Subjective Performance Evaluation in a Multi-tasking Environment: a Firm-level Experiment in China *

Xiaocheng Hu,[†] Thomas Gall,[‡] and Michael Vlassopoulos[§]

January 2019

Abstract

We examine a multitasking problem where one task is to produce private goods while the other is to create public goods which is hard to measure. Such problems can be found in organisations that make use of multitasking leaders. Group leaders take responsibility for organising teams (public goods) and contribute as a member (private goods). Presenting evidence from a natural field experiment, we shed light on the impact of a high-powered remuneration system regarding leaders' organisational behaviours. In particular, we designed a monitoring system which subjectively evaluates leaders' organisational inputs, and we offered each leader a new bonus scheme that is depending on her relative performance in organising teams among other group leaders within the factory. Using individual daily production records, we find an overall 6% increase in workers' productivity, excluding the leaders. In line with our theoretical model, strengthening incentives on organising teams does not necessarily have a negative impact on leaders' production performance. We show that leaders' production performances increase as they invest more time on the job.

Keywords: Multitasking, public good, subjective evaluation, incentive, productivity, leadership

JEL: J24, J33, M52, C93.

---

*We greatly benefited from comments and suggestions received from seminar participants at Southampton, East Anglia Economics PhD Workshop, the 13th Nordic Conference in Behavioral and Experimental Economics and the 2018 Advances with Field Experiments Conference. Financial assistance from $ESRC$ is gratefully acknowledged. All remaining errors are our own.

†Department of Economics, University of Southampton.

‡Department of Economics, School of Social Sciences, University of Southampton, Southampton SO17 1BJ, UK; email: t.gall@soton.ac.uk.

§Department of Economics, University of Southampton; IZA, Bonn; email: m.vlassopoulos@soton.ac.uk.

1 Introduction

In the modern business world, performing more than one task is virtually unavoidable: at the office, during the regular meeting, or in the boardroom. The general problem is how to incentivise multitasking employees on multiple dimensions. When incentives fail to encourage employees to meet the employer's goals and objectives, the outcomes are inefficient. For instance, a fixed payment system may be extremely inefficient as it overcompensates poor performers and under-compensates top performers while a performance-based compensation scheme can encourage workers to work harder on the measured dimension to earn a higher wage but the quality of job and creativity which are also important but hard to measure are usually disregarded. When a group-based pay is used for producing or improving a public good - such as team cooperation and cohesion - free-riding comes down the pipeline.[1] Hence, an important and complex problem arises when outputs have different dimensions and vary in externality and quantifiability.

In organisations, this problem is particularly daunting at the first line management level such as team leaders. A team leader usually works within a group, as a member, carrying out the same roles as others but with additional 'leader' responsibilities - such as providing supervision, motivation, and maintaining the quality and quantity of production and/or service - as opposed to higher level management who often have a separate job role altogether. A 2016 Deloitte study reported that only 26% of 7,000 large companies from 130 countries are functionally organised, 90% of these companies cite leadership as a significant issue, and 70% of them rate it as urgent.

As multitasking models predict, leaders will not invest effort in tasks that are not covered by the pay-for-performance scheme. For example, leaders in a US law firm shifted their billable hours to non-billable hours after the firm reduced the individual-based billable performance pay and introduced a bonus scheme which is determined by objective and subjective evaluation of both billable and non-billable activities (Bartel et al., 2017). However, this study provided limited insights for how efficiency can be achieved within a firm by employing a compensation system that monitors multiple dimensions as the complexity of their remuneration design.

Understanding how multidimensional incentive schemes affect leaders' effort choice in individual performance (private good) and subordinates' performance (public good) is essential for the contract design of team leaders. In particular, our study focuses on two dimensions: perfectly quantifiable performance (production output) and hard-to-measure

[1] Griffith and Neely (2009) find that moving from a group-based incentive - profit-related pay - to a Balanced Scorecard system which includes a variety of measures increased the organisational performance. They argue that the information content on the Balanced Scorecard is the key for the improvement, not the incentive per se.

inputs (organising teams). When performance is hard to measure, firms tend to use subjective performance evaluations to solve the multitasking problem.[2] Information about the effort leaders invest in organising groups can be gathered from the senior management.[3] However, while subjective evaluations may enable the firm to reward leaders' organisational inputs and bring efficiency to the team, they may have some pitfalls. First, subjectively evaluating and rewarding leaders' organisational inputs may be costly; the firm needs to invest enormous resources, such as time, training, and money. Second, the employee being evaluated will have an incentive to invest effort in developing a good rapport with those evaluating them, rather than spending time on the job. Third, they can have a counterproductive impact on employee morale and satisfaction, as employees may question the fairness and accuracy of the evaluation process. We developed a novel system to evaluate leaders' organisational behaviours while circumventing these issues.

This study presents evidence from a natural field experiment designed to measure the effect of an incentive system which subjectively evaluates forewomen (group leaders) regarding their hard to measure effort - organising teams. By taking advantage of an opportunity that a Chinese manufacturing company was avid for increasing production efficiency, we conducted a field experiment in a natural setting with factory workers. This collaboration enables us to explore how forewomen respond to multidimensional incentives in a multitasking environment. We choose work where the forewomen perform the same manufacturing task as the workers (group members), and the quantity of output is carefully recorded.

We constructed an incentive scheme and trained the production management to implement it, which helped the company to minimise the cost of introducing a new evaluation system. The new scheme is composed of two elements: rank incentives and monetary prizes, both are linked to the subjective evaluation results of forewomen's organisational activities. In doing this, we can assess how multitasking forewomen respond to an incentive, which based on partially observable inputs - organising teams. We decompose the total effect on team productivity into that caused by changes in workers' productivity (as the group becomes more organised) and that caused by changes in forewomen's production effort (due to multitasking).

The firm we study has two independent factories located on the northern side of Jiangxi, a south-eastern province of China with a high concentration of manufacturers of

[2]A 2015 Global Partner Compensation System Survey of the processes law firm partnerships use to determine partner compensation found that 52% of US and Canadian law firms use a purely subjective system while 38% use a modified subjective or combination system. See https://www.edge.ai/2018/05/edge-global-partner-compensation-system-survey/ for details.

[3]Evaluation can also be given by those who are evaluated if the interaction is not repeated and collusion can be prevented.

medical products. The field experiment was designed and implemented in collaboration with the production management in both factories, and they allowed us to introduce the treatment in one of the factories and left the other one operationally unchanged throughout (the control group). Ours is not a randomised controlled trial, namely we do not randomly assign teams to treatment and control groups. The choice of between-factory experimental design is determined by the fact that teams from the same factory can easily observe each other and react to the incentives offered to their colleagues so that the comparison of contemporaneously assigned treatment and control groups would yield biased treatment effect estimates.

Importantly, before our intervention, the salary structure of forewomen in both factories is a flat monthly payment for organising teams, an individual piece rate for manufacturing, and other fixed bonus schemes. Our experiment offers every forewoman in the treated factory an opportunity to compete for an extra bonus per month. The senior management ranks the forewomen regarding their organisational behaviours. The higher the ranking, the higher the reward. The monetary prize for the lowest ranked forewoman is the same size as the payment she receives from the company for organising the team, while the highest ranked forewoman receives a prize which is more than twofold of the flat rate paid by the company.[4]

During the treatment period, the ranking is public information and updated on a weekly basis in the treated factory, whereas in the controlled factory both forewomen and workers were not aware of the evaluation process and the results. Publicly educating the treated forewomen regarding the evaluation system and their weekly ranking results alleviate the concern of fairness and accuracy.[5] Moreover, in each factory, more than one manager was assigned to evaluate the forewomen. This can also prevent the manager's misreport due to personal perceptions. Hence forewomen are less likely to question the fairness and accuracy. Employing multiple assessors can increase forewomen's costs for collusion as well. Last but not least, this also allows the more productive forewomen to motivate other forewomen through contagious enthusiasm or through embarrassment over the unfavourable direct performance comparison. Peer pressure may force forewomen to internalise their spillovers. If pressure is sufficiently strong, it could push forewomen toward higher organisational performance, as illustrated in Eugene Kandel (1992).

By comparing the output quantity of workers and forewomen between these two factories, we can estimate the treatment effects of our interventions. To the best of our knowl-

[4]The field evidence on tournament incentives tests whether individual behaviour changes with various schemes can be found in Erev et al. (1993), Bandiera et al. (2013), and Delfgaauw et al. (2013, 2015).

[5]Rankings can also effectively provide information on relative earnings and which is a key determinant of happiness (Layard, 2011; Kahneman et al., 2006).

edge, this paper is the first field experiment designed to evaluate the effect of subjective evaluations along with monetary incentives in a multitasking environment. However, our experimental design does not allow us to separately identify the effect of rank incentives from the effect of monetary prizes. In the literature, empirical studies show that the provision of relative performance feedback can have mixed effects on individual performance. Blanes i Vidal and Nossol (2011), Bradler et al. (2016), Delfgaauw et al. (2013), and Kosfeld and Neckermann (2011) find that it can induce higher performance, while Bandiera et al. (2013), Barankay (2012), and Eriksson et al. (2009) obtain an opposite result.[6]

We develop a simple theoretical framework in which a forewoman's manufacturing effort and organising effort are complements because of the team efficiency spillover. The rationale underlying this assumption is that a forewoman brings efficiency to the workers by organising the team production, which can make the forewoman more productive as efficient workers are less likely to create problems on the production line relative to the inefficient ones. The model makes precise how the intervention affects the forewoman's and the worker's performance.

It is expected that forewomen can shape team productivity by facilitating mutual learning or by influencing the group production norm. Mutual learning suggests that forewomen may be able to encourage more able workers (e.g., those who are more productive under individual piece rates) to teach the less able workers to be more productive, thereby enhancing team productivity. On the other hand, peer pressure may be used to achieve a productive group norm, as modelled in Eugene Kandel (1992). A forewoman maximises but takes into account the effect of her actions on the views of her line workers, which enter the forewoman's utility function. Norms can be established as the equilibrium outcome of a process where deviations from any given (say, mean) level of effort result in direct or indirect sanctions. When a forewoman departs from the team norm and other workers impose disutility on her for the extent of her departures, peer pressure arises, and the equilibrium effort is higher than it would be without peer pressure.

Using more than 5,000 individual-daily observations across 59 workers and 13 forewomen, we first show that a sufficiently large increase in incentive power motivates forewomen to exert more efforts to organise the team, and this increases both the average production output and productivity of workers by 8% and 6% respectively. Second, we show that the average production output of forewomen unambiguously increases while the effect on their average productivity is ambiguous. On the one hand, this can be ex-

[6]Theorists highlight that relative performance feedback has an impact on individual performance if individuals have concerns for their relative status (Lizzeri et al., 2002; Ertac, 2005; Moldovanu et al., 2007; Besley and Ghatak, 2008; Ederer, 2010), and supported by research in psychology (Kluger and Denisi, 1996), and neuroscience (Fliessbach et al., 2007).

plained by the fact that the positive spillover effect may vary across production lines. As precisely predicted in our model, a forewoman's output would fall when this spillover effect is particularly small. On the other hand, average forewomen's productivity can decrease if they frequently organise the team production. This is because switching between different tasks can slow them down, and multitasking can sometimes cause unnecessary stresses.[7] Another reason that we are unable to measure the precise change in forewomen's productivity is we do not observe the exact time a forewoman invested in the manufacturing task.

More subtly, the introduction of subjective evaluations and monetary prizes lead to significant changes in forewomen's working time. Relative to the forewomen worked in the controlled factory the treated forewomen spent roughly an extra 30 minutes a day on the job during the treatment period. In fact, this result is consistent with our expectations as the monthly prize for the highest (lowest) ranked forewomen is equivalent to the product between the highest (lowest) piece-rate wage per hour and the average number of days worked in a month. It was the senior management's aspiration to have the forewomen's engagements in organisation and management for about an hour in a day.

Furthermore, we find that the ranking of forewomen's organisational performance does positively correlate with workers' productivity but not perfect, suggesting that workers do benefit from a better-organised forewoman and the senior management was not simply taking the workers' production outcome into account when evaluating the forewomen. For the forewomen, the ranking of their organisational performance is positively associated with their productivity, suggesting that a better-ranked forewoman is also more productive as the production line is better-organised and efficient which is consistent with our assumption on team efficiency spillover.

Finally, we take the difference between forewoman's weekly organisational ranks to help us understand the underlying mechanism of the treatment effects. We find that forewomen had to increase the rankings by marginally sacrificing their productivity. This is consistent with juggling multiple tasks slows the forewomen down. Nevertheless, due to a small sample size this result is not statistically significant at the conventional levels.

The results from the scant literature that analyses the effect of incentives in a multitasking environment using experimental data are mixed. Shearer (2004), Bandiera et al. (2005), Hossain and List (2012), and Englmaier et al. (2017) do not find that incentives focusing on one dimension (e.g. productivity) affected the performance in the other dimension (quality). On the other hand, Kishore et al. (2013) find that multitasking concerns are modest when workers reached their targets and they are paid bonus-based incentive

[7]Psychologists find that multitasking can lead to distraction and stress (e.g. Drews et al., 2008; Mark et al., 2012).

schemes. Similarly, Al-Ubaydli et al. (2015) and Hong et al. (2018) find that workers under a piece-rate wage produce high-quality work while workers under a flat wage rate do not. In this paper, we do not investigate the quantity-quality trade-off but a group leader's choice between individual performance and group performance. There have been very few empirical studies on how multitasking issues affect the structure of incentive schemes, one notable exception is Slade (1996). Similarly, Manthei et al. (2018) shows that workers' efforts are distorted towards the more profitable tasks when managers have no access to objective measures but assess worker's performance subjectively. Once the managers have access to objective performance measures, both worker's and firm's performance increase significantly. To the best of our knowledge, this paper is the first to provide field evidence on the productivity effect of incentivising group leaders regarding their (hard-to-measure) organisational inputs under a subjective performance evaluation system.

Using field experimental data, Bandiera et al. (2013), Casas-Arce and Martínez-Jerez (2009), and Delfgaauw et al. (2013) have studied tournaments among fruit pickers and retailers. They find a positive effect of tournament incentives on performance, but none of them varies the prize spread. Lim et al. (2009) varies both the number and the distribution of prizes in contests among fundraisers. They find that performance is higher in tournaments with multiple prizes in comparison with single-prize tournaments, but there is no further effect on performance by differentiating prizes by rank.

There is a large literature, both theoretical (see Eugene Kandel, 1992) and empirical (see Armin and Andrea, 2006; Mas and Moretti, 2009) have studied the effects of workers on their peers and team members (see Ichniowski and Shaw, 2003, for teams and complementarities). Peer effects may be important, but the relationship with one's superior is likely to be as important as or more important than that to any other worker. Using data from a service firm, Lazear et al. (2015) find that a higher quality manager increases the output of the supervised team over that supervised by a lower quality manager by about as much as adding one member to the team. Different from the forewomen studied in our experiment these managers are not multitasking.

A similar relationship can be found in other settings, such as education. Kremer et al. (2010) and Muralidharan and Sundararaman (2013) conduct experiments to show that performance pay to teachers increases student performance in the dimensions along which teachers are incentivised, and there are no adverse effects in the unrewarded dimensions. If one assumes that students do not know their production functions, adverse effects may be found for poor-performing students (see Fryer et al., 2012).

The remainder of the paper proceeds as follows. The next section outlines the theoretical framework. Section 3 describes the setting and our experimental design. Sections

4 and 5 discuss our main results and the underlying causes of these results, respectively. Finally, section 6 concludes.

2 Theoretical Framework

We develop a simple model to demonstrate the impact of changing incentive schemes for multitasking foremen regarding their hard-to-measure inputs - organising team production. In the context of our experiment, the firm hires two types of employees: a worker and a foreman. The worker only performs production task while the foreman is responsible for both production and organisation tasks. The firm observes employees' production output and offers them a piece-rate payment scheme. While the piece-rate remains constant, the firm replaces a fixed bonus scheme which compensates foreman's organisational activities with a new bonus depending on the foreman's relative position in the leadership ranking within the firm.

We first derive the results of employees' optimal effort choices when the firm offers the foreman a fixed bonus for organising the team. We then derive the results in the context of the new bonus scheme where the foreman receives a bonus depending on her relative rank in the subjective organisational input distribution as perceived by the firm management. By comparing these results, we are able to illustrate the effect of introducing the new bonus scheme on the worker's effort provision and the foreman's. We interpret effort choices in our model as intensity. It is important to notice that we are not aiming to derive an optimal incentive scheme from the firm's or the social planner's perspective.

A. Basic Model

Production Function and Team Efficiency Spillover.— First and foremost, the production function of worker w who produces individual output $y_w(e_w)$ can be written as follows:

$$y_w(e_w) = e_w(1 + \lambda g_f),$$

where $y_w(e_w)$ is depending on individual production effort e_w and foreman f's organisational input g_f. In this production function, the worker's production effort and the foreman's organisational input are complements, meaning that the worker benefits from the outcome of the foreman's organisational activities only if she expends production efforts. That is to say, the return of production effort is increasing in foreman's organisational input, and the greater this increase, the more important the team efficiency spillover captured by the parameter $\lambda > 0$. The worker benefits from the positive team efficiency when she exerts more effort since we define effort as intensity, meaning that working harder yields a better outcome when the team is efficient.[8]

On the other side, the multitasking foreman f who produces individual output $y_f(e_f, g_f)$ and organises the team according to the following production function:

$$y_f(e_f, g_f) = e_f(1 + \lambda g_f),$$

where this is to say the foreman's production output $y_f(e_f, g_f)$ is depending on individual production effort e_f and organisational input g_f. The complementarity assumption between the foreman's organisational input and production effort implies that the foreman benefits from the outcome of organising the team production only if she exerts more production efforts. The idea is that when the foreman invests more organisational inputs the worker becomes more efficient. The foreman, therefore, earns herself time to concentrate on her own production task, which provides her with a higher wage, as she encounters less disturbance (e.g. informal conversations) from the worker than if the worker is not working very hard. For simplicity, we assume that the team efficiency spillover here is identical to the parameter λ in the worker's production function, and it is complementary to the foreman's production effort. In the general case, the effect of team efficiency may be different across employees. Removing this assumption does not affect our results.

Cost of Efforts.— Because exerting effort is costly to the foreman regardless the work type, the cost of effort function is quadratic in both production and organisational inputs: $C(e_f, g_f) = \frac{(e_f+g_f)^2}{2}$. In this cost function, the positive cross derivative with respect to e_f and g_f implies that increasing effort in one dimension increases the marginal cost in the other. When the foreman increases her effort in organising the team, it leads to some negative externality on her production effort. For the worker, the cost of production effort is $C(e_w) = \frac{(e_w)^2}{2}$.

B. Pre-Intervention: Firm Offers the Foreman a Fixed Bonus for Organising Teams

The Forman's and The Worker's Maximisation Problems Before Interventions.— During the pre-intervention period, production output is compensated by piece rate w and the firm

[8] As in existing studies, this formulation abstracts from the dynamic implications of contemporaneous spillover through support and cooperation between the worker and the foreman on the job. The underlying rationale is that workers with well-organised foremen are more productive because they have much better access to resources than if they were in an unorganised group. In addition, this equation assumes that worker's production effort is required to "unlock" the potential of team efficiency. This assumption of complementarity between team efficiency spillover and production effort provision is one of the drivers of why team efficiency translates into worker's wage in our model: workers in efficient teams exert higher effort, for which they are compensated with higher wages.

offers the foreman a constant bonus b for organising the team to optimise its profit. We assume that both the firm and the employees are risk neutral. This assumption simplifies our analysis without being a necessary condition for our general argument. Because of risk neutrality, the foreman maximises her expected wage minus the combined cost of effort:

$$\max_{\{e_f,g_f\}} U_f(e_f,g_f) = W_f(e_f,g_f) - C(e_f,g_f) = wy_f(e_f,g_f) + b - C(e_f,g_f)$$
$$= w[e_f(1+\lambda g_f)] + b - \frac{(e_f+g_f)^2}{2},$$

This leads to the first order condition with respect to e_f:

$$e_f = w + (\lambda w - 1)g_f,$$

and the first order condition with respect to g_f:

$$g_f = (\lambda w - 1)e_f.$$

For the worker w, she maximises her expected wage minus the cost of effort:

$$\max_{\{e_w\}} U_w(e_w) = W_w(e_w) - C(e_w) = wy_w(e_w) - C(e_w)$$
$$= w[e_w(1+\lambda g_f)] - \frac{(e_w)^2}{2}.$$

the first order condition with respect to e_w is given as follows:

$$e_w = w(1+\lambda g_f).$$

Solving these first order conditions we get the optimal effort levels of the foreman:

$$e_f^* = \frac{w}{1-(\lambda w - 1)^2}, \tag{1}$$

$$g_f^* = \frac{w(\lambda w - 1)}{1-(\lambda w - 1)^2}. \tag{2}$$

For the worker, the optimal production effort level is given by:

$$e_w^* = \frac{\lambda w^2}{1-(\lambda w - 1)^2}. \tag{3}$$

Wage Contracts and The Firm's Maximisation Problems Before Interventions.— Now, we move on to solve the firm's profit maximisation problem. Taking into account both the worker's and the foreman's optimal effort levels, the firm chooses the piece rate w to maximising its expected profit, $E\pi = (p - w)(y_f(e_f, g_f) + y_w(e_w)) - b$. In line with the context of our experiment that the firm's marketing team usually sets the price of products at the beginning of each year, we assume that the market price for per unit of output is given exogenously at $p > 0$.

As detailed in Appendix A.1, we should assume $0 < \lambda w < 2$ and there are two distinct solutions for the firm's profit maximisation problem. When $\lambda w \leq 1$, $\overline{g}_f = 0, \overline{e}_f = \overline{e}_w = w$. If $\lambda w > 1$, g_f^*, e_f^*, and e_w^* are expressed as in equations 1, 2, and 3. We can then derive the firm's first order condition and an expression of the piece rate for each case. By comparing the firm's expected profits across these two cases, we predict that for a given λ there exists a p^* such that the profit maximising firm will choose w^* which yields g_f^*, e_f^*, and e_w^* and a higher profit than choosing $\overline{w}$ for $p > p^*$. On the other hand, the firm prefers $\overline{w}$ which yields $\overline{g}_f = 0, \overline{e}_f = \overline{e}_w = \overline{w}$ for all positive p that $p < p^*$. These results are reported in Lemma 1 (see Appendix A.2 for details).

Lemma 1 *For a given λ, the firm facing a market price where $p > p^*$ will set $w = w^*$ when it maximises its expected profit, and the foreman responds to it by choosing $g_f = g_f^*$. However, if $p < p^*$ the firm favours another piece rate scheme $\overline{w}$ in which the foreman exerts 0 effort in organising team production.*

Lemma 1 implies that when the market price is dramatically low selling products is not profitable for the firm. Thus, the firm would not value the foreman's organisational behaviour and sets the piece rate at $\overline{w}$. The foreman, therefore, has no intention to organise the team production. On the other hand, when the market price is sufficiently high producing products is beneficial to the firm. Hence, the firm attempts to expand its production by offering a piece rate w^* which also encourages the foreman to organise the team due to the complementarity between her production effort and organisational input imposed by the production technology.

C. Post-Intervention: Firm Offers the Foreman a Performance-Related Bonus for Organising Teams

Wage Contracts and The Maximisation Problems After Interventions.— When the management starts to evaluate the foremen's organisational activities subjectively, the foreman receives a bonus depending on her rank in the subjective organisational input distribution as perceived by the manager. If the management believes that the foreman is more engaged in organising the team's production relative to her counterparts, the foreman receives a higher rank which provides her with a higher bonus. For simplicity, we opt for a random variable $B(g_f)$ to capture the incentive scheme which is based on an individual's relative position in the firm (see Lazear and Rosen (1981) for details of modelling rank-order tournament incentives). We impose a standard set of conditions on $B(g_f)$ as below:

- $B(g_f)$ is strictly convex and is continuously differentiable on its domain and
- $B'(g_f)$ is strictly positive where the superscript denotes the derivative with respect to g_f.

Thus, a foreman expects a higher bonus that is paid for her organisational activities as she increases effort in organising team production.

Furthermore, we assume that the market price in our case is high (i.e. w^* is offered to the employees and $1 < \lambda w^* < 2$) because the firm intends to increase its compensation paid for foreman's organisational activities. Since the firm does not readjust the piece rate scheme or re-maximise its expected profit after introducing the new incentive scheme, individual piece rate is given at w^*.[9] Thus, the foreman's maximisation problem is now changed to:

$$\begin{aligned}\max_{\{e_f, g_f\}} EU_f(e_f, g_f) &= E[B(g_f) + w^* y_f(e_f, g_f) - C(e_f, g_f)] \\ &= EB(g_f) + w^*[e_f(1 + \lambda g_f)] - \frac{(e_f + g_f)^2}{2}.\end{aligned}$$

This leads to the first order condition with respect to e_f:

$$e_f = w^* + (\lambda w^* - 1)g_f,$$

and the first order condition with respect to g_f:

$$g_f = \frac{dEB(g_f)}{dg_f} + (\lambda w^* - 1)e_f,$$

For the worker w, her maximisation problem is unchanged which implies that the first order condition with respect to her effort choice is given as below:

$$e_w = w^*(1 + \lambda g_f).$$

Solving these first order conditions we get the optimal effort levels of the foreman:

[9]The management told us that it is extremely difficult to adjust the piece rate in the workplace, especially decrease the rate, as the employees are very defensive about changes in their performance-related compensations.

$$\widehat{e}_f = \frac{(\lambda w^* - 1)}{1 - (\lambda w^* - 1)^2} * \frac{dEB(\widehat{g}_f)}{dg_f} + \frac{w^*}{1 - (\lambda w^* - 1)^2}, \tag{4}$$

$$\widehat{g}_f = \frac{1}{1 - (\lambda w^* - 1)^2} * \frac{dEB(\widehat{g}_f)}{dg_f} + \frac{w^*(\lambda w^* - 1)}{1 - (\lambda w^* - 1)^2}. \tag{5}$$

For the worker, the optimal production effort level is given by:

$$\widehat{e}_w = \frac{\lambda w^*}{(1 - (\lambda w^* - 1)^2} * \frac{dEB(\widehat{g}_f)}{dg_f} + \frac{\lambda (w^*)^2}{1 - (\lambda w^* - 1)^2}. \tag{6}$$

C. Optimal Effort levels: Pre-Intervention vs Post-Intervention

In this subsection, we aim to show whether the firm does successfully increase the optimal effort levels of both the foreman and the worker by subjectively evaluating the foremen's organisational activities and offering the foreman a bonus depending on her relative position within the firm. As the piece rate scheme is constant throughout, taking the differences of the optimal effort levels between the pre-intervention and the post-intervention gives us:

$$\begin{aligned} \widehat{e}_f - e_f^* &= \frac{(\lambda w^* - 1)}{1 - (\lambda w^* - 1)^2} * \frac{dEB(\widehat{g}_f)}{dg_f} + \frac{w^*}{1 - (\lambda w^* - 1)^2} - \frac{w^*}{1 - (\lambda w^* - 1)^2} \\ &= \frac{(\lambda w^* - 1)}{1 - (\lambda w^* - 1)^2} * \frac{dEB(\widehat{g}_f)}{dg_f}, \end{aligned} \tag{7}$$

$$\begin{aligned} \widehat{g}_f - g_f^* &= \frac{1}{1 - (\lambda w^* - 1)^2} * \frac{dEB(\widehat{g}_f)}{dg_f} + \frac{w^*(\lambda w^* - 1)}{1 - (\lambda w^* - 1)^2} - \frac{w^*(\lambda w^* - 1)}{1 - (\lambda w^* - 1)^2} \\ &= \frac{1}{1 - (\lambda w^* - 1)^2} * \frac{dEB(\widehat{g}_f)}{dg_f}, \end{aligned} \tag{8}$$

$$\begin{aligned} \widehat{e}_w - e_w^* &= \frac{\lambda w^*}{(1 - (\lambda w^* - 1)^2} * \frac{dEB(\widehat{g}_f)}{dg_f} + \frac{\lambda (w^*)^2}{1 - (\lambda w^* - 1)^2} - \frac{\lambda (w^*)^2}{1 - (\lambda w^* - 1)^2} \\ &= \frac{\lambda w^*}{(1 - (\lambda w^* - 1)^2} * \frac{dEB(\widehat{g}_f)}{dg_f}. \end{aligned} \tag{9}$$

The assumptions $1 < \lambda w^* < 2$ and $\frac{dEB(\widehat{g}_f)}{dg_f} > 0$ imply that optimal effort levels are higher when the foreman is offered a performance-related incentive concerning her organisational activities. As this higher-powered incentive increases the foreman's organisational inputs, production efforts of the foreman and the worker also increase because of the team efficiency spillover λ.

It is important to note that the first order condition for the foreman's maximisation problem with respect to her organisational inputs g_f has an extra positive term $\frac{dEB(g_f)}{dg_f}$ during the post-intervention period relative to the pre-intervention period. This implies that the foreman's organisational inputs would increase from zero to above zero after the intervention even in the case of $\lambda\overline{w} < 1$ provided that $\frac{dEB(g_f)}{dg_f} + (\lambda\overline{w} - 1)\overline{w} > 0$. However, the foreman's production effort would then fall below $\overline{w}$ in this regard as it is equal to $\overline{w} + (\lambda\overline{w} - 1)g_f$ and the second term is negative. In the case of $\lambda\overline{w} = 1$, the foreman's organisational input is guaranteed to increase while her production effort remains unchanged.

Recall that the production function is increasing in g_f, e_f, and e_w. If the foreman and the worker increase their effort levels, the production output grows after the intervention. In the case that the foreman's g_f increases while e_f decreases, the foreman's output would fall if $\lambda\overline{w} + (\lambda\overline{w} - 1)(1 + \lambda g_f) < 0$ where $\lambda\overline{w} < 1$.[10]

Furthermore, taking the difference between equation 7 and equation 9, we get:

$$\begin{aligned}\widehat{e}_w - e_w^* - (\widehat{e}_f - e_f^*) &= \frac{\lambda w^*}{(1-(\lambda w^*-1)^2} * \frac{dEB(\widehat{g}_f)}{dg_f} - \frac{(\lambda w^*-1)}{1-(\lambda w^*-1)^2} * \frac{dEB(\widehat{g}_f)}{dg_f} \\ &= \frac{1}{(1-(\lambda w^*-1)^2} * \frac{dEB(\widehat{g}_f)}{dg_f} > 0,\end{aligned}$$

this indicates that the worker increases more production effort than the foreman. This result is rather trivial since it is directly imposed by the production technology. The first order conditions for both the foreman's and the worker's maximisation problems with respect to their production effort indicate that an increase in the foreman's organisational inputs has a larger impact on the worker's production effort (multiplied by λw^*) than the foreman's (with multiplier equals to $\lambda w^* - 1$).

Proposition 1 (Predictions) *To summarise the theoretical model would predict the following:*

(i) The introduction of a performance-related bonus scheme regarding foreman's organisational activities (weakly) increases the foreman's g_f, holding w constant, and strictly so for some types of bonuses.

(ii) When the market price and λ are sufficiently high, introducing a higher-powered incentive scheme concerning the foreman's organisational activities increases the foreman's

[10]When $\lambda\overline{w} < 1$, the foreman produces $\overline{w}$ in the pre-intervention period as $\overline{g}_f = 0$ and $\overline{e}_f = \overline{w}$. After the introduction of the new bonus scheme, g_f becomes positive if $\frac{dEB(g_f)}{dg_f} + (\lambda\overline{w} - 1)\overline{w} > 0$ and the foreman produces $\overline{w} + [\lambda\overline{w} + (\lambda\overline{w} - 1)(1 + \lambda g_f)]g_f$ given that $\overline{w}$ remains constant. Therefore, there would be a decrease in the foreman's output if $\lambda\overline{w} + (\lambda\overline{w} - 1)(1 + \lambda g_f) < 0$.

e_f even though she is multitasking. However, if the market price or λ is extremely low and the slope of the new bonus scheme is large enough, the foreman's choice of e_f would decrease.

(iii) e_w increases when g_f increases.

(iv) For a given w, the increase in g_f as a result of the new bonus scheme has a larger impact on e_w than on e_f.

(v) Production output increases after the introduction of the new incentive scheme if effort level increases. The foreman's output may fall if she increases g_f while reducing e_f.

3 The Firm and Experimental Design

3.1 Production Setting

We conducted an experiment in two sister medical-device companies between June 2017 and September 2017. Both companies are located on the northern side of Jiangxi, a south-eastern non-coastal province of China. Each sister company has its own personnel and branding, they are not closely related and have limited interactions with each other below the top-level management. One company is located in Fuzhou prefecture while the other is in Fengcheng county, the driving distance between Fuzhou and Fengcheng is about seventy-five miles. For simplicity, we refer the former company as Fuzhou and the latter as Fengcheng below.

In our sample, 70 regular employees (all females, 27 from Fuzhou and 43 from Fengcheng respectively) produce disposable infusion sets in an assembly-line fashion but noncontinuous (see Online Appendix).[11][12] Each production line is composed of numerous workers and one foreman. Seven lines (consist with one foreman and averagely four workers per line) operate regularly in Fuzhou, while six lines (consist with one foreman and averagely six workers per line) operate regularly in Fengcheng. Their backgrounds are mostly local farmers. The manufacturing task for both workers and foremen is a supporting work, product packaging, which requires relatively little training or human capital. Salary schemes for such task are identical to the compensation schemes offered to other

[11] This sample excludes newly hired workers because their compensation schemes are different from those who work more than three months, a few workers who we do not have records during the status quo hence they are not valid for our difference in difference estimation, and some workers who left before the experiment ended as they may respond differently to our treatment.

[12] The disposable infusion set is a major source of revenue for this company, accounting for approximately 50% of its total revenue based on the data in 2016.

tasks within the same production unit such as assembling, leak testing, or pressure testing. Thus, it is unlikely that workers in our experiments have sorted themselves out to a specific kind of base salary structure by their choice of profession.

In addition to the manufacturing job, the foreman in each line is also responsible for monitoring workers' performance, organising and distributing materials, and assisting production managers on production matters. According to the production managers, foremen are internally promoted only, and a successful candidate should be able to demonstrate her loyalty to the company, reliability, and modest leadership. However, qualitative evidence from interviewing the workers and forewomen reveals that no one craves the foreman position because it requires more effort, sidetracks them from the primary task, and the corresponding compensation is relatively low. All existing forewomen have worked in the company for more than two years. They had established a good rapport with the production managers over the years. They accepted the foreman appointment mainly because they ran out of excuses to reject it again. This suggests that we would not expect certain types of worker deliberately stand out to be a foreman. The sorting effect is negligible.

Both factories offer an individual-based multiple piece rate payment scheme (i.e., producing more outputs yields higher rates) to employees for packing the products. In each month, in addition to the piece rate, their base salary comprises some other bonuses and vary across factories. Table 1 summaries the various piece rates and bonuses by factory. This salary structure is consistent throughout our experiment.

Table 1: Summary of Wage Structure by Factory

Daily Average Output	Piece Rate (per unit) (1)	Performance Bonus (2)	Attendance Bonus (3)	Tenure Bonus (4)	Lunch Subsidy (5)	Foreman Subsidy (6)
Panel A. Fengcheng						
Less than 2,400	.0195	200	30	50	42	90
2,400 - 2,600	.0205	200	30	50	42	90
2,600 - 2,800	.0210	200	30	50	42	90
2,800 - 3,000	.0225	200	30	50	42	90
3,000 - 3,200	.0230	250	30	50	42	90
3,200 - 3,400	.0235	250	30	50	42	90
More than 3,400	.0240	300	30	50	42	90
Panel B. Fuzhou						
Less than 3,100	.0188	60	40	65	60	40
3,100 - 3,500	.0193	80	40	65	60	40
More than 3,500	.0196	100	40	65	60	40

A forewoman is given an additional flat rate on a monthly basis for her services regardless of outputs. This rate is higher (more than doubled) in Fengcheng than in Fuzhou. In particular, it equals 40 RMB ($\approx$ 6 dollars) in Fuzhou as shown in Column 6 of Table 1, which is roughly 2% of a forewoman's monthly income. In Fengcheng, a forewoman receives an extra 90 RMB ($\approx$ 13.5 dollars) per month, which is about 3% of her monthly income on average. Similarly, the piece rates offered in Fengcheng are higher than those in Fuzhou. A fast-packaging worker, who can make more than 3,500 units averagely in a day, is given 0.0044 RMB more per unit in Fengcheng in comparison to Fuzhou. The daily average output is calculated by dividing the total production output in a month by the number of days worked during that month. This implies a difference of 430 RMB ($\approx$ 65 dollars) in 28 working days. Therefore, either the worker or the forewoman working at the Fengcheng factory earns 20 percent more income than in Fuzhou.

These differences do not necessarily raise a concern. The two factories are independently operated and organised. The decision, with regard to the rates, made by each factory manager was unassociated. It is mainly determined by the condition of the local labour market. Furthermore, employees work at the production level in one factory barely know any information about the other factory. They do not choose which factory to work.

In a manufacturing setting like ours where employee turnover rate is high, it is unlikely that the management allocates workers into groups randomly. In general, there are two types of relocations of workers. One type of relocations only applies to the newly hired employees. The management usually separates the newly hired employees from the regular workers and assigns them to work on a different line as they call "probation line". Based on the turnover rate of regular workers, the management assigns these newly hired ones to fill the vacancies in the regular operating production lines. Some new workers may stay at the probation line for more than 5 months while others may be relocated to the regular operating lines in 1- or 2-months' time. The other type of relocations is an extreme case and takes place when there are no new employees at stock and the turnover rate is high (e.g. before the Chinese New Year), or when there is a technical breakdown at the downstream of the production. The management may disband one regular operating line and randomly assign these workers or allow them to self-select into other lines.

Nevertheless, the forewoman of each line does not change in general unless the management decides to disband a production line for more than a month and the forewoman of the dismantled line is assigned to work on another line. This didn't occur throughout our experiment. A change in forewomen also occurs if a forewomen decides to quit the job or leave for a long time because of sickness, in this case a new forewoman will be appointed. This type of forewoman change did occur during our experimental period, in our analysis we exclude the observations of workers who had worked with the new forewoman.

In the first relocation scenario which I described above, it should not raise concerns for our identification strategy. First of all, workers who were hired during our experimental period or three months (probation period) before our experiment started were excluded in our analysis. Secondly, according to the managers, new workers are very unlikely to affect group norms (if any group specific norms are present due to line composition) as they are already taking a highly demanding production task. It is unlikely that they have time and energy to change the group norm or establish a rapport with the line foreman in a short time. Lastly, all forewomen in our sample have at least one-year experience of being the group leader. Any relationships/norms that have been established between the regular line workers and the forewomen should be captured by the individual fixed effects.

For the extreme case, no regular operating lines had been disbanded by the management during our experimental period. But, there were a couple of incidences took place at the downstream of the production lines during the experiment, and some workers were assigned to other lines. Nevertheless, these relocations do not decrease the effectiveness of organisation. The relocated workers only changed their working locations. They continued to report daily outputs or personal issues to their initial line forewoman. Therefore, such relocations can potentially have an impact on workers' productivity (because of the location of lines, some production lines may have easier access to raw materials or better illumination and temperature conditions), but there is no impact on the effectiveness of forewomen's organisation. Each forewoman continued to be responsible for her line workers who changed working locations. This may increase the cost of organisation, but managers would have taken this into account when they were evaluating the forewomen. Hence, we include an indicator which captures the incidences of workers changing their locations to control for these effects.

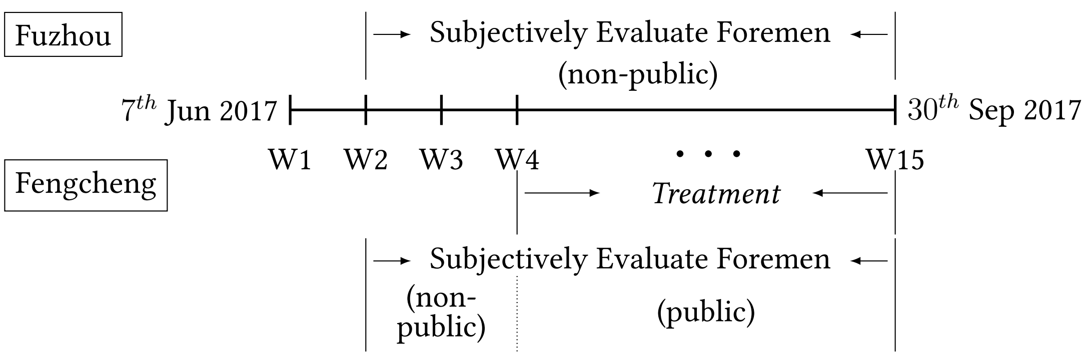
3.2 The Field Experiment

During our experiment (between 7^{th} June 2017 and 30^{th} September 2017), both workers and forewomen performed their tasks individually within their natural work environment. They were unaware of an experiment was taking place. The treatment was introduced to them by the production manager via the internal communication channel in Fengcheng factory. The reason Fengcheng was selected for implementing the treatment is that the factory manager in Fuzhou had unexpectedly submitted his resignation in June. This makes the Fuzhou factory a natural setting as a control group since the workplace condition is most likely to remain constant during the transition period. Most importantly, the board of the company shared the same notion and agreed to introduce the treatment in Fengcheng while keeping Fuzhou constant for this period of time. Therefore, we denote our experiment as a field experiment, following the terminology of Harrison and List (2004).

3.2.1 Timeline

The experiment was designed to generate exogenous variation in the trade-off between performing the perfectly measurable task and partially quantifiable job. The timeline of the experiment is shown in Figure 1. Starting from 7^{th} June 2017 individual daily production records were collected and monitored by our research team.[13] During the first experimental week (W1), production managers from both factories were trained to use the evaluation system we designed to assess the forewomen subjectively (discussed in the next subsection). On the last day of the second week (W2), forewomen's organisational activities during W2 were subjectively evaluated by the production managers in both factories while neither the workers nor the forewomen were aware of this evaluation. In each following week, the production manager performed a weekly independent evaluation of the forewomen's organisational activities. Notice that the evaluation process and the corresponding results were never made public in Fuzhou throughout the experiment. Whereas both workers and forewomen in Fengcheng were well informed about the evaluation process, how the forewomen are evaluated, and their evaluation results after the treatments were introduced (weeks 4-15). For instance, ranking results for W4 were posted on the factory floor at the end of the fourth week.

Figure 1: Timeline



Notes: W denotes the experimental week.

On the last day of week 3 (30^{th} June 2017), the production manager in Fengcheng had

[13]The factories were not recording worker's daily productivity before our intervention. They only collected the output data. Therefore, historical production data from these factories cannot be used in this study.

a regular monthly meeting with all workers and forewomen from the packaging unit. The manager ventilated production issues raised during that month and outlined plans for the upcoming months including our treatment. We instructed the manager to announce our treatment as follows. The production managers will subjectively evaluate forewomen's organisational activities each week. The evaluation starts from 1^{st} July 2017. Four criteria regarding management and organisation will be assessed. At the end of each week, the ranking for each evaluated criterion and the weekly overall ranking will be updated on the whiteboards located next to the production lines. All weekly rankings within a month are important as they will be used to compute the ranking of the month. On the last day of each month, every forewoman will receive a pecuniary reward in cash based on her monthly ranking.[14] A higher ranking yields a higher payment. The monthly ranking is then reset at the beginning of next month. In other words, both weekly rankings and monthly rankings are intra-independent.[15]

A detailed instruction was handed to each forewoman after the meeting. It illustrated the four criteria assessed with brief examples, detailed the incentive scheme, and outlined other organisation-related information. Notably, we explicitly stated that this monetary prize is independent of the current foreman subsidy. Hence the forewomen would not consider this as a replacement of the current subsidy.

3.2.2 Subjective Evaluation

In each factory, more than one manager was asked to perform the evaluation task. Two managers in Fengcheng were assigned to perform the evaluation task, while there are three evaluators in Fuzhou. The reason for appointing at least two direct managers to assess the forewomen subjectively was threefold. It is consistent with other evaluations that were organised in the factories, such as the 5S system which was assessed by five managers.[16] This prevents manager's personal perceptions and biases to influence the evaluation results. Hence forewomen are less likely to question the fairness and accuracy. Moreover, employing multiple examiners increases forewomen's costs for collusion. Last but not the least, on some occasions one manager can still provide reliable evaluations

[14] Even though the employees in this firm are paid one month in arrears, receiving prizes in advance is not fresh. There were policies rewarded employees in advance such as bringing in new recruits.

[15] There is a reason for introducing the treatment by the end of June. Individual piece rate is not constant. It is determined by the worker's monthly output. The more they produce, the higher the piece rate. Hence it affects workers' manufacturing behaviour over time. A worker may feel sluggish at one point when she realises that she can no longer reach the next higher piece rate, and vice versa. Therefore, the production manager advised us to not interfere in the middle of a month.

[16] 5S is a workplace organisation system designed to improve manufacturing efficiency. For details, see https://en.wikipedia.org/wiki/5S_(methodology).

when the other is absent due to illness or on holiday.

We consulted the management to list all the organisational activities they demand the forewomen to perform. The management proposed and exemplified four criteria which we embedded in the evaluation system, including: maintain an efficient production process (e.g. make sure the raw materials are sufficiently and unerringly distributed on the line for workers to work); increase the productivity of the line workers (e.g. manage the team effectively so that workers work efficiently, such as talk to the workers and motivate them to focus on working); reduce line defect rates (e.g. constantly remind workers to use standardised operating procedure in order to reduce the number of faulty products); and team building (e.g. provide support and communication to foster a friendly and positive work environment). Each indicator along with its associated example is clearly elucidated in the instructions given to the forewomen.

For efficiency purposes, we designed a novel spreadsheet to minimise the time required for the production managers to perform the subjective evaluation (see Figure B1 in the appendix as an example). Production managers were asked to provide relative performance evaluation by positioning sliders instead of giving exact scores to avoid a tie. We underlined that the positions of sliders within each assessed criterion should be unalike across forewomen since these are relative measurements. After positioning the sliders under each criterion, the overall ranking of each forewoman is automatically calculated and displayed. The examiners were then asked to verify whether the overall rankings are authentic. If not, they were instructed to repositioning the sliders without altering the ranking of each criterion until the valid overall rankings were reached.

In each week, both the ranking for each evaluated criterion and the overall week-ranking were posted on the Fengcheng factory floor in the form of a scoreboard and displayed in descending order. The management was instructed to put up this ranking board on the wall next to the production lines as shown in the Online Appendix. For consistency, the scoreboard only provides information for each week. At the end of each month, a hard copy which summarises four weekly rankings and the aggregated rankings of that month was posted next to the scoreboard.

3.2.3 Monetary Prizes

The management insisted to reward all forewomen rather than the highest ranked one(s). They do not have pleasant experiences with only rewarding the best employee(s) and find this type of incentive structure divisive. Therefore, we designed an incentive scheme which rewards every forewoman as presented in Table 2. Both Table 2 and the reasoning of this design (as we discuss below) were documented in the instructions given to the forewomen.

Table 2: Monetary Prizes

	Initial Foreman Fee (RMB/M)	Tournament Reward (RMB/M)	Difference from the next lower rank	Change in Total Foreman Fee (%)
	(1)	(2)	(3)	(4)
#.1 ranked forewoman	90	205	45	228%
#.2 ranked forewoman	90	160	25	178%
#.3 ranked forewoman	90	135	15	150%
#.4 ranked forewoman	90	120	10	133%
#.5 ranked forewoman	90	110	10	122%
#.6 ranked forewoman	90	100	10	111%
#.7 ranked forewoman	90	90		100%

Notes: 0 RMB will be paid to the forewoman if she is eliminated by the management. RMB/M denotes Chinese yuan per month.

The lowest ranked forewoman is paid 90 RMB per month, which is identical to the amount of money the company paid to the forewoman for her leadership role. To further determine the amount of payment given to the highest ranked forewoman, we first calculate the highest piece-rate wage per hour a forewoman can possibly get. For a forewoman to be eligible for the highest piece rate .024 as shown in Table 1, she has to produce at least 3,400 units every day. It implies that the hourly output is 310 units for a forewoman works 11 hours a day. Therefore, packing products for an hour earns her 7.44 RMB. This suggests an opportunity cost of spending one hour per day on organising teams for 28 working days is 208 RMB.[17] Similarly, for the least productive forewomen her opportunity cost of spending one hour per day on organising teams for 28 working days is 94 RMB since our data indicate that the mean of the daily output of forewomen in Fengcheng is 2,600 units, the fastest forewoman can produce 4,400 units while the slowest forewoman only produces 1,900 units in a single day.[18] Notice that before our experiment not more than two forewomen could get the highest piece rate.

By intentionally making the top three ranks more attractive, the highest ranked forewoman receives 205 RMB per month as shown in Table 2. In particular, the highest ranked forewoman is paid 45 RMB more relative to the second highest ranked forewoman, 70 RMB more than the third highest ranked forewoman, and 85 RMB more than the fourth highest ranked forewoman. For the lower ranked forewomen (rank 4-7), the payment difference between individuals adjacent in rank is parallel which equals 10 RMB. In the case of a tie, a standard competition ranking is applied, i.e. all forewomen will be paid 205 RMB if they

[17] In general, workers from Fengcheng factory work 11 hours per day and 28 days per month.

[18] The corresponding piece rate for 1,900 daily output is .0195, hence $94 = 1900 \div 11 \times .0195 \times 28$.

share the same score.

Therefore, the incentive for spending one hour per day to organise the team is as attractive as the hourly piece-rate wage. The highest ranked forewoman in the monthly tournament receives 205 RMB while the most productive forewoman is paid 208 RMB for 28 hours (i.e. one hour per day for 28 days). As for the lowest ranked forewoman, if she also turns out to be unproductive, spending one hour in a day to organise the team for 28 days gives her 90 RMB which is identical to the amount of payment for packaging the products for 28 hours - 94 RMB. Since the ranking is based on subjective evaluation results and forewomen's relative performance, the corresponding cost is perceived to be lower than packing products as reviewed in the interviews which we conducted with the forewomen after the experiment. If the most productive forewoman is given the lowest rank, she will have a strong motivation to invest more efforts in organising the team in the following weeks for a better rank. Therefore, by providing equivalent incentives and incrementally increasing the prizes when the ranking increases we believe that this scheme can offer sufficient incentives to the forewomen to spend around one hour per day to perform organisational tasks as the management craves.

Last but not the least, forewomen were also informed that the management was given the authority to eliminate their eligibility for the prizes. If the direct managers conclude that a forewoman had exerted zero efforts on any of the assessed criteria (i.e. maintain an efficient production process, increase the productivity of the line workers, reduce line defect rates, or team building), the forewoman will not be given the bonus in that month. Nonetheless, no forewoman was eliminated for the prize throughout the experiment.

4 Empirical Analysis

To test whether the introduction of an incentive scheme regarding foremen's organisational behaviours affects either the workers' or the foreman's productivity, we exploit the fact that workers and forewomen from both factories are observed over time. We estimate the following Difference-in-Difference (DiD) specification for individual i in factory f and day t:

$$log(Y)_{i,f,t} = \beta \mathit{FACTORY}_f + \gamma \mathit{POST}_t + \eta' Z_{i,f,t} + \rho I_i + \delta D_{f,t} + \epsilon_{i,f,t}, \tag{10}$$

where $log(Y)_{i,f,t}$ is the logged production-related outcomes of individual i (we analyse workers and forewomen separately) in factory f and day t, in particular, we are interested in production output and productivity. The productivity in our case is defined as a measure of the output per hour. The line forewomen are responsible for recording the data of every individual from the same production line including the daily output, time work started, and time left the factory. These figures are further scrutinised by the production manager with little measurement error and used to compute the daily productivity.

FACTORY$_f$ and *POST*$_t$ are dichotomous variables indicating the treatment factory f and the treatment period t, respectively. To take account of the natural trends in production process we control for the time-varying determinants $Z_{i,f,t}$: (i) the production line individual i worked on day t (i.e. line fixed effects), workers are assigned to work on the other production line when the downstream of her own line is disordered, to capture the variation in team efficiency spillover between production lines and account for unobserved and permanent differences in productivity across lines (e.g. distance to raw materials); (ii) an indicator variable for whether individual i is recorded sick or if there is an organisational error, which may cause negative sentiments and therefore reduce performance; (iii) a vector of variables captures the time effects including experimental weeks and the day of each week (e.g. Monday), to control for the time trend which influences individual i's productivity. Individual fixed effects I_i account for unobserved and time-invariant heterogeneities in productivity among individuals.

Finally, $D_{f,t}$ is the interaction of *FACTORY*$_f$ and *POST*$_t$ which equals to 1 if individual i is working at the factory where the treatment is already taking place. Therefore, δ is the coefficient of interest. It estimates our treatment effect. The disturbance term $\epsilon_{i,f,t}$ is individual specific. We present estimates under the fixed effects framework while the results are robust qualitatively or quantitatively under the random effects specifications. Heteroskedasticity-robust standard errors clustered at the individual level are used in all regression specifications. This allows us to address the concern that observations for an individual are not independent over time.[19]

The most critical assumption of the DiD is that the treatment and control factories have pre-treatment parallel trends in the outcome. In principle, the treatment factory and control factory are a good match. They are two sister companies which share the same board and corporate culture. Workers' incentive structure does not differ qualitatively but varies quantitatively. The marginal variations in quantity are mainly driven by the condition of the local labour market which is exogenous to the workers' outcomes. Therefore, working patterns in these two factories should be comparable.

[19]In general, the variation in worker production outcome over time across workers should be independent. There are cases where workers are likely to be dependent when the upstream production unit is short-handed. This is because all lines acquire manufacturing materials from the same upstream, the faster a productive worker (or a line collectively) can consume the materials the more likely the less productive ones have to wait for materials. This wastes the less productive workers' time on production and leads to a fall in productivity. For robustness, we applied the wild cluster bootstrap (see Cameron et al., 2008, for details) while clustering at both individual and line levels. The main results remain unchanged qualitatively.

4.1 Descriptive Evidence

To form a consistent sample throughout our analysis, we exclude workers who were new recruits and those who were working in their three-month probation period when the experiment started.[20] Employees under their probation period are offered a different payment scheme - hourly rate - in comparison to regular workers who are paid by piece rate. This implies that employees who were hired after the first day of March 2017 are excluded from the sample. A few workers from the treated factory were on holidays when we started the experiment in June, and they returned to work after the treatment was introduced. They are also not valid for the difference-in-difference estimation. Hence we leave them out of the analysis. Moreover, there are some workers left the factory during the experiment. It is reasonable to assume that they may not respond to our treatment because they intend to leave. Therefore, we exclude them from the analysis and opt for a clear measure of the treatment effect. In total, 70 workers constitute the final sample with 27 workers from the controlled factory Fuzhou and 43 in the treated factory Fengcheng.

In addition, the types of the infusion sets packed are different across factories. Products sold in the domestic market are easier to pack than those sold in the international market. When Fuzhou factory mainly focuses on the local market, Fengcheng factory produces goods for both markets. Luckily, the management of Fengcheng factory has developed methods to calculate piece rates for different types of products based on the level of difficulty. We used the same technology to standardise individual outputs in Fengcheng factory so that figures are comparable across factories.

Table 3 presents summary statistics for each factory during the pre-treatment period (June) and the post-treatment period (July, August, and September), including number of employees, number of production lines (which is also the number of forewomen, as there is only one forewoman assigned to each line), worker's daily output, worker's productivity (output per hour), forewoman's daily output, and forewoman's productivity.

Hourly productivity is the ratio of daily output to the total hours worked in that day. The total number of hours worked per day is derived by the difference between the time when the individual started her work and the time when she left the production line. We do not observe the precise time an individual had spent on the manufacturing task. Nevertheless, when we use the difference-in-difference estimation and assume that individual's work behaviour is constant, the treatment effects estimated are valid.

As shown in column 4 of Table 3, in August, one worker in factory Fuzhou was absent for the whole month because of illness, and one worker in factory Fengcheng was assigned to another production unit which is not included in our sample. In September (column 5), one forewoman from factory Fengcheng got sick and left the job for two weeks

[20] All forewomen had more than two years working experience in the factories.

Table 3: Summary Statistics

	June	Jul-Sep	Jul	Aug	Sep
	(1)	(2)	(3)	(4)	(5)
Panel A. Fengcheng					
Number of Employees	43	41.10	43	42	38
Number of Lines	6	6	6	6	5
Worker Daily Output	1,125.9	1,176.8	1,179.4	1,161.3	1,190.5
	(242.2)	(248.6)	(236.0)	(268.8)	(239.0)
Worker Hourly Productivity	91.74	96.86	93.94	97.25	99.67
	(17.87)	(17.45)	(16.57)	(18.08)	(17.22)
Forewoman Daily Output	1,027.8	1,093.9	1,082.0	1,092.4	1,109.6
	(210.4)	(250.2)	(222.7)	(274.0)	(252.3)
Forewoman Hourly Productivity	85.65	89.98	86.67	90.40	93.34
	(20.03)	(20.10)	(19.24)	(20.52)	(20.13)
Panel B. Fuzhou					
Number of Employees	27	26.71	27	26	27
Number of Lines	7	7	7	7	7
Worker Daily Output	1,082.9	1,049.7	1,039.4	1,032.7	1,072.4
	(221.9)	(242.4)	(228.5)	(284.8)	(216.6)
Worker Hourly Productivity	93.04	91.98	90.95	91.94	92.96
	(16.68)	(13.60)	(14.24)	(14.23)	(12.39)
Forewoman Daily Output	1,121.4	1,087.3	1,067.1	1,073.5	1,119.5
	(172.9)	(197.7)	(179.6)	(237.7)	(171.4)
Forewoman Hourly Productivity	94.72	94.97	93.51	95.00	96.39
	(12.62)	(10.35)	(10.07)	(11.08)	(9.809)

Notes: Productivity is a measure of the output per hour. June indicates the pre-treatment period and Jul-Sep implies the post-treatment period. The top number in each cell denotes the mean and the number in parentheses denotes the standard deviation.

whose foreman responsibility was soon succeeded by a line supervisor. However, a line supervisor does not perform manufacturing tasks. We have neither production records to determine her productivity nor the relative performance in organising teams. Observations of workers from this line and the newly appointed line supervisor are not included in our sample.[21]

[21]Summary statistics for employees' other characteristics which we collected after the experiment in factories Fengcheng and Fuzhou are reported in Tables C1 and C2 in the appendix, respectively.

The contrasting patterns in Table 3 are that performance was increasing from June onwards in Fengcheng but decreasing in Fuzhou. In particular, workers' daily output, workers' hourly productivity, and forewomen's daily output in Fuzhou were all lower in July-September in relation to June, and forewomen's hourly productivity in July is lower than their productivity in June. According to the company, this decline in Fuzhou is a normal pattern that applies to both factories. It is mainly due to the weather. Temperature reaches the peak of each year during July and August. Because the factories produce medical appliances, the workplaces are sterile, clean, and purified plants. Workers must wear impervious gowns in the workplace to reduce the risk of contamination. When the temperature gets high, the environment becomes too uncomfortable to work, and the production performance falls sharply around this time of the year. Therefore, in general, performance is expected to fall in both factories during this period. The fall in Fuzhou is following this normal trend, while the contrasting increase in Fengcheng is due to the introduction of our treatment.[22]

Furthermore, as the number of days a worker worked in each month vary across workers and months, the summary statistics of monthly production outcomes aggregating observations from all workers are not particularly informative. We discuss the results with the assistance of other analyses in the next subsections.

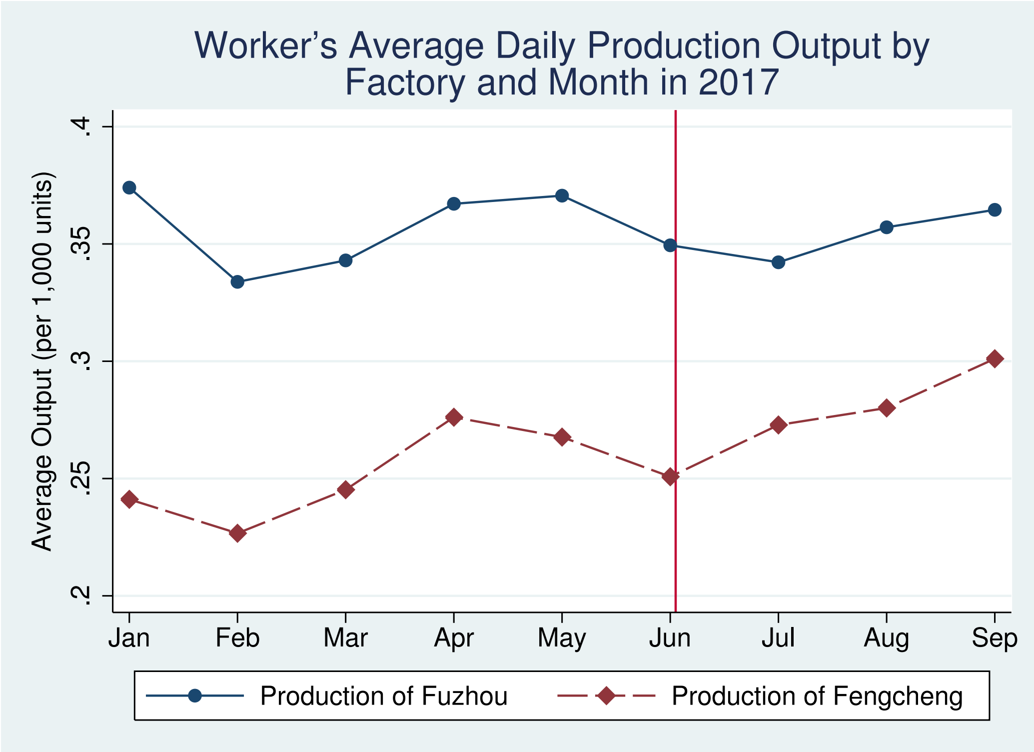
We are not able to show the trends in workers' productivity beyond our intervention because the company only has workers' output data to determine their salary. But, we are able to show the evidence of a parallel pre-trend based on the firm's administrative data. Figure 2 shows the average daily production output of workers for both the treatment and the control factory in each month in 2017 before our experiment ended (by September).

It indicates that there is a comparable trend in workers' daily output before our treatment was introduced in July 2017, with the exception of January. The variations between January and February are subject to the Chinese New Year. Depending on the condition of local labour markets and the turnover rate of workers, for instance, the output would fall deeper in February if the factory experienced a hard time to retain its workers and to hire new ones (such as Fuzhou). Therefore, the first trustworthy data point is February. The movements of these two factories between February and June are indeed parallel. This is not surprising because, as I discussed above, both factories share the same company culture, and their remuneration systems are qualitatively identical.

Furthermore, this administrative data can also illustrate our treatment effect. While the daily output of workers in Fuzhou (the solid line) fluctuates at its normal interval, the figures in Fengcheng (the dashed line) started to rise after the introduction of our

[22]In the Online Appendix, I provide the evidence from WorldWeatherOnline.com which shows the maximum temperature in both cities is above 35-degree Celsius between July and August.

Figure 2: Production Trend in Both Factories



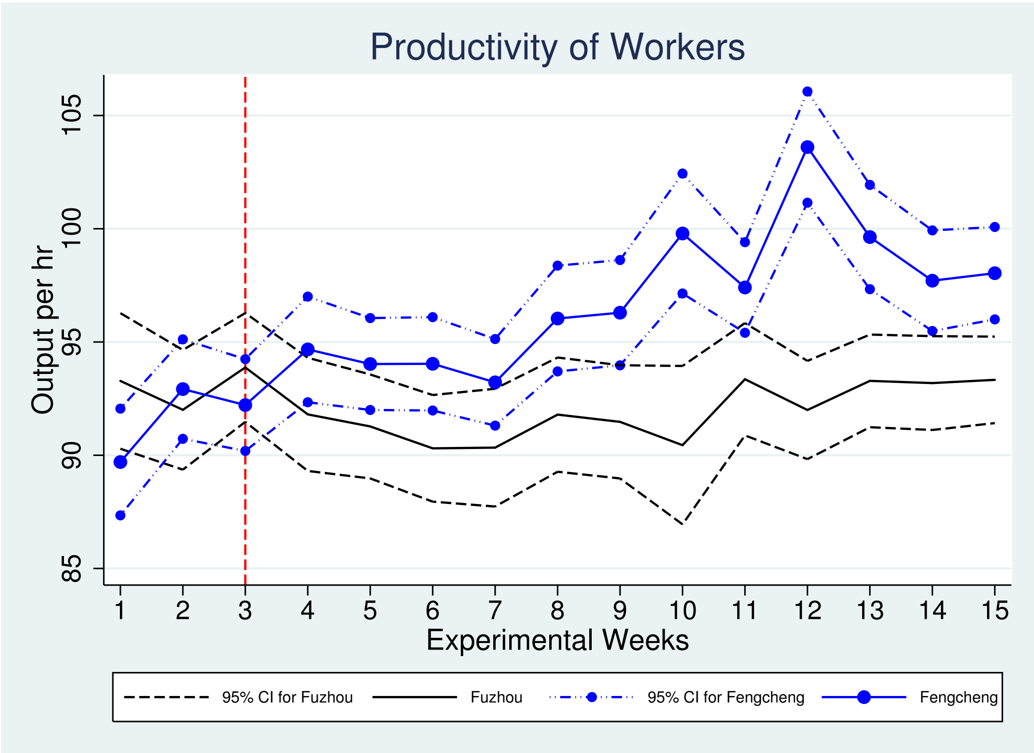
Notes: The vertical line indicates the introduction of our treatment in Fengcheng.

treatment (in July) and further exceeded the peak of the year (on April 2017) since August.

4.2 Performance of Workers

First, we graphically compare worker's hourly productivity in the treated factory with the one in the controlled factory before and after the treatment was introduced. Recall that the treatment was introduced in Fengcheng by the end of the third experimental week, and it lasted from week 4 to the last week of our experiment. For each factory, we calculated the average output per hour of all workers in each week. Figure 3 depicts the mean of worker's productivity, averaged across all workers, in each experimental week. The area between two dashed lines corresponds to the 95% confidence intervals. The figure suggests that worker's productivity increased marginally during the first four weeks of treatment while worker's productivity in the controlled factory is somewhat flat. The treatment effect started to rise dramatically in the second treatment month. It did not decrease over time albeit there was a decline during the last few experimental weeks. This drop is mainly due to an unexpected incident took place in the treated factory which we will discuss below. On the other hand, in the status quo (the first three experimental weeks), we do observe a similar trend between factories. This suggests that the parallel trends assumption for the difference-in-difference estimation is satisfied.[23] [24]

Figure 3: Worker Mean Productivity



Notes: The vertical dashed line indicates the last week before our treatment was introduced in Fengcheng.

To present formal evidence on the effect of the introduction of subjective evaluations and monetary prizes regarding forewomen's organisational behaviours on worker's performance, we estimate specification 10.

Columns (1) and (5) of Table 4 shows that compared to workers who had no additional incentives to perform organisational tasks in Fuzhou factory the subjective evaluations and monetary prizes increased workers' production output and productivity in Fengcheng

[23]One way to formally test the parallel trends assumption is to drop the *POST* from specification 10 and augment it with the experimental week variable and its interaction with the treatment indicator. To allow for weekly fluctuations and variations, we put every two weeks into a group. Regression results for both productivity and production output are shown in Table C3 and Table C4 in appendix, respectively. An alternative way is to include linear factory specific time trends among other regressors in specification 10. Our conclusion does not change. Results are available upon request.

[24]Figure 3 does not change qualitatively if we substitute productivity with worker's production output. However, outputs have more noises as workers were late for work or left work early sometimes (e.g. due to sickness).

Table 4: The Treatment Effect on Worker's Performance

	Log(Output)				Log(Productivity (output per hour))			
	Jul-Sep	Jul	Aug	Sep	Jul-Sep	Jul	Aug	Sep
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fengcheng	-0.426***	0.299***	0.043	0.069**	-0.286***	0.208***	-0.094***	-0.055**
	(0.101)	(0.009)	(0.062)	(0.030)	(0.043)	(0.008)	(0.033)	(0.023)
Post	-0.038**	-0.056***	-0.086***	-0.018	0.027	0.003	0.031**	0.022
	(0.018)	(0.017)	(0.022)	(0.021)	(0.018)	(0.015)	(0.015)	(0.020)
Fengcheng*Post	0.091***	0.099***	0.105***	0.079***	0.066***	0.044***	0.060***	0.082***
	(0.017)	(0.016)	(0.019)	(0.022)	(0.017)	(0.014)	(0.017)	(0.020)
Observations	5,655	2,770	2,647	2,712	5,655	2,770	2,647	2,712
Clusters	57	57	57	57	57	57	57	57
R^2	0.528	0.670	0.483	0.610	0.780	0.811	0.761	0.791
Controls	YES	YES	YES	YES	YES	YES	YES	YES

Notes: The unit of observation is worker i. The dependent variables in Columns 1-4 and Columns 5-8 are the log of worker's daily output and the log of worker's productivity, respectively. Columns 1 and 5 show the results for the full sample includes observations from June 7th until September 30th while Columns 2-4 and 6-8 compare the observations from the pre-treatment period (June) to each post-treatment month separately. Productivity is a measure of the output per hour. Worker fixed effects, factory-line fixed effects, week fixed effects, day of the week fixed effects (e.g. Monday), and an indicator variable for whether the worker is recorded sick or if there is an organisational error are included in all regressions. Robust standard errors clustered at the worker level are reported in brackets below the estimates. *** Significant at 1% level, ** significant at 5% level, * significant at 10% level.

factory by 9% and 7% respectively. This improvement is statistically significant at 1% level.[25]

All remaining workers in our sample had more than three months of experience on the job. Thus, learning by doing should not be significant during our experimental period. Nevertheless, as the incentive scheme for the forewomen was new and the forewomen exerted more efforts in organising the production line, the workers may still need to learn how to work efficiently on a better-organised line. On the other hand, foreman-worker ties might strengthen as time passes when the forewoman invests more time in organising the production line. As a result, this may develop team cohesion and further increase workers' productivity. To test this, we divide the post-treatment period into three months. We can then estimate whether the treatment effects during the first, second, and third post-treatment month vary. Columns (6), (7), and (8) show the results regarding productivity

[25]Robust standard errors clustered at the individual level are reported in brackets below the estimates.

for July, August, and September, respectively. The estimates are indeed increasing over time, and they are statistically significant. Our treatment increased worker's productivity significantly by 4.4% in July and the figure further increased to 8.2% in September. Thus, we conclude that a high-powered incentive scheme regarding foreman's organisational performance has a significant impact in increasing worker's productivity.

Furthermore, in line with Figure 3 Column (5) indicates that the treatment failed to further increase worker's output in September. The tumble was due to the fact that a large number of defective products were returned from the market. Workers who participated in our experiment were responsible for unpacking these products for remaking. This task was not incentivised monetarily. Hence, workers who spent some time on this unpaid job were displeased, and their outputs were diminished. However, this is not reflected in workers' productivity figures as shown in Column (8). This is because the management requested the forewomen to record the time workers worked on this task, which enables us to deduct the time the workers spent on this task when we calculate their productivity.[26]

By design, we can observe the real scores the managers gave to each forewoman regarding their organisational performance, although these scores are unobservable to both the managers and forewomen. Hence, we can further add the subjective evaluation scores into the DiD specification 10 to test whether subjective evaluation scores are correlated with workers' productivity. Because the evaluations took place on a weekly basis, the sample is now aggregated to the week level. Some individual-level controls are no longer available. We only control for the forewoman and week fixed effects here. Table 5 shows that the coefficient of *FACTORY* * *SCORES* * *POST* in Column (1) is indeed positive and statistically significant at 5% level, where *SCORES* indicates the logged evaluation scores of the forewomen. It indicates that a one percent increase in a forewoman's evaluation score is associated with a 0.24% increase in the (week) average productivity of workers during the treatment period. This positive association can be found in all three treatment months as shown in Columns (2)-(4). It reached its peak at 0.4% in August, and the coefficient becomes statistically significant at 10% level in September.

4.3 Performance of Forewomen

Similarly, we analyse how the subjective evaluation and monetary prizes affect forewomen's production performance as we did with the workers above. This is a formal test of the prediction we derived from the theoretical model. First, we test whether forewoman's productivity in the treated factory changed after the introduction of the treat-

[26] Recording the time workers spent on this particualr task was not too intricate for the forewoman because, in general, workers who were assigned to the unpacking job did not return to their regular job on the same day.

Table 5: Subjective Evaluation Scores of Forewomen and Workers' Productivity

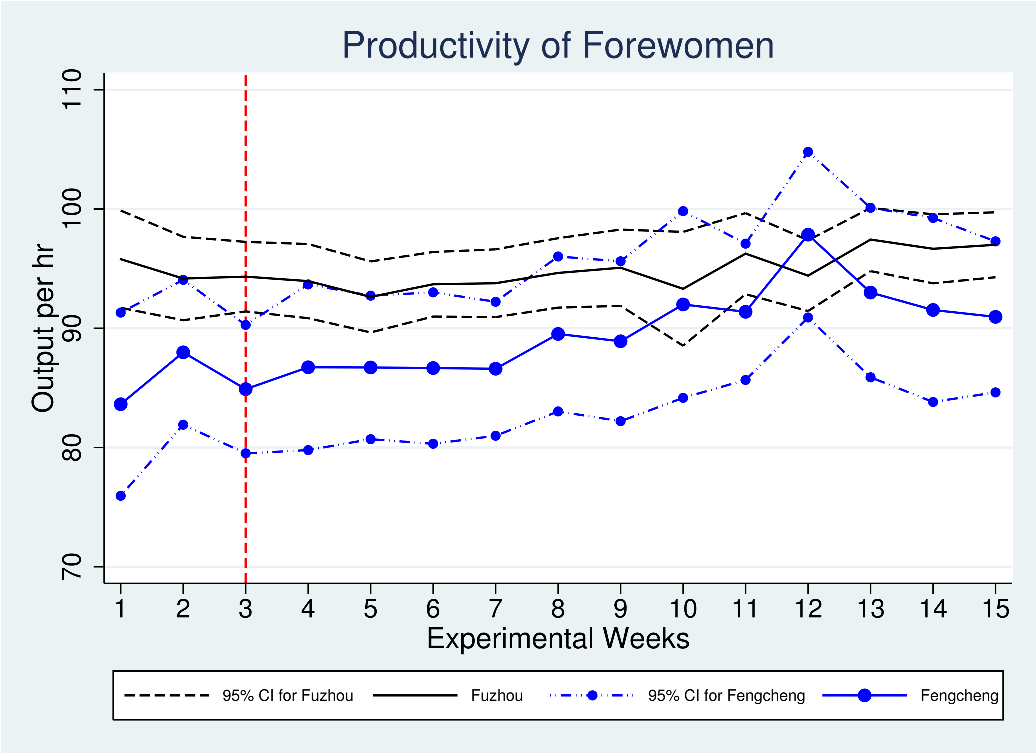
	Productivity (output per hour)			
	Jun-Sep	Jun vs. Jul	Jun vs. Aug	Jun vs. Sep
	(1)	(2)	(3)	(4)
Fengcheng	1.065*	0.784	0.846	0.975
	(0.537)	(0.511)	(0.589)	(0.713)
Scores	0.053	0.054	0.064	0.063
	(0.063)	(0.066)	(0.065)	(0.068)
Post	0.450	0.521	0.411	0.470
	(0.344)	(0.426)	(0.359)	(0.387)
Fengcheng*Post	-1.436**	-1.444**	-2.366**	-1.565*
	(0.523)	(0.555)	(0.908)	(0.814)
Post*Scores	-0.077	-0.095	-0.070	-0.083
	(0.061)	(0.075)	(0.064)	(0.069)
Fengcheng*Scores	-0.185*	-0.147	-0.157	-0.175
	(0.088)	(0.084)	(0.096)	(0.116)
Fengcheng*Post*Scores	0.247**	0.245**	0.400**	0.273*
	(0.086)	(0.092)	(0.150)	(0.133)
Observations	174	76	74	72
R^2	0.878	0.900	0.893	0.833
Experimental Week FE	YES	YES	YES	YES
Forewoman FE	YES	YES	YES	YES

Notes: The unit of observation is forewoman i per week. The dependent variables are the log of forewoman's productivity. Scores indicates the logged evaluation scores of the forewomen. Column 1 shows the results for the full sample includes observations from June 7th until September 30th while Columns 2-4 compare the observations from the pre-treatment period (June) to each post-treatment month separately. Forewoman fixed effects and week fixed effects are included in all regressions. Robust standard errors clustered at the forewoman level are reported in brackets below the estimates. *** Significant at 1% level, ** significant at 5% level, * significant at 10% level.

ment. Figure 4 shows the mean of forewomen's weekly productivity, averaged across all forewomen, in each experimental week. Forewomen's productivity started to increase gradually in the treated factory after week 7 but descended in the last three weeks, while forewomen's productivity in the controlled factory is exceptionally consistent throughout the experiment.

We also observe a significant difference in forewomen's productivity between the treated factory and the controlled factory since the status quo. This can be mainly driven by the individual-specific heterogeneity as we only have 7 forewomen in the controlled

Figure 4: Forewomen Mean Productivity



Notes: The vertical dashed line indicates the last week before our treatment was introduced in Fengcheng.

factory and 6 in the other. Nevertheless, there is no reason to expect that the average employee performance must be identical across factories since the two factories are independent of each other. Since the parallel trends assumption holds our difference-in-difference estimations are valid.[27]

One explanation for the convergence later in the experiment is that we exclude a forewoman, who was replaced by a line supervisor, from the sample after week 9. Therefore, the aggregated observations are not informative. If we plot Figure 4 and exclude this forewoman who was replaced in the mid of the experiment throughout, the convergence is moderated considerably.

Columns (5-8) of Table 6 confirms this result by regressing log of forewoman's productivity on the regressors of specification 10. The coefficients for the variable of interest (Fengcheng*Post) in Columns (5-8) are positive but statistically insignificant. On the other hand, Columns (1-4) show that the treatment leads to an increase in forewomen's production outputs. In particular, the subjective evaluation and monetary prizes statistically significantly increased forewoman's output by 8% overall, while the coefficients

[27]Formal tests are used to confirm the parallel trends assumption. Regression results for forewomen's productivity and production output are shown in Table C5 and Table C6 in appendix, respectively.

Table 6: The Treatment Effect on Forewoman's Production Performance

	Log(Output)				Log(Productivity (output per hour))			
	Jul-Sep	Jul	Aug	Sep	Jul-Sep	Jul	Aug	Sep
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fengcheng	0.180***	0.168***	0.194***	0.160***	0.258***	0.252***	0.259***	0.238***
	(0.024)	(0.014)	(0.023)	(0.023)	(0.017)	(0.012)	(0.012)	(0.014)
Post	-0.056*	-0.088**	-0.122***	-0.028	0.027	-0.003	0.030	0.031
	(0.031)	(0.039)	(0.039)	(0.032)	(0.026)	(0.030)	(0.027)	(0.028)
Fengcheng*Post	0.082**	0.100***	0.098**	0.058	0.034	0.023	0.042	0.038
	(0.031)	(0.028)	(0.040)	(0.043)	(0.021)	(0.023)	(0.023)	(0.025)
Observations	1,312	644	621	621	1,312	644	621	621
Clusters	13	13	13	13	13	13	13	13
R^2	0.350	0.491	0.314	0.497	0.845	0.873	0.832	0.840
Controls	YES	YES	YES	YES	YES	YES	YES	YES

Notes: The unit of observation is forewoman i. The dependent variables in Columns 1-4 and Columns 5-8 are the log of forewoman's daily output and the log of forewoman's productivity, respectively. Columns 1 and 5 show the results for the full sample includes observations from June 7th until September 30th while Columns 2-4 and 6-8 compare the observations from the pre-treatment period (June) to each post-treatment month separately. Productivity is a measure of the output per hour. Forewoman fixed effects, factory-line fixed effects, week fixed effects, day of the week fixed effects (e.g. Monday), and an indicator variable for whether the forewoman is recorded sick or if there is an organisational error are included in all regressions. Robust standard errors clustered at the forewoman level are reported in brackets below the estimates. *** Significant at 1% level, ** significant at 5% level, * significant at 10% level.

are extremely large during the first two months as shown in Columns (2) and (3). It increased forewoman's output by around 10% in July and 9.8% in August, and the estimates are statistically significant at the 1% level and 5% level respectively. However, the coefficient for September is relatively small (by 5.8%) and statistically insignificant. Like the workers, forewomen were sometimes assigned to work on the products returned from the market during September which can somewhat explain why the statistical significance disappeared.[28]

Overall, the evidence indicates that strengthening incentives in the organisational dimension encouraged forewomen to exert more effort in organising teams as workers became more productive which affirms the efficiency improvements among team members. On the other hand, the production outputs of forewomen also increased after the intro-

[28]As the number of clusters in our case is small, we perform the wild cluster bootstrap as suggested in Cameron et al. (2008). With more than 200 replications, the results remain unchanged qualitatively.

duction of subjective evaluations and monetary prizes. Both changes are in line with the predictions listed in Proposition 1 that the introduction of a new bonus scheme increases worker's production efforts, and forewomen's organisational inputs and production efforts. As a result, their outputs increase.

5 Additional Evidence

5.1 Team Efficiency and Forewoman's Production Performance

The underlying assumption of our model is that there is a positive spillover effect among employees (including the forewoman) on the production line. For instance, in our case, one of the assessed organisational criteria is to maintain the efficiency of the production process, whereas forewomen have to invest effort and time in allocating raw materials effectively for workers to use. This effort increases not only workers' production performance but also the forewomen's because when the resources on the production line are systematically organised the forewoman is unlikely to be interrupted by the workers for this issue frequently, and the forewoman herself also gains easy access to the resources. Therefore, a well-organised forewoman is able to work on her own rhythm and maintain productive. To check for this, we use the average production performance of workers, excluding the forewoman, and test its association with forewoman's production performance by estimating the following specification:

$$log(Y)_{i,l,f,t} = \beta log(\overline{Y})_{-i,l,f,t} + \eta' Z_{i,l,f,t} + \rho I_i + u_{i,l,f,t}, \tag{11}$$

where $log(Y)_{i,l,f,t}$ denotes the logged production-related outcomes (either output or productivity) of forewoman i from production line l in factory f and day t. $\overline{Y}_{-i,l,f,t}$ is the logged average performance of other coworkers (excluding the line forewoman) $-i$ from the same production line l in factory f and day t and all other variables are as previously defined.

Columns (1) and (2) of Table 7 reports the estimates from specification 11 for the subsample of forewomen who are observed throughout the experiment. In line with our assumption we find a positive association between forewoman's production performance and her coworkers', and the coefficients are statistically significant at 1% level. The spillover effect of team efficiency is extremely large: a 10% increase in a forewoman's coworkers' average production (productivity) will lead to a 10% (5%) increase in her own production (productivity).

While the productivity of workers can be linked to the team efficiency spillover, it can also be dependent on the experience of workers in a fabricating setting like ours. Experi-

Table 7: The Spillover Effect of Team Efficiency

	Log(Output)	Log(Productivity (output per hour))		
	(1)	(2)	(3)	(4)
Log(Ave. production of coworkers)	0.962***			
	(0.043)			
Log(Ave. productivity of coworkers)		0.589***	0.602***	
		(0.088)	(0.088)	
Log(Ave. experience of coworkers)			-0.054	
			(0.032)	
Assigned to a new line				-0.017
				(0.012)
Observations	1,305	1,305	1,305	3,276
Clusters	13	13	13	66
R^2	0.741	0.888	0.844	0.801
Controls	YES	YES	YES	YES
Sample	Only Foremen	Only Foremen	Only Foremen	Workers & Foremen

Notes: The unit of observation is forewoman i in Columns 1-3 while Column 4 uses the sample includes both forewomen and workers. The dependent variables in Columns 1 is the log of individual daily output while the dependent variables of Columns 2-4 are the log of individual productivity. Productivity is a measure of the output per hour. Individual fixed effects, factory-line fixed effects, week fixed effects, day of the week fixed effects (e.g. Monday), and an indicator variable for whether the forewoman is recorded sick or if there is an organisational error are included in all regressions. Robust standard errors clustered at the individual level are reported in brackets below the estimates. *** Significant at 1% level, ** significant at 5% level, * significant at 10% level.

enced workers may have developed advanced techniques to perform the task, which they can share with their teammates. Hence any sorting into teams based on individuals' experience or productivity might lead to overestimating the positive spillover effect of team efficiency proxied by worker's productivity. To provide evidence on this, Column (3) tests the relationship between forewoman's productivity and her coworkers' seniority. Experience is defined as the number of days the worker worked in the factory. The association is statistically indistinguishable from zero. Furthermore, to test whether the management's decision to assign a worker to a production line depends on the worker's productivity, we use the sample includes both forewomen and workers and adjust specification 11 by regressing individual productivity (including both forewomen and workers) on a dummy variable for individuals who were relocated to another production line during our experiment. In line with the management's statement, the estimates reported in Columns (4) provide no evidence that the management's decision of relocating workers is associated with worker's productivity.[29]

5.2 Forewomen's Trade-off

Taken together, our findings in Section 4.3 and Section 5.1 indicate that incentives which encourage forewomen to undertake the organisational task indeed improve the production efficiency on the assembly line, for instance by rearranging the raw materials so that workers have easier access to the resources. While workers become more productive, the spillover effect of team efficiency also increases forewomen's production performance. The introduction of subjective evaluations and monetary prizes to forewomen regarding their organisational performance increased both workers' and forewomen's output by 9% and 8%, respectively. However, we do find evidence that workers' productivity had been increased by roughly 7%, the effects on forewomen's productivity cannot be economically and statistically distinguished from zero. A quick answer to this is the sample size. There are only 13 forewomen work in these factories comparing to 57 workers. Thus, the standard errors reported in Table 6 may be not informative.

Another way to explain an increase in the daily output but not output per hour is that the number of hours worked has also increased. This is consistent with the results reported in Table 8 which presents regressions using specification 10 with the number of minutes worked on the job per day as the dependent variable.

The results illustrate that the introduced incentive scheme increases the time forewomen spent on the job by roughly 28 minutes per day during the post-treatment period, but it is not significantly different from zero. If we decompose the treatment effects into each post-treatment month, in Column (2), we find that the incentive scheme has a significantly positive impact on the time forewomen invested per day during July. Forewomen from Fengcheng factory spent an additional 51 minutes per day on the job after the introduction of subjective evaluations and monetary prizes. This is consistent with the design of our monthly prizes which is precisely aimed to motivate the forewomen to spend one hour per day on organising the team instead of packing the products. It suggests that forewomen exert more effort in production by extending their hours of work when they invest more time in coordinating with workers. Columns (3) and (4) suggest that the treatment effects fade away over time. This is not particularly surprising because, according to the management, after the introduction of the subjective evaluations and monetary prizes forewomen are more frequently engaging in organisational tasks. This helped them to

[29]Performing the wild cluster bootstrap does not change our results qualitatively.

Table 8: The Treatment Effect on Forewoman's Working Time

	Number of Minutes Worked in a Day			
	Jul-Sep	Jul	Aug	Sep
	(1)	(2)	(3)	(4)
Fengcheng	-56.179**	-58.518***	-50.622**	-53.720**
	(18.471)	(11.001)	(18.044)	(17.638)
Post	-50.008***	-49.066**	-60.461***	-36.427**
	(13.999)	(22.448)	(16.016)	(15.044)
Fengcheng*Post	27.716	50.698**	28.924	12.794
	(23.804)	(18.852)	(32.352)	(31.925)
Observations	1,312	644	621	621
Clusters	13	13	13	13
R^2	0.316	0.419	0.295	0.371
Controls	YES	YES	YES	YES

Notes: The unit of observation is forewoman i. The dependent variables in Columns 1-4 are the working time (number of minutes) a forewoman worked in a day. Columns 1 shows the results for the full sample includes observations from June 7th until September 30th while Columns 2-4 compare the observations from the pre-treatment period (June) to each post-treatment month separately. Forewoman fixed effects, factory-line fixed effects, week fixed effects, day of the week fixed effects (e.g. Monday), and an indicator variable for whether the forewoman is recorded sick or if there is an organisational error are included in all regressions. Robust standard errors clustered at the forewoman level are reported in brackets below the estimates. *** Significant at 1% level, ** significant at 5% level, * significant at 10% level.

develop different styles of leadership and further equipped them with a variety of organisational skills. With more organisational experience forewomen were able to organise the workers more efficiently. An organisational task which costs the forewoman half an hour in July might only take ten minutes in September.[30]

Therefore, these findings suggest that multitasking forewomen spent more time on the job when they were given a higher-powered incentive on the organisational dimension.[31] Nonetheless, the impact of subjectively evaluating forewomen's organisational behaviour and monetary prizes on their productivity remains ambiguous since we do not observe the precise time forewomen invest in either manufacturing task or organising task.

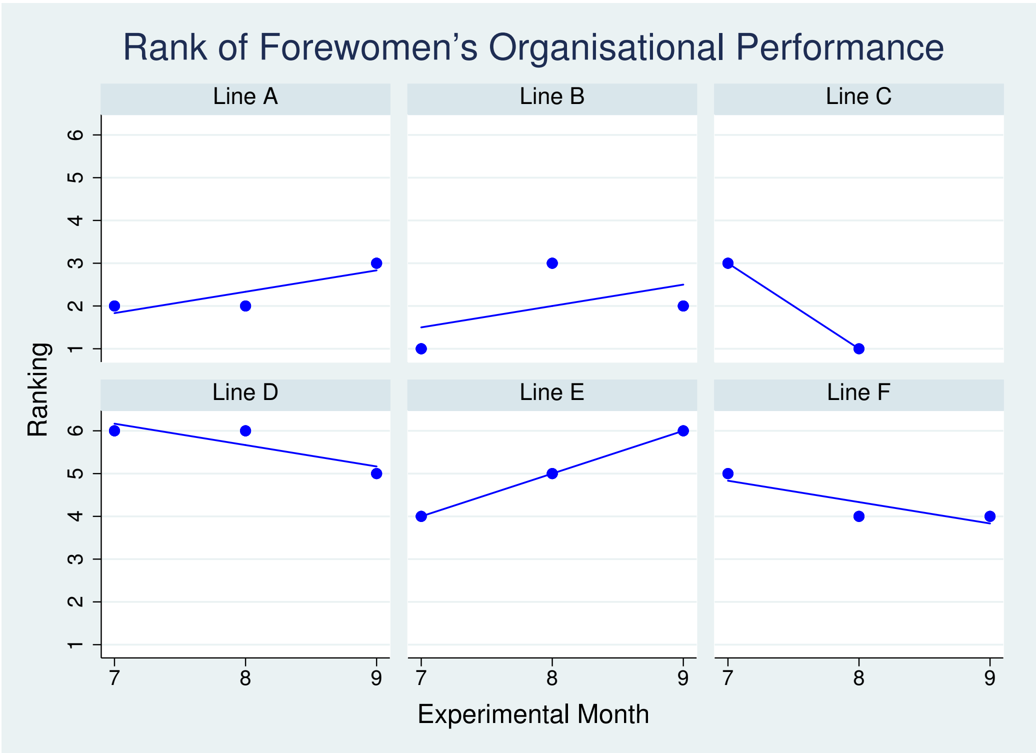
[30]The number of clusters in our case is small. Therefore, we perform the wild cluster bootstrap. Results do not change qualitatively.

[31]For the workers, they have also increased their time on the job during the treatment period as shown in Table C7 in the appendix. But, the effect is relatively small, the coefficients are more than 50% smaller than the ones for the forewomen.

5.3 Forewomen's Ranking

Finally, we use information on forewomen's rankings to provide evidence on whether the effect of subjective evaluations and monetary prizes are heterogeneous across forewomen's productivity distribution when the firm offers multiple piece-rate schemes (higher individual productivity yields a higher piece rate). In line with the complementarity assumption made in our production function that in Fengcheng during the treatment period the ranking of forewomen's organisational performance positively associate with their productivity ranking (correlation coefficient = 0.621 and statistically significant at 1% level), suggesting that a forewoman is more productive if she is ranked higher in leadership. It is also worth noting that the ranking distribution is heterogeneous as shown in Figure 5. Three forewomen (from Line A, Line B, and Line C) remained at the bottom of the ranking distribution, while other three forewomen competed for the top three rankings throughout the experiment.

Figure 5: Forewomen's Leadership Ranking in Fengcheng during the Treatment Period



Given the nature of the data, we are unable to identify how much time the forewoman have spent on organising teams in each day. However, it is still informative to compare the shift in a forewoman's weekly organisational ranking with the change in her weekly productivity because a forewoman who invests more time on organisation may increase her ranking but hurt productivity. The result suggests a negative correlation but relatively small and statistically insignificant (correlation coefficient = -0.077).

The evidence also suggests that in the treated factory Fengcheng during the treatment period, the forewoman's overall ranking regarding her organisational performance is positively correlated with the ranking of workers' average productivity but not perfect (coefficient = 0.651). It indicates that the management evaluated the forewomen as we instructed. They considered workers' productivity as one of the criteria to determine forewoman's ranking but not entirely rely on it. This is crucial as the ranking incentive and monetary prizes provided to forewomen are entirely dependent on how the management subjectively evaluate the forewomen. It also implies that workers indeed benefit from a well-organised forewoman.

However, the correlation between forewoman's organisational ranking and workers' productivity ranking is extremely weak and negative (coefficient = -0.047) when the evaluation is private information (in the controlled factory and during the pre-treatment period in the treated factory). This does not worry us from measuring the treatment effect because both forewomen and workers were not aware of this evaluation, it would not affect their performance. But, this result may raise a concern to the firm that the quality of assessment is poor when examiners are not monitored by either the firm or the examinees. Transparency is vital for the success of subjective evaluations.

Taken together, these results are consistent with our assumptions that forewoman's organisational inputs and production efforts are complements, and there exists a positive team efficiency spillover effect. Workers are more productive when the forewoman is a better organiser, and an efficient team results in a more productive forewoman.

6 Conclusions

Group leaders are usually responsible for organising the groups and contribute to the goal as a member. In the workplace, when one dimension of output is perfectly observable and quantifiable and the other is not, the classic multitasking theory applies. We address the issue by providing empirical evidence on the effects of multidimensional incentives and subjective evaluations. Through our interactions with managers at two Chinese factories who are struggling with this problem, we implement a natural field experiment to evaluate the impact of subjective evaluations and monetary prizes regarding foremen's organisational performance on two outcomes: worker's production performance, and foremen's production performance. When the former should be undoubtedly benefited from a better-organised group, the latter faces trade-offs. Specifically, by introducing the treatment in one factory while keeping the other factory as constant for three months, we provide a clean difference-in-difference test of the effects in a natural setting. This is important given the increased popularity of teams in industries such as manufacturing, academia, and healthcare.

Our results provide some meaningful insights: first, as we incent foremen to invest their time (not more than one hour per day) in organising the production process by introducing subjective evaluations and monetary prizes (which is equivalent to the hourly wage loss from manufacturing), their organisational inputs increase. As a result, the workers become more productive. We also find that a shift in a foreman's organisational ranking is negatively associated with the change in her productivity. A policy implication is that an incentive scheme which is based on the subjective evaluation results of group leaders' organisational activities is able to encourage leadership behaviour. But, there is a caveat to this: we do not find a positive correlation between foremen's organisational ranking and workers' productivity ranking when the subjective evaluation is not public information. This suggests that a subjective evaluation system may be ineffective when it is not under public scrutiny.

Second, we find that foremen's daily production output does not fall even when they spend more time on organising teams. This is because forewomen increased their working time on the job. We further show that there is a strong and positive spillover effect of team efficiency in the workplace, and foremen's organisational inputs and manufacturing efforts are complements. Nonetheless, we do not observe the specific time foremen invest in either production or organisation. The change in foremen's productivity is ambiguous.

Further, it is possible that peer pressure also plays a role in our setting as we do find a positive association between a forewoman's coworkers' productivity and her own productivity. The peer pressure which Chinese usually refer to as "Face" represents a person's reputation and feelings of prestige in the workplace. It may force a comparison of oneself versus her colleagues' performance. In our context, a forewoman whose productivity falls behind of line workers, or falls short of the local norm, may feel disgraced or dishonoured. This may propel them to increase efforts (e.g. Mas and Moretti, 2009). Absent such peer pressure, we might expect a relatively weak spillover effect of team efficiency, i.e. smaller λ, so that firms should be cautious when they are considering increasing the compensation for leaders' organisational activities to achieve a substantial effect like ours as this might lead to an adverse effect as shown in the model.

With all that being said, it provides a broad research agenda to learn about leadership and how multitasking leaders respond to multidimensional incentives.

References

Al-Ubaydli, O., Andersen, S., Gneezy, U., and List, J. A. (2015). Carrots that look like sticks: Toward an understanding of multitasking incentive schemes. *Southern Economic Journal*, 81(3):538–561.

Armin, F. and Andrea, I. (2006). Clean evidence on peer effects. *Journal of Labor Economics*, 24(1):39–57.

Bandiera, O., Barankay, I., and Rasul, I. (2005). Social preferences and the response to incentives: Evidence from personnel data*. *The Quarterly Journal of Economics*, 120(3):917–962.

Bandiera, O., Barankay, I., and Rasul, I. (2013). Team incentives: Evidence from a firm level experiment. *Journal of the European Economic Association*, 11(5):1079–1114.

Barankay, I. (2012). Rank incentives: Evidence from a randomized workplace experiment. *mimeo, Wharton School, University of Pennsylvania*.

Bartel, A. P., Cardiff-Hicks, B., and Shaw, K. (2017). Incentives for lawyers: Moving away from "eat what you kill". *ILR Review*, 70(2):336–358.

Besley, T. and Ghatak, M. (2008). Status incentives. *The American Economic Review*, 98(2):206–211.

Blanes i Vidal, J. and Nossol, M. (2011). Tournaments without prizes: Evidence from personnel records. *Management Science*, 57(10):1721–1736.

Bradler, C., Dur, R., Neckermann, S., and Non, A. (2016). Employee recognition and performance: A field experiment. *Management Science*, 62(11):3085–3099.

Cameron, A. C., 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.

Casas-Arce, P. and Martínez-Jerez, F. A. (2009). Relative performance compensation, contests, and dynamic incentives. *Management Science*, 55(8):1306–1320.

Delfgaauw, J., Dur, R., Non, A., and Verbeke, W. (2015). The effects of prize spread and noise in elimination tournaments: A natural field experiment. *Journal of Labor Economics*, 33(3):521–569.

Delfgaauw, J., Dur, R., Sol, J., and Verbeke, W. (2013). Tournament incentives in the field: Gender differences in the workplace. *Journal of Labor Economics*, 31(2):305–326.

Drews, F. A., Pasupathi, M., and Strayer, D. L. (2008). Passenger and cell phone conversations in simulated driving. *Journal of Experimental Psychology: Applied*, 14(4):392.

Ederer, F. (2010). Feedback and motivation in dynamic tournaments. *Journal of Economics and Management Strategy*, 19(3):733–769.

Englmaier, F., Roider, A., and Sunde, U. (2017). The role of communication of performance schemes: Evidence from a field experiment. *Management Science*, 63(12):4061–4080.

Erev, I., Bornstein, G., and Galili, R. (1993). Constructive intergroup competition as a solution to the free rider problem: A field experiment. *Journal of Experimental Social Psychology*, 29(6):463 – 478.

Eriksson, T., Poulsen, A., and Villeval, M. C. (2009). Feedback and incentives: Experimental evidence. *Labour Economics*, 16(6):679 – 688. European Association of Labour Economists 20th annual conference University of Amsterdam, Amsterdam, The Netherlands 18–20 September 2008.

Ertac, S. (2005). Social comparisons and optimal information revelation: Theory and experiments. *mimeo, University of California, Los Angeles.*

Eugene Kandel, E. P. L. (1992). Peer pressure and partnerships. *Journal of Political Economy*, 100(4):801–817.

Fliessbach, K., Weber, B., Trautner, P., Dohmen, T., Sunde, U., Elger, C. E., and Falk, A. (2007). Social comparison affects reward-related brain activity in the human ventral striatum. *Science*, 318(5854):1305–1308.

Fryer, R. G., Holden, R., et al. (2012). Multitasking, learning, and incentives: a cautionary tale. *NBER Working Paper*, 17752.

Griffith, R. and Neely, A. (2009). Performance pay and managerial experience in multitask teams: Evidence from within a firm. *Journal of Labor Economics*, 27(1):49–82.

Harrison, G. W. and List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4):1009–1055.

Holmstrom, B. and Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership and job design. *Journal of Law, Economics and Organization*, 7:24–52.

Hong, F., Hossain, T., List, J. A., and Tanaka, M. (2018). Testing the theory of multitasking: Evidence from a natural field experiment in chinese factories. *International Economic Review*, 59(2):511–536.

Hossain, T. and List, J. A. (2012). The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science*, 58(12):2151–2167.

Ichniowski, C. and Shaw, K. (2003). Beyond incentive pay: Insiders' estimates of the value of complementary human resource management practices. *Journal of Economic Perspectives*, 17(1):155–180.

Kahneman, D., Krueger, A. B., Schkade, D., Schwarz, N., and Stone, A. A. (2006). Would you be happier if you were richer? a focusing illusion. *science*, 312(5782):1908–1910.

Kishore, S., Rao, R. S., Narasimhan, O., and John, G. (2013). Bonuses versus commissions: A field study. *Journal of Marketing Research*, 50(3):317–333.

Kluger, A. N. and Denisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, pages 254–284.

Kosfeld, M. and Neckermann, S. (2011). Getting more work for nothing? symbolic awards and worker performance. *American Economic Journal: Microeconomics*, 3(3):86–99.

Kremer, M., Glewwe, P., and Ilias, N. (2010). Teacher incentives. *American Economic Journal: Applied Economics*, 2(3):205–227.

Layard, R. (2011). *Happiness: Lessons from a new science*. Penguin UK.

Lazear, E. P. and Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, 89(5):841–864.

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

Lim, N., Ahearne, M. J., and Ham, S. H. (2009). Designing sales contests: Does the prize structure matter? *Journal of Marketing Research*, 46(3):356–371.

Lizzeri, A., Meyer, M. A., and Persico, N. (2002). *The incentive effects of interim performance evaluations*. University of Pennsylvania, Center for Analytic Research in Economics and the Social Sciences.

Manthei, K., Sliwka, D., et al. (2018). Multitasking and subjective performance evaluations: Theory and evidence from a field experiment in a bank. Technical report, Institute for the Study of Labor (IZA).

Mark, G., Voida, S., and Cardello, A. (2012). A pace not dictated by electrons: an empirical study of work without email. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 555–564. ACM.

Mas, A. and Moretti, E. (2009). Peers at work. *American Economic Review*, 99(1):112–45.

Moldovanu, B., Sela, A., and Shi, X. (2007). Contests for status. *Journal of Political Economy*, 115(2):338–363.

Muralidharan, K. and Sundararaman, V. (2013). Contract teachers: Experimental evidence from india. Working Paper 19440, National Bureau of Economic Research.

Shearer, B. (2004). Piece rates, fixed wages and incentives: Evidence from a field experiment. *The Review of Economic Studies*, 71(2):513–534.

Slade, M. E. (1996). Multitask agency and contract choice: An empirical exploration. *International Economic Review*, 37(2):465–486.

A Appendix: Model Details

A.1 Assumptions on *w*

We impose the agent's limited liability constraint on the piece rate w in the firm's profit function $E\pi(.)$, therefore, it must be positive. We also impose that individual efforts cannot be negative.

First, we assume that w is large enough. By looking at equations 1, 2, and 3 we know that all effort levels are negative if $\lambda w > 2$. This means that no one exerts any effort and the total production output equals zero, so the firm's profit is negative ($-b$) which would not occur in practice. Similarly, if $\lambda w = 2$ output is infinite and profit is infinite. Hence, we assume $\lambda w < 2$ to rule out these cases.

Now, the inspection of λw in equation 2 reveals that g_f equals to zero if $\lambda w \leq 1$ because individual efforts cannot be negative. Both the foreman and the worker solve their maximisation problems considering $\overline{g}_f = 0$, and we have $\overline{e}_f = \overline{e}_w = w$.

Furthermore, if $1 < \lambda w < 2$, g_f^*, e_f^*, and e_w^* as expressed in equations 1, 2, and 3 are all positive. Thus, all three first order conditions hold, and we would predict that the foreman and the worker who are offered a piece rate such that $1 < \lambda w < 2$ would choose the interior levels of effort g_f^*, e_f^*, and e_w^*.

A.2 Proof of Lemma 1

We first analyse the case when $1 < \lambda w < 2$, substituting the optimal effort levels e_w^*, e_f^*, and g_f^* into the output functions the firm's optimisation problem can be written as below:

$$\begin{aligned}\max_{\{w\}} E\pi &= (p-w)(y_f(e_f^*, g_f^*) + y_w(e_w^*)) - b \\ &= (p-w)\frac{\lambda w^2(1+\lambda w)}{(1-(\lambda w - 1)^2)^2} - b.\end{aligned}$$

The first order condition with respect to w is therefore given by:

$$\lambda(4+\lambda w)(p-w) - (2 + \lambda w - \lambda^2 w^2) = 0 \tag{12}$$

solving this equation we get:

$$w^* = \frac{4p\lambda - 2}{5\lambda - p\lambda^2}. \tag{13}$$

Substituting firm's offer w^* into the foreman's and worker's optimal effort levels, we have:

$$g_f^* = \frac{5p\lambda - 7}{12\lambda - 6p\lambda^2},$$
$$e_f^* = \frac{5 - p\lambda}{12\lambda - 6p\lambda^2},$$
$$e_w^* = \frac{4p\lambda - 2}{12\lambda - 6p\lambda^2}.$$

Taking together, the firm's expected profit when $1 < \lambda w < 2$ can be written as:

$$\begin{aligned}
E\pi^* &= (p - w^*)\frac{\lambda(w^*)^2(1+\lambda w^*)}{(1-(\lambda w^* - 1)^2)^2} - b = (p - w^*)\frac{1+\lambda w^*}{\lambda(2-\lambda w^*)^2} - b \\
&= (p - \frac{4p\lambda - 2}{5\lambda - p\lambda^2})\frac{1 + \lambda\frac{4p\lambda-2}{5\lambda-p\lambda^2}}{\lambda(2-\lambda\frac{4p\lambda-2}{5\lambda-p\lambda^2})^2} - b \\
&= (p - \frac{4p\lambda - 2}{5\lambda - p\lambda^2})\frac{(p\lambda+1)(5-p\lambda)}{12\lambda(2-p\lambda)^2} - b \\
&= \frac{p(p\lambda+1)(5-p\lambda)}{12\lambda(2-p\lambda)^2} - \frac{(4p\lambda-2)(p\lambda+1)}{12\lambda^2(2-p\lambda)^2} - b \\
&= \frac{[p\lambda(5-p\lambda) - (4p\lambda-2)](p\lambda+1)}{12\lambda^2(2-p\lambda)^2} - b \\
&= \frac{(p\lambda+1)(2-p\lambda)(p\lambda+1)}{12\lambda^2(2-p\lambda)^2} - b = \frac{(p\lambda+1)^2}{12\lambda^2(2-p\lambda)} - b,
\end{aligned}$$

note that from equation 13 we know that the underlying assumption for π^* (i.e. $1 < \lambda w < 2$) only holds if $1.4 < p\lambda < 2$.

If $\lambda w \leq 1$, substituting $\overline{g}_f = 0, \overline{e}_f = \overline{e}_w = w$ into the firm's expected profit function, we have:

$$E\overline{\pi} = (p - w)2w - b.$$

Solving this maximisation problem for the firm, the firm would prefer to set its piece rate at $\frac{p}{2}$. The optimal effort levels of the foreman and the worker are therefore given by: $\overline{g}_f = 0, \overline{e}_f = \frac{p}{2}$, and $\overline{e}_w = \frac{p}{2}$. These results only exist if $\lambda\overline{w} = \lambda * \frac{p}{2} < 1 \Rightarrow p\lambda < 2$, and we can rewrite the firm's expected profit as:

$$E\overline{\pi} = (p - \overline{w})2\overline{w} - b = \frac{p^2}{2} - b.$$

To see the firm's choice between these two potential outcomes, given the fact that π^* only exists if $1.4 < p\lambda < 2$, taking the difference between $E\pi^*$ and $E\overline{\pi}$ gives us:

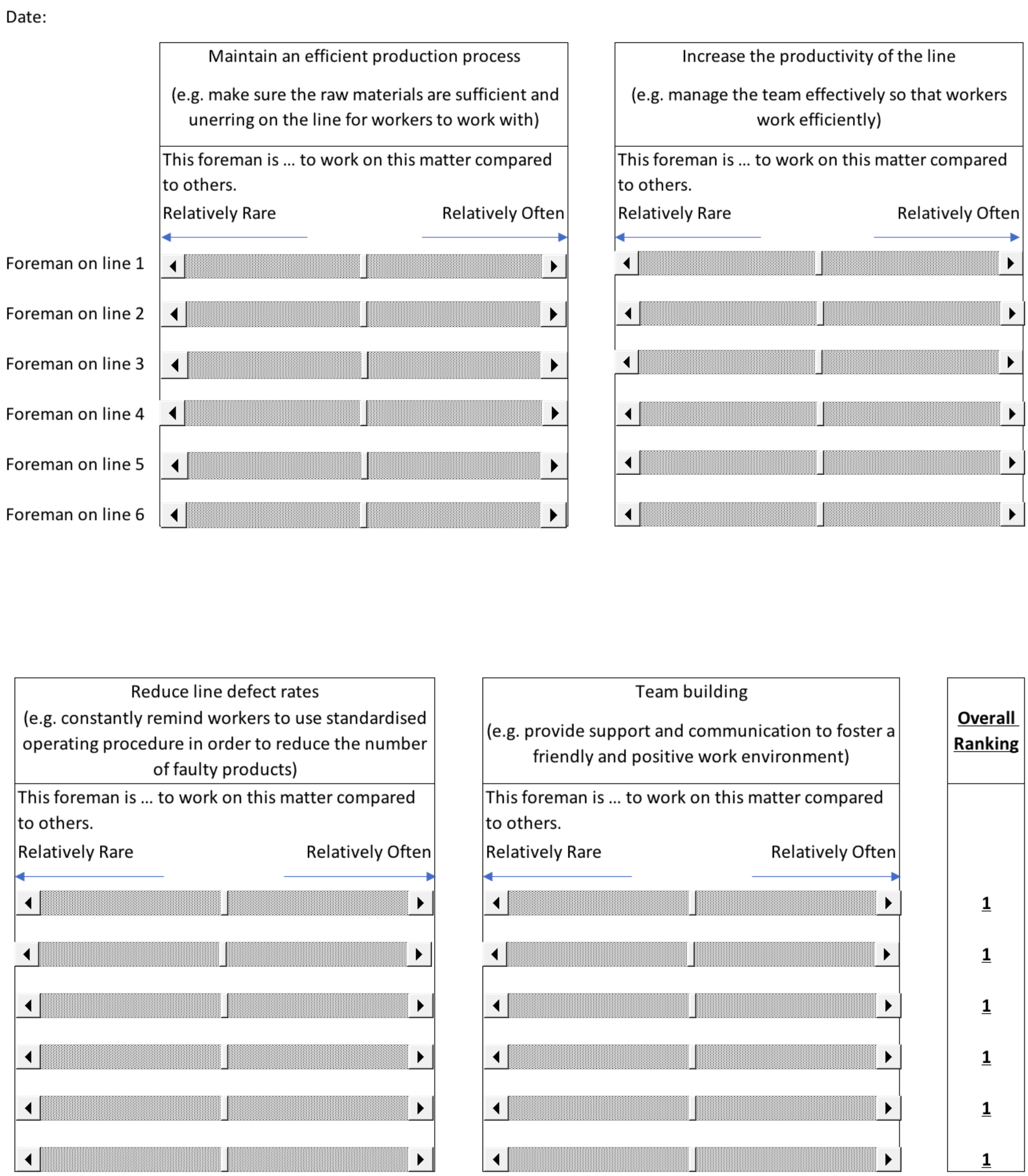
$$
\begin{aligned}
E\pi^* - E\overline{\pi} &= \frac{(p\lambda+1)^2}{12\lambda^2(2-p\lambda)} - b - (\frac{p^2}{2} - b) \\
&= \frac{(p\lambda+1)^2}{12\lambda^2(2-p\lambda)} - \frac{p^2}{2} \\
&= \frac{(p\lambda+1)^2 - 6p^2\lambda^2(2-p\lambda)}{12\lambda^2(2-p\lambda)} \\
&= \frac{p^2\lambda^2 + 2p\lambda + 1 - 12p^2\lambda^2 + 6p^3\lambda^3}{12\lambda^2(2-p\lambda)} \\
&= \frac{(2p\lambda-1)(3(p\lambda)^2 - 4p\lambda - 1)}{12\lambda^2(2-p\lambda)},
\end{aligned}
$$

the solutions of this function when it is equal to zero are $p\lambda = \frac{2-\sqrt{7}}{3} \approx -0.22, 0.5, and \frac{2+\sqrt{7}}{3} \approx 1.55$. $E\pi^* - E\overline{\pi} > 0$ if $\frac{2-\sqrt{7}}{3} < p\lambda < 0.5$ or $p\lambda > \frac{2+\sqrt{7}}{3}$. $E\pi^* - E\overline{\pi} < 0$ for $0.5 < p\lambda < \frac{2+\sqrt{7}}{3}$. Because both the market price and the team efficiency spillover are greater than zero, we have $E\pi^* > E\overline{\pi}$ if $0 < p\lambda < 0.5$ or $p\lambda > \frac{2+\sqrt{7}}{3}$, $E\pi^* < E\overline{\pi}$ if $0.5 < p\lambda < \frac{2+\sqrt{7}}{3}$, and $E\pi^* = E\overline{\pi}$ if $p\lambda = 0.5$ or $\frac{2+\sqrt{7}}{3}$.

As we already know that $\overline{\pi}$ exists if $p\lambda < 2$, and π^* is attainable if and only if $1.4 < p\lambda < 2$. Taking together, the profit maximising firm is indifferent between $\overline{w}$ and w^* if $p = p^* = \frac{2+\sqrt{7}}{3\lambda}$. It chooses $w = w^*$ if $p^* < p < \frac{2}{\lambda}$. Under w^*, the foreman and the worker both invest positive effort levels and the expected profit of the firm equals $\frac{(p\lambda+1)^2}{12\lambda^2(2-p\lambda)} - b$. On the other hand, if $0 < p < p^*$ the firm sets the piece rate at $\overline{w}$. The foreman's organisational effort is zero while both the foreman and the worker choose production effort equals to the piece rate $\overline{w}$. The firm receives $\frac{p^2}{2} - b$ as a return.

B Appendix: Other Figures

Figure B1: Sliders for Ranking the Foremen



C Appendix: Other Tables

Table C1: Summary Statistics for Other Individual Characteristics (Fengcheng)

	N	mean	sd	min	max
Female	43	1	0	1	1
Married	43	0.977	0.152	0	1
Live in the factory	43	0.279	0.454	0	1
Commute by factory bus	43	0.698	0.465	0	1
Commute by bike	43	0.047	0.213	0	1
Commute by motorbike	43	0.140	0.351	0	1
Number of years worked in the factory	43	2.930	2.005	0	7
Number of different types of products worked per day	43	1.919	0.288	1.630	2.439
Number of different products worked per day	43	2.284	0.374	1.917	3.030
Number of temporary coworkers from other lines	43	1.416	1.094	0	3.041
Education level:					
Illiterate	42	0.214	0.415	0	1
Primary school	42	0.405	0.497	0	1
Secondary school	42	0.333	0.477	0	1
High school	42	0.048	0.216	0	1

Table C2: Summary Statistics for Other Individual Characteristics (Fuzhou)

	N	mean	sd	min	max
Female	27	1	0	1	1
Married	27	1	0	1	1
Live in the factory	27	0	0	0	0
Commute by factory bus	24	0.375	0.495	0	1
Commute by bike	24	0.125	0.338	0	1
Commute by motorbike	24	0.500	0.511	0	1
Number of years worked in the factory	27	8.111	3.105	1	13
Number of different types of products worked per day	27	1.024	0.016	1.01	1.049
Number of different products worked per day	27	1.047	0.028	1.01	1.086
Number of temporary coworkers from other lines	27	0	0	0	0
Education level:					
Illiterate	27	0.037	0.192	0	1
Primary school	27	0.333	0.480	0	1
Secondary school	27	0.593	0.501	0	1
High school	27	0.037	0.192	0	1

Table C3: The Parallel Trend Assumption Test for Worker's Productivity

	DiD
Fengcheng × Weeks 1-2	-0.001
	(0.014)
Fengcheng × Weeks 3	0
	(0)
Fengcheng × Weeks 4-5	0.042**
	(0.011)
Fengcheng × Weeks 6-7	0.052**
	(0.014)
Fengcheng × Weeks 8-9	0.062**
	(0.014)
Fengcheng × Weeks 10-11	0.066**
	(0.016)
Fengcheng × Weeks 12-13	0.104**
	(0.018)
Fengcheng × Weeks 14-15	0.058**
	(0.018)
Observations	5,655
Clusters	57
R^2	0.782
Controls	YES

Notes: The unit of observation is worker i. The dependent variable is the log of worker's productivity. Productivity is a measure of the output per hour. Worker fixed effects, factory-line fixed effects, week fixed effects, the day of the week fixed effects (e.g. Monday), and an indicator variable for whether the worker is recorded sick or if there is an organisational error are included in all regressions. Robust standard errors clustered at the worker level are reported in brackets below the estimates. *** Significant at 1% level, ** significant at 5% level, * significant at 10% level.

Table C4: The Parallel Trend Assumption Test for Worker's Production Output

	DiD
Fengcheng × Weeks 1-2	-0.015
	(0.015)
Fengcheng × Weeks 3	0
	(0)
Fengcheng × Weeks 4-5	0.066**
	(0.015)
Fengcheng × Weeks 6-7	0.123**
	(0.020)
Fengcheng × Weeks 8-9	-0.006
	(0.018)
Fengcheng × Weeks 10-11	0.216**
	(0.020)
Fengcheng × Weeks 12-13	0.050*
	(0.023)
Fengcheng × Weeks 14-15	0.077**
	(0.021)
Observations	5,655
Clusters	57
R^2	0.538
Controls	YES

Notes: The unit of observation is worker i. The dependent variable is the log of worker's production output. Worker fixed effects, factory-line fixed effects, week fixed effects, the day of the week fixed effects (e.g. Monday), and an indicator variable for whether the worker is recorded sick or if there is an organisational error are included in all regressions. Robust standard errors clustered at the worker level are reported in brackets below the estimates. *** Significant at 1% level, ** significant at 5% level, * significant at 10% level.

Table C5: The Parallel Trend Assumption Test for Forewoman's Productivity

	DiD
Fengcheng × Weeks 1-2	0.001
	(0.024)
Fengcheng × Weeks 3	0
	(0)
Fengcheng × Weeks 4-5	0.023
	(0.018)
Fengcheng × Weeks 6-7	0.022
	(0.022)
Fengcheng × Weeks 8-9	0.041**
	(0.013)
Fengcheng × Weeks 10-11	0.035*
	(0.019)
Fengcheng × Weeks 12-13	0.072***
	(0.019)
Fengcheng × Weeks 14-15	0.021
	(0.020)
Observations	1,312
Clusters	13
R^2	0.846
Controls	YES

Notes: The unit of observation is forewoman i. The dependent variable is the log of forewoman's productivity. Productivity is a measure of the output per hour. Forewoman fixed effects, factory-line fixed effects, week fixed effects, the day of the week fixed effects (e.g. Monday), and an indicator variable for whether the forewoman is recorded sick or if there is an organisational error are included in all regressions. Robust standard errors clustered at the forewoman level are reported in brackets below the estimates. *** Significant at 1% level, ** significant at 5% level, * significant at 10% level.

Table C6: The Parallel Trend Assumption Test for Forewoman's Production Output

	DiD
Fengcheng $\times$ Weeks 1-2	-0.045
	(0.029)
Fengcheng $\times$ Weeks 3	0
	(0)
Fengcheng $\times$ Weeks 4-5	0.044
	(0.038)
Fengcheng $\times$ Weeks 6-7	0.121***
	(0.035)
Fengcheng $\times$ Weeks 8-9	-0.033
	(0.040)
Fengcheng $\times$ Weeks 10-11	0.195***
	(0.048)
Fengcheng $\times$ Weeks 12-13	-0.016
	(0.061)
Fengcheng $\times$ Weeks 14-15	0.051
	(0.037)
Observations	1,312
Clusters	13
R^2	0.361
Controls	YES

Notes: The unit of observation is forewoman i. The dependent variable is the log of forewoman's production output. Forewoman fixed effects, factory-line fixed effects, week fixed effects, the day of the week fixed effects (e.g. Monday), and an indicator variable for whether the forewoman is recorded sick or if there is an organisational error are included in all regressions. Robust standard errors clustered at the forewoman level are reported in brackets below the estimates. *** Significant at 1% level, ** significant at 5% level, * significant at 10% level.

Table C7: The Treatment Effect on Worker's Working Time

	Number of Minutes Worked in a Day			
	Jul-Sep	Jul	Aug	Sep
	(1)	(2)	(3)	(4)
Fengcheng	-71.597	57.206***	99.724***	109.510***
	(46.606)	(5.448)	(28.232)	(20.904)
Post	-36.565***	-35.215***	-47.421***	-24.140**
	(10.198)	(9.938)	(10.760)	(10.634)
Fengcheng*Post	11.333	35.951***	12.837	-2.043
	(9.391)	(10.299)	(10.060)	(11.370)
Observations	5,655	2,770	2,647	2,712
Clusters	57	57	57	57
R^2	0.422	0.577	0.411	0.507
Controls	YES	YES	YES	YES

Notes: The unit of observation is worker i. The dependent variables in Columns 1-4 are the working time (number of minutes) a worker worked in a day. Columns 1 shows the results for the full sample includes observations from June 7th until September 30th while Columns 2-4 compare the observations from the pre-treatment period (June) to each post-treatment month separately. Worker fixed effects, factory-line fixed effects, week fixed effects, the day of the week fixed effects (e.g. Monday), and an indicator variable for whether the worker is recorded sick or if there is an organisational error are included in all regressions. Robust standard errors clustered at the worker level are reported in brackets below the estimates. *** Significant at 1% level, ** significant at 5% level, * significant at 10% level.