Learning How to Choose or Learning How to Lead? Experiments on Selecting and Training Female Managers in Bangladesh's Garment Industry

Hannah Uckat and Christopher Woodruff *

Abstract

We report the results of field experiments designed to understand the importance of the selection of and training for new female supervisors in Bangladesh's garment factories. Participating factories have little prior experience with promoting women. We show that formal diagnostic tests lead factories to select candidates that are more likely to be promoted and who, according to their subordinates, perform better as supervisors. Diagnostics measuring attitudes and soft skills are particularly relevant for factories and predictive of later outcomes. Supervisory training for the selected candidates leads to higher rates of promotions, but has only marginal effects on performance. In none of our results do we find that training in technical skills has an additional effect when compared to a training that focuses on attitudes and soft skills. These results indicate the importance of hard measures of soft skills and attitudes in the process of selecting female supervisors, and suggest that training in non-cognitive skills could be a promising avenue to increase the participation of women in managerial positions.

**The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent. The paper has benefited greatly from discussions with and comments by Abi Adams, Elwyn Davies, Kate Orkin, and from audiences at the EDI Annual Conference 2020, the Monash Conference on RCTs 2019, and OxFLO 2017-18. The authors thank Gunjita Gupta and Ferran Vega for research assistance, and the International Finance Corporation, the Urban Services Initiative (USI) at J-PAL, and the Economic Development and Institutions research programme (EDI), funded by UK aid from the UK government, for financial support. Woodruff's time was supported by ERC Advanced Grant RMGPP. The surveys and interventions were supported by the Bangladesh office of Innovations for Poverty Action (IPA). The project would not have been possible without the support and cooperation of participating factories and workers, which is greatly appreciated. This study, together with Uckat (2020), was registered with the AER RCT Registry as AEARCTR-0001843, and received ethical approval from the University of Oxford and Innovations for Poverty Action.

1 Introduction

Supervisors matter. Lazear et al. (2015) find that replacing a supervisor in the 10^{th} percentile with one in the 90^{th} percentile increases output of a work team by more than 10 %. However, identifying the best candidates for supervisory positions is challenging. Benson et al. (2019), using detailed internal records from a large US firm, show that the best workers do not necessarily make the most effective supervisors. For higher-level managers, making promotion decisions is essentially a prediction problem. The difficulty is to identify, measure, and base promotions upon characteristics that are correlated with more successful performance after promotion. Evidence from similar settings suggests that this type of prediction exercise is hard (McKenzie and Sansone, 2017; Fafchamps and Woodruff, 2017). Firms may focus on a suboptimal set of characteristics when selecting employees or leaders, especially when they have limited prior experience in making these decisions (Ashraf et al., 2020; Bandiera et al., 2020).

The challenge of selecting the best supervisors arises at least in part because, compared with workers, managers carry out a much more diverse and complex set of tasks. This increase in complexity and variety of tasks implies that new managers need to learn new skills, either on-the-job or through training. However, moulding managers through education is fraught with difficulties. Training programs for managers and entrepreneurs across a wide range of contexts have produced only limited improvement in firm outcomes (McKenzie and Woodruff, 2014).

We conduct experiments to understand the importance of manager selection and training in a particularly challenging context: promoting women to supervisors in Bangladesh's garment industry. Garment factories in Bangladesh are increasingly interested in promoting women. The factories face increasing competition for male employees in domestic labour markets and pressure from international brands to provide opportunities for women. More than three quarters of workers in the sewing sections of factories are women, but female supervisors remain very rare; less than 10 % of supervisors are female (Menzel and Woodruff, 2019).

Previous research highlights several factors which suggest that factories have limited information both on how to select and on how to train the best female candidates for supervisory positions. Candidates for promotions in the factories are typically recommended by existing supervisors and higher-level managers. Beaman et al. (2018) report experimental evidence that men are more likely to refer other men for jobs, even when they are aware of better-qualified women. Moreover, the characteristics of the best female candidates may differ from those of the most qualified males, which impedes learning. Managers may overlook "obvious" characteristics because they lack experience in choosing female candidates. In addition, the feedback to decision-makers whether the decisions were correct is slow. In this context, all new supervisors typically spend many weeks in a trial period, and then often underperform initially before growing into the job. Outside of this on-the-job-training during a trial period, formal training for new supervisors is rare. Macchiavello et al. (2020) show that formal training can increase the share of female supervisors in Bangladeshi garment factories, who perform as well as new male supervisors after an adjustment period.

With this in mind, we conduct experiments in two acts with 27 large garment factories in Bangladesh. We aim to investigate how predictive initial skill diagnostics are for the success of new female supervisors, and whether a supervisory training can produce successful supervisors. The factories were asked to nominate women for a training program designed to prepare them for supervisory positions. At the beginning of the project, one-third of the participating factories had no female supervisors, and there

was no factory where more than 20 % of supervisors were female. We asked managers to rank the candidates in order of their expected performance as supervisors.

In the first intervention, we implemented a selection experiment with higher-level management staff in a randomly selected half of the factories. The participants of the selection experiment received a two-hour session introducing them to diagnostic tests as a means of selecting candidates for promotions. At the end of the training, managers were shown scores from seven diagnostic tests conducted with the women they had nominated for the supervisory training program. Four of the diagnostic tests related to cognitive ability or technical skills: tests of literacy, numeracy, fluid intelligence and garments knowledge. We aggregate these scores into an index of initial aptitude. The other three scores were measures of non-cognitive skills and attitudes that previous research showed were important to the success of female supervisors: confidence in the ability to work as a supervisor, interest in the position, and support from the family for being a supervisor.[1] We aggregate these into an initial attitude index. The attitude index therefore combines aspects of non-cognitive skills which are often termed soft skills in economics (Heckman and Kautz, 2012), as well as attitudes of the individual and the family that quantitative and qualitative research has shown to be predictive of women's promotion rates in this context (Macchiavello et al., 2020). After reviewing the diagnostic scores for all of the factory's nominees, the managers that participated in the selection experiment were invited to re-rank the candidates if they desired to do so. The factories that did not participate in the selection experiment were not able to see their nominees' diagnostic scores nor could they change their ranking from the one they had initially provided.

In the second intervention, we provided supervisory training to the selected trainees in all factories. We randomised those selected for training into three groups. The first received a four-day training program focusing on attitudes and soft skills – stress management, assertive communication and leadership – and a five-day training programme in aptitude and hard skills required of a successful supervisor – line balancing, quality control, etc. – immediately. The second group received the attitude training immediately and the aptitude training about six months later. The third group received both training modules six months later. All of the nominees were assigned to trial immediately as assistant supervisors for at least two months, which is the status quo method of training candidates for supervisory positions.[2]

Using the variation created by these experiments, we document four results with a combination of survey and administrative data. First, factories do not take into account the initial attitude and aptitude measures of candidates when they nominate women for the promotion training. That is, the initial ranking of nominees before the diagnostic tests are taken is orthogonal to the initial attitude and aptitude indices we measure. Second, the factories that are provided information about their candidates' skills and can change their ranking as part of the selection experiment react to this information. They

[1] Data from the supervisor training programme in Macchiavello et al. (2020) show that baseline interest in promotion to supervisor is associated with completion of the training program (rather than dropping out), and continuing to work as a supervisor at the time of the final survey. The claim on family support is based on qualitative research with factory management. None of the factories reported using similar standardised tests in their promotion decisions prior to the project. Qualitative research suggests that, in the absence of the project, the identification of potential new managers is an informal process. For example, it appears to be common that established line supervisors suggest workers from their lines for supervisor vacancies, often those who have shown leadership ability in the group or who have approached the supervisors to express their interest in a promotion.

[2] Of course, the trainees either in the status quo or in our programme may drop off of the supervisory ladder either because they themselves decide they do not want to be a supervisor, or because managers decided they are not qualified for the position.

select candidates who have higher baseline skills, especially on the attitude dimension. We posit that the attitude diagnostics have a large effect on managers' ranking because they are characteristics that are likely to be less relevant for male supervisory candidates. Similar to Hanna et al. (2014), the attitude diagnostics lead the managers to notice characteristics that they otherwise are not conditioned to notice. This also ties in with evidence on self-selection of workers into jobs. For example, Ashraf et al. (2020) find that emphasising career prospects in advertisements for health worker positions in Zambia leads to a more talented but less pro-social applicant pool, and better-performing hires.

Third, we find that the initial skills – and especially attitudes and soft skills – matter for the outcomes of the promotion programme. They are predictive of a gain in skills, the promotion rates to official line supervisor, higher wellbeing of their subordinates and – to some degree – higher performance evaluations as judged by subordinates. We do not find that they matter for production outcomes such as efficiency and alteration rates. Fourth, attitude and aptitude training for the participants of the supervisory training leads to some improvements in skills, promotions, and performance, though the results are not always significant. In none of our results do we find that training in technical skills has an additional effect when compared to a training that only focuses on attitudes and soft skills.

The evidence we present suggests that both the selection process and the supervisory training matters. We find that initial skills predict the new female supervisors' success, and observe some evidence that training can make the supervisors more successful. For the new female low-level managers that are the focus of this paper, attitudes and soft skills appear to be particularly important. The results of the selection experiment suggest that factories had some idea that attitudes and soft skills would matter for their trainees' success, but the little experience in choosing female supervisors likely meant that they lacked the ability to convincingly measure these skills. In the case of female career advancement, this paper therefore not only provides evidence for the importance of hard measures of soft skills and attitudes, but also indicates that training in non-cognitive skills is a promising avenue to increase the participation of women in managerial positions.

In addition to the potential to increase the pool of management talent in factories, increasing the rate of promotion of women may also have transformative effects on women's lives. Female labour force participation generates changes in development dynamics in lower-income countries, increasing female empowerment and educational attainment of children (Heath and Jayachandran, 2018; Duflo, 2012; Qian, 2008; Heath and Mobarak, 2015). The garment sector provided an entry into wage work for women in Bangladesh. Even in 2017, garment factories were the workplace for 40 % of female wage workers with less than tertiary education. As in other countries, employment in the sector is responsible for higher educational attainment and later age of marriage for women (Heath and Mobarak, 2015). Uckat (2020) examines the impact of a promotion on the women's position in the household, both for the women selected for training and for those working under their direction, and finds significant increases in the women's bargaining power.

The paper proceeds as follows. We discuss the contribution to the literature in Section 2, and explain the two interventions in Section 3. We discuss our empirical strategy and present results in Section 4. Section 5 discusses the results and concludes.

2 Contribution to the literature

This paper is contributing to the literature that describes challenges in worker and manager selection, and investigates whether and how managerial success can be predicted. For example, Benson et al. (2019) for US firms find that prioritising current job performance over observable characteristics that predict managerial success in promotion decisions entails high costs – which, however, could be justified by the benefits of promotion-based incentives for workers. In a similar vein, Hoffman et al. (2018) and Autor and Scarborough (2008), again in the US, find that using skills testing in recruitment decisions increases the quality of new hires compared to relying on human judgement.

In low- and middle-income countries, however, the literature on managerial selection has largely focused on managers of micro and small enterprises, which represent the vast majority of firms in these economies.[3] Fafchamps and Woodruff (2017) in Ghana, McKenzie and Sansone (2017) for Nigeria, and Hussam et al. (2020) in India find that baseline survey data of entrepreneurs or firms can predict firm growth. Even though these key characteristics outperform the predictions of expert judges in business competitions in both Fafchamps and Woodruff (2017) and McKenzie and Sansone (2017), Hussam et al. (2020) observe that truthfully reported peer predictions can outperform baseline characteristics. As Quinn and Woodruff (2019) point out, hard skills of entrepreneurs are typically found to be more important than soft skills in this strand of the literature. Overall, however, the variation in future outcomes explained by the characteristics the authors look at is small.

Compared to the literature on managers of micro or small enterprises, our extensive diagnostic tests allow us to focus on predicting success for a very different set of managers. Line supervisors are the lowest level of managers in highly hierarchical, large garment factories. Compared to the owner-managers of micro and small firms discussed in most of the literature, who are responsible for leading and running every aspect of their enterprise, line supervisors have very different responsibilities. As will be further discussed below, line supervisors are assigned to a production line or part of a production line and are tasked with ensuring that their line is running well and meeting its production target. This typically involves motivating the subordinate line operators, solving simple problems on the line, and communicating more complex problems to higher-level managers.

Compared to the literature on manager selection in high-income countries, we present the first experimental evidence on the selection of managers, by allowing a random subsample of factories to reconsider the selection of trainees for a promotion programme after presenting the results of the diagnostic tests to them. This allows us to test whether factories are well-informed about candidates for management positions in advance, and whether they respond to the baseline data presented to them. If we do see a response, we can analyse whether factories respond to the characteristics that we identify as predictive.

We also contribute to a large literature on managerial training in low- and middle-income countries, which has found little evidence that standard business trainings for micro and small firms have large impacts on business performance or growth. This literature is reviewed in detail in Quinn and Woodruff (2019) and McKenzie and Woodruff (2014). While there is a lot of variation in the content and the length of the trainings investigated in the literature, they typically aim to improve entrepreneurs' business practices, mainly by teaching hard skills such as accounting, financial planning, or marketing. These practices have been shown to be correlated with business performance for both small and large

[3] See Quinn and Woodruff (2019) for a recent overview of experimental work on entrepreneurship in developing countries.

firms.[4] However, the typical business training does not appear to lead to large improvements in business practices over the longer term, which likely explains their limited effects on business performance or growth (Quinn and Woodruff, 2019).

Two newer approaches have shown more promise and are now being researched more widely. First, consulting services tailored to each firm's circumstances have shown success for small, medium and large firms. Bloom et al. (2013) offered intensive consulting services to large Indian textile plants. They found improvements in management practices and productivity in the first year, and an increase in new plant openings after three years, compared to a control group. Similarly, Bruhn et al. (2018) find that providing consulting to micro and small firms in Mexico leads to large increases in employment in the long term and some positive effects on firm profitability in the short term.

Second, trainings focusing on soft rather than hard skills have shown potential. For example, Campos et al. (2017) and Glaub et al. (2014) investigate so-called personal initiative trainings in Togo and Uganda, which aim to develop a proactive mindset by teaching "self-starting behavio[u]r, innovation, identifying and exploiting new opportunities, goal-setting, planning and feed-back cycles, and overcoming obstacles" (Campos et al., 2017). Compared to a standard business training, these papers show that personal initiative trainings have larger, positive effects on profits, input use, product diversification and access to finance.

By assigning women to different training groups, the supervisory training we implemented allows us to investigate the impacts of training in attitudes and soft skills only and the additional effects of training in technical skills. The training contents focused on the specific skills needed in a supervisor position in the garment industry (e.g. communication skills, managing bottlenecks on a production line), as compared to the standard business skills (e.g. accounting, financial planning) contained in the trainings described above. It is a shortened and improved version of the training in Macchiavello et al. (2020). However, Macchiavello et al. (2020) compare between female and male candidates for promotion and conduct attitude and aptitude training for all selected candidates. This paper focuses on female candidates and randomly assigns the training group to distinguish between the effects of building attitude versus aptitude.

The information on the effects of attitude and aptitude training on official promotions, skills, performance and production outcomes enable us to conduct a horse-race type analysis between baseline skills and a training in the same skills for the female candidates. We focus on female candidates for line supervisor positions since women continue to be underrepresented in managerial roles in the Bangladeshi garment industry (Menzel and Woodruff, 2019). Women only represent about 5 % of all supervisors in the industry (Macchiavello et al., 2020). The evidence in this paper therefore helps to illuminate how women's participation in management positions in this labour-intensive manufacturing industry could be improved – by better selection of suitable female candidates according to skills that predict managerial success, or by training women in these skills.[5]

3 Study design

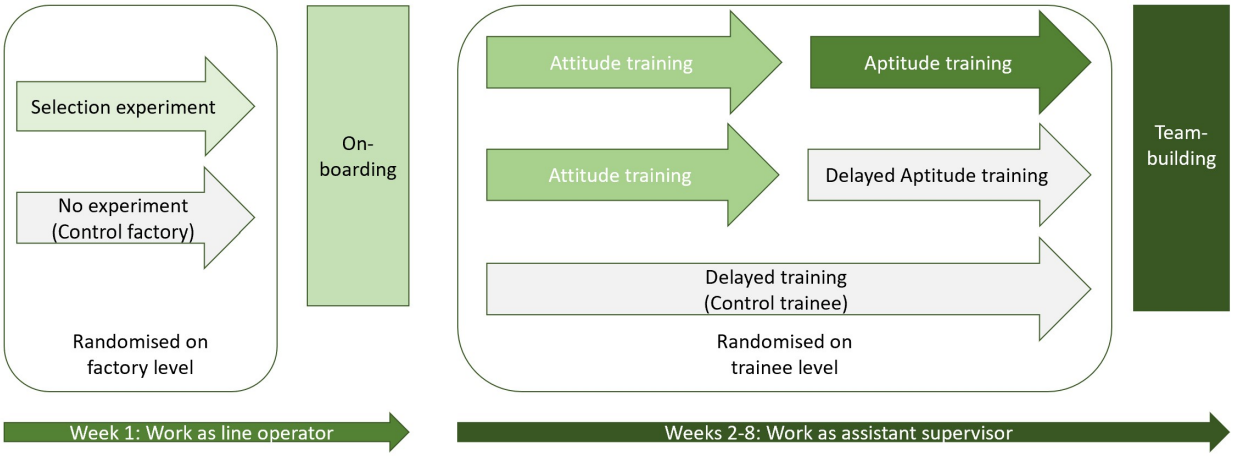
We implemented two distinct interventions sequentially. In the first, we conducted extensive diagnostic tests with all candidates for a promotion programme and implemented a selection experiment. During

[4] See, for example, McKenzie and Woodruff (2017) and Bloom et al. (2014).

[5] With this, we also contribute to the literature on female leadership, which is discussed in detail in Uckat (2020).

the first experiment, a random subsample of factories was able to reconsider their candidate selection based on the results of the diagnostic tests. In the second intervention, we trained female operators to become line supervisors. Figure 1 provides an overview of the design.

Figure 1: Overview of study design



At the beginning of the project, we determined the number of operators to train in each factory in conversations with the management of the 30 factories that initially agreed to participate. We advised them to select this number based on the number of openings for supervisory positions they expected to have in the following few months.[6] We then asked managers to nominate 1.67 times the number of workers they intended to train, and to rank all nominees based on their expected performance as supervisors. Thus, a factory that intended to train six women nominated ten for the programme, and provided us with a ranked list of the ten candidates. A smaller rank indicated a higher expected performance as supervisor.

Approximately two weeks after receiving this information from the factory, we conducted baseline surveys and diagnostics with all nominees. The survey elicited information on education, labour history, and other demographic data from each of the nominees. We also asked questions about relationships in their household, including measures of their participation in household decision-making.

For each nominee in all factories, we conducted diagnostic tests with seven components. The diagnostic, described in more detail in Appendix A, included tests of literacy, numeracy, fluid intelligence (measured by processing speed tests from the Wechsler Adult Intelligence Test), and technical knowledge related to garments production. In the analysis, we show results for the individual diagnostics and for a standardised index following Anderson (2008) that combines these four diagnostics into a single index measuring aptitude. We also asked questions which we group into three scores reflecting non-cognitive skills and attitudes: a measure of how much the nominee's family would support her working as a supervisor, a measure of the nominee's own interest in the position, and a measure of the nominee's confidence. In the analysis, we will group these three diagnostics into a single attitude index following Anderson (2008).

Our choice of diagnostic tests was informed by previous work in Macchiavello et al. (2020) and by qualitative research with factory management. Taking first the three measures of attitude, we included

[6]The variance in this expected demand led to variance in the number of trainees across factories. For example, some factories anticipated opening new production lines, and hence anticipated the need for a larger number of new supervisors.

the measures of confidence and interest because of earlier evidence indicating that female supervisor trainees had lower confidence levels than male trainees, and that the training closed this confidence gap (Macchiavello et al., 2020). Discussions with factory managers and focus groups research also revealed the importance of family support. Regarding the aptitude measures, previous work reports data from baseline surveys of workers at all level of the factory showing a wide-spread belief that men are better at "Understanding machines: which machines are appropriate for which tasks, and knowing when machines are not functioning properly" (Macchiavello et al., 2020). The garments knowledge diagnostic measures exactly this. Literacy was included in the diagnostics because it is required for successful participation in the training. Discussions with factory managers indicated that numeracy and fluid reasoning are widely believed to be important for supervisor performance. Note that we do not claim that these diagnostic measures are an exhaustive list of either aptitude or attitude. The measures almost certainly could be improved.

3.1 Selection experiment

Our first intervention allowed factory managers to use these quantitative diagnostic tools to select the best candidates for the promotion training. We refer to this as the "selection experiment". Factories were randomised into two equally sized groups.[7] The first received a two-hour introduction to the diagnostic tests, and the second served as a control group without any further intervention. The intervention in the treated factories provided data from previous projects that supported the case for promoting female supervisors. It also introduced the managers to the seven different diagnostic scores and allowed them to take part of the test themselves. This session was aimed at mid- and high-level managers – those involved in nominating the female operators and in making decisions about whether they would be promoted.

None of the factories reported using similar diagnostics in their promotion decisions prior to the management training. At the end of the session, managers were shown the scores for the seven diagnostics that we conducted with the nominees at the same time as the session with managers was run. Managers from factories that were randomised to the selection experiment were offered the opportunity to revise their rankings of the operators they had nominated for the supervisor training program.

For the factories that were in the control group for the selection experiment, the original ranking of nominees was changed only if nominees were ruled ineligible due to low scores on the literacy or numeracy tests (see Appendix A). In that case, all nominees were moved up in the original order. The nominees with the best ranks in the final ranking were selected for the supervisory training, up to the

[7]The intervention took place sequentially, as did the randomisation for the selection experiment. 26 of the initial factories were part of the International Labour Organizations BetterWork programme. For the BetterWork factories, the randomisation method was a stratified randomisation with rebalancing in two sequences. We used two stratification variables, 1) participation in a previous training project described in Macchiavello et al. (2020) and 2) a dummy for producing only knit garments. We re-balanced on three re-randomisation variables: 1) the number of workers in a factory, 2) the share of female workers and 3) the date of joining the BetterWork project. This information came from administrative data provided by BetterWork. We conducted the randomisation 1000 times in the first and 20 times for the second, smaller sequence. For each iteration, we identified the minimum p-value of the three two-sided t-tests of mean equality for the balancing variables between treatment and control (for the second sequence, the balance tests included all factories, including those assigned in the previous sequence). We re-randomised according to a clear criterion: Out of the assignments with a minimum p-value of >0.20, we randomly chose one. For the four participating non-BetterWork factories, we implemented a simple randomisation and allocated half to treatment and half to control. This, again, took place in two sequences. In Table 0.B.1 in the appendix, we show randomisation checks for the selection experiment for a wide range of factory characteristics and the average diagnostic scores of nominees. Out of 22 variables for which we test balance, we find one rejection using the p-value based on randomisation inference in the last column. This is fewer than one would expect to find by chance.

number determined in advance with the factory.

3.2 Promotion training

The best-ranked operators represented the pool for the second intervention, which was aimed at training the selected operators to be line supervisors. The training program was an improved and streamlined version of the program developed initially by GIZ and also used in Macchiavello et al. (2020). For this project, the training was reduced to nine days, divided into four days of training on attitudes and soft skills and five days on apttitude or hard skills. The attitude training focused on stress management, assertive communication, and leadership, elements which earlier research identified as being particularly important for female supervisors (Macchiavello et al., 2020).[8] The aptitude training focused on the technical skills required for line supervisors, e.g. production processes, sewing machines, quality control, cutting, finishing, printing, embroidery, and the responsibilities of the supervisory role. Appendix **??** of Uckat (2020) presents more details.[9]

In each factory, we conducted a public lottery to randomly allocate trainees into one of three groups. The first received both the attitude and aptitude training immediately, while the second group received only the attitude training immediately (and the aptitude training around six months later). The third did not receive any training initially and functions as our control group; they received both the attitude and aptitude training around six months later.[10] The training sessions were held on consecutive weeks at a local training centre on two days per six-day work week.[11]

In addition to the classroom-based training, all trainees were assigned to work as assistant supervisors on the line for two months from the start of the programme. We asked factories to choose a set of trial lines, and to assign the selected trainees to those lines. These choices were made before anyone was aware of which operators would be randomised into which of the three training groups. Hence, while the assignment of individual trainees to lines is not random, the assignment of training to the lines is random.

The standard practice in the factories is to train supervisors on the job by having them work for a period as assistant supervisors alongside an experienced supervisor. The assistant supervisors gradually take responsibility for an increasing number of sewing machine operators until they are managing a full section of the line. The group receiving both the soft- and hard-training sessions with six months delay thus mimicked the standard practice at factories.[12]

[8]This part of the training also included sessions on understanding harassment, developing integrity and fairness, workers' rights and responsibilities, and human resources management, including types of management styles.

[9]Compared to the full six-weeks GIZ training programme, the attitude and aptitude training in this project puts relatively more weight on attitude training. We chose to emphasise the attitude training because earlier work showed that the biggest effect of the GIZ training was to increase the confidence level of female trainees (Macchiavello et al., 2020). Of course, it is possible that the increased confidence comes as a result of an increase in technical skills, but we chose to focus the attention on building confidence more directly.

[10]In Table 0.B.2 in the appendix, we show randomisation checks for the promotion training for a wide range of demographics, non-cognitive measures and diagnostic scores of the trainees. Out of 27 variables for which we test balance, we find one rejection using the p-value based on a wild percentile-t cluster bootstrap in the last column. This is fewer than one would expect to find by chance.

[11]At the end of the classroom training, all trainees, supervisors, line chiefs, and industrial engineering officers in charge of the respective trial lines participated in a team-building session at the factory. This was designed to increase collaboration among those working on the same line to ease the incorporation of the trainee in the management team. This was implemented in the same manner for all three treatment groups.

[12]To increase acceptance of the new female supervisors, on the first day each trainee began working as an assistant supervisor, the project trainers conducted an "onboarding training" that involved line chiefs and supervisors from the lines where the trainees were assigned to trial as assistant supervisors. Lower level managers were provided with a short

For logistical reasons, the 30 participating factories were divided into five sessions of roughly equal size, with the first beginning in November 2016 and the last in March 2017. Three factories dropped out of the project after we conducted the baseline survey, but before the promotion training began, leaving 27 factories that completed the program. We use the 27 factories for the main analysis, but – where we have the data – show robustness to including the dropout factories in the appendix.

The initial training for the attitude and aptitude training and the attitude-only groups finished in May 2017. Attitude and aptitude training began for the control group, and aptitude training for the attitude-only group, in August 2017; all training was completed in October 2017. We conducted a follow-up survey in each factory just before the control group training started in each factory. The follow-up survey included samples of subordinate operators working on lines on which trainees were assigned to work as assistant supervisors. We use these data to conduct intent-to-treat regressions on the effectiveness of trainees as supervisors.

4 Results

4.1 Nominee rankings and baseline skills

We first evaluate the effect of the selection experiment on the rankings of the nominees and the initial skills of the selected trainees. These, and all other, outcome variables are defined in Table 0.B.3 in the appendix. As noted, in factories that were assigned to the selection experiment, we offered managers the chance to re-rank their nominees.[13] Managers in 11 of the 13 factories randomised into the treatment revised their lists after viewing the diagnostic scores of their nominees. We begin by testing whether the factories were implicitly taking into account the information contained in the diagnostic scores in their initial ranking of nominees. We do so by regressing the nomination rank of the candidates in all 27 factories before any potential re-ranking on the indices of nominees' initial attitude and aptitude. The indices are created as explained in Section 3. For clarity of exposition, we reversed the order of the rankings for the analysis such that a higher rank indicates a better position. Recall that the factories provided the nomination rankings before the nominated women even took the diagnostic tests, so none of the factories had seen the scores at this point. The regression takes the following form for nominee i in factory f,

$$y_{if} = \alpha + \beta' \cdot \boldsymbol{S}_{if} + \theta_f + \epsilon_{if}, \tag{1}$$

where y is the nomination ranking as defined above[14] and $\boldsymbol{S}$ is a vector of the attitude and aptitude indices. We use within-factory variation by including factory fixed effects θ and cluster standard errors at the factory level.[15] Note that we are not arguing for a causal interpretation here, but test whether

training on the effectiveness of female line supervisors and the best way to support female line supervisors in succeeding in their new role. The onboarding exercise also included a session in which higher-level management introduced the trainee to the workers on the production line. The onboarding was implemented in the same manner for all three treatment groups.

[13] See Appendix Tables 0.B.5 and 0.B.6 for results of the re-ranking exercise for the sample including factories that dropped out. Those results are very similar to what we report here for the sample of the factories that continued through the full program.

[14] Because the number of nominees varies across factories, both the initial cardinal ranking and movements in the ranking may be disproportionately affected by a few factories with a large number of nominees. Regressions using percentile rankings, which reduces the influence of factories with more nominees (though arguably leads to excess influence of factories with a very small number of nominees), produce similar results.

[15] Standard errors from wild cluster percentile-t bootstraps are shown in squared brackets in this paper, to account for a potentially small number of clusters (following Cameron et al. (2008)).

Table 1: Nominee rankings and initial skills

	(1)	(2)
	Nomination rank	Movements in rank (after Selection experiment)
Aptitude (Index)	0.25	1.10**
	(0.41)	(0.05)
	[0.43]	[0.06]
Attitude (Index)	0.16	1.85***
	(0.66)	(0.00)
	[0.66]	[0.00]
Outcome mean	4.84	0.00
P-value (Soft=Hard skills)	0.87	0.30
Factory FE	Yes	Yes
Observations	243	138
Number factories	27	13
Sample	All factories	Selection experiment only

Notes: P-values clustered at the factory level in parentheses, p-values from wild cluster percentile-t bootstraps with 10000 repetitions in Column (1) and 8192 repetitions in Column (2) in squared brackets. * p<0.1, ** p<0.05, *** p<0.01

the initial skills predict the nomination rank.

The result of this regression is shown in the first column of Table 1. Both the aptitude and the attitude indices are positively related to a better nomination rank, with a one standard deviation higher index associated with a rank that is about a quarter and a sixth of a rank higher for aptitude and attitude, respectively. However, the coefficients are not significant and not significantly different from each other. In Table 0.B.4 in the appendix, we show the same regression but include all seven diagnostic scores (as fractions out of 1, where 1 represents a perfect score) separately as explanatory variables in the first column. We find positive but insignificant coefficients on all scores except numeracy, for which we find a negative and significant coefficient. Since this is likely due to the high positive correlations among the diagnostic scores, and especially among the two subgroups making up the attitude index and the aptitude index, we focus on the results for the indices.

This indicates that, at baseline, there was little correlation between the nominees' initial skills and the factories' preference for training them. We can think of several explanations for this. First, factories may have lacked good measures of these skills. Alternatively, they may not have thought these skills were important. Finally, this may reflect the highly selected nature of the pool of nominees to begin with, which could have made distinguishing between the candidates difficult. A typical factory selected only one woman for every two or three production lines, that is, around one in 40 female operators.

We now turn to ask whether the selection experiment changed how factories ranked their nominees for the promotion programme. This helps us to understand which initial skills presented in the results of the diagnostic tests, if any, the factories respond to. For this analysis, we only include the factories that were assigned to the selection experiment since only these factories were allowed to see the nominees' seven diagnostic scores and could then revise the ranking of their nominees. We capture how much nominees moved in the ranking by taking the difference between a nominee's ranking after the selection experiment and that nominee's initial ranking prior to the experiment. A positive difference indicates that a nominee moved to a better rank, whereas a negative difference indicates that a nominee moved to a worse rank. We again estimate equation 1, but with the nominees' movements in the rankings as

defined above as the outcome variable. Thus, a positive coefficient implies that nominees with high initial attitude or aptitude were more likely to be moved up in the ranking. Note that the net movement as a result of re-ranking is zero – every move up is associated with an equivalent move down. Hence, a positive coefficient implies that managers placed a higher weight on the characteristic represented by that diagnostic. For this analysis, we include the 138 nominees in 13 factories that completed the selection experiment.

Column (2) in Table 1 shows the results. We find that nominees with both higher initial aptitude and attitude indices were moved to better positions in the rankings. On average, a one standard deviation higher index of aptitude or attitude is related to a movement up 1.10 ranks for aptitude and 1.85 ranks for attitude. We cannot reject that the coefficients for the aptitude and attitude indices are the same, however.[16] Note that the participating factories in the selection experiment only saw the seven diagnostic scores, not the attitude and aptitude indices aggregated from these scores. In Column (2) of Table 0.B.4 in the appendix, we include all seven diagnostic scores in the regression instead of the indices. Despite the correlation among the scores, we see that the numeracy, family support and confidence test scores are significantly associated with movements up the rankings.

The previous analysis suggests that factories re-ranked the nominees for the promotion programme based on their initial aptitude and attitude, once they were given the information. In the next step, we ask whether the selection experiment therefore leaves the participating factories with better trainees as measured by their initial skills. We address this in Table 2, where we report the results of a treatment effects regression of the form,

$$S_{if} = \alpha + \beta \cdot selection\_experiment_f + \gamma' \cdot \boldsymbol{X}_f + \epsilon_{if}, \tag{2}$$

where S is one of the indices of initial attitude and aptitude, $selection\_experiment$ indicates that factory f participated in the selection experiment, and $\boldsymbol{X}$ is a vector of variables used in the randomisation of the selection experiment. We again cluster standard errors at the factory level. The coefficient on the treatment variable tells us the difference between the average diagnostic scores of the selected trainees – those nominated at a sufficiently high enough level to be included in the training, after any re-ranking – in factories randomised to receive the selection experiment, compared with trainees in factories randomised out of the selection experiment.

Panel A of Table 2 reports the results for the diagnostics measuring aptitude – literacy, numeracy, fluid intelligence and technical knowledge related to garment production. All are defined as fractions out of 1, where 1 represents a perfect score. We see that factories participating in the selection experiment have trainees with marginally significantly higher fluid intelligence (as measured by processing speed) and technical knowledge, though the effect size is small and the coefficients are insignificant once we implement bootstrapping to account for a potential small number of clusters (p-values shown in square brackets). However, the effect on the aggregate aptitude index is larger at close to 0.3 SD and remains significant after bootstrapping. This indicates that the selected trainees in the factories that participated in the selection experiment have higher initial technical skills than the trainees in the non-participating factories.

[16] The regression does not control for the initial ranking of the nominee. When we add the initial ranking, we find that it is insignificant and does not change the magnitude or significance of the coefficients on either the indices or diagnostic measures.

Table 2: Trainees' initial skills

Panel A: Aptitude

	(1) Literacy	(2) Numeracy	(3) Processing speed	(4) Garment knowledge	(5) Aptitude (Index)
Selection experiment	0.06	0.04	0.01*	0.02*	0.29**
	(0.11)	(0.13)	(0.08)	(0.07)	(0.02)
	[0.16]	[0.22]	[0.10]	[0.16]	[0.05]
Control mean	0.55	0.45	0.32	0.53	0.10
Randomisation controls	Yes	Yes	Yes	Yes	Yes
Observations	199	199	199	199	199
Number factories	27	27	27	27	27

Panel B: Attitude

	(1) Family support	(2) Interest	(3) Confidence	(4) Attitude (Index)
Selection experiment	0.11***	0.15***	0.09***	0.62***
	(0.00)	(0.01)	(0.00)	(0.00)
	[0.01]	[0.04]	[0.00]	[0.00]
Control mean	0.68	0.68	0.68	−0.10
Randomisation controls	Yes	Yes	Yes	Yes
Observations	199	199	199	199
Number factories	27	27	27	27

Notes: P-values clustered at the factory level in parentheses, p-values from wild cluster percentile-t bootstraps with 10000 repetitions in squared brackets. The test of the cross-equation restriction that the coefficient for the Attitude and the Aptitude index is the same has a p-value of 0.02. * p<0.1, ** p<0.05, *** p<0.01

Panel B of Table 2 shows the differences in average scores for the three diagnostics measuring attitudes as well as for the attitude index combining the three measures. The effects on the attitude diagnostics are larger than on any of the aptitude measures, and highly significant.[17] This is also true for the aggregate attitude index of the selected trainees, which is 0.6 SD higher in the factories who participated in the selection experiment than in the non-participating group. When testing the cross-equation restriction whether the coefficient is the same for the attitude and aptitude indices, we can reject equality with a p-value of 0.02.

Table 2 thus indicates that the factories completing the selection experiments selected trainees with both higher initial attitude and aptitude indices, though the effect is twice as large for attitude. The baseline balance shown in Appendix Table 0.B.1 indicates that the family support score is marginally imbalanced at baseline (p = 0.10), so some portion of the results on Table 2 may be driven by this imbalance. However, the combined results from Tables 1 and 2 do suggest that managers moved toward nominees with higher diagnostic scores following the selection experiment.

In sum, we find that factories value both the initial attitudes and aptitude of candidates for promotion, though they value attitude more. Factories who are given detailed information about the candidates' diagnostic scores choose trainees who have stronger initial attitude and aptitude.

4.2 Endline skills

The previous analysis showed that factories value the initial skills – and especially attitude measures – of their candidates for a promotion. We now investigate whether the initial skills that factories are selecting on are predictive of our outcomes of interest, and whether training in these skills can produce successful managers. We analyse this for the trainees' endline skills, promotions, as well as their performance according to evaluations of their subordinates and according to production data.

To measure these outcomes, we use a combination of survey and administrative data. Survey data collected just prior to the start of the delayed training sessions allows us to assess the short-term effect on endline skills and on official promotions. Regular telephone follow-up surveys during the initial training period allow us to measure the percentage of days the trainees worked as assistant supervisors rather than operators. Finally, surveys of operators working under the direction of the trainees and administrative data provide measures of trainee performance.

We start by investigating the trainees' endline skills, i.e. the skills that were enumerated at the end of the trainees' eight-week trial period as line supervisors. Recall that only the nominees that were ranked the highest by the factories – after a potential re-ranking in the factories that participated in the selection experiment and the removal of ineligible nominees based on the literacy and numeracy tests in all factories – were selected to receive training that aimed to prepare them for a promotion to supervisor. The selected operators were randomised into one of three groups, one receiving both the attitude and aptitude training immediately, one receiving only the attitude training immediately, and one receiving neither training immediately. All three groups were assigned to trial as assistant line supervisors for an eight-week trial period.

We use the randomisation to the different training regimes and the indices of the trainees' initial attitude and aptitude to investigate their impact on the skills trainees gained after the trial period.

[17] Fluid intelligence and technical knowledge have lower variances than the other five diagnostic scores, with standard deviations around 0.10 rather than 0.18-0.23 for the other five scores.

We use a regression of the following form:

$$\begin{aligned} y_{ift=post} = \alpha &+ \beta \cdot attitudeonly\_training_{if} + \gamma \cdot aptitude\&attitude\_training_{if} \\ &+ \delta \cdot attitude_{ift=pre} + \mu \cdot aptitude_{ift=pre} \\ &+ \rho \cdot y_{ift=pre} + \omega' \boldsymbol{I}_{ift=pre} + \theta_f + \pi_e + \epsilon_{ift=post}, \end{aligned} \tag{3}$$

where y is a skill measure for nominee i in factory f at time t, which is post the trial period for $t = post$ and at baseline for $t = pre$. The variables $attitudeonly\_training$ and $aptitude\&attitude\_training$ indicate the random assignment to the different training groups, and $attitude$ as well as $aptitude$ are the indices measuring the initial conditions as defined above. We include controls for the one baseline imbalance we find (see Table 0.B.2 in the appendix), factory fixed effects θ, enumerator fixed effects π and cluster standard errors ϵ at the factory level.[18]

We are interested in the coefficients β and γ, which capture the effects of the trainings, as well as δ and μ, which show how the trainees' initial skills – that the factories selected on in the selection experiment – relate to the endline skills. Note that the randomisation of the training means that the variables capturing the treatment effects in expectation are orthogonal to the initial skills indices. The randomisation checks in Table 0.B.2 in the appendix confirm that this also holds in practice. The Analysis of Covariance (ANCOVA) specification in equation 3 also ensures that δ and μ capture the explanatory power of the initial attitude and aptitude indices once we have accounted for the outcomes' stability over time, i.e. for their gain in skills.[19]

As an alternative specification, we also show results of treatment effect regressions that include the assignment to the selection experiment as a dummy variable in equation 3, instead of the initial skill indices. We focus on the main specification in equation 3 since (1) this approach has greater statistical power compared to only relying on the across-factory variation of the assignment to the selection experiment, and (2) because this allows us to investigate the explanatory power of the initial attitude and aptitude indices separately.

Table 3 shows the results for five skills measures at endline. The first measure is the score in percentages on a test of knowledge about garment processes and production, measured after the trial period. This is the garments knowledge diagnostic that is described in Appendix A, repeated after the trial period. The second column is a self-assessment of the trainee's expected performance as a supervisor. We asked trainees to rate both the typical supervisor and their own expected performance on a scale of one to ten. The self-assessment measure takes the difference between the own rating and the rating of the typical supervisor as reported by the trainee. A third measure is the self-efficacy score, which aggregates responses from the Generalized Self-Efficacy scale (Schwarzer and Jerusalem, 1995).[20] It captures an individual's belief in their own abilities to deal with new situations and to cope with any associated setbacks. The fourth is a measure of the trainees' internal locus of control, which we measure by asking

[18] Note that we do not have the statistical power to investigate the interactions between the training and initial skill levels.

[19] Table 0.B.7 in the appendix presents results from a non-ANCOVA specification, i.e. equation 3 estimated without including $y_{ift=pre}$. As can be expected, the coefficients for the initial aptitude and attitude indices are larger than in the ANCOVA specification.

[20] Scholz et al. (2002) show that the questions produce reliable responses in 25 countries. Recent work by Laajaj et al. (2019) suggests that psychological measures may be measured with more noise in lower-income countries, and that there may also be concerns with response bias. The responses to the Generalized Self Efficacy scale have reasonable internal consistency, with Cronbach's alpha of 0.75 in our data.

Table 3: Effect on trainee skills

	(1) Garment knowledge	(2) Self-assessment	(3) Self-efficacy	(4) Internal LOC	(5) Stress
Attitude only training	−0.52	1.02**	0.12	0.35	0.88
	(0.77)	(0.03)	(0.10)	(0.20)	(0.47)
	[0.76]	[0.04]	[0.06]	[0.16]	[0.45]
Aptitude & Attitude training	1.50	0.52*	0.07	0.25	1.09
	(0.39)	(0.09)	(0.34)	(0.32)	(0.30)
	[0.34]	[0.07]	[0.31]	[0.29]	[0.28]
Aptitude (Index)	1.90	−0.05	−0.04	0.19**	−0.44
	(0.11)	(0.83)	(0.18)	(0.04)	(0.41)
	[0.11]	[0.82]	[0.28]	[0.04]	[0.36]
Attitude (Index)	−0.74	0.56***	0.14***	0.26	−0.17
	(0.32)	(0.01)	(0.00)	(0.10)	(0.77)
	[0.25]	[0.02]	[0.01]	[0.06]	[0.76]
Control mean	52.57	−1.68	3.27	4.18	11.30
P-value(Attitude=Aptitude&Attitude training)	0.19	0.22	0.45	0.75	0.86
Imbalance controls	Yes	Yes	Yes	Yes	Yes
Factory FE	Yes	Yes	Yes	Yes	Yes
Enumerator FE	Yes	Yes	Yes	Yes	Yes
Observations	188	188	188	188	154
Number factories	27	27	27	27	25

Notes: ANCOVA specification. P-values clustered at the factory level in parentheses, p-values from wild cluster percentile-t bootstraps with 10000 repetitions in squared brackets. * p<0.1, ** p<0.05, *** p<0.01

respondents to choose one of two statements in each of seven different pairs of statements (based on Rotter (1966)). In each pair, the two statements either represent the view that an individual's life is controlled by their own actions (i.e. an internal locus of control) or that it is externally determined (i.e. an external locus of control). The final measure is the score on the Generalized Anxiety Disorder 7-item scale (Spitzer et al., 2006). This measures stress and anxiety over the preceding two weeks, where a higher score indicates higher stress.

Among these five measures, we find little effect on the trainees' garment knowledge at endline in column (1) of Table 3. While the coefficients of both the aptitude & attitude training and the aptitude index are positive, and the coefficients of the attitude only training and the attitude index are negative, they are not significant. In comparison, we do find some effects on the soft skills measures in columns (2) to (4). The supervisory trainings have a positive and significant effect on the trainee's self-assessment of their expected performance as a supervisor. Their effects on self-efficacy and the internal locus of control are also positive but insignificant. The initial skills indices, and especially the attitude measure, are significant predictors of the trainees' endline attitudes – over and above the baseline value of the outcome variable, which is also included in the regressions. For example, a one standard deviation increase in the initial attitude index is related to an increase in the trainees' self-assessment of about half a point, compared to a similar effect of half a point for the aptitude & attitude training and a whole point for the attitude only training. We do not find any significant relation of either the trainings or the initial skills measures on stress.

These results seem to suggest that both the initial skills that factories select on and the supervisory training matter for the new managers' gain in skills. Interestingly, there does not seem to be an additional value of the aptitude training, as we can in no case reject equality of the effects of attitude training only and of the training that focused on both attitude and aptitude.[21] That both training and initial skills matter is also confirmed if we estimate the alternative specification, by including a dummy

[21] Of course, as in any training intervention, this could also mean that our skills training was not fit for purpose.

for the selection experiment instead of the initial skills indices in equation 3.[22] We show the results in Table 0.B.8 in the appendix. The variable for the selection experiment has a positive coefficient for all five outcomes, but the effect is only significant for self-efficacy. The coefficient's magnitude is in a similar ballpark as the supervisory training variables for the trainees' self-assessment, self-efficacy and internal locus of control.

4.3 Promotions

The next outcome of interest is whether the trainees complete the training programme and are promoted to supervisor. At one extreme, some operators dropped out of the program shortly after being selected as a participant; at the other extreme, some were officially promoted to a supervisory position by the time of the follow-up survey. Just over four-fifths (81 %) of the trainees reported working as an assistant supervisor at least once in weekly phone surveys. Table 4 reports the effects of the supervisory trainings and the coefficients of the initial skill indices on two outcome measures. The first is the percentage of days during the eight-week trial period that each trainee worked as an assistant supervisor. This data was collected in high-frequency phone surveys during the trial period.[23] The second outcome is whether trainees report having been officially promoted to a supervisory position. This was measured at two points in time, the follow-up survey at the end of the trial, and the household surveys described in more detail in Uckat (2020). The household survey was conducted between January and March of 2018, six to ten months beyond the follow-up survey, and after all trainees had received the full training in attitude and aptitude.

We estimate equation 3 for these outcomes in Table 4, but do not include the baseline value in the regressions. In columns (1) and (2), we find that the percentage of workdays the trainee worked in a supervisory role is significantly higher among trainees receiving at least one of the formal supervisory training sessions when they began their trial as an assistant supervisor. This is particularly relevant given that the normal procedure in most factories is to start the trial period without any formal training. The second column of Table 4 adds the initial attitude and aptitude indices. We find no evidence that trainees with higher technical scores worked more often as an assistant supervisor, but the attitude index is very strongly associated with working as a supervisor. Nevertheless, the magnitude of the training effect is larger, since the coefficients on the training variables are roughly equivalent to increasing the attitude index by two standard deviations.

The results in columns (3) to (6) in Table 4 indicate that the supervisory training increased the likelihood that the participating women were offered a promotion and were promoted. However, the difference to the control group is significant only for those in the aptitude & attitude training group when we measure promotions at the follow-up survey in column (3). Recall that at this point, the control group had not yet received any training. By the time the household survey was conducted, when we measured promotions for columns (5) and (6), trainees in all groups had completed training. The data show that around two-thirds of the trainees in each of the three randomisation groups were offered a promotion at that point. However, only those in the group that received early attitude only

[22] Since the selection experiment was randomised on the factory level, we cannot include factory fixed effects. Instead, we include the variables used in the randomisation as controls.

[23] Management often returned workers to machines during peak production periods, or when there was particular production pressure on a given day. Still, compliance in this regard was reasonably high: among those working at least one day as an assistant supervisor, 75 % worked more than half of the available days, and one-third worked every day in a supervisory capacity.

Table 4: Effect on promotions

	Share of trial days as SV		Promoted (at end of trial)		Promoted (after control training)	
	(1)	(2)	(3)	(4)	(5)	(6)
Attitude only training	0.22***	0.23***	0.03	0.03	0.17*	0.17*
	(0.00)	(0.00)	(0.48)	(0.42)	(0.06)	(0.05)
	[0.00]	[0.00]	[0.45]	[0.39]	[0.04]	[0.03]
Aptitude & Attitude training	0.18**	0.19**	0.10*	0.10*	0.06	0.07
	(0.03)	(0.02)	(0.08)	(0.07)	(0.47)	(0.41)
	[0.03]	[0.01]	[0.06]	[0.06]	[0.42]	[0.36]
Aptitude (Index)		−0.02		0.00		−0.02
		(0.57)		(0.83)		(0.74)
		[0.56]		[0.82]		[0.77]
Attitude (Index)		0.10***		0.02		0.10*
		(0.00)		(0.29)		(0.06)
		[0.01]		[0.27]		[0.09]
Control mean	0.51	0.51	0.00	0.00	0.19	0.19
P-value(Attitude=Aptitude&Attitude training)	0.44	0.41	0.15	0.15	0.25	0.28
P-value(Attitude=Aptitude)		0.00		0.61		0.20
Imbalance controls	Yes	Yes	Yes	Yes	Yes	Yes
Factory FE	Yes	Yes	Yes	Yes	Yes	Yes
Enumerator FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	175	175	188	188	178	178
Number factories	27	27	27	27	27	27

Notes: P-values clustered at the factory level in parentheses, p-values from wild cluster percentile-t bootstraps with 10000 repetitions in squared brackets. * p<0.1, ** p<0.05, *** p<0.01

training were more likely to have accepted the promotion and hence be working as a supervisor at the time of the household surveys, compared to the delayed group.[24]

The initial skill indices do not appear to be predictive of promotions at the end of the trial in column (4), though the attitude measure again is predictive for promotions measured in the longer term in column (6). This is confirmed in Table 0.B.9 in the appendix, when we regress the promotion outcomes of interest on the supervisory training variables and the selection experiment. The selection experiment in column (5) leads to a 14 percentage point increase in promotions in the longer term, though this effect is not significant after bootstrapping. This magnitude is nearly identical to the significant effect of the early attitude only training of 15 percentage points. Table 0.B.9 in the appendix also provides further evidence that the effect of the selection experiment indeed operates through factories selecting trainees with a higher initial attitude index, since the magnitude of the coefficient of the selection experiment reduces by two thirds once we add the initial attitude and aptitude indices to the regression in column (6).

Similar to the results on endline skills, overall we find that both the supervisory training and the initial skills are related to higher promotion rates, though the results are sometimes not significant. Again, the initial attitude measure seems to be more predictive than the aptitude measure, and we do not discern a difference between training focusing on attitudes only and teaching additional hard skills.

[24]Though we cannot reject that the coefficients for the attitude only and the aptitude & attitude training are equal.

4.4 Performance as supervisor

Do the formal training or the initial skills affect the performance of trainees as supervisors?[25] We use two sources of data: production data – further discussed below – and surveys with 708 subordinates of the trainees, i.e. operators working on lines where the trainees were assigned to work during the trial period. These surveys were conducted at follow-up, when the early training was completed but the late training not yet started. To obtain a performance evaluation for each trainee as viewed by their subordinates, we ask the operators to rate, on a scale of one to ten, a typical supervisor in the factory, and then to rate the trainee on the same scale. We then subtract the typical supervisor rating from the rating for the trainee to obtain the subordinate rating for each trainee. This construction is analogous to the self-assessment measure in Table 3. In addition to this performance evaluation, we also investigate how the subordinates' wellbeing is affected. We use a wellbeing index, which combines the GAD-7 diagnostic discussed above (recoded such that higher numbers mean less stress), questions about verbally and physically abusive behaviour on the line, aspirations to be a supervisor and a question on general happiness over the previous two weeks. These questions are detailed in Table 0.B.3 in the appendix.

The regression specification we estimate is similar to equation 3, except that this time the unit of observation is subordinate s on production line l for the two outcomes y just discussed:

$$\begin{aligned} y_{slft=post} = \alpha &+ \beta \cdot attitudeonly\_training_{lf} + \gamma \cdot aptitude\&attitude\_training_{lf} \\ &+ \delta \cdot attitude_{lft=pre} + \mu \cdot aptitude_{lft=pre} \\ &+ \omega' \boldsymbol{I}_{lft=pre} + \theta_f + \pi_e + \epsilon_{slft=post}. \end{aligned} \tag{4}$$

The other variables are defined as above. Note that β and γ capture the effect of the trainee's assignment to a training group on the outcomes of the subordinates working on the production line to which the trainee was assigned to work as assistant supervisor during the trial period. Similarly, δ and μ capture how the trainee's initial attitude and aptitude measures relate to the outcomes from the subordinate surveys.

We report on these intent-to-treat regressions in Table 5.[26] In the first two columns of Table 5, we show results for the ratings the subordinates gave the new supervisors. On average, trainees are rated about a half point (one-third of a standard deviation) lower than typical supervisors. We see that both the variables capturing the training effects as well as the initial skill indices are positively related to the ratings, but they are far from statistical significance. When we investigate the subordinate wellbeing index in columns (3) and (4), we find positive coefficients of the supervisory training, which is significant only for the aptitude & attitude training group. However, we cannot reject the equality of coefficients for the training groups. When we add the two initial skill indices in column (4), we find

[25] We note at the start that the design of the project is better suited to answer the question of continued interest and promotion than performance, given that the experimentally generated variation lasts only a short time after the normal trial period ends. Only one fifth of the trainees had been offered a promotion to full supervisor at the time of the follow-up, though an additional 40 % were still working as assistant supervisors.

[26] There are two types of non-compliance relevant for interpreting the results. First, as we have noted, almost one fifth of the trainees never work as assistant supervisors. For the survey questions, we therefore begin by asking operators if they recall the trainee working as a supervisor. Those responding "not at all" or reporting that the trainee never worked in a supervisory capacity on their line, we ask for a generic comparison between a typical supervisor and a "typical female supervisor." Just over one third of operators (38 %) answer the generic question. Second, some of the trainees work on a line other than the line where they were assigned. Of course, these movements may be endogenous, so we present the intent-to-treat regressions.

Table 5: Effect on subordinate outcomes

	Subordinate ratings		Subordinate wellbeing	
	(1)	(2)	(3)	(4)
Attitude only training	0.18	0.22	0.08	0.08
	(0.42)	(0.35)	(0.34)	(0.38)
	[0.44]	[0.37]	[0.34]	[0.39]
Aptitude & Attitude training	0.07	0.06	0.17*	0.17*
	(0.73)	(0.77)	(0.07)	(0.06)
	[0.73]	[0.77]	[0.06]	[0.06]
Aptitude (Index)		0.14		−0.04
		(0.14)		(0.34)
		[0.26]		[0.36]
Attitude (Index)		0.07		0.09**
		(0.40)		(0.03)
		[0.43]		[0.02]
Control mean	−0.50	−0.50	−0.01	−0.01
P-value(Attitude=Aptitude&Attitude training)	0.46	0.34	0.51	0.45
P-value (Attitude=Aptitude)		0.61		0.04
Imbalance controls	Yes	Yes	Yes	Yes
Factory FE	Yes	Yes	Yes	Yes
Enumerator FE	Yes	Yes	Yes	Yes
Observations	707	707	708	708
Number factories	26	26	26	26

Notes: P-values clustered at the factory level in parentheses, p-values from wild cluster percentile-t bootstraps with 10000 repetitions in squared brackets. * p<0.1, ** p<0.05, *** p<0.01

that the initial attitude score is significantly associated with higher subordinate well-being, and find an insignificant negative coefficient of the aptitude index. We can also reject equality between the effects of the initial attitude and aptitude indices.

These patterns are confirmed in Table 0.B.10 in the appendix, where we include the selection experiment indicator instead of the initial skills indices. The coefficients on this variable are positive and insignificant for the subordinate ratings in column (1) and the wellbeing index in column (3). Similar to the results on promotions, we find that the treatment effect of the selection experiment becomes smaller once we control for the initial skills of the trainees in the even columns of Table 0.B.10. There, we see that the initial aptitude and attitude indices are marginally significant predictors of the subordinate ratings, and that only the initial attitude index is a highly significant predictor for subordinate wellbeing. These results from the subordinate surveys suggest that both the initial skills and training matter, to an extent.

In addition to the survey data, we use administrative records from the factories measuring daily productive efficiency, quality defect rates (in percent) and absenteeism (in percent) at the production-line level. These measures are also described in 0.B.3. We report how the trainings and the initial skills indices relate to these performance measures in Table 6.[27] We estimate an intent-to-treat ANCOVA

[27]Note that, in addition to the two sources of non-compliance discussed above with regard to the operator opinions, there is a third measurement issue that is relevant in interpreting these results. Our productivity measures are made at the production line level. More complex products are produced on lines that typically have two or even three line supervisors, each responsible for only a part of the line. The line-level measures will therefore reflect the combined effort of more than one supervisor. Moreover, even on lines with only a single supervisor, our trainees were almost always working as an assistant supervisor, and hence responsible for only a part of the line.

regression for production line l for the outcomes y in the eight-week trial period as follows:

$$\begin{aligned} y_{lfdt=post} = \alpha &+ \beta \cdot attitudeonly\_training_{lf} + \gamma \cdot aptitude\&attitude\_training_{lf} \\ &+ \delta \cdot attitude_{lft=pre} + \mu \cdot aptitude_{lft=pre} \\ &+ \rho \cdot \bar{y}_{lft=pre} + \omega' \boldsymbol{I}_{lft=pre} + \theta_f + \tau_d + \epsilon_{lfdt=post}. \end{aligned} \tag{5}$$

By including factory-fixed effects θ and date fixed effects τ, we only rely on variation within each factory and day. By including the mean of the outcome for each line in the pre-trial period, $\bar{y}$, we ensure that a potential assignment of the trainees to more or less productive lines is not driving results.[28]

As Table 6 shows, the data give little indication that the training or the initial skills predict the production outcomes. The coefficients for both the supervisory trainings and the initial skill indices are always insignificant and small compared to the control mean, though the initial attitude index does have signs in the direction of improvement for all three measures.[29] The lack of an effect may not be surprising given that the trainees on most lines represented one of two or three supervisors on each line managing 20-80 line operators. Also, only a small percentage of the trainees are working as full supervisors at the time of the follow-up, and even those would have accumulated little experience at the time.[30]

[28]Note that we only keep lines that were working on specific days, and also drop the few lines where several trainees were assigned to trial with different trial start dates or different training assignments.

[29]When we include the selection experiment variable instead of the initial skill indices in the regression in Table 0.B.11 in the appendix, we find that it is associated with significantly worse outcomes for quality defect rates. However, since we do not find these associations for the initial skill indices in Table 6, we posit that this is primarily due to not being able to control for factory fixed effects in Table 0.B.11 in the appendix.

[30]When we compare the production outcomes of lines to which trainees were assigned with outcomes of lines with no assigned trainee during the factory's trial period (again controlling for factory fixed effects, date fixed effects, and baseline means of outcomes), we find that lines with trainees have significantly lower absenteeism, insignificantly higher efficiency and insignificantly lower alteration rates. These results seem to suggest that the trainees – compared to typical lines in the factory – positively contribute to their lines' productivity, and support the conclusions we drew from the subordinate survey outcomes. They are also in line with evidence from a similar context, the garment sector in India, for which Adhvaryu et al. (2019) show that managerial skills are important determinants of line productivity.

Table 6: Effect on production outcomes

	Efficiency		Alteration rates (%)		Absenteeism (%)	
	(1)	(2)	(3)	(4)	(5)	(6)
Attitude only training	0.20	0.23	0.15	0.16	0.30	0.30
	(0.89)	(0.88)	(0.64)	(0.63)	(0.17)	(0.19)
	[0.89]	[0.88]	[0.65]	[0.64]	[0.14]	[0.17]
Aptitude & Attitude training	0.37	0.37	−0.29	−0.29	0.29	0.27
	(0.73)	(0.73)	(0.37)	(0.36)	(0.28)	(0.32)
	[0.73]	[0.73]	[0.38]	[0.39]	[0.27]	[0.31]
Aptitude (Index)		0.15		0.11		0.15
		(0.82)		(0.59)		(0.12)
		[0.82]		[0.61]		[0.33]
Attitude (Index)		0.07		−0.16		−0.03
		(0.90)		(0.28)		(0.88)
		[0.90]		[0.23]		[0.93]
Control mean	52.21	52.21	6.85	6.85	4.65	4.65
P-value(Attitude=Aptitude&Attitude training)	0.89	0.91	0.25	0.23	0.96	0.89
P-value (Attitude=Aptitude)		0.93		0.38		0.50
Imbalance controls	Yes	Yes	Yes	Yes	Yes	Yes
Factory FE	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes
Pre-treatment line mean	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5769	5769	5911	5911	5680	5680
Number factories	17	17	23	23	17	17

Notes: Intent-to-treat ANCOVA comparison within the 8 weeks trial period. P-values clustered at the factory level in parentheses, p-values from wild cluster percentile-t bootstraps with 10000 repetitions in squared brackets. * p<0.1, ** p<0.05, *** p<0.01

5 Conclusion

After four decades of rapid growth, the garment sector in Bangladesh now represents around one eighth of GDP and more than 80 % of exports. Importantly, the sector remains the primary source of employment for women working full time outside agriculture. The 2017 Bangladesh Labor Force Survey indicates that among women without tertiary education and working full time, 40 % of those employed outside agriculture are employed in the garment sector. However, women's role in the sector is limited almost exclusively to production-worker positions; fewer than one in ten factory managers are women. Facing new pressures from several sources, factories are increasingly interested in promoting women to supervisory positions. The shift toward increased interest in promoting women represents an important cultural shift in the factories.

The results in this paper should be viewed in the context of this recent interest in promoting women. We document four results. First, factories do not take into account the initial aptitude and attitudes of candidates when they nominate women for a promotion training. Second, the factories that are provided information about their candidates' skills and are able to change their selection as part of an experiment react to this information. They select candidates who have higher baseline skills, especially on the attitude and soft skills dimension. Third, we find that the initial skills – and especially attitudes and soft skills – matter for the outcomes of the promotion programme. They are related to higher endline skills, to the promotion rates to official line supervisor, to higher wellbeing of their subordinates and – to some degree – to higher performance evaluations as judged by subordinates. Fourth, training in these attitudes and aptitude supervisory skills leads to improvements in some of these same outcomes, though the results are not always significant. In none of our results do we find that training in technical

skills has an additional effect when compared to a training only targeting attitudes and soft skills.

Returning to our initial question, we therefore find that both factories learning how to choose supervisors and women learning how to lead matters. For the new female low-level managers that are the focus of this paper, attitudes and soft skills appear to be particularly important. The initial attitudes and soft skills are predictive of the trainees' success, and training in these attitudes and soft skills also enables trainees to be more successful. The results of the selection experiment suggest that factories had some idea that attitudes and soft skills would matter for their trainees' success, but the little experience in choosing female supervisors likely meant that they lacked the ability to convincingly measure these skills. This is also important from the perspective of cost-effectiveness, since we find effects of similar magnitude from a short, cheap intervention with higher-level management that aims to base manager selection on formal skills tests as compared to a longer-term, expensive training programme for new female line supervisors. In the case of female career advancement, this paper therefore not only indicates that training in non-cognitive skills is a promising avenue to increase the participation of women in managerial positions, but also provides evidence for the importance of hard measures of soft skills and attitudes.

References

Adhvaryu, A., Nyshadham, A., and Tamayo, J. (2019). Managerial Quality and Productivity Dynamics. *Working Paper*.

Anderson, M. L. (2008). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, 103(484):1481–1495.

Ashraf, N., Bandiera, O., Davenport, E., and Lee, S. S. (2020). Losing Prosociality in the Quest for Talent? Sorting, Selection, and Productivity in the Delivery of Public Services. *American Economic Review*, 110(5):1355–1394.

Autor, D. and Scarborough, D. (2008). Does Job Testing Harm Minority Workers? Evidence from Retail Establishments. *The Quarterly Journal of Economics*, 123(1):219–277.

Bandiera, O., Prat, A., Hansen, S., and Sadun, R. (2020). CEO Behavior and Firm Performance. *Journal of Political Economy*, 128(4):1325–1369.

Beaman, L., Keleher, N., and Magruder, J. (2018). Do Job Networks Disadvantage Women? Evidence from a Recruitment Experiment in Malawi. *Journal of Labor Economics*, 36(1):121–157.

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

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

Bloom, N., Lemos, R., Sadun, R., Scur, D., and Van Reenen, J. (2014). JEEA-FBBVA Lecture 2013: The New Empirical Economics of Management. *Journal of the European Economic Association*, 12(4):835–876.

Bruhn, M., Karlan, D., and Schoar, A. (2018). The Impact of Consulting Services on Small and Medium Enterprises: Evidence from a Randomized Rrial in Mexico. *Journal of Political Economy*, 126(2):635–687.

Cameron, C. A., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics*, 90(3):414–427.

Campos, F., Frese, M., Goldstein, M., Iacovone, L., Johnson, H. C., McKenzie, D., and Mensmann, M. (2017). Teaching Personal Initiative Beats Traditional Training in Boosting Small Business in West Africa. *Science*, 357(6357):1287–1290.

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

Fafchamps, M. and Woodruff, C. (2017). Identifying Gazelles: Expert Panels vs. Surveys as a Means to Identify Firms with Rapid Growth Potential. *World Bank Economic Review*, 31(3):670–686.

Glaub, M. E., Frese, M., Fischer, S., and Hoppe, M. (2014). Increasing Personal Initiative in Small Business Managers or Owners Leads to Entrepreneurial Success: A Theory-based Controlled Randomized Field Intervention for Evidence-based Management. *Academy of Management Learning and Education*, 13(3):354–379.

Hanna, R., Mullainathan, S., and Schwartzstein, J. (2014). Learning through Noticing: Theory and Evidence from a Field Experiment. *The Quarterly Journal of Economics*, 129(3):1311–1353.

Heath, R. and Jayachandran, S. (2018). The Causes and Consequences of Increased Female Education and Labor Force Participation in Developing Countries. In Averett, S. L., Argys, L. M., and Hoffman, S. D., editors, *The Oxford Handbook of Women and the Economy*. Oxford University Press, New York, NY.

Heath, R. and Mobarak, A. M. (2015). Manufacturing Growth and the Lives of Bangladeshi Women. *Journal of Development Economics*, 115:1–15.

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

Hoffman, M., Kahn, L. B., and Li, D. (2018). Discretion in Hiring. *The Quarterly Journal of Economics*, 133(2):765–800.

Hussam, R., Rigol, N., and Roth, B. (2020). Targeting High Ability Entrepreneurs Using Community Information: Mechanism Design In The Field. *Harvard Business School Working Paper 20-082*.

Laajaj, R., Macours, K., Pinzon Hernandez, D. A., Arias, O., Gosling, S. D., Potter, J., Rubio-codina, M., and Vakis, R. (2019). Challenges to Capture the Big Five Personality Traits in Non-WEIRD Populations. *Science Advances*, 5(7):eaaw5226.

Lazear, E. P., Shaw, K. L., and Stanton, C. T. (2015). The Value of Bosses. *Journal of Labor Economics*, 33(4):823–861.

Macchiavello, R., Menzel, A., Rabbani, A., and Woodruff, C. (2020). Challenges of Change: An Experiment Promoting Women to Managerial Roles in the Bangladeshi Garment Sector. *Working Paper*.

McKenzie, D. and Sansone, D. (2017). Man vs. Machine in Predicting Successful Entrepreneurs: Evidence from a Business Plan Competition in Nigeria. *World Bank Policy Research Working Paper 8271*.

McKenzie, D. and Woodruff, C. (2014). What are We Learning from Business Training and Entrepreneurship Evaluations around the Developing World? *World Bank Research Observer*, 29(1):48–82.

McKenzie, D. and Woodruff, C. (2017). Business Practices in Small Firms in Developing Countries. *Management Science*, 63(9):2967–2981.

Menzel, A. and Woodruff, C. (2019). Gender Wage Gaps and Worker Mobility: Evidence from the Garment Sector in Bangladesh. *NBER Working Paper No. 25982*.

Qian, N. (2008). Missing Women and the Price of Tea in China: The Effect of Sex-specific Earnings on Sex Imbalance. *The Quarterly Journal of Economics*, 123(3):1251–1285.

Quinn, S. and Woodruff, C. (2019). Experiments and Entrepreneurship in Developing Countries. *Annual Review of Economics*, 11(1):225–248.

Rotter, J. B. (1966). Generalized Expectancies for Internal versus External Control of Reinforcement. *Psychological Monographs: General and Applied*, 80(1):1–27.

Scholz, U., Gutiérrez Doña, B., Sud, S., and Schwarzer, R. (2002). Is General Self-Efficacy a Universal Construct? Psychometric Findings from 25 Countries. *European Journal of Psychological Assessment*, 18(3):242–251.

Schwarzer, R. and Jerusalem, M. (1995). Generalized Self-efficacy Scale. In Weinman, J., Wright, S., and Johnston, M., editors, *Measures in Health Psychology: A User's Portfolio. Causal and Control Belief*, pages 35–37. NFER-NELSON, Windsor, England.

Spitzer, R. L., Kroenke, K., Williams, J. B., and Löwe, B. (2006). A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine*, 166(10):1092–1097.

Uckat, H. (2020). Leaning in at Home: Women's Promotions and Intra-household Bargaining in Bangladesh. *Working Paper*.

Appendices

Appendix A Description of diagnostic tests

After the factory nominated workers, the IPA team visited the factory to assess the abilities of the nominees. Nominees were only disqualified based on numeracy and literacy scores, see below. The areas that were tested and their respective tests are detailed below. For each of the seven diagnostics, the score is calculated as a fraction out of 1, where 1 is a perfect score.

1. **Literacy**: Multiple choice questions testing reading comprehension, vocabulary, basic grammar, paragraph/letter structure, and writing. The maximum is 20 points.
2. **Numeracy**: Multiple choice questions testing calculations, fractions, percentages, number patterns, angles, and visual patterns. The maximum is 20 points.
3. **Processing speed**: This score consisted of two tests of fluid intelligence, detailed below.
   (a) Coding: This test is modelled after the Wechsler Adult Intelligence Scale is called 'Digit Symbol' (WAIS-R), 'Digit-Symbol-Coding' (WAIS-III), 'Coding' (WAIS-IV), also known as the Digit symbol substitution test. It is a neuropsychological test sensitive to brain damage, dementia, age and depression. It consists of digit-symbol pairs followed by a list of digits. Under each digit the subject should write down the corresponding symbol as fast as possible. The number of correct symbols within the allowed time is measured. The maximum is 133 points.
   (b) Symbol search: Symbol Search is a subtest of the Wechsler Adult Intelligence Scale (WAIS). The Symbol Search subtest is designed to assess information processing speed and visual perception. High scores require rapid and accurate processing of nonverbal visual information. During Symbol Search, the examinee is asked to mark either the yes or no checkbox with a pencil in response to as many items as possible within 2 min. The maximum is 60 points.
4. **Garments Knowledge**: Contains multiple choice, and open-ended questions testing what machine to use for what operation, names of processes, cause of mechanical issues, cause of quality issues, identifying quality issues in photographs, identifying working condition issues in photographs, and understanding an operation breakdown. The maximum is 84 points.
5. **Family Support**: Gives 5 statements about family support, and asks the respondent to respond on a four-point scale from agree to disagree. An additional three questions ask about the level of support given to other women in the family who work in garment factories. The answers to all questions are recoded so that higher numbers represent more support, and are summed to give a maximum of 24 points.
6. **Interest**: The survey instrument includes 2 questions about whether the nominee would want to be promoted to supervisor or line chief. In addition, four questions indirectly probe whether nominees are interested in the supervisor position, and asks the respondent to respond on a four-point scale from agree to disagree. The answers to all questions are recoded so that higher numbers represent more interest, and are summed to give a maximum of 18 points.

Table 0.A.1: Overview of diagnostic test scores

Testing area	Maximum points
Literacy	20
Numeracy	20
Processing speed	193
Garment knowledge	84
Family support	24
Interest	18
Confidence	7

7. **Confidence**: Consists of asking how they would rate their performance compared to a typical supervisor on a 5-point scale. In addition, three questions indirectly ask whether the respondent is confident. The respondent is asked to choose between two statements. One statement that says "I am confident" using various words, and a dummy statement about the factory. The answers are recoded so that higher numbers represent more confidence and are summed to give a maximum of 7 points.

Disqualification rule: Nominees were only disqualified based on the scores in the numeracy and literacy tests. The rule is:

- If the nominee scores below 25 % on both literacy and numeracy tests, she fails.
- If the nominee scores 0 % in either literacy or numeracy tests, she fails irrespective of her score in the other test.

Appendix B Additional tables

Table 0.B.1: Factory balance, Selection experiment

	Control		Selection		Mean equality	Mean equality
	Obs	(Mean)	Obs	(Mean)	p-value	p-value (RI)
Participated in prev project	14	0.36	13	0.46	0.60	0.27
Produces knit garments	14	0.50	13	0.46	0.85	0.84
Number of workers	14	2332.21	13	2860.77	0.46	0.34
Share of female workers	14	0.57	13	0.59	0.69	0.63
Date of joining programme	14	20346.43	13	20386.15	0.72	0.62
Number of supervisors	14	48.14	13	52.69	0.78	0.77
Share of female supervisors	14	0.04	13	0.06	0.44	0.43
Number of lines	14	24.43	13	30.46	0.49	0.44
(A)PM's age (avg)	14	37.94	13	37.14	0.59	0.59
(A)PM's education years (avg)	14	10.40	13	10.45	0.92	0.90
(A)PM's spouse works (avg)	14	0.08	13	0.06	0.81	0.81
(A)PM exposure (avg)	14	0.98	13	0.97	0.81	0.76
(A)PM nr. male SV are better (avg)	14	4.68	13	5.03	0.55	0.50
Actual to calculated trainees	14	3.94	11	4.12	0.89	0.90
SV/Line Chief IAT (avg)	14	−0.25	13	−0.29	0.58	0.59
Literacy score	14	0.49	13	0.54	0.22	0.22
Numeracy score	14	0.42	13	0.43	0.80	0.80
Processing speed score	14	0.31	13	0.33	0.19	0.21
Garments knowledge score	14	0.54	13	0.55	0.36	0.38
Family support score	14	0.69	13	0.74	0.15	0.10*
Interest score	14	0.70	13	0.74	0.39	0.33
Confidence score	14	0.69	13	0.74	0.21	0.20

Notes: p-value from a regression with robust standard errors in penultimate column, and from randomisation inference with 10,000 permutations in last column. Scores are the average by factory for all operators nominated by that factory. Missing values in Actual to calculated trainees are due to calculated number of trainees being 0 for 2 factories.
* p<0.1, ** p<0.05, *** p<0.01

Table 0.B.2: Balance, Promotion training

	Control (N=65) (Mean)	**Attitude only (N=67)** (Mean)	**Aptitude & Attitude (N=67)** (Mean)	**Joint test p-value**	**p-value** (bootstrapped)
Age	25.26	25.88	25.79	0.28	0.30
Married	0.77	0.76	0.72	0.79	0.81
Household members	3.18	3.51	3.12	0.22	0.24
Household head	0.22	0.25	0.22	0.89	0.90
Migrant	0.63	0.69	0.60	0.55	0.55
Education years	8.38	8.36	8.42	0.98	0.98
Experience in garment sector	5.60	5.84	6.36	0.12	0.15
Tenure	3.14	3.40	3.84	0.12	0.13
Nr. of factories worked in	1.28	1.34	1.31	0.96	0.96
Exposure to female SV	0.68	0.46	0.51	0.04**	0.06*
Nr. male SV are better	3.34	3.37	3.57	0.78	0.79
Internal locus of control	4.43	4.57	4.45	0.82	0.83
Grit	2.90	2.98	2.92	0.66	0.67
Self-efficacy	3.58	3.66	3.60	0.46	0.50
Emotional competence	3.05	3.15	3.11	0.17	0.19
Multi-factor Leadership	3.67	3.72	3.78	0.19	0.21
Life satisfaction	7.72	7.33	7.52	0.45	0.46
Numeracy	1.12	0.87	1.19	0.12	0.15
Self-assessment	0.12	0.25	−0.15	0.31	0.35
Ambition	2.35	2.18	2.12	0.30	0.31
Numeracy score	0.49	0.47	0.46	0.57	0.59
Literacy score	0.59	0.55	0.60	0.44	0.48
Processing speed score	0.34	0.32	0.33	0.37	0.40
Garment knowledge score	0.55	0.53	0.56	0.21	0.23
Family support score	0.74	0.72	0.74	0.91	0.91
Interest score	0.74	0.77	0.73	0.47	0.50
Confidence score	0.75	0.71	0.71	0.24	0.26

Notes: p-value of the joint test Attitude only = Aptitude & Attitude = 0 from a regression with standard errors clustered on factory level in penultimate column, and p-value of the same test from wild percentile-t cluster bootstrap with 10,000 repetition in the last column. * p<0.1, ** p<0.05, *** p<0.01

Table 0.B.3: Outcome variables

Variable	Definition
Nomination rank	Rank of nominated woman in first ranking provided by factories, after removal of ineligible candidates but before a potential re-ranking. A higher number indicates a better rank.
Movements in rank	Number of ranks that a nominee moved between the first ranking and the final ranking after a potential re-ranking. Calculated as (nominee's final rank after a potential re-ranking - nominee's nomination rank). A positive number indicates a movement to a better rank. Only defined for nominees in factories that participated in the selection experiment.
Literacy score	Score out of 1 on the diagnostic tests detailed in Appendix A, calculated as (points achieved divided by maximum points possible).
Numeracy score	
Processing speed score	
Garment knowledge score	
Family support score	
Interest score	
Confidence score	
Aptitude (Index)	Standardised index of the Literacy, Numeracy, Processing speed, and Garment knowledge scores, created following Anderson (2008).
Attitude (Index)	Standardised index of the Family support, Interest and Confidence scores, created following Anderson (2008).
Self-assessment	The variable is created from two survey questions. Respondents were first asked to rate the overall supervisor ability of a typical supervisor in their factory on a scale from 1 to 10, where a higher rating is better. Second, they were then asked how they think they perform or would perform as supervisor on the same scale. Self-assessment is calculated as (own rating - typical supervisor rating). A higher rating indicates a better assessment compared to a typical supervisor.
Self-efficacy	Self-efficacy score of 10 items following Schwarzer and Jerusalem (1995). After re-coding reverse-coding items, the score is calculated as the mean of all 10 items. A higher scores indicates higher self-efficacy.
Internal LOC	Internal locus of control score of 7 items that are a subset of Rotter (1966). The scores is calculated as the number of items on which the respondent chooses the internal option.
Stress	Stress is measured by the Generalized Anxiety Disorder 7-item scale (Spitzer et al., 2006), which respondents completed on paper with emojis representing response options from 1 to 4. The stress score is calculated as the total of all 7 responses. A higher number represents higher reported stress.
Share of trial days as supervisor	Variable is derived from weekly phone surveys during the trial period. For each weekly phone survey, we calculate the share of work days that the respondent worked as supervisor or assistant supervisor (taking into account holidays and training days). For each respondent, the share of trial days as supervisor is calculated as the average across the completed phone surveys of the share of work days that the trainee worked as supervisor or assistant supervisor.
Promoted	Dummy variable indicating whether respondent accepted a promotion to line supervisor since the on-boarding session (if measured at end of trial) or since December 2016 (if measured after control training).

(Table 0.B.3 continued.)

Subordinate wellbeing	Standardised index created following Anderson (2008), capturing respondent wellbeing and composed of • Stress: Created as defined above, but the score was reversed such that a higher number captures less stress. • Aspiration: Dummy variable indicating whether respondents would someday accept an offer for a promotion to line supervisor. • No verbal abuse: Dummy variable indicating that, in a phone survey, respondents replied "Not at all" to the question "Please think about the other operators on your line. Over the last two weeks, how often have they needed to put up with shouting or abusive language at work?" • No physical abuse: Dummy variable indicating that, in a phone survey, respondents replied "Not at all" to the question "Some people experience situations at work that make them feel uncomfortable. I am going to read some examples: People staring persistently or winking, coming very close to them or calling them to get close, flirting, singing, making sounds or whistling, making gestures, bumping into or rubbing against them, grabbing their hand or other parts of their bodies, tickling, etc. Please think about the other operators on your line and answer this question. Over the last two weeks, how often did at least one of these things happen to them at work?" • Very happy: Dummy variable indicating that, in a phone survey, respondents reported that they were very happy over the last two weeks.
Subordinate ratings	Defined in analogy to the self-assessment variable. The variable is created from two survey questions. Respondents were first asked to rate the overall supervisor ability of a typical supervisor in their factory on a scale from 1 to 10. Second, they were then asked how they think the trainee assigned to trial on their production line performs or would perform as supervisor on the same scale. Subordinate ratings is calculated as (trainee rating - typical supervisor rating). If the respondents did not remember the trainee well or very well, they were asked to rate a female line supervisor in their factory, and this rating then replaced the trainee rating.
Efficiency	Measures how the daily output of each line in garment units, adjusted for the standard minute value of the garment, compares to the total output possible that day with the workers present on the line and the hours worked. Calculated as $$efficiency = \frac{(Pieces\,produced \cdot SMV)}{Workers\,on\,the\,line \cdot Hours\,operated \cdot 60} \cdot 100. \quad (6)$$
Alteration rate	Percentage of all garments that needs to be altered out of all produced units
Absenteeism rate	Percentage of workers absent amongst all workers (present and absent) on the line

Table 0.B.4: Nominee rankings and initial diagnostic scores

	(1)	(2)
	Nomination rank	Movements in rank (after Selection experiment)
Literacy	1.08	1.94
	(0.45)	(0.50)
	[0.45]	[0.49]
Numeracy	−3.16 **	5.17**
	(0.03)	(0.03)
	[0.04]	[0.07]
Processing speed	2.43	−5.26
	(0.53)	(0.22)
	[0.54]	[0.27]
Garment knowledge	4.83	4.76
	(0.12)	(0.33)
	[0.13]	[0.32]
Family support	1.19	3.56*
	(0.52)	(0.08)
	[0.50]	[0.19]
Interest	0.09	0.64
	(0.97)	(0.68)
	[0.97]	[0.67]
Confidence	0.05	5.70**
	(0.96)	(0.01)
	[0.96]	[0.02]
Outcome mean	4.84	0.00
Factory FE	Yes	Yes
Observations	243	138
Number factories	27	13
Sample	All factories	Selection experiment only

Notes: P-values clustered at the factory level in parentheses, p-values from wild cluster percentile-t bootstraps with 10000 repetitions in Column (1) and 8192 repetitions in Column (2) in squared brackets. * p<0.1, ** p<0.05, *** p<0.01

Table 0.B.5: Nominee rankings, and initial skills and scores – including dropout factories

	Nomination rank		Movements in rank (after Selection experiment)	
	(1)	(2)	(3)	(4)
Aptitude (Index)	0.25		1.03*	
	(0.37)		(0.05)	
	[0.40]		[0.06]	
Attitude (Index)	0.16		1.82***	
	(0.62)		(0.00)	
	[0.63]		[0.00]	
Literacy		0.92		1.72
		(0.50)		(0.54)
		[0.50]		[0.53]
Numeracy		−2.91 **		5.21**
		(0.03)		(0.03)
		[0.04]		[0.06]
Processing speed		2.62		−5.39
		(0.48)		(0.21)
		[0.48]		[0.25]
Garment knowledge		4.54		4.60
		(0.13)		(0.34)
		[0.15]		[0.32]
Family support		1.00		3.57*
		(0.58)		(0.07)
		[0.56]		[0.16]
Interest		0.36		0.45
		(0.87)		(0.75)
		[0.86]		[0.75]
Confidence		0.02		5.67**
		(0.99)		(0.01)
		[0.99]		[0.02]
Outcome mean	4.76	4.76	0.01	0.01
Factory FE	Yes	Yes	Yes	Yes
Observations	257	257	143	143
Number factories	28	28	13	13
Sample	All factories	All factories	Selection experiment only	Selection experiment only
Includes dropout factories	Yes	Yes	Yes	Yes

Notes: P-values clustered at the factory level in parentheses, p-values from wild cluster percentile-t bootstraps with 10000 repetitions in Columns (1)-(2) and 8192 repetitions in Columns (3)-(4) in squared brackets. * p<0.1, ** p<0.05, *** p<0.01

Table 0.B.6: Trainees' initial skills – including dropout factories

Panel A: Aptitude

	(1) Literacy	(2) Numeracy	(3) Processing speed	(4) Garment knowledge	(5) Aptitude (Index)
Selection experiment	0.06	0.03	0.01	0.02**	0.28**
	(0.11)	(0.19)	(0.19)	(0.04)	(0.02)
	[0.18]	[0.30]	[0.23]	[0.11]	[0.06]
Control mean	0.54	0.45	0.32	0.54	0.13
Randomisation controls	Yes	Yes	Yes	Yes	Yes
Observations	233	233	233	233	233
Number factories	29	29	29	29	29

Panel B: Attitude

	(1) Family support	(2) Interest	(3) Confidence	(4) Attitude (Index)
Selection experiment	0.11***	0.14**	0.08***	0.60***
	(0.00)	(0.01)	(0.00)	(0.00)
	[0.01]	[0.07]	[0.00]	[0.00]
Control mean	0.67	0.67	0.68	−0.11
Randomisation controls	Yes	Yes	Yes	Yes
Observations	233	233	233	233
Number factories	29	29	29	29

Notes: P-values clustered at the factory level in parentheses, p-values from wild cluster percentile-t bootstraps with 10000 repetitions in squared brackets. The test of the cross-equation restriction that the coefficient for the Attitude and the Aptitude index is the same has a p-value of 0.01. * p<0.1, ** p<0.05, *** p<0.01

Table 0.B.7: Effect on trainee skills – no ANCOVA

	(1) Garment knowledge	(2) Self-assessment	(3) Self-efficacy	(4) Internal LOC	(5) Stress
Attitude only training	−0.45	1.03**	0.16*	0.35	1.22
	(0.80)	(0.02)	(0.06)	(0.22)	(0.30)
	[0.79]	[0.03]	[0.04]	[0.17]	[0.26]
Aptitude & Attitude training	1.58	0.50	0.09	0.24	1.27
	(0.35)	(0.10)	(0.22)	(0.34)	(0.21)
	[0.30]	[0.08]	[0.19]	[0.30]	[0.18]
Aptitude (Index)	2.33**	−0.07	−0.02	0.19*	−0.55
	(0.02)	(0.77)	(0.40)	(0.07)	(0.30)
	[0.00]	[0.76]	[0.45]	[0.09]	[0.24]
Attitude (Index)	−0.72	0.60***	0.20***	0.30**	−0.20
	(0.34)	(0.00)	(0.00)	(0.04)	(0.71)
	[0.29]	[0.01]	[0.00]	[0.02]	[0.70]
Control mean	52.57	−1.68	3.27	4.18	11.30
P-value(Attitude=Aptitude&Attitude training)	0.19	0.20	0.35	0.73	0.96
Imbalance controls	Yes	Yes	Yes	Yes	Yes
Factory FE	Yes	Yes	Yes	Yes	Yes
Enumerator FE	Yes	Yes	Yes	Yes	Yes
Observations	188	188	188	188	154
Number factories	27	27	27	27	25

Notes: P-values clustered at the factory level in parentheses, p-values from wild cluster percentile-t bootstraps with 10000 repetitions in squared brackets. * p<0.1, ** p<0.05, *** p<0.01

Table 0.B.8: Effect on trainee skills – with Selection experiment

	(1) Garment knowledge	(2) Self-assessment	(3) Self-efficacy	(4) Internal LOC	(5) Stress
Attitude only training	−1.16	0.88**	0.08	0.33	0.93
	(0.46)	(0.03)	(0.28)	(0.22)	(0.37)
	[0.47]	[0.04]	[0.26]	[0.20]	[0.36]
Aptitude & Attitude training	0.55	0.49*	0.05	0.22	1.07
	(0.76)	(0.09)	(0.49)	(0.37)	(0.27)
	[0.75]	[0.08]	[0.49]	[0.37]	[0.28]
Selection experiment	0.24	0.40	0.14**	0.19	0.25
	(0.91)	(0.38)	(0.03)	(0.34)	(0.68)
	[0.92]	[0.55]	[0.06]	[0.43]	[0.71]
Control mean	52.57	−1.68	3.27	4.18	11.30
P-value(Attitude=Aptitude&Attitude training)	0.29	0.27	0.69	0.68	0.90
Imbalance controls	Yes	Yes	Yes	Yes	Yes
Randomisation controls	Yes	Yes	Yes	Yes	Yes
Factory FE	No	No	No	No	No
Enumerator FE	Yes	Yes	Yes	Yes	Yes
Observations	188	188	188	188	154
Number factories	27	27	27	27	25

Notes: ANCOVA specification. P-values clustered at the factory level in parentheses, p-values from wild cluster percentile-t bootstraps with 10000 repetitions in squared brackets. * p<0.1, ** p<0.05, *** p<0.01

Table 0.B.9: Effect on promotions – with Selection experiment

	Share of trial days as SV		Promoted (at end of trial)		Promoted (after control training)	
	(1)	(2)	(3)	(4)	(5)	(6)
Attitude only training	0.26***	0.28***	0.04	0.05	0.15**	0.16**
	(0.00)	(0.00)	(0.24)	(0.18)	(0.04)	(0.02)
	[0.00]	[0.00]	[0.24]	[0.17]	[0.03]	[0.01]
Aptitude & Attitude training	0.18**	0.19***	0.10*	0.10*	0.04	0.04
	(0.02)	(0.01)	(0.07)	(0.06)	(0.61)	(0.55)
	[0.01]	[0.01]	[0.07]	[0.06]	[0.58]	[0.51]
Selection experiment	0.01	−0.07	−0.05	−0.08 *	0.14	0.04
	(0.91)	(0.37)	(0.26)	(0.08)	(0.22)	(0.68)
	[0.92]	[0.43]	[0.29]	[0.12]	[0.35]	[0.73]
Aptitude (Index)		0.00		0.01		0.03
		(0.95)		(0.52)		(0.67)
		[0.95]		[0.54]		[0.69]
Attitude (Index)		0.14***		0.04*		0.14***
		(0.00)		(0.05)		(0.00)
		[0.00]		[0.05]		[0.03]
Control mean	0.51	0.51	0.00	0.00	0.19	0.19
P-value(Attitude=Aptitude&Attitude training)	0.14	0.08	0.25	0.30	0.15	0.13
P-value(Attitude=Aptitude)		0.00		0.42		0.21
Imbalance controls	Yes	Yes	Yes	Yes	Yes	Yes
Randomisation controls	Yes	Yes	Yes	Yes	Yes	Yes
Factory FE	No	No	No	No	No	No
Enumerator FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	175	175	188	188	178	178
Number factories	27	27	27	27	27	27

Notes: P-values clustered at the factory level in parentheses, p-values from wild cluster percentile-t bootstraps with 10000 repetitions in squared brackets. * p<0.1, ** p<0.05, *** p<0.01

Table 0.B.10: Effect on subordinate outcomes – with Selection experiment

	Subordinate ratings		Subordinate wellbeing	
	(1)	(2)	(3)	(4)
Attitude only training	0.15	0.19	0.07	0.06
	(0.48)	(0.39)	(0.43)	(0.50)
	[0.49]	[0.41]	[0.44]	[0.51]
Aptitude & Attitude training	0.05	0.04	0.15	0.16*
	(0.77)	(0.82)	(0.11)	(0.08)
	[0.78]	[0.83]	[0.10]	[0.08]
Selection experiment	0.08	−0.05	0.01	−0.03
	(0.66)	(0.77)	(0.83)	(0.66)
	[0.76]	[0.82]	[0.84]	[0.67]
Aptitude (Index)		0.15*		−0.06
		(0.10)		(0.10)
		[0.20]		[0.13]
Attitude (Index)		0.12*		0.10***
		(0.07)		(0.01)
		[0.11]		[0.01]
Control mean	−0.50	−0.50	−0.01	−0.01
P-value(Attitude=Aptitude&Attitude training)	0.51	0.35	0.50	0.39
P-value (Attitude=Aptitude)		0.77		0.01
Imbalance controls	Yes	Yes	Yes	Yes
Randomisation controls	Yes	Yes	Yes	Yes
Factory FE	No	No	No	No
Enumerator FE	Yes	Yes	Yes	Yes
Observations	707	707	708	708
Number factories	26	26	26	26

Notes: P-values clustered at the factory level in parentheses, p-values from wild cluster percentile-t bootstraps with 10000 repetitions in squared brackets. * p<0.1, ** p<0.05, *** p<0.01

Table 0.B.11: Effect on production outcomes – with Selection experiment

	Efficiency		Alteration rates (%)		Absenteeism (%)	
	(1)	(2)	(3)	(4)	(5)	(6)
Attitude only training	0.09	0.21	0.19	0.25	0.18	0.17
	(0.95)	(0.88)	(0.64)	(0.55)	(0.50)	(0.54)
	[0.95]	[0.88]	[0.68]	[0.61]	[0.51]	[0.56]
Aptitude & Attitude training	0.11	0.13	−0.35	−0.32	0.23	0.21
	(0.91)	(0.91)	(0.21)	(0.26)	(0.45)	(0.49)
	[0.92]	[0.91]	[0.21]	[0.26]	[0.46]	[0.50]
Selection experiment	−0.88	−1.30	1.13**	1.07**	−0.45	−0.40
	(0.51)	(0.30)	(0.02)	(0.01)	(0.38)	(0.41)
		[0.35]		[0.05]		[0.44]
Aptitude (Index)		0.69		0.25		0.07
		(0.32)		(0.39)		(0.51)
		[0.36]		[0.55]		[0.57]
Attitude (Index)		0.18		−0.04		−0.30
		(0.68)		(0.81)		(0.14)
		[0.68]		[0.82]		[0.14]
Control mean	52.21	52.21	6.85	6.85	4.65	4.65
P-value(Attitude=Aptitude&Attitude training)	0.98	0.95	0.17	0.15	0.73	0.84
P-value (Attitude=Aptitude)		0.60		0.37		0.17
Imbalance controls	Yes	Yes	Yes	Yes	Yes	Yes
Factory FE	No	No	No	No	No	No
Date FE	Yes	Yes	Yes	Yes	Yes	Yes
Randomisation controls	Yes	Yes	Yes	Yes	Yes	Yes
Pre-treatment line mean	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5769	5769	5911	5911	5680	5680
Number factories	17	17	23	23	17	17

Notes: Intent-to-treat ANCOVA comparison within the 8 weeks trial period. P-values clustered at the factory level in parentheses, p-values from wild cluster percentile-t bootstraps with 10000 repetitions in squared brackets. * p<0.1, ** p<0.05, *** p<0.01