadership exercises were intended to measure important aspects of confidence and preparedness to lead a production line. We also elicited a direct measure of self-confidence of the trainees. We first asked each trainee to rate the average supervisor in her/his factory on a scale of 1-10. We then asked them to rate themselves as a supervisor, two months after beginning the job. The take the gap between the typical supervisor and the trainee’s own expected performance as a measure of self-confidence. At baseline, we find that the male trainees express more confidence in their ability. In the raw data, they rate themselves 0.33 points lower than a typical supervisor, while the female trainees rate themselves 0.79 points lower. The top row of column 6 in Table 3 shows a similar deficit for women of 0.47 points, controlling for factory fixed effects, significant at the .10 level. Thus, while we find no significant differences between the female and male trainees in the technical assessment or the teaching and leadership exercises, we do see differences in their self-reported confidence levels.

4.2 Training and Trialling effects

We repeated the teaching, communication and leadership assessments at the end of the six-week training period and again at the end of the factory trial period. For the

latter assessment, the trainees returned to the training centre for a review day during which we conducted these assessments as well. Rows 2 and 4 of Table 3 show the various post-training measures, all measured relative to baseline. At the bottom of the table, we show the p-value for tests of equivalence of female and male trainees at each point in time. The table is an unbalanced panel, as there were several nominated trainees who either failed the literacy / numeracy exam or dropped out for other reasons, and there was further attrition before the post-trial review day. However, results from regressions using the balanced panel are very similar, suggesting that the patterns we observe are not driven by selection.

Looking first at the scores on the assessment of technical knowledge, we see insignificant improvements in both males and females after the training, and no change in the relative performance across gender. Relative males at baseline – the base group for the regressions – females perform 0.5 points better and males 0.2 points better following the training. After the factory trial (rows 3 and 5), however, we see the performance of females appears to deteriorate somewhat. Indeed, comparing female and male trainees, the post-trial technical assessment is the only measure showing a significant difference by gender. We are unsure what might explain this dip in performance, but we find the same effect in the balanced panel. The other outcome worth noting is that the self-confidence increases significantly after training for both males (by 0.5 points) and females (by 0.71 points). The magnitude of the increase is slightly larger for females, but not significantly so. However, the self-confidence gap between females and males shrinks and become statistically insignificant.

5. How are operator perceptions changed by experience?

The skills diagnostics indicate that the female trainees have only a very small and statistically insignificant gap in technical skills. On the other hand, there are more significant gaps in self-confidence and in the outcomes of the teaching and leadership exercises. The training closes these gaps. But the outcomes on the production floor are of more interest than the outcomes of the diagnostics. We examine these using both surveys of operators working for the trainees and using administrative data on productivity of the lines where the trainees are assigned (ITT) or work (OLS).

We conducted a first follow-up survey in the factory just after the end of the initial trial period. During the six- to eight weeks between the end of the training and

this first follow-up survey, trainees were meant to be working as assistant line supervisors, together with their mentor. Compliance with this agreement was very high. Of the 135 operators completing training 129 were trialled as an assistant supervisor. Four of the six not trialled (three females and one male) left the factory before the trial started. Recall that we randomly allocated the trainees to one of the lines selected by the factory for the trials. We can measure non-compliance with this assignment either at the individual level or at the gender level. Some individuals were trialled on a different line than the one assigned. In many cases, the factory switched two females or two males, leading to non-compliance at the individual level but compliance at the gender level. We are primarily concerned with compliance at the gender level – that is, that factories placed a female (male) trainee on a line assigned to a female (male). Within gender non-compliance, there are two types. The majority of the non-compliance involved the trainee being placed on a line not included in the original pool of trial lines. Indeed, 34 percent of the trainees were trialled on a line not selected as a trial line. But in the 77 cases where trainees were trialled on the lines designated for trials, there was compliance on the gender level – i.e., females on lines assigned to females – in 71 cases.

At the first follow-up, conducted at the end of the trial period, we surveyed randomly selected operators working on all of the lines where the trainees were assigned (ITT lines) and on all of the lines where the trainees were actually working. However, at the second follow-up, we were not able to survey all of the ITT lines, and hence have information only on the lines where the trainees were working at that time. The reason for this is logistical. Recall that for each factory, the training was conducted in two rounds approximately two months apart. The first follow-up survey was then conducted on two different days in each factory, at the end of the factory trial for each training round. The second follow-up survey, however, was conducted on both training groups at the same time. This meant that we were surveying twice as many lines on the day of the second follow-up, limiting our flexibility with regard to ITT lines. The simultaneous implementation of the second follow-up survey also implies that the time gap between the end of the trial and the second follow-up survey was about two months longer for the trainees in the first training round than for those in the second round.

As a result, we are able to report both ITT and OLS regressions for the first follow-up survey data, but only OLS regressions for the second follow-up survey. At each follow-up survey, we selected three operators at random from each of the

surveyed lines. We focus on two outcomes. First, we asked the operators to rank on a scale of 1-10 both a typical supervisor in the factory and the trainee on their line based on their knowledge of her/him. We regress the ranking of the trainee on an indicator for his or her gender and the gender of the surveyed operator, controlling for the ranking of the typical supervisor by the operator. Second, we asked the operators whether they prefer to work for a female or male supervisor, and as before code the responses as 1 for “prefer female", 0 for “prefer male", and 0.5 for “indifferent". For the first of these outcomes, we are interested in the ranking of female trainees relative to male trainees, and for the second, we are interested in whether exposure to a female trainee affects the preference for supervisors.

The first three columns of Table 4 below show the ITT regressions for the relative ranking (columns 1 and 2) and the preference for female supervisors (column 3). We find that the female trainees are rated almost a point - about 0.4 standard deviations - lower than the male trainees. In column 2, we allow the relative ranking to differ for female and male operators. We find, if anything, males rate the females more harshly, though the difference is not statistically significant (p=0.26). In column 3, we see that exposure to female trainees has the effect of making male operators significantly less opposed to working with female supervisors. While female operators are more inclined than male operators to say they prefer to work for female supervisors, their opinion is not influenced by exposure to the female trainees.

Columns 4-7 of Table 4 repeat the same regressions using the actual placement of the trainees. We find almost identical effects in the ranking regressions (columns 4 and 5), but slightly weaker effects in the preference regressions (column 7). Finally, columns 8-10 show the results of OLS regressions using the second follow-up survey data. Because we use the sample of trainees working as assistant supervisors or full supervisors at the time of the second follow-up, in column 6 we show the first follow-up results using the sample of trainees working as supervisors at the time of the second follow-up. We see that the results for male operators are very similar to those in the full sample (compare column 6 with column 5), though the smaller sample yields higher standard errors and an insignificant effect. The results for female operators appear slightly different, and less negative, for the sample of trainees that continue to work as supervisors at the second follow-up. This indicates that the weaker female trainees may be those who do not continue as supervisors.

In the second follow-up survey, the deficit for female trainees is erased completely (See columns 8 and 9 of Table 4). Female trainees are rated as equal to male trainees, by both female and male operators. Moreover, operators of either genders who are exposed to the female trainees express higher preferences for working with female supervisors. Note as well that the trainees as a whole are now rated as slightly better than the typical supervisor in the factory. The improvement in the relative ranking of the trainees is consistent with statements by production managers that new supervisors require four to six months of experience to reach their full potential.

6. Trainee Performance measured by Production Data

6.1 Production outcomes

The literature measuring the effects of job training programs has typically relied on outcome measures such as employment or earnings of trainees.[11] This is reasonable if wages equal the value of the marginal product of labour. In our context, we believe this approach has drawbacks. First, the factories typically have very specific wages for each worker grade. Many or most of these are determined by minimum wage levels, which are set nationally and vary at the worker grade level. Thus, wages may not reflect marginal products. Second, factories will attempt to make promotion training decisions based on their beliefs about actual productivity of the workers rather than the changes in wages. Managers were uniformly puzzled by the claim that we could learn anything looking at changes in earnings following training.

With this in mind, we have attempted to gather very detailed production data for each of the factories. For the second phase of the project, we have daily, line-level data for 12 or 13 months, typically starting two months prior to the beginning of training and extending seven to nine months after the end of the training (see Appendix B for a more detailed description of the data and its collection process). There are three outcomes of interest: productivity, quality defects, and absenteeism. By focusing on sewing, we are able to capture a measure of output which is very close the pure quantity measure. A trained industrial engineer can take any garment and estimate the number of minutes a fully-efficient worker will take to produce the garment. These calculations come from summing the required time for each stich to make the garment. The times come from a

[11] Much of this literature focuses on programs aimed at individuals who are out of work. See, for example Card et al 2011; Attanasio et al 2011.

combination of international databases and in-factory time-and-motion studies. By multiplying these 'standard minute values' - SMVs (or standard allowable minutes - SAMs) by the number of units of a given garment which are produced during the day, we obtain a measure of output - output minutes - which is highly comparable across products. For example, a line producing 1,000 shirts with an SMV of 15 minutes has production of 15,000 output minutes.

For productivity, we divide the output minutes by input minutes - the sum of minutes worked by operators and helpers on the line over the same time period[12] - to obtain the industry standard measure of efficiency. This is essentially a measure of Q/L:

$$\text{Output} * \text{SMV} / [(\text{Operators} + \text{Helpers}) * \text{hours} * 60^{\text{mins}}] \quad (1)$$

The average efficiency in the sample we are currently using is 53 per cent, which is higher than the 38-40 per cent that those in the industry typically quote.[13] A second measure of interest is the number of quality defects. Factories typically report both the number or percentage of garments that require some re-work and the number or percentage that must be rejected. Reject rates are typically very low, averaging less than 0.5 percent in our sample. Rework rates are much higher, averaging around 7 per cent (with a median of almost 5 per cent). Because the re-work time is included in the measure of "input minutes", the efficiency measure incorporates improvements in quality.

We construct a panel at the line level, with dummy variables indicating the presence of a trainee working on the line either as an assistant supervisor or a full supervisor. We begin with an ITT specification, using the gendered assignment of a trainee on the line during the trail period, and then assuming this initial assignment predicts the line on which the trainee will be promoted.

[12] We could improve the input minutes measure by a step if we had the wage bill for the whole line. However, the industry typically uses three different wage grades for operators, and we most often know only the total number of operators, not the number by grade.

[13] The higher efficiency in our sample may come from having a more efficient sample of factories. However, the data across factories are not always comparable because the international SMV values are often adjusted upwards by factories to account for some expected level of inefficiency. We are currently working to ensure the data are comparable across factories, but we include factory fixed effects in all of the regressions using production data, which will absorb systematic measurement differences across factories.

$$y_{gfld} = \alpha_l + \beta_{fd} + \sum_{g\in 0,1} \gamma_g TRIAL_{fld} + \sum_{g\in 0,1} \delta_g POST\_TRIAL_{fld} + \varepsilon_{gfld} \qquad (2)$$

where g ={0,1} represents male or female trainees, *f* is factory, *l* line, *d* the week of production, and *y* the outcome of interest. *TRIAL* reflects the assignment of the line to a female /male trainee during the trial weeks and *POST_TRIAL* the assignment of the line to a female/male trainee during the period after the trial.

We also present OLS results on the actual placement and roles of trainees. These may suffer from both the endogenous placement of trainees and the endogenous decisions to promote. As with the ITT regressions, we include both line and factory/week fixed effects, which mitigates to some degree the issue of endogenous placement. However, some of the trainees leave the factory and some return to being operators after the trial. Since these outcomes are more frequent for females than for males, we should clearly be concerned with the endogenous promotion decisions in interpreting the OLS regressions. We nevertheless think that the OLS results are potentially interesting in spite of these selection issues, because promotion of almost any females represents a change relative to what likely would have happened in the absence of the experiment.

The first three columns of Table 5 report the ITT regressions for efficiency, absenteeism and defect rates. The samples for each of the regressions vary somewhat because data on some measures are not available in some factories.[14] The cleanest results relate to efficiency. Compared to lines without trainees, we see that lines where male trainees were assigned are about 2.3 percentage points - roughly 5 percent - more efficient during the trial period. During the trial, the trainees represent extra supervisory labour on the line. Hence, even though they are least experienced at this point, it is perhaps not surprising that they have a positive effect on efficiency. There is no increase in efficiency during the trial period on the lines assigned a female trainee, suggesting that even though the female trainees are additional supervisory labour, they are not effective in increasing efficiency. However, the situation changes during the post-trail period. Those trainees remaining as supervisors may either be classed as Assistant Supervisors or as full Line Supervisors during this period. In the latter case,

[14] The sample size drops by about 75 percent when we use lines for which all three variables are not missing.

and perhaps even in the former, they are replacing an existing line supervisor, and hence no longer represent incremental supervision. During this period, the female trainees catch up to the males. We see that both female and male trainees have very similar effects on efficiency, with positive coefficients which are economically important but statistically insignificant at conventional levels.

Columns 4 through 7 present OLS results based on actual assignment. We use actual assignment because the initial line assignment was agreed to only for the trial period. We did not necessarily expect the factories to promote the trainees to the same lines.[15] The patterns are very similar to the ITT regressions, though the coefficients are generally of slightly larger magnitude. The regression in column 4 shows that females perform significantly worse during the trial period, perform equally well as males when both are assistant supervisors, and perform insignificantly better than male trainees when both have been promoted to full supervisor. In column 5, we limit to sample to observations from days when the trainee was working on one of the original ITT lines. The patterns are similar, though now the better performance of female trainees as full supervisors is marginally significant (p=.096). Columns 6 and 7 report results for absenteeism and defect rates, respectively. Again the patterns are similar to the ITT regressions except that underperformance of female trainees relative to male trainees on quality issues is almost significant when working as assistant supervisors (p=0.111; see bottom of table).

The efficiency results mirror the opinions of operators working on the lines. Female trainees start slower; they perform significantly worse than males during the trial period. However, they catch up in the months after the trial period. We see the same pattern in the ITT and OLS regressions. In the ITT regressions, the gain made by female trainees relative to male trainees is significant at the 0.10 level, while the gain in defect rates is marginally insignificant (p=0.125).[16] In the OLS data, we find significant relative performance gains between the trial and promotion to line supervisors for efficiency, and between the period working as an assistant supervisor and promotion to line supervisor for defect rates.

[15] In a future version of the paper, we will match lines based on pre-trial characteristics to obtain a somewhat cleaner comparison.

[16] This is the p-value comparing the gap between female performance in the post-trial period and female performance in the trial period with the same gap for males.

6.2 *Do attitudes adjust?*

Both the survey data and the production data suggest that the female trainees start more slowly than their male counterparts, but catch up three to five months after returning from training. The attitudes of operators with direct exposure to the female trainees adjust over this time. We might ask whether there is any evidence that the attitude adjustment is more general. That is, does the increase in female supervision in the factory have indirect effects on operator attitudes towards female supervisors? The data suggest there is no change to attitudes of other workers: The sum of the generic female / male rankings - coded, as before, 1/0/-1 - is -5.06, -4.97 and -5.19 for the male operators surveyed at baseline, first follow-up and second follow-up, respectively, and -3.33, -3.46 and -3.28 for female operators at the same surveys. None of the differences across survey periods are statistically significant.

Direct exposure to female trainees, on the other hand, has a significant effect on these rankings by the time of the second follow-up survey: Female (male) operators working on lines with female trainees have a cumulative ranking across the eight tasks of -2.86 (-4.71), compared with -3.71 (-5.28) for those on lines with a male trainee. The female operator gap is significant with a p-value of 0.07. Thus, generic attitudes show some evidence of movement with direct exposure, but there is no evidence of any broader effect in the factory.

Survey data available from the first phase of the project indicate that direct exposure leads to a different sort of attitude change, particularly among male operators. We asked operators how long they expected to continue to work at the factory where they were currently working, and whether they expected to be promoted to line supervisor one day. Two-thirds (69%) of male operators working on lines without a female trainee said they excepted to remain at the factory for more than five years, and 95 percent said they expected to be promoted to supervisor.[17] These percentages both fall significantly among males working on lines with a female trainee; only 38 percent expect to stay at least five years and 83 percent expect to become a supervisor (in some factory). The attitudes of females move in the opposite direction, but the movements

[17] Certainly these numbers are exaggerated, but perhaps not by as much as it seems. Given supervisor / operator ratios and the growth of the sector, it is likely that half of the male operators would become supervisors in the "almost all male" world. It is reasonable that some who will not make it still expect, ex ante, to become a supervisor.

are much smaller and statistically insignificant. The percentage expecting to stay five years or more increases from 39 percent to 44 percent among females without and with female trainees on the line. The self-reported likelihood they will become a supervisor increases from 58 percent to 64 percent with exposure to a female trainee.

These two sets of results both underscore the potential costs to individual factories of making the transition from an (almost) all-male supervisory force to a mixed-gender supervisory force. The attitudes toward ability of the female supervisors are changed only with direct exposure, implying that each line needs to be exposed to female supervision before beliefs about women's skills are increased. And the initial change leads to a re-assessment of prospects for male operators, and potentially the loss of talent.

7. Initial Underperformance: beliefs vs. skills

Both the operator opinions and the production data suggest that the female trainees underperform initially. This initial under-performance might arise for either of two reasons. The female trainees may simply be weaker supervisors after training when they are assigned to a line. They might be either because their skills lag or because they lack self-confidence, and hence are less effective as leaders. Alternatively, they may perform equally well as supervisors, but those they work with – either the operators working under them, their peers as supervisors, or their superiors – may be believe they are weaker supervisors, and hence be more likely to question their leadership. In other words, they make lack authority not because of skills but because others believe they lack authority, for example because of their own prior beliefs that women are weaker supervisors than men. We lack definitive evidence which would clearly separate these two possibilities, but two pieces of data provide at least a suggestion that the underperformance arises from a lack of authority arising from the beliefs of co-workers. The first data point is simply that the skills assessment conducted after training revealed no differences between female and male trainees. (See Table 3.) Moreover, although there was an initial confidence gap, that gap was also closed following training.

A somewhat stronger piece of evidence comes from a management simulation exercise we conducted during the first phase of the project. The exercise was conducted during a follow-up survey around four months after the completion of training. The simulation involved the trainees and eight randomly selected operators. The operators

were placed into four teams of two each and played two “production" games, one involving Legos and one involving buttons. We randomized the order in which the games were played at the factory level. Each team was assigned a leader whose job was to explain the particular exercise and manage the operators as they performed their tasks. For the results we present here, the team leader was either a female or male trainee.[18] Each pair of operators played the production game twice, once with Legos and once with buttons. Each team leader played only one session - either Legos or button - so there were eight team leaders in each factory, and each pair of workers played with two different team leaders.

For each of the Lego and button exercises, the teams played five separate sessions. The first was a simple sorting exercise in each case, sorting either buttons or Legos by colour. For Legos, the second, third and fourth sessions involved constructing chains of Legos with a particular colour pattern - blue, yellow, green, blue, yellow, green, etc. The three games were differentiated by their payoffs: the first summed the length of the chains produced by the two operators, the second paid based on the length of the longest chain produced by either worker, and the third paid based on the shortest chain produced by either operator. The team leaders were given incentive payments according to the payoff function.

Here we assess the performance of teams led by female trainees with that of teams led by male trainees measured by the payoffs. We combine each of the five individual games into a single regression by standardizing the payoffs on the game-round level. We then run regressions with the standardized payoffs on the left-hand side and a set of controls for characteristics of the team leader on the right hand side. We focus the discussion here on the subset of games where trainees are team leaders, comparing the performance of female and male team leaders.

Each pair of operators plays the set of five games twice, with one team leader in the first session and a different team leader in the second session. The order of the games (Lego - buttons, or buttons - Lego) is random, and within a round the assignment of team leaders to operator pairs was random. But the assignment of team leaders to session 1 or round 2 depended on the (non-random) order in which leaders were

[18] The full exercise also involved team leaders who were operators from the control group, the most recently promoted supervisor who was not a trainee, another supervisor from the same production line as one of the trainees selected at random (a “matched" supervisor), or another supervisor selected at random (a ”random" supervisor).

provided by the factories. Logistical complexities working in the factory prevented us from randomizing the session in which any team leader participated. In particular, because factories anticipated that we wanted to talk with trainees, the trainees were more likely to be assigned to the first session, and the existing supervisors and control operators were more likely to be assigned to the second. This matters, because even controlling for the team leader and game types (Lego vs. buttons), operators were significantly more productive during the second session. This is logical because we expect some learning by the operators from the first to the second session - even though they play different games in each session. We control for the session order effects in regressions.

Table 6 shows how the standardized payoffs vary with the gender of the team leader in the sample of games involving female and male trainees. The specification in column 1 includes controls for factory, session (first or second) and game fixed effects. We find that teams led by female trainees have payoffs which are 0.29 standard deviations higher than teams led by male trainees, a difference which is highly significant. In other words, female trainees appear to be more effective as team leaders than male trainees. Column 2 adds team leader demographics - age, education, industry experience and factory tenure - and Column 3 adds operator team fixed effects. Note that the third regression then isolates the cases where a single team was led by both a male and a female trainee. Only 19 teams had this pair of team leaders, so while the table shows a sample size of 600, the effective size is much smaller. Nevertheless, the additional production by female-led teams is statistically the same, increasing to 0.42 standard deviations.

In columns 4 and 5 we examine whether trainees who before the survey visit had been tried out as supervisors or promoted to supervisor perform better than those not tried out or promoted. We find that those promoted to supervisor perform significantly better than those not promoted. Since promotion is not random, we are unable to say whether this is due entirely to selection - more able trainees are promoted, while less able ones are not - or whether the experience as a supervisor also makes the individual more effective as a leader. But the result does provide some validation for the exercise itself, showing that those with more ability or experience perform significantly better in the game.

Finally, in column 6 we explore whether the gender composition of the operator team interacts with the gender of the team leader. We compare the performance of mixed or all-male teams with those of all-female teams.[19] The superior performance of the female-led teams is significant only for the all-female team. When both operators are women, their output is 0.41 standard deviations higher with a female leader than with a male leader. But female leaders obtain no higher production from mixed team than male leaders do.

We find, then, that the female trainees were significantly more effective in generating payoffs than were the male trainees. Trainees who were promoted before the time of the first follow-up survey perform significantly better than those not promoted. Crucially, female trainees only perform better when they are matched with a pair of female operators, and perform no better than male trainees when they lead mixed-gender or all-male teams. Since the team leaders were randomly assigned to teams, the quality of leadership provided by the female trainees is independent of the composition of the team, even though the outcomes differ significantly.

There are two further outcomes from the games. The first involves the strategy choices of the team leaders. Recall that the payoffs changed from one game to the next. In the second round, payoffs were for the sum of the output of the two operators, but the third (fourth) game, we paid on the maximum (minimum) output of either worker. After the second game, we asked each team leader which of the two operators was better at the game. We then recorded whether the team leader focused attention on the stronger operator in game 3 and on the weaker operator in game 4, as these are the operators whose performance is expected to determine the payoffs in these games.

The fifth game involved a complex figure that was most efficiently made in a “line", with each operator specializing on one component. We record whether the team leader organized production in that manner. We then sum the number of times the team leader adopted the “correct" strategy in each of these three games. Column 7 regresses this sum on the gender of the team leader, demographics of the team leader and the operators, and factory fixed effects. We find that the male leaders adopted the correct strategy significantly more often, in spite of the female leaders inducing higher output.

Finally, after the second session, the operators on the production team were asked to compare the management style of the two team leaders they worked with.

[19] Since only 20 percent of operators are male, all-male teams are very rare.

They were asked whether the first or second team leader they worked with was better at explaining the game, better at answering questions, better at motivating them, always pressuring them, and so forth. Focusing on the responses of the 19 teams led by both a female and a male trainee, we find that operators are more likely to say that the male trainees were better at answering question, at motivating, and at encouraging, though the gaps across gender are never quite significant at the .10 level. Female trainees were selected more often only as "always pressuring", and effect which is significant at the .05 level. The last two outcomes, on strategy and operator opinions, are interesting in the light of the superior performance of female trainees.

In sum, the post-training skill assessment and the management simulation exercise both indicate that female trainees were as skilled as male trainees when these exercises were conducted. We read these as suggestive that the prior beliefs of operators, and perhaps of other co-workers as well, explain at least partly the initial under-performance of female trainees.

8 **Conclusion**

We examine the imbalance between the percentage of females among production workers and the percentage of females among supervisors in a set of large garment factories in Bangladesh. Survey responses show that, at baseline, factory employees at all levels perceive males to be more effective supervisors. But a detailed skills assessment conducted with the women and men selected by the factories for training shows that perceptions sometimes deviate from reality. This is most striking with regard to machine and technical knowledge, which shows the largest perceived advantage for males and no actual difference between females and males.

We partnered with local training centres to provide training for female and male operators selected by the factories. Our purpose for implementing the training was to induce factories to promote more females to supervisory positions, to allow us to compare the actual performance of females and males as supervisors. The project was successful in this regard, increasing the number of females working as assistant or full supervisors in the participating factories. During a six- to eight-week trail period, the female and male trainees were assigned lines randomly.

The initial underperformance of female trainees raises a crucial question: Was the underperformance caused by a lack of skills, or by a lack of cooperation on the part

of operators being managed by the supervisors? Supervisors may lack authority because they less skills, or simply because those they supervise do not believe they possess the authority, and hence they do not take their instruction. The strong baseline beliefs in the relative ineffectiveness of females as supervisors raises the possibility that operators would respond differently to exactly the same level of supervision given by males and by females.

We present two pieces of evidence that suggest the channel from baseline perceptions to the reaction of operators and other managers is an important one. First, the post-training skills assessment showed no differences by gender in the measured skills and only a small and insignificant difference in the self-reported confidence of the trainees. Second, the controlled management simulation showed that female trainees (from the first phase of the project) outperformed male trainees, but only when working with all-female teams of operators. Since the worker pairs in the simulation were randomly assigned, this suggests that the same quality of supervision delivered by females was more effective with female production workers. As males express much stronger preferences to work with male supervisors, this is consistent with the 'prior beliefs / resistance' story.

Roughly four months after the first assessment, we find that the performance of female trainees has caught up to that of male trainees. Since we do not find that the female trainees significantly outperform male trainees, we can not give a definitive answer to the question of whether factories are, from the perspective of efficiency, under-promoting women. But we note several reasons to think that we understate the effectiveness of female supervisors relative to the long-run steady state. First, we compare the top males with the top females. The proper comparison would be the top females against the marginal males, those who just missed being selected for training. Second, as we noted, managers have less experience selecting females as supervisors, and express less confidence in their predictions about the performance of the females selected. More experience might be expected to lead to more efficient selection. Third, males enter the factory expecting to become supervisors with very high probability. They therefore have an incentive to invest in the skills required to be a supervisor. Females, with little prospect for promotion, lack that incentive. Women should be expected to increase the investment in the necessary skills if their prospects for future promotion are increased. Finally, the women promoted to supervisor are in some sense

pioneers, or at least very early entrants, in their factories. Many qualified women may not want to go against such a strong current, and so may decline to participate.

What we can say with more certainty is that the experiment points to a substantial cost to an individual factory in making the transition from an (essentially) all male supervisory staff to a mixed-gender supervisory staff. The initial under-performance of females, whatever the cause, is a part of that cost. But another substantial part of the cost is the shift in attitudes of male operators who are exposed to female supervisors. They report lower probabilities of being promoted and shorter expected tenure at the factory. Training may be a third cost. Practicalities did not allow a design which would have allowed us to separate training from promotion of females. But we do note that training had an important effect on the self-confidence of female trainees. Presumably over time, with more role models in the position, women would start with higher levels of self-confidence. But in the initial transition, the training may be important.

Anecdotally, we have had conversations about a year after the final survey with nine of the 24 factories participating in the project. Among those nine, two have reported promoting significant number of additional female supervisors. In both cases, factory management stated that an important motivation for the change is that male workers cause more trouble and unrest in the factories, and hence they are now hiring almost exclusively women for all positions. Whether justified or not, these beliefs provide a reason for being willing to pay the costs associated with transitioning to an equilibrium with female supervisors.

References

Achyuta, A., N. Kala, and A. Nyshadham (2014). Management and Shocks to Worker Productivity: Evidence from Air Pollution Exposure in an Indian Garment Factory. Working Paper, University of Michigan.

Atkin, D., A. Chaudhry, S. Chaudry, A. Khandelwal, and E. Verhoogen (2015). Organizational Barriers to Technology Adoption: Evidence from Soccer-Ball Producers in Pakistan. NBER Working Paper 21417.

Backus, M. (2014). Why is Productivity Correlated with Competition? Working Paper, Columbia Business School.

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.

Bernard, A., B. Jensen, S. Redding, and P. Schott (2007). Firms in International Trade. Journal of Economic Perspectives 21 (3), 105-130.

Bertrand, M., S. Black, S. Jensen, and A. Lleras-Muney (2014). Breaking the Glass Ceiling? The Effect of Board Quotas on Female Labour Market Outcomes in Norway. NHH Dept. of Economics Discussion Paper No. 28/2014.

Bertrand, M. and K. Hallock (2000). The Gender Gap in Top Corporate Jobs. NBER Working Papers 7931.

Bloom, N., B. Eifert, D. McKenzie, A. Mahajan, and J. Roberts (2013). Does Management Matter: Evidence from India. Quarterly Journal of Economics 128 (1), 1-51.

Bloom, N., K. Manova, J. V. Reenen, and Z. Yu (2015). Management, Product Quality and Trade: Evidence from China. Work in Progress, Stanford University.

Bloom, N. and J. V. Reenen (2007). Measuring and Explaining Management Practices Across Firms and Countries. Quarterly Journal of Economics 122 (4), 351-1408.

Bloom, N., R. Sadun, and J. V. Reenen (2012). The Organization of Firms across Countries. Quarterly Journal of Economics 127 (4), 1663-1705.

Bruhn, M., D. Karlan, and A. Schoar (2012). The Impact of Consulting Services on Small and Medium Enterprises: Evidence from a Randomized Trial in Mexico. C.E.P.R. Discussion Paper 8887.

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

Dezso, C. and D. G. Ross (2012). Does Female Representation in Top Management improve Firm Performance? A Panel Data Investigation. Strategic Management Journal 33 (9), 1072-1089.

Foster, L., J. Haltiwanger, and C. Syverson (2008). Reallocation, Firm Turnover and Efficiency: Selection on Productivity or Profitability. American Economic Review 98(1).

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

Glover, D., A. Pallais, and W. Pariente (2015). Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores. Working Paper, Harvard.

Gneezy, U. and A. Rustichini (2004). Gender and Competition at a Young Age. American Economic Review 94 (2), 377-381.

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

Ichniowski, C., K. Shaw, and G. Prennushi (1997). The Effects of Human Resource Management Practices on Productivity: A Study of Steel Finishing Lines. American Economic Review 87(3), 291-313.
Matsa, D. and A. Miller (2013). A Female Style in Corporate Leadership? Evidence from Quotas. American Economic Journal: Applied Economics 5 (3), 136-69.
McKenzie, D. and C. Woodruff (2015). Business Practices in Small Firms in Developing Countries. NBER Working Paper 21505.
McKinsey (2011). Bangladesh's ready made garments landscape: The challenge of growth. McKinsey&Company, Apparel, Fashion & Luxury Practice.
Niederle, M., C. Segal, and L. Vesterlund (2013). How costly is diversity? Affirmative action in light of gender differences in competitiveness.
Niederle, M. and L. Vesterlund (2007). Do Women Shy Away From Competition? Do Men Compete Too Much? Quarterly Journal of Economics 122 (3), 1067-1101.
Ruggles, S., T. Alexander, K. Genadek, R. Goeken, M. Schroeder, and M. Sobek (2010). Integrated public use microdata series: Version 5.0 [machine-readable database]. Minneapolis, MN: Minnesota Population Center [producer and distributor].
Schoar, A. (2011). The Importance of Being Nice: Evidence from a Supervisory Training Program in Cambodia. mimeo, MIT.
Syverson, C. (2004). Product Substitutability and Productivity Dispersion. Review of Economics and Statistics.

Appendix A: Description, Project Phase 1

A.1 Differences in Design

The first phase of the project began in November 2011. The training program was designed with the goal of increasing the number of female supervisors in factories, and GIZ 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 females and one male 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. The first training session began in November 2011. After six rounds of training, we stepped back in January 2013 to assess the design.

Our aim was to select a sample of factories capable of selling directly to large international buyers. We obtained an initial sample using data from transaction-level trade data obtained from the Bangladeshi National Bureau of Revenue. These data provide volume (net weight) and value of exports at the shipment level. The data have identifiers which allow data from individual exporters to be aggregated. We aggregated data by exporter and calculated the unit value (USD per kilogram) for each exporter / product / year. We also summed total exports by exporter. Using these two measures, we selected a sample of firms with annual shipment volumes large enough to sell directly to large foreign buyers, with unit values in the range of mid-level buyers. This selection process yielded an initial sample of 665 exporters. We then selected the group of around 20 suppliers to one particular mid-range buyer based in the UK. For each of the 665 exporters on the initial list, we created a score based on export volume and unit values indicating how close the exporter was to the 20 suppliers of the initial UK-based buyer. We selected around 400 exporters, and searched local directories and the internet for contact information. This yielded a sample of 230 factories, which we began contacting in August 2011.

By November 2011, after contacting about 200 of the factories on the list, we had received an initial commitment to participate in the project from 96 units of 85 distinct factories. Table A1 shows how the characteristics of the 85 factories differ from the initial list. The table shows both a comparison characteristic by characteristic, and the p-values from a probit regression including several of the characteristics. We find that that those factories agreeing to participate sell to more buyers, and sell to higher-end

buyers. The quality of buyers is measured by the average unit price paid by each buyer. For each seller, we then ordered the buyers by unit price, and measured the unit value paid by the buyer at the 90th percentile in the ranking. 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 eight treatment rounds of 12 factories each. In practice we allowed factories to defer participation to a later round once, and in the end, several factories decided not to participate. By December 2012, when training round 6 began, we had exhausted the initial list of 96 factories. Note that all of the comparisons we will make with trainees control for factory fixed effects, so we view the factory-level attrition issue as mainly one of external, but not internal, validity. During the second round of the program, discussions with the local office of the International Finance Corporation led to inclusion of seven factories located in the Dhaka EPZ in the project. These factories were added in training rounds 4 and 5.

Table A1: Take-up of the Program

	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	--
Median Buyer	560	631	0.18	0.44
90th Percentile Buyer	183.6	283	0.03	0.01

Notes: * On a sample of 80 and 135 exporters respectively.

Table A2 shows characteristics of the factories participating in rounds 1-6. 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. Given the rapid growth of the sector, this is very likely older than the industry average.

Table A2: Description, Factories Phase 1

	Mean	Medi-an
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	48	36
Percentage female supervisors	10.80%	5.60%
Percent conducting training	68.10%	NA
Percent training outside factory	8.90%	NA
Year factory established	1999	2001

A.1.1 Selection of Trainees

Our aim was to select from each factory four female and one male operator for training, and a valid comparison group against which to measure the trainees. The details of selecting workers evolved a bit across training rounds, as we describe below, but in all rounds the process started with factories selecting a pool of potential trainees to which we administered a diagnostic test. The test was based on one designed by GIZ to measure literacy, a requirement for the training, and technical knowledge. We also gave the potential trainees a short non-verbal reasoning test and asked them questions about aspirations to work as a line supervisor. Because women were sometimes forbidden to participate in the training by their families, we also asked the potential trainees if their families would allow and support them to attend the training. Potential trainees were excluded if they did not pass the literacy test or said their families would not allow them to participate in the training.

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. We then took the top four females after excluding the two selected by the factory and randomly selected two for treatment and two for control. We also altered the method of replacing trainees when the selected individuals declined to participate. In round 5, we modified the process further by reducing the number of operators the factory identified as candidates to eight females and four males. The factory then selected two of the eight females for training; the remaining two females and the male were selected randomly in the same manner as the previous rounds.

We further modified the method for selecting “replacement" trainees, as described below. There was a non-trivial amount of noncompliance. Over the six rounds, 50 workers assigned to training did not attend at all, and an additional eight attended for less than one full week. Factories most often reported that these workers either had decided they did not want to attend, or their families had said they could not attend. However, the family was most likely to intervene in the case of female trainees, while we note that the percentage of non-complying males assigned to training (21.2 percent) was higher than the percentage of non-complying females assigned to training (15.2 percent).[20] These non-compliers were replaced by 40 workers receiving training even though they were not assigned to training including 19 workers assigned as controls. Thus, non-compliance is a concern in the Phase I data when we compare the outcomes of those assigned to treatment against the controls.

As with the selection of trainees, the protocol for selecting replacements also evolved over the training rounds. In training round 1, the factories chose the

[20] We interpret this as suggesting that factories cared more about which males received training than they did about which females received training, perhaps because they did not plan to promote all of the females.

replacements themselves, as we had not anticipated the severity of this non-compliance. Beginning in round 2, we insisted that the factory send the next female or male on the diagnostic ranking if a selected trainee declined to attend. Then, beginning in round 5, we altered the initial selection process to add a third female control - selecting 2 of the females ranked 3 to 7 by diagnostic score - and a second male control - selecting one of the males in the top three diagnostic scores as the trainee. Replacements were then selected at random from among the controls.

Over the first six training rounds, 271 operators (213 females and 58 males) received training. We exclude from this total eight workers who attended for five days or fewer. Conditional on attending at least one week, attendance was very high. Out of the 36 training days, males attended 34.4 days on average and females 34.5 days. All but two of the men attended at least four of the six training weeks, as did 96 percent of the women. After the sixth training round, we decided to suspend the training temporarily. Having already gathered a substantial amount of data and information, we felt we would gain by analysing those data and perhaps adjusting the design for the remaining factories. We resumed the training with the start of Phase 2 of the project in February 2014, which's details are described in the main body of the paper.

A.2 Comparison of samples and outcomes

We report first-phase results from the management simulation and the promotion prospects of operators exposed to female and male trainees in the main body of the paper. We did not conduct the management simulation in the second phase of the project because it was seen as costly to participating factories given the amount of time required. We did not ask about promotion prospects in the second phase in the interest of time and because we had not analysed those results when we designed the second phase. However, there are several questions which were repeated in both phases of the project, and we report here on comparisons of outcomes in the two phases where we have comparable data in each phase.

Table A3: Phase 1 Demographic Characteristics

	Mean			Comparisons	
	Trainee Pool	Operators	Supervisors	Trainee vs. Operators	Trainee vs. Supervisors
	N = 539	N = 301	N = 292		
Gender (female =1)	0.75	0.76	0.12	-0.02	0.63***
Age	23.54	23.53	28.21	0.01	-4.67***
Migrant	0.71	0.75	0.86	-0.04	-0.15***
Married	0.59	0.64	0.75	-0.05*	-0.16***
Number of Children	0.46	0.60	N/A	-0.14***	N/A
Education (= years in school)	8.29	6.46	9.48	1.83***	-1.18***
Raven Score	3.84	2.02	2.74	1.83***	1.10***
Experience in Garments	3.47	2.82	4.64	0.65***	-1.18***
Tenure in Factory	3.10	2.52	4.15	0.57***	-1.06***
Ever Received Training	0.15	0.08	0.15	0.07***	0.00
Received Training in Factory	0.06	0.12	0.13	-0.058***	-0.013

Table A4: Outcomes 4 Months after Training

Dep. Variable: Working after 4 months	[I] **OLS:attrition at FU1**	[II] **OLS: In factory at FU1**	[III] **OLS: Tried out by FU1**	[V] **OLS: working as SV at FU1**
Training	.008	014***	0.560***	0.592***
	(0.031)	(0.03)	(0.086)	(0.083)
Female	0.064**		-0.160**	-0.175**
	(0.026)		(0.068)	(0.073)
Training X Female	-0.084*		-0.0002	-0.029
	(0.038)		(0.090)	(0.093)
Dep. Variable, male controls	0.1	0.87	0.43	0.25
Dep. Variable, female controls	0.18	0.8	0.168	0.114
Demographic Controls	No	No	No	No
Factory Fixed Effects	yes	yes	yes	yes
Sample	All	All	All	All
N. Observations	523	523	454	360

Notes: Attrition measure by having information on ever being tried as an SV. Infactory is set to zero when we have no information on the worker after baseline. Where interaction effects are now shown, they are highly insignificant. Controlling for demographic characteristics of the operators makes very little difference, but reduces sample size. The IV regressions are on the treatment/control sample.

Table A5: Promoted Trainees vs. Existing Supervisors: Productivity

	ANCOVA Specification			LINE FE Specification		
Outcome Variable	Efficiciency	Quality defects	Absenteeism	Efficiciency	Quality defects	Absenteeism
Male Trainee as SV	-0.0350	0.0108	-0.0009	-0.0049	0.0066	0.0029
	(0.0337)	(0.0088)	(0.005)	(0.022)	(0.0063)	(0.0046)
Female Trainee as SV	-0.0181	0.0156	-0.0065	-0.0046	0.0041	-0.0056*
	(0.0113)	(0.0154)	[0.0036]	(0.009)	(0.0060)	(0.0032)
Y_pretraining period	0.253**	0.209***	0.348***			
	(0.090)	(0.030)	(0.070)			
Y_Training period	0.361***	0.118***	0.281***			
	(0.032)	(0.039)	(0.035)			
Female vs. Male Trainee	0.017	0.0048	-0.0056	0.0003	-0.0025	-0.0085
Dependent Variable Mean	0.464	0.099	0.07	0.459	0.096	0.068
Factory-Month Fixed Effects	yes	yes	yes	yes	yes	yes
Line Fixed Effects	No	No	No	yes	yes	yes
Number of Lines	295	263	151	344	337	247
Number of Observations	77376	44963	32109	85862	52691	38663

Note: day level specification. Standard Errors clustered at factory_id level. Day fixed effects not included (but they don't matter). *** 1%, ** 5%, * 10

Appendix B: Production Data, Description and Collection

As part of both Phase 1 and 2 of the project, we collected daily production data from all factories on the sewing line level. The data is similar in its format and organization across the two project rounds. However, in Phase 1 of the project we collected data in a two week interval every other month, while in Phase 2 we collected data for each day between January 2014 and March 2015. Given the continuity and greater amount of data, we base the analysis in the main part of the paper on the data from Phase 2, which we describe in more detail in this appendix.

We collected the data with three main outcome variables in mind: line-level productivity, the quality defect rate, and worker absenteeism. We asked factories to share all internal data needed to construct these variables. The standard measure of productivity in the Bangladeshi garment industry is (piecewise output * SMV)/(workers * daily hours * 60^{mins}), where SMV is the Standard Minute Value of the garment being produced. The SMV is the time industrial engineers estimate a fully efficient production line would take to produce one unit of the garment. When estimated to a common standard, the SMV thus allows us to compare the efficiency of production of different products - e.g., the efficiency of a line producing a tank top with an SMV of six minutes

can be compared with the efficiency of a line producing a dress shirt with an SMV of 18 minutes.

We asked the factories to provide productivity records for each sewing line and day detailing on daily output, the number of defective units, the SMV of garment being produced, the number of hours each line operated, and daily number of workers present and absent on the line. Not all factories record information on all of the variables. In some instances, the factories record data, but declined to provide it for certain outputs. For example, one factory declined to provide SMV data, and a few others do not have industrial engineering departments, and hence do not estimate SMVs by product. For other variables, there are sometimes differences in the specific data the factories record, though often these differences are not consequential. For example, for defects, we sometimes received defect rates (defective units / output) and sometimes the number of defective garments. Records on absenteeism would sometimes contain information on the numbers of workers assigned to the line, allowing to standardize the absenteeism numbers. At factories where this information was not included, we instead standardized the number of absent workers by the number of present workers provided in the productivity data.

In almost all factories, the three types of data (on productivity, defects, and absenteeism) was provided by different departments within the factories (usually the production, quality, and HR departments), and thus came in different formats, which required to enter the data separately and subsequently merge them to one document. Likewise, in most factories, the data we requested was provided in a digital format, usually a spreadsheet maintained by the factories, which allowed for easy collection and entering. At some factories however, data was provided as copies of paper files, requiring the data be digitised. Ultimately, though, we harmonise the data so that variables are comparable across factories.

As we noted, the data from some factories did not contain the information necessary to calculate all of the outcomes of interest. This is especially the case for efficiency, where our standard calculation relies on the availability of the SMV data. Some of the factories that do not measure SMV have other data which can be used to estimate a roughly equivalent measure of efficiency. For example, four factories in the Phase 2 sample have information on daily targets for their sewing lines. By assuming

that the targets are set such that line efficiency would be 100%, we can back out a 'synthetic SMV' by setting Daily Target * SMV = workers * daily hours * 60mins.[21]

From 17 of the 19 factories remaining in the project throughout, we were able to collect data for at least one of our three outcome variables of interest; productivity, defects, and absenteeism. Table B1 shows from how many of these 17 factories we could collect enough data to construct each of the three variables, and for how many we could construct all three. While the availability of defects data is most widespread, productivity data is reduced by a number of factories recording neither SMVs nor targets. Finally, the availability of absenteeism data for our analysis is limited by a number of factories recording only daily absenteeism numbers for the whole factory (or sometimes the sewing floor), but not recording data on the sewing lines on which workers are assigned.

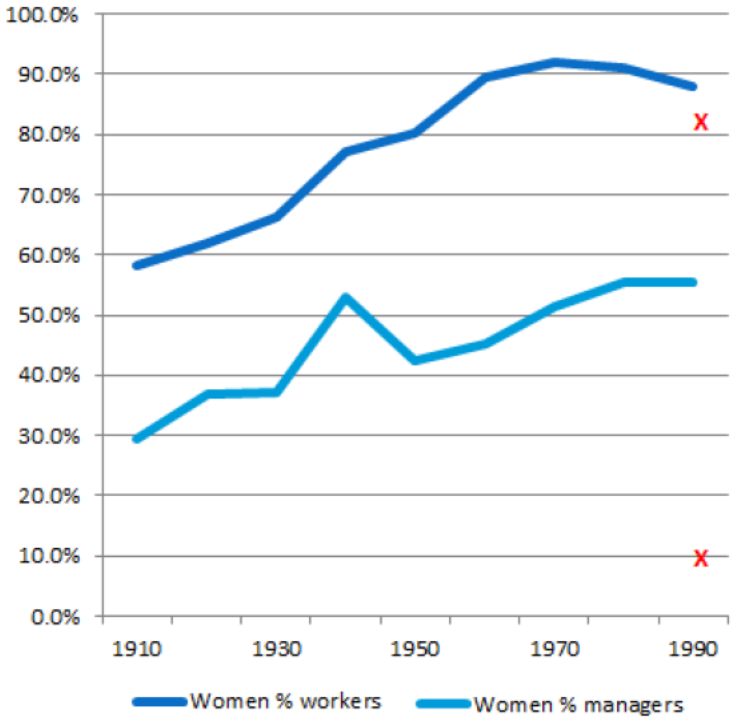
Table B1: Production Data Availability

Outcome variable	# Factories
Productivity	14
Productivity, excluding synthetic SMV	10
Defects rate	16
Absenteeism	10
Productivity + Defects + Absenteeism	7
Productivity (excl. synth. SMV) + Defects + Absentee-ism	4
Total Number Factories with some Prod. Data	17

Notes: Table shows for how many factories participating in Phase 2, usable daily data on line-wise productivity, defects rates, and absenteeism could be collected.

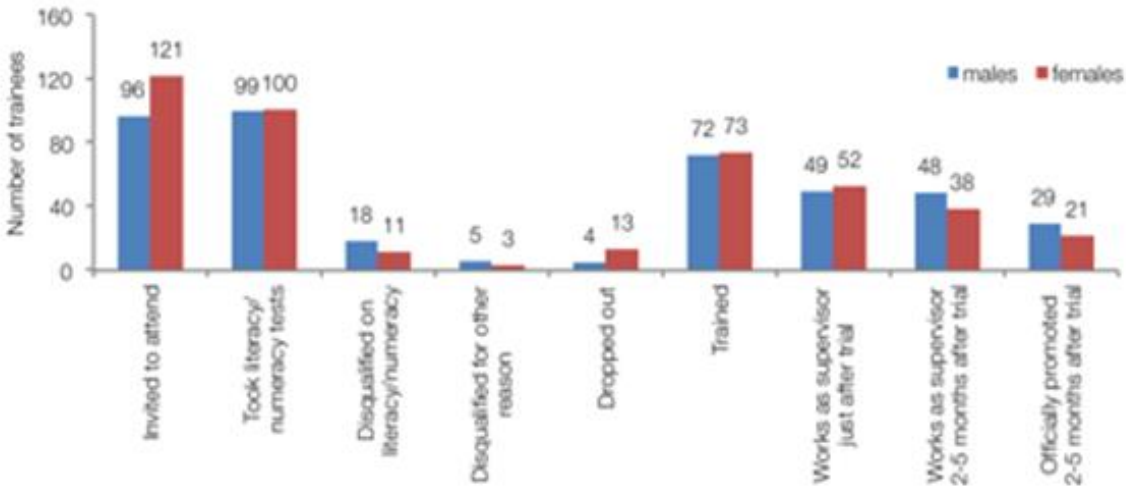
[21] Factories from which both SMV and targets are available show that targets are usually not set such that efficiency, in case the target is met, is 100%. Rather efficiency in these cases would be around 50%, which is in line with the typical average efficiency in almost all Bangladeshi garment factories. Thus, the 'synthetic SMVs' which we back out using targets are likely to overstate actual SMVs by a factor of two. And indeed, efficiency values at those factories where we use 'synthetic SMVs' are on average twice as high as in the other factories (93% vs 47%). However, note that all analysis we conduct with the production data uses factory fixed effects, therefore relying only on within factory variation in productivity. Given that for each factory we use either only productivity based on original or synthetic SMVs, the productivity data is consistent within each factory.

***Figure 1*: Female Worker and Manager in Garment Industry, U.S. vs. Bangladesh**



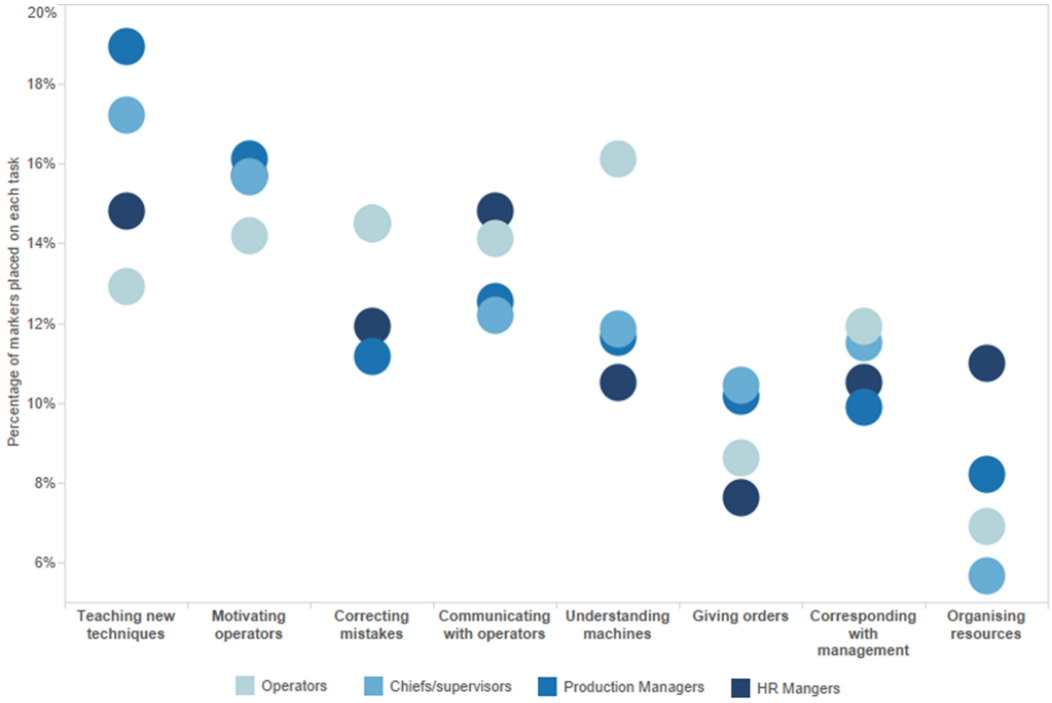
Notes: Figure shows the historical evolution of the share of female workers and managers in the US garment industry, and compares it against the current shares in the Bangladeshi industry. US Data from Ruggles et al. [2010], Bangladeshi data own calculations.

***Figure 2*: Selection, Training, Trial and Promotion of Trainees**



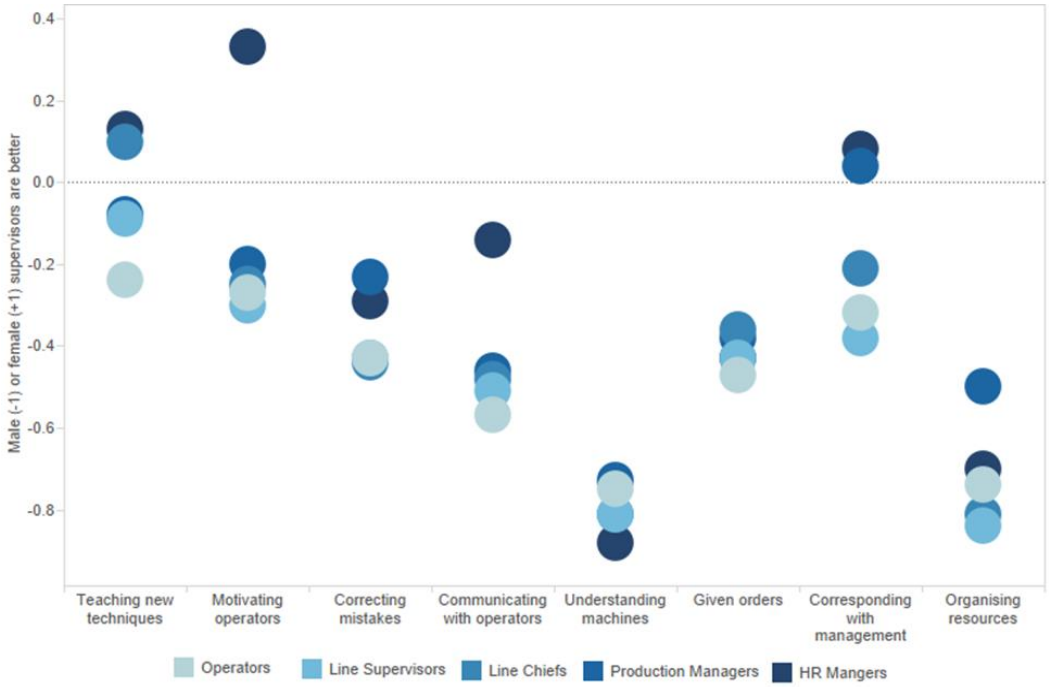
Notes: Figure shows the number of female (red bars) and male (blue bars) trainees which participated or dropped out in different stages of the project, and which got promoted to supervisor levels in their factories.

Figure 3: Tasks of Supervisors, Attached Importance



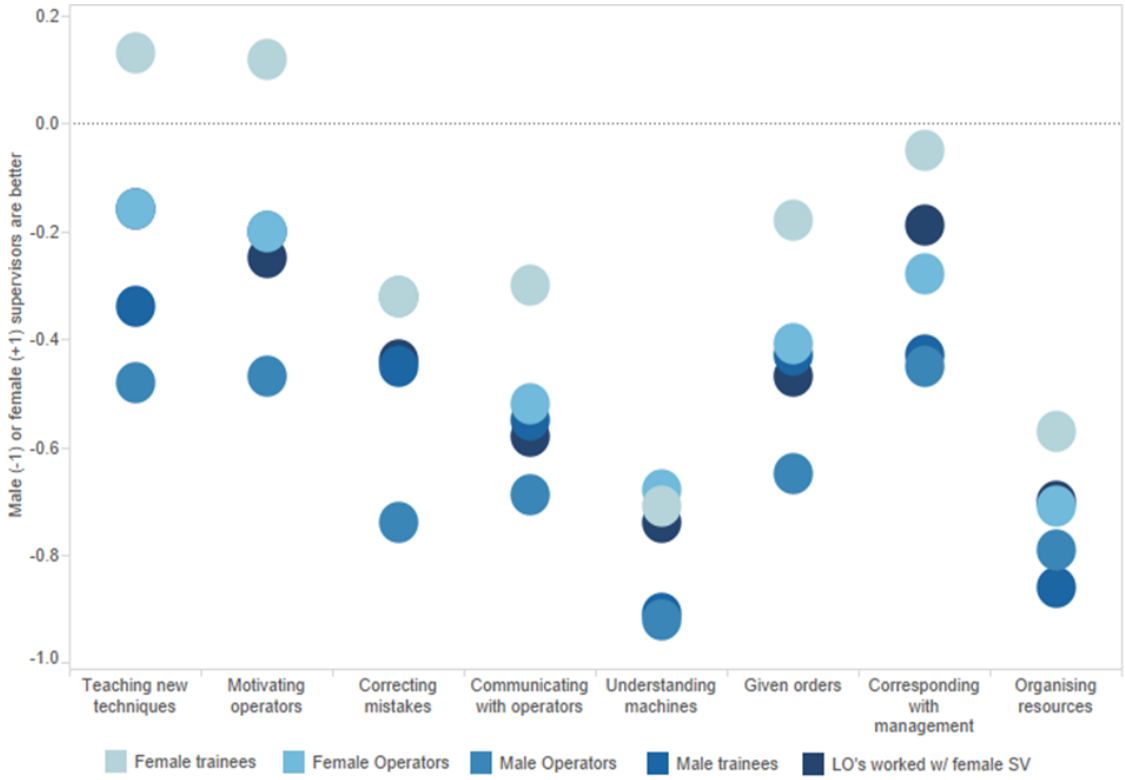
Notes: Workers on various levels in 26 factories were asked to place 10 tokens on a list of eight general tasks of line supervisors (generated after open ended conversations with several factory managers), according to the relative importance they attach to the task.

Figure 4: Tasks of Supervisors: Perceived Ability by Gender



Notes: Workers on various levels in 26 factories were asked for each of the eight main supervisor tasks, whether they perceive female or male supervisor as more capable. Answers were aggregated on the task and designation of respondent level, with answers being coded as -1 for “males are more capable", 0 for “both are equally capable", and 1 for “females are more capable".

***Figure 5*: Tasks of Supervisors: Perceived Ability by Gender, Extended**



Notes: Workers on various levels, of different gender, and with varying experience of working under female supervisors in 26 factories were asked for each of the eight main supervisor tasks, whether they perceive female or male supervisor as more capable. Answers were aggregated on the task and group of respondent level, with answers being coded as -1 for “males are more capable", 0 for “both are equally capable", and 1 for “females are more capable".

Table 1: Demographic Characteristics

Panel A: Trainees vs. Operators and Supervisors

	Mean				Comparisons	
	Trainee Pool	Operators	Random SVs	Mentor SVs	Trainee vs. Operators	Trainee vs. Random SVs
	N = 199	N = 430	N = 92	N=142		
Gender (female =1)	0.50	0.73	0.04	0.04	-0.23***	0.46***
Age (current)	24.4	24.1	29.1	29.3	0.30	-4.66***
Age at time of promotion to SV	24.4	NA	25.3	24.3	NA	0.84
Married	0.71	0.77	0.85	0.89	0.06	-0.14**
Education (= years in school)	7.83	5.80	8.77	9.56	2.03***	-0.95***
Experience in Garments (years)	6.53	6.10	8.83	9.21	0.43	-2.30***
Tenure in Factory (years)	3.41	2.78	3.30	3.80	0.63***	0.11***

Panel B: Trainees vs. Mentor Supervisors

	Mean		
	Trainee Pool	Mentor SVs	Trainee vs. Mentor SVs
	N = 197	N=113	
Literacy	7.14	9.54	-2.39***
Numeracy	3.4	5.2	-1.72***
Non-verbal reasoning	2.75	2.78	-0.03

Notes: The table reports mean characteristics of trainee and random sewing operators, random line supervisors, and the supervisors selected as mentor supervisors, all selected from the lines where the trainees were to be assigned to work as assistant supervisors following training. Statistical differences in comparisons: *** p<.01; ** p<.05; * p<.10

Table 2: Attitudes toward female SVs: Baseline data

	(1) Females better than males, 8 tasks	(2)	(3) Prefer female SV to male SV	(4) Prefer female SV to male SV
Operator is female	0.122***	0.142***	0.175***	0.208***
	(0.017)	(0.022)	(0.043)	(0.046)
Experience working for female SV	0.011		0.122***	
	(0.017)		(0.039)	
Experience * female operator		-0.002		0.099*
		(0.019)		(0.051)
Experience * male operator		0.047		0.182***
		(0.028)		(0.057)
Obs	428	428	426	426
R-squared	0.19	0.19	0.18	0.18
Factory FE	YES	YES	YES	YES
Mean	0.25	0.25	0.32	0.32

Notes: Coefficients show the effect of each variable on the probability of saying females are more skilled or preferred as supervisors. Standard errors clustered at the production line level; regressions include age, education and marital status of the respondent. Statistical differences: *** p<.01; ** p<.05; * p<.10

Table 3: Training and Trial Effects by Gender

VARIABLES	(1) Aptitude score	(2) Percentage drawings correct	(3) Drawing, "soft" score	(4) Communica-tion, "soft" score	(5) Leadership, "soft" score	(6) Self confi-dence
Female trainee, baseline	-1.19	-0.09	0.22	-0.27	-0.60	-0.47*
	(0.90)	(0.06)	(0.56)	(0.72)	(0.53)	(0.24)
Female trainee, after training	-0.69	0.13**	0.22	-0.70	-0.80	0.24
	(0.86)	(0.06)	(0.60)	(0.81)	(0.67)	(0.25)
Male trainee, after training	0.20	0.18***	0.10	0.21	0.09	0.50**
	(0.80)	(0.06)	(0.64)	(0.79)	(0.79)	(0.25)
Female trainee, after factory trial	-4.93***	0.09	-0.43	-0.01	-1.09	0.32
	(1.82)	(0.08)	(0.58)	(0.85)	(0.69)	(0.24)
Male trainee, after factory trial	-0.37	0.05	0.24	-0.75	-0.07	0.58**
	(1.25)	(0.07)	(0.69)	(0.96)	(0.76)	(0.24)
Baseline mean, male trainees	55.4	0.33	-0.15	0.16	0.35	-0.33
Observations	470	420	421	423	329	470
R-squared	0.191	0.125	0.098	0.068	0.125	0.133
Factory FE	YES	YES	YES	YES	YES	YES
Female vs. Male, after training (p)	0.28	0.50	0.86	0.30	0.30	0.32
Female vs. Male, after trial (p)	0.02	0.66	0.32	0.48	0.20	0.30

Notes: Unbalanced panel of trainees measured on first day of training, last day of training, and at the end of the factory trial period. Statistical differences: *** p<.01; ** p<.05; * p<.10.

Table 4 Comaprison of Female and Male Trainees

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Midline							Follow-up		
	ITT			OLS						
VARIABLES	Rank (1-10) for Trainee		Prefer Male SV	Rank (1-10) for Trainee Full sample		SV at FU	Prefer Male SV	Rank (1-10) for Trainee		Prefer Male SV
Female trainee on line	-0.90***			-0.91***				0.06		
	(0.25)			(0.22)				(0.29)		
Female trainee on line * Female operator		-0.65**	-0.01		-0.79***	-0.38	0.02		0.04	0.16***
		(0.31)	(0.06)		(0.26)	(0.30)	(0.05)		(0.31)	(0.05)
Female trainee on line * Male operator		-1.39***	0.18***		-1.19**	-1.04	0.12**		0.13	0.17*
		(0.52)	(0.06)		(0.46)	(0.65)	(0.06)		(0.57)	(0.10)
Operator is female	0.13	-0.25	0.24***	0.15	-0.03	-0.20	0.16***	0.48*	0.51	0.16**
	(0.32)	(0.52)	(0.13)	(0.27)	(0.39)	(0.50)	(0.05)	(0.28)	(0.34)	(0.07)
Ability (1st PC)	0.00	-0.01	0.003	0.14	0.14	0.23**	-0.01	0.20*	0.20*	0.01
	(0.10)	(0.10)	(0.014)	(0.11)	(0.11)	(0.11)	(0.01)	(0.11)	(0.11)	(0.014)
Rank for typical supervisor	0.41***	0.41***		0.35***	0.35***	0.47***		0.35***	0.35***	
	(0.09)	(0.09)		(0.07)	(0.07)	(0.08)		(0.08)	(0.08)	
Observations	238	238	357	295	295	203	410	206	206	215
R-squared	0.221	0.226	0.116	0.232	0.233	0.314	0.093	0.279	0.279	0.201
Mean of male operators, prefernce			**0.22**				**0.22**			**0.25**
Avg ranking Typical SV				**7.41**	**7.41**	**7.41**		**7.58**	**7.58**	
Avg ranking of trainees				**6.86**	**6.86**	**6.86**		**7.82**	**7.82**	
Line FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

Notes: Ability is the first principal component of years of schooling and scores on the numeracy, literacy, non-verbal reasoning and technical aptitude test. Statistical differences: *** p<.01; ** p<.05; * p<.10.

Table 5: Productivity of Trainees on the Line

MODEL	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	---------- ITT ----------			---------- Actual Assignments ----------			
VARIABLES	Efficiency	Absenteeism Rate	Defect Rate	Efficiency	Efficiency ITT lines	Absenteeism Rate	Defect Rate
Trial period, Female trainee	-0.0093	-0.0010	-0.0020	-0.0237	-0.0174	0.0007	-0.0013
	(0.0118)	(0.0054)	(0.0037)	(0.0165)	(0.0180)	(0.0035)	(0.0029)
Trial period, Male trainee	0.0236**	0.0050	-0.0042	0.0190	0.0112	0.0070	-0.0043
	(0.0108)	(0.0046)	(0.0036)	(0.0146)	(0.0154)	(0.0046)	(0.0032)
Post-trial, Female trainee	0.0176	-0.0004	-0.0055*				
	(0.0150)	(0.0039)	(0.0032)				
Post-trial, Male trainee	0.0196	-0.0025	-0.0013				
	(0.0128)	(0.0030)	(0.0035)				
Assistant SV, Female Trainee				0.0250	0.0220	-0.0049	0.0043
				(0.0177)	(0.0170)	(0.0083)	(0.0042)
Assistant SV, Male Trainee				0.0240**	0.0302	0.0082**	-0.0057
				(0.0121)	(0.0238)	(0.0036)	(0.0049)
Line SV, Female trainee				0.0101	0.0369	0.0039	-0.0036
				(0.0153)	(0.0244)	(0.0043)	(0.0037)
Line SV, Male trainee				0.0015	-0.0060	-0.0045	0.0006
				(0.0113)	(0.0147)	(0.0043)	(0.0035)
Observations	78241	70814	70814	78241	18888	70814	87952
R-squared	0.54	0.39	0.39	0.53	0.56	0.39	0.59
Number of factory_Lines	419	264	431	495	495	495	495
Line FE	YES	YES	YES	YES	YES	YES	YES
Factory/week FE	YES	YES	YES	YES	YES	YES	YES
TEST: Female vs. male trainee, trial period	**0.027**	0.375	0.663	0.045	0.179	0.293	0.475
TEST: Female v.s male trainee, Asst Supervisors	0.916	0.636	0.336	0.962	0.792	0.141	0.111
TEST: Female v.s male trainee, Supervisors				0.631	0.096	0.185	0.435
Nr. Lines assigned Male Trainee:	49	32	51				
Nr. Lines assigned Female Trainee:	43	28	44				

Notes: Efficiency is defined as the output sewing minutes divided by the input labor minutes. The defect rate is the number of products requiring re-work divided by total output. See textand appendix for more details. Robust standard errors clustered on the line level; *** p<0.01, ** p<0.05, * p<0.1

Table 6: Management Simulation Exercises

Trainees Females Vs. Males	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Outcome: Standardized Pay-Off in Games						
Outcome: Pay-Off in Games (Standardized)	OLS	OLS	Team FE	OLS	OLS	OLS	Correct Strategy
Female Team Leader	0.290***	0.255**	0.420**	0.305**	0.332***	0.412***	-0.371**
	(0.109)	(0.122)	(0.190)	(0.131)	(0.124)	(0.521)	(0.167)
Mixed gender / male team						-0.092	
						(0.725)	
Mixed * Female TL						-0.433*	
						(0.240)	
Tried as Line Supervisor				0.329			
				(0.212)			
Promoted to Line Supervisor					0.508**		
					(0.206)		
Team Fixed Effects	no	no	yes	no	no	no	no
Game Fixed Effects	yes	yes	yes	yes	yes	yes	yes
Team Leader Demogr.	no	yes	yes	yes	yes	yes	yes
Number of Observations	676	612	612	612	608	612	612

Notes: The dependenet variable is the standardized payoff from the game. Standard errors are clustered at the game level. *** p<0.01, ** p<0.05, * p<0.1.