Gall, Thomas; Hu, Xiaocheng; Vlassopoulos, Michael

Working Paper

Incentivizing Team Leaders: A Firm-Level Experiment on Subjective Performance Evaluation of Leadership Skills

IZA Discussion Papers, No. 16123

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Gall, Thomas; Hu, Xiaocheng; Vlassopoulos, Michael (2023) : Incentivizing Team Leaders: A Firm-Level Experiment on Subjective Performance Evaluation of Leadership Skills, IZA Discussion Papers, No. 16123, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/272750

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16123

Incentivizing Team Leaders: A Firm-Level Experiment on Subjective Performance Evaluation of Leadership Skills

Thomas Gall
Xiaocheng Hu
Michael Vlassopoulos

MAY 2023

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16123

Incentivizing Team Leaders: A Firm-Level Experiment on Subjective Performance Evaluation of Leadership Skills

Thomas Gall
University of Southampton

Xiaocheng Hu
University of Exeter

Michael Vlassopoulos
University of Southampton and IZA

MAY 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Incentivizing Team Leaders: A Firm-Level Experiment on Subjective Performance Evaluation of Leadership Skills*

In teamwork settings, providing effective leadership can be challenging for team leaders due to multitasking and the difficulty in measuring and rewarding leadership input. These challenges might lead to underprovision of leadership activities, which can ultimately impede the productivity of the team. To address this problem, we conduct a field experiment at a manufacturing firm, introducing a relative subjective performance evaluation of team leaders' leadership activities by their managers, coupled with bonuses based on their leadership rank among all leaders. Our intervention increased worker productivity by approximately 7%, while leaving team leaders' productivity unchanged, and was profitable for the firm. During the intervention, we observe a positive correlation between the evaluations of team leaders and the productivity of team members, suggesting that the subjective evaluation indeed increased leadership activities and thus productivity.

JEL Classification: J24, J33, M52, C93

Keywords: multitasking, subjective evaluation, teamwork, incentive schemes, productivity, leadership

Corresponding author:
Michael Vlassopoulos
Department of Economics
University of Southampton
Southampton, SO17 1BJ
United Kingdom
E-mail: m.vlassopoulos@soton.ac.uk

* We greatly benefited from comments and suggestions received at Southampton, East Anglia Economics PhD Workshop, 13th Nordic Conference in Behavioral and Experimental Economics, 2018 Advances with Field Experiments Conference, 2019 Colloquium on Personnel Economics, RES 2019 Annual Conference, 2019 North American ESA meeting, 2020 Cedefop-Eurofound-IZA virtual conference on workplace and management practices. Financial assistance from ESRC is gratefully acknowledged. All remaining errors are our own.

1 Introduction

In today's economy, team production is widespread in various sectors, such as manufacturing, healthcare, finance, and technology. As a result, organizations recognize the importance of having effective management practices in place to foster efficient team production. In practice, teams are typically managed either by a manager or supervisor who is external to the team and specialises in management and leadership activities, or by a team leader who is a member of the team and provides both leadership and contributes to the team's output. Indeed, both team leaders and supervisors have been shown to have significant impacts on the productivity of their supervisees as well as on firms' profits and productivity (e.g., Lazear et al., 2015; Adhvaryu et al., 2022; Englmaier et al., 2021).

However, the role of a team leader is subject to the familiar problem of multitasking. Team leaders are expected to allocate their resources between two different and possibly competing tasks: leadership activities that benefit all team members, for instance, through raising productivity and output, and direct output generation that increases their own productivity and output. The first task amounts essentially to exerting a positive externality on the other team members, while the second one generates mainly private benefits. Since output tends to be comparatively easy to observe, especially in manufacturing settings involving low skilled work, while leadership activities are less easily measured, standard economic reasoning would suggest a potential underprovision of leadership, adversely affecting team productivity and output (see e.g., Holmstrom and Milgrom, 1991; Drago and Garvey, 1998). The challenge of adequately incentivising team leaders to perform multiple tasks of differing measurability is not limited to unskilled work on the factory floor. For instance, in a study of US law firms, Bartel et al. (2017) find that when senior partners were rewarded based on a combination of output-based and subjective performance evaluation, they reduced their billable hours and increased non-billable activities to the benefit of other team members.

In this paper, we study a management intervention aimed at incentivizing team leaders to invest more in leadership activities. We conduct a natural field experiment in collaboration with a company manufacturing medical equipment in Southeast China. The setting is one where teams of low skilled workers perform packaging tasks in a line production. Teams are led by a team leader who is expected to both produce output and to manage the team. Prior to our intervention, the company was operating piece rate payment schemes for all workers, including team leaders, who also received a fixed bonus. However, the company's management felt that team leaders did not optimally fill their role of managing teams and were open to trying a new incentive design. Our intervention involves the introduction of a relative subjective performance evaluation of team leaders' leadership activities by their managers and paying leaders monthly bonuses depending on their

ranking in leadership performance among all leaders in the plant. This type of intervention reflects received economic wisdom and long-standing practice (see e.g., Baker et al., 1994) and would theoretically allow implementation of the first best effort allocation in a simple formal model of teamwork in the spirit of Itoh (1991), where leaders can support other team members.[1]

The company operated two manufacturing plants in different towns, producing the same products using the same technology. This offered us the opportunity to implement a research design where teams in one plant served as the treatment group and those in the other plant as the control group, an approach that has been previously employed in various studies (see e.g., Griffith and Neely, 2009; Song et al., 2018; Krueger and Friebel, 2022). To minimize individual biases, leadership activities were assessed across four dimensions and aggregated into a leadership score, with multiple managers acting as evaluators.[2] Furthermore, to reduce the impact of common shocks, such as weather conditions, on the performance of team leaders, we adopted relative scores to determine bonuses. Leadership score rankings were publicly displayed on the factory floor.

Following the implementation of our intervention, we observe that the productivity of workers on the team, measured as output per hour, increased in the treatment plant compared to the control plant, relative to productivity measured before the intervention. A difference-in-difference regression approach with an extensive set of fixed effects confirms this observation: the treatment effect on worker productivity is about 6-7%, and statistically significant. However, the treatment effect for team leaders' productivity is lower and not statistically significant. There is also little evidence for heterogeneous effects across teams in terms of initial productivity. These results are in line with the predictions of a simple model in which a team leader faces the trade-off of allocating effort between two tasks, producing own output and raising team productivity. The relative performance evaluation of team leaders was also economically significant in that the company's profit increased. In fact, the company expressed a strong desire to introduce our intervention also in the control plant, which they did after three months, thereby limiting our observation period. We also find an increase in actual work time, particularly during the first month of the intervention, which is more pronounced for team leaders. This is consistent with team leaders redirecting their efforts from production towards investing in the organisation of their team immediately after the start of the intervention.

To assess whether the positive effect of the intervention on worker productivity is due

[1] See Dewatripont et al. (2000) for a survey of influential models and theoretical findings for multitask agency problems.

[2] See, e.g., Bol (2011) for evidence on evaluators' biases. de Janvry et al. (2023) find that using multiple evaluators mitigates the problem. Demeré et al. (2019) report that the use of calibration committees also helps in mitigating biases.

to team leaders increasing their leadership activities, we examine whether team leaders' leadership scores correlate with the average productivity of workers in their team. We find a significant positive correlation, but only in the treatment plant during the intervention period. This is consistent with two main, not mutually exclusive interpretations: first, for subjective performance evaluation to work evaluators need to be adequately incentivised, which can take the form of making public their evaluation scores, inviting public scrutiny. The second interpretation is information transmission to team leaders who were shown the different dimensions on which their leadership activities were assessed and thus learned, possibly for the first time, the exact objectives they were meant to achieve.[3]

Anecdotal evidence from post-intervention interviews indicates that at least some of the team leaders started to seriously engage in managing their team and came up with innovations in managing their teams. For instance, one team leader designed their own incentive scheme for their team, paid out of their salary, while another one introduced team-building practices. In this sense, the intervention appears to have achieved a productivity increase by inducing appointed team leaders to become true leaders and actively manage their teams.

To the best of our knowledge, this paper is the first to provide field experimental evidence on combining objective and subjective performance measures to induce team leaders to raise the productivity of their team, that is, to engage in a profit-relevant aspect of leadership. We contribute to the literature assessing incentive schemes in multitasking environments. Most closely related is the work of Bartel et al. (2017) who study partners in law firms and their trade-off between allocating their time toward billable hours, attributable to work for a client, and non-billable hours, which includes acquisition of new clients and similar activities. Exploiting a change in the law firm's reward policy for all team leaders, explicitly incentivising non-billable activities, they find a shift toward these activities after the introduction of the new reward scheme. By contrast, our intervention is evaluated through a natural field experiment.

A number of experimental studies focus on a possible quantity-quality trade-off, showing mixed results. 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) affect the performance on the other dimension (quality). Kishore et al. (2013) report modest multitasking concerns when workers reached their targets and they are paid bonus-based incentive schemes. 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

[3]This relates to a literature studying effects of combining performance evaluation with information transmission and learning, (see e.g. Manthei et al., 2023; Song et al., 2018).

flat wage rate do not. Jones et al. (2018) find that the introduction of pay-for-performance on the quantity dimension has more muted effects when the quality dimension has a prosocial element. This paper focuses on a different multitasking trade-off, that between production and providing leadership.

Our study is also connected to a literature on subjective performance evaluation, often pointing out possible pitfalls, such as evaluators' biases (see e.g. Bol, 2011; Bol and Smith, 2011; Rosaz and Villeval, 2012; Manthei and Sliwka, 2019). We aggregate scores both across different dimensions and across multiple evaluators to mitigate these concerns, and indeed performance evaluations do correlate strongly with team productivity in our data set.

Finally, our paper adds to the evidence from field experiments on how tournament and rank incentives affect performance in organizations (Casas-Arce and Martínez-Jerez, 2009; Kosfeld and Neckermann, 2011; Bandiera et al., 2013; Delfgaauw et al., 2013, 2015; Hong et al., 2015; Boudreau et al., 2016; List et al., 2020; Englmaier et al., 2023). While the previous evidence shows that designing a tournament to raise the performance of individuals or teams can be effective, we show that a tournament aiming to incentivize difficult to measure activities of team leaders can also enhance the performance of teams.

The remainder of the paper proceeds as follows. The next section provides some background information. Section 3 lays out a simple theoretical model to illustrate our expectations for the possible effects of the intervention. Section 4 describes our intervention and Section 5 our empirical approach. We present our results in Section 6 and our conclusions in Section 7. All tables and figures not in the text can be found in the appendix.

2 Background

To carry out our field experiment we partnered with a company manufacturing medical devices located in the province of Jiangxi, China. This company had two factories that produced the same product line using the same production technology, but operated under different brand names. The factories operated as independent entities, each with their own management, and had limited interactions with each other, except at the top-level management. The driving distance between the two factories situated in two different cities is more than 70 miles. For clarity, we will refer to these factories as the control and treatment plants, respectively.

During our field experiment, employees in both factories were tasked with packaging disposable infusion sets.[4] Workers were organised in teams, called "production lines", each comprising about five workers and a team leader appointed by the factory's manage-

[4]The disposable infusion set was a major source of revenue for this company, accounting for approximately 50% of its total revenue in 2016.

ment. Although packaging itself did not require any specific skills or teamwork, the team leader was responsible for the team's work environment, including the flow of inputs, parts and outputs, for monitoring workers' performance, organising and distributing materials, and assisting the factory management on production matters. Team leaders were also responsible for packaging products, like other members in the team. According to the company's management, team leaders were internally promoted only, and a successful candidate would demonstrate loyalty to the company, reliability, and modest leadership abilities.

Both factories employed a multiple piece-rate payment scheme, where producing more output resulted in a higher rate. Workers could earn additional bonuses on top of the piece rate each month, as summarized in Table 1. This salary structure remained constant throughout our experiment. A team leader received an additional monthly payment, independent of output, to recognise their role, which amounted to 2-3% of their average monthly income.

Table 1: Summary of Payment Structure, by Factory

Daily Average Output	Piece Rate (per unit)	Performance Bonus	Attendance Bonus	Tenure Bonus	Lunch Subsidy	Team Leader Bonus
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Treatment Plant						
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. Control Plant						
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

Notes: Daily average output is measured in physical units, while payments are in Chinese yuan (RMB).

3 Theoretical Model

The environment outlined above is best described by a technology where workers produce output based on individual effort, but where individual productivities may be affected by a team leader. The team leader allocates time and effort between two tasks, producing

own output and raising team productivity. This is a multitask problem, in which one task yields private benefits, and its output is easy to measure, and the other one has a positive externality on the team, for which inputs and output are relatively hard to measure.

Standard economic theory would posit that piece rates on output will lead to inefficiently low provision of the positive externality. Adding a reward system that increases in provision of the externality may thus improve efficiency, raise output and possibly also profits. Our intervention involved rewarding team leaders for inputs to organising production, as evaluated by the factory management. Hence, we would expect an increase of worker productivity, total output and possibly profits, but not necessarily an increase of a team leader's output as productive effort may be diverted into providing more of the externality. To give this reasoning some formal underpinnings consider the following multitask agency model in the spirit of Itoh (1991).

A Multitask Team Problem

Suppose individuals in a team produce output by exerting individual labor effort. One individual can choose to exert effort on two different tasks, one increasing only own output, and the other one having a positive externality on the team members' productivity of effort. To simplify matters consider a team of two individuals, A, the agent, and B, the team leader. The general reasoning is robust to adding team members and results generalise without qualification when adding identical agents.

Individual A exerts effort e at utility cost $e^2/2$ to produce output $y^A = (1 + \alpha z)e$, where z is a variable representing organisational capital, increasing A's productivity of effort. Individual B chooses both organisational effort z and individual productive effort x, at a combined cost of $x^2/2 + z^2/2 + \kappa(x + z)$, where κ captures the rivalry of the two tasks. B's output is $y^B = (1 + \beta z)x$. That is, B's organisational effort z affects both own productivity, at rate β, and the other team member's productivity, at rate α.

In our analysis, we assume the parameters to obey the following assumption:

Assumption 1 *$\alpha > \kappa \geq \beta$ and $\alpha^2 + \kappa^2 < 1$.*

The main reason for this assumption is to guarantee an interior first best solution (i.e. both x and z are strictly positive). Its first part implies that organisational effort increases A's productivity by more than the marginal cost of B's effort κ, but that the effect on B's own productivity is relatively limited. The second part implies that the positive externality of effort z on A, net of the increase of the marginal cost of B's effort, is small enough for a strictly positive optimal production effort x (otherwise B would specialise in z in the first best, i.e., become a manager, not a team leader).

This well describes a situation where the team leader can sacrifice some of their own production effort to make the worker more productive, e.g., by monitoring, sharing information or optimally designing workplace practices, but where the externality is sufficiently weak so that full specialisation of the team leader into organisational effort is not efficient (i.e. a team leader is a productive member of the team rather than a specialised manager).

First Best

First best efforts will solve:

$$\max_{e,x,z} p(1+\alpha z)e + p(1+\beta z)x - e^2/2 - (x^2+z^2)/2 - \kappa xz.$$

Assuming $p = 1$, in optimum:

$$e = 1+\alpha z \text{ and } x = 1-(\kappa-\beta)z \text{ and } z = \alpha e - (\kappa-\beta)x.$$

That is, $z^* = \frac{\alpha+\beta-\kappa}{1-\alpha^2-(\kappa-\beta)^2}$. Our assumptions guarantee an interior solution with both $z^* > 0$ and $x^* > 0$ – otherwise either could hit the zero bound. This implies $e^* = (1+\alpha)z^* = \frac{(1+\alpha)(\alpha+\beta-\kappa)}{1-\alpha^2-(\kappa-\beta)^2}$ and $x^* = 1-(\kappa-\beta)z = \frac{1-\alpha^2-\alpha(\kappa-\beta)}{1-\alpha^2-(\kappa-\beta)^2}$. Hence, the higher α the higher z^* and the lower x^*.

Piece Rate Contracts

Labor contracts with piece rates w_A and w_B will mean that individuals solve

$$\max_{e} w_A(1+\alpha z)e - e^2/2 \text{ and}$$
$$\max_{x,z} w_B(1+\beta z)x - (x^2+z^2)/2 - \kappa xz.$$

Therefore, optimally

$$e = w_A(1+\alpha z) \text{ and } x = w_B - (\kappa - w_B\beta)z \text{ and } z = -(\kappa - w_B\beta)x.$$

That is, if $w_B \leq p = 1$, x and z are substitutes, the (implicit) non-negativity constraint on effort z binds and optimal efforts are $z^w = 0$, $x^w = w_B$ and $e^w = w_A$. Note that to induce $z^w > 0$ requires $w_B > \kappa/\beta$. Note also that it is impossible to implement first best efforts x^* and z^* using w_B alone (since $-(\kappa - w_B\beta)x = \alpha e - (\kappa - w_B\beta)$ only has a solution for $e = 0$).

Assuming $w_B \leq \kappa/\beta$ the principal chooses w_A and w_B to maximise $\pi = (1-w_A)w_A + (1-w_B)w_B$, yielding $w_A = 1/2 = w_B$, which implies $z^w = 0$ is indeed optimal for B.[5]

The following proposition summarises these observations.

Proposition 1 *Using piece rates w_A and w_B (i) optimal effort z^w falls short of first best effort z^*, indeed $z^w = 0$ for $\kappa \geq \beta$, and (ii) first best efforts e^*, x^* and z^* cannot be implemented.*

For given w_A and w_B, the principal's profit can be written as

$$\pi = (1-w_A)w_A(1+\alpha z)^2 + (1-w_B)(1+\beta z)(w_B - (\kappa - w_B\beta)z).$$

This function increases for $w_A = w_B = 1/2$ in z at $z = 0$, so that given the profit maximising piece rates w_A and w_B the principal's profit increases in z, ignoring the cost of inducing positive z for the moment.

Differentiating output y^A and y^B given optimal effort choices with respect to z yields $\frac{\partial y^A}{\partial z} \geq w_A > 0$ and $\frac{\partial y^B}{\partial z} < 2w_B\beta - \kappa$, which is zero or strictly negative for $w_B = 1/2$. That is, increasing z will increase A's effort at the expense of B's productive effort x. Total effort of B, $x + z = w_B + (1 - (\kappa - w_B\beta))z$ increases in z.

Incentivising Organisational Effort

Let now the principal have use of an informative signal s of z, and offer the agent a payment dependent on the signal. Suppose that the agent is risk-neutral, again for the sake of tractability. Denote the expected payment as a function of z by $R(z)$, and assume that

Assumption 2 *$R(z)$ strictly increases in z.*

This is consistent e.g., with a signal such that $E[s|z] > E[s|z'] \Leftrightarrow z > z'$ (i.e. higher z induces a move to a first order stochastic dominant posterior distribution) and payment that strictly increases in the signal. Assuming the expected payment $R(z)$ is differentiable, B's optimal effort choices are now

$$x = w_B - (\kappa - w_B\beta)z \text{ and } z = R'(z) - (\kappa - w_B\beta)x.$$

Note here that setting $w_B = 1$ and $R'(z) = \alpha(1+\alpha z)$, or $R(z) = \alpha z + \alpha^2 z^2/2 - F$, where F is a constant, will implement the first best efforts x^* and z^* for individual B.

The following proposition summarises these observations.

[5] It is, however, possible to set $w_B > \kappa/\beta$, inducing strictly positive z^w, and making x and z complements. This potentially could achieve higher profit than does $w_A = w_B = 1/2$. Under our assumption $\kappa \geq \beta$, the principal's profit is $\pi = 1$ for $w_B = \kappa/\beta$, which is smaller than for $w_B = 1/2$, and decreases in w_B for $w_B \geq \kappa/\beta$. For $\beta > \kappa$ this may no longer be true and profit is maximised for strictly positive z, although the optimal z is smaller than in the first best.

Proposition 2 *Using an reward scheme $R(z)$ (i) first best efforts e^*, x^* and z^* can be implemented, (ii) the principal's profit can increase compared to the profit maximising piece rates, and (iii) if $R'(z) > 0$ optimal effort z strictly increases, B's total effort $x + z$ strictly increases, output y_A increases, output y_B decreases.*

For (ii), note that e.g. setting $R(z) = w_z z$ will imply that the principal's profit increases in z at $z = 0$ for $w_B = 1/2$ and $w_A = 1/2$.

Predictions

Interpreting y_A and a_B as output per hour, and effort levels e, x and z as unobservable intensities, then the model implies the following predictions.

Prediction 1 *Given and maintaining piece rates w^A and w^B, adding a reward scheme $R(z)$ with $R'(z) > 0$ will*

1. *strictly increase individual A's output per hour y^A,*
2. *weakly decrease B's output per hour,*
3. *increase total output per hour if β, $R'(z)$, or piece rates w_A and w_B are sufficiently high,*
4. *increase the principal's profit if $R(z)$ is adequately chosen.*

Our empirical analysis will therefore focus on assessing possible effects on measures of labor productivity, given by output per hour, and profitability. The model remains silent on the extensive margin, but predicts that output per worker and total output (if the reward scheme is convex enough) will increase, but output per team leader will decrease, if hours worked remain constant.

4 The Intervention

Our field experiment took place over a period of 4 months between June and September 2017. During this period, both workers and team leaders performed their tasks individually within their natural work environment, without being aware that an experiment was taking place. The intervention was introduced to them by the production manager through the usual internal communication channels in the treatment plant. We selected the treatment plant as the intervention factory, because the factory manager in the control plant unexpectedly resigned for personal reasons in early June. This made the control

plant an ideal choice as the control setting, as during a transition of management typically no changes of management practices are implemented and thus management practices and working conditions in the control plant would very likely remain constant for the duration of our experiment. Indeed, the Board of the company owning both factories agreed to introduce the intervention in the treatment plant, while holding policies and practices in the control plant constant until the end of September 2017.

4.1 Subjective Evaluation

To determine the criteria for subjective evaluation we asked the factory management to list all organisational activities they expected team leaders to perform. The management team agreed on the following four evaluation criteria:

- Maintain an efficient production process (e.g., by ensuring that raw materials are sufficient and appropriately distributed in the workplace).
- Increase the productivity of the workers (e.g., by managing the team effectively, such as motivating workers to focus on their work).
- Reduce the rate of defective outputs in the team (e.g., by reminding workers to use appropriate standardised operating procedures).
- Team building (e.g., by providing support and communication to foster a friendly and positive work environment).

In each factory multiple managers were asked to perform the evaluation task (two managers in the treatment plant and three in the control plant), to prevent any single manager's personal perceptions and biases to influence the evaluation results and to increase acceptance of the practice by team leaders. This was consistent with existing management practice in the factories, such as the 5S workplace organisation system which was assessed by five managers.[6] Moreover, employing multiple evaluators increases the cost of collusion for the evaluated.

To minimize the time required by managers to perform the evaluation, we designed a spreadsheet for the evaluators to use (see Figure A3 in the Appendix for an example). We used sliders for input instead of numerical values and emphasized that evaluations were meant to be relative and to allow ranking team leaders to reduce the probability of ties. After positioning the sliders under each criterion, the overall ranking of each team leader is automatically calculated and displayed. The evaluators then had the opportunity

[6] 5S is a workplace organisation system designed to improve manufacturing efficiency. For details, see https://en.wikipedia.org/wiki/5S_(methodology).

to verify whether the overall ranking of team leaders on the spreadsheet corresponded to their intention and could alter the sliders if necessary.

In each week of the intervention, the rankings for each criterion and the overall ranking were posted on the factory floor in the form of a scoreboard and displayed in descending order. We asked the management to display the scoreboard on the wall next to the production lines, as shown in the Appendix. The scoreboard only provided information for the most recent week. At the end of each month, a printout of the four weekly rankings and the aggregated rankings of that month was posted next to the scoreboard.

4.2 Reward Scheme

Team leaders were paid a bonus based on their rank in the management's subjective performance evaluation, as shown in Table 2.

Table 2: Reward Scheme

	Original Bonus (RMB/M) (1)	Intervention Reward (RMB/M) (2)	Difference to next lower (3)	Increase in Bonus (%) (4)
Ranked first	90	205	45	228%
Ranked second	90	160	25	178%
Ranked third	90	135	15	150%
Ranked fourth	90	120	10	133%
Ranked fifth	90	110	10	122%
Ranked sixth	90	100	10	111%
Ranked seventh	90	90		100%

Notes: RMB/M denotes Chinese yuan per month.

We chose a convex payment scheme to increase marginal incentives. In case of a tie, all tied team leaders are paid the same bonus, according to their rank.

To determine the payment for the highest ranked team leader, we computed the opportunity cost of spending one hour per day on organising teams (instead of producing output) for 28 working days as 208 RMB for the most productive workers.[7] The lowest ranked team leader is paid 90 RMB per month, which is the same as the bonus they would have earned without the intervention. This is because the management felt that prior

[7]Generally, workers in the treatment plant work 11 hours per day and 28 days per month. To obtain the highest piece rate of .024, as shown in Table 1, a worker has to produce at least 3,400 units every day, i.e. 310 units per hour, given an 11 hours workday. Hence, one hour of packing products yields 7.44 RMB.

experience strongly indicated that all team leaders would need to be rewarded for reasons of fairness and team cohesion. Incidentally, the least productive team leader had an opportunity cost of 94 RMB of spending one hour per day in organising her team.[8]

The management reserved the right to remove a team leader from the reward scheme, if they concluded that the team leader had exerted zero effort toward any of the four assessed criteria. In actual fact no team leader was removed from the payment scheme. This was corroborated by interviews with the team leaders after the experiment, which indicated that time spent on the task of organising their team was perceived to be less onerous than time spent on packaging products.

4.3 Timeline

The timeline of the field experiment is shown in Figure 1. Starting from 7^{th} June 2017, individual daily production records were collected and monitored by our research team.[9] During the first experimental week (W1), production managers from both factories were trained to use the evaluation system designed to subjectively evaluate team leaders (see subsection 4.1). On the last day of the second week (W2), production managers in both factories evaluated team leaders' organisational activities during W2, but neither the workers nor the team leaders were aware of this evaluation. This evaluation was repeated in each of the remaining weeks.

In the control plant, neither the evaluation procedure and criteria nor the results were made public during the duration of the experiment. In the treatment plant, however, both workers and team leaders were informed about the evaluation procedure, criteria, and the results once the intervention had started (from week 4 to 15). The first ranking results, for week 4, were posted on the factory floor in the treatment plant at the end of that week.

4.4 Communication

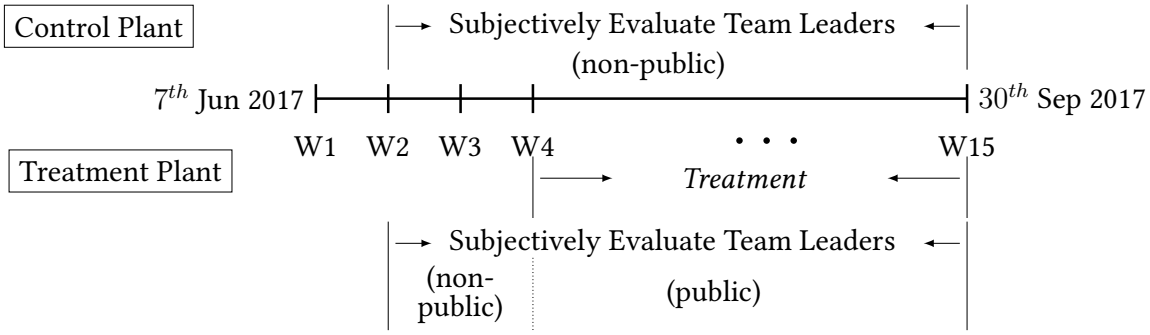
On the last day of week 3 of the experiment (30^{th} June 2017), the production manager in the treatment plant had a regular monthly meeting with all workers and team leaders from the packaging unit. The manager discussed current production issues and outlined plans for the upcoming months, including our treatment. We instructed the manager to announce our treatment as follows:

> The production managers will subjectively evaluate team leaders' organisational activities each week. The evaluation starts on 1^{st} July 2017. Four cri-

[8] The least productive team leader produced $1,900$ units per day. The corresponding piece rate for $1,900$ daily output is $.0195$, so that $1900 \div 11 \times .0195 \times 28 = 94$.

[9] Before our intervention, the factories were collecting output data, but not hours worked.

Figure 1: Timeline



Notes: W denotes the experimental week.

> teria 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 monetary reward in cash based on her monthly ranking. A higher ranking yields a higher payment. The monthly ranking is then reset at the beginning of next month.

After the announcement, a detailed instruction was handed to each team leader. It illustrated the four evaluation criteria with brief examples, detailed the incentive scheme, and outlined other organization-related information. The information explicitly stated that the new reward scheme was independent of the existing constant team leader bonus. Hence, team leaders would not consider it as a replacement for the existing bonus.

5 Empirical Approach

5.1 Data

Our sample includes 70 regular employees (all females, 27 from the control plant and 43 from the treatment plant respectively).[10] A team (production line) consists of several workers and one team leader. Seven lines (with team sizes of four on average) operate regularly in the control plant, while six lines (with team size of seven on average) operate regularly in the treatment plant.

[10]We excluded newly (after 1^{st} March 2017) hired workers because their compensation schemes are different, a few workers who were on holidays in June and thus lack baseline observations and some workers who quit during our experiment, but to the best of our knowledge not because of it.

The factories recorded and shared data on employees' daily hours and output, as well as the weekly subjective performance evaluations. Team leaders recorded daily data for every member of their team, including daily output, the time work started and the time the worker left the factory. These reports are checked by the factory production manager with little measurement error and used to compute payments to workers. The types of goods packed slightly differ between factories: while production in the control plant mainly focused on the local market, the treatment plant produced goods also for export. Products sold in the domestic market are easier to pack than those sold in the international market. However, the management of the treatment plant developed a method to calculate standardized piece rates for different types of products accounting for the level of difficulty. We used their method to obtain individual output that is comparable across factories.

In addition, we obtained administrative data from the human resource department, which included individual demographic characteristics such as gender, ethnicity, education, marriage status, age, residential area, mode of transportation to work, and recruitment channel (e.g., introduced by a current employee, job market advertisement, etc.).

5.2 Descriptives

Employees in our sample predominantly come from local, farming backgrounds.[11] The manufacturing task involves product packaging, which requires little training or human capital. The salary scheme for this task is identical to those for other tasks within the same production unit such as assembling, leak testing, or pressure testing.

All existing team leaders had worked in the company for more than two years and had established a good rapport with the production managers over the years. According to the factory managers, they had accepted the team leader appointment mainly because they ran out of excuses to reject it again. Qualitative evidence from interviewing the workers and team leaders reveals that the foreman position is not desirable because it requires more effort, sidetracks them from the primary task, and the corresponding compensation is relatively low.

Table 1, which shows the pre-intervention payment schemes, indicates differences between the two factories. The treatment plant offers higher piece rates than does the control plant. A fast-packaging worker, who can make more than 3,500 units averagely in a day, earns 0.0044 RMB more per unit in the treatment plant than in the control plant (the daily average output is computed by dividing total production output in a month by the number of days worked during that month). This yields a difference of 430 RMB ($\approx$

[11] Summary statistics for employees' characteristics, collected after the experiment, are reported in Tables B1 and B2 in the Appendix, respectively.

65 dollars) in 28 working days. Indeed, both workers and team leaders in the treatment plant earn 20 percent more than those in the control plant. This difference mainly reflects differences in local labour markets; factory management decided independently on the wages they pay. Individual workers did not know the pay in the other factory and there were no transfers of production workers between the two factories.

Table B3 presents summary statistics for each factory during the pre-treatment period (June) and the post-treatment period (July, August, and September), including the number of employees, number of production lines (which is also the number of team leaders, as there is only one team leader assigned to each line), worker's daily output, worker's productivity (output per hour), team leader's daily output, and team leader's productivity.

5.3 Estimation Strategy

To test whether the introduction of the incentive scheme intended to foster team leaders' organisational efforts indeed affected outcomes, we estimate the following Difference-in-Difference (DiD) specification:

$$log(Y)_{i,f,t} = \beta T_{f,t} + \theta_i + w_t + d_t + \epsilon_{i,f,t}, \tag{1}$$

where $log(Y)_{i,f,t}$ is the logarithm of an outcome for individual i in factory f on day t. Our main outcomes of interest are output, hours and productivity (output per hour worked).[12]

$T_{f,t}$ is an indicator variable that takes the value 1 for individuals in the treatment factory during the intervention period (i.e. 1^{st} July or after) and 0 otherwise. β is our main coefficient of interest.

We include individual fixed effects (θ_i) to account for unobserved and time-invariant heterogeneity in productivity among individuals. We also include week fixed effects (w_t) and day of the week fixed effects (d_t) to capture seasonal variation and shocks to production. We also include an indicator variable capturing whether individual i was assigned to work in another production line and an indicator variable for whether individual i was recorded sick or if there was an organisational error.

Heteroskedasticity-robust standard errors clustered at the individual level are used in all regression specifications.[13]

[12]Hours worked are computed as the difference between the time when an individual started 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.

[13]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.

Identification

The most critical assumption of our approach is that workers in the treatment and control factories have parallel trends pre-treatment in the outcomes of interest. This seems very plausible, as both factories operate under the same management board of a larger company, and share the same corporate culture. The workers' incentive structure does not differ qualitatively, and quantitative (i.e. level) differences reflect differences between the two local labour markets.

To formally test the parallel trends assumption we estimate a specification in which we interact the experimental week with $T_{f,t}$. To account for weekly fluctuations and variations, we use two-week averages. Regression results for productivity and production output, for workers and team leaders, are shown in Figures A5, A6, A7 and A8 in the Appendix, respectively.

Another concern may be that assignment to teams (i.e. production lines) is non-random and may differ between the factories and over time. Indeed, assignment is directed by management. For instance, newly hired employees are assigned separately to a particular team, called "probation line", which is used as a reserve. Workers stay there until vacancies in regular teams become available, through turnover. New workers may stay in the probation line for up to six months. To address these issues, we exclude all workers who were hired during our experimental period or three months (probation period) before our experiment started from our analysis.

It is noteworthy that there was relatively low turnover among team leaders during our experiment; only one team leader quit during the duration of the experiment, for personal reasons, and we exclude observations of her workers under the new team leader.

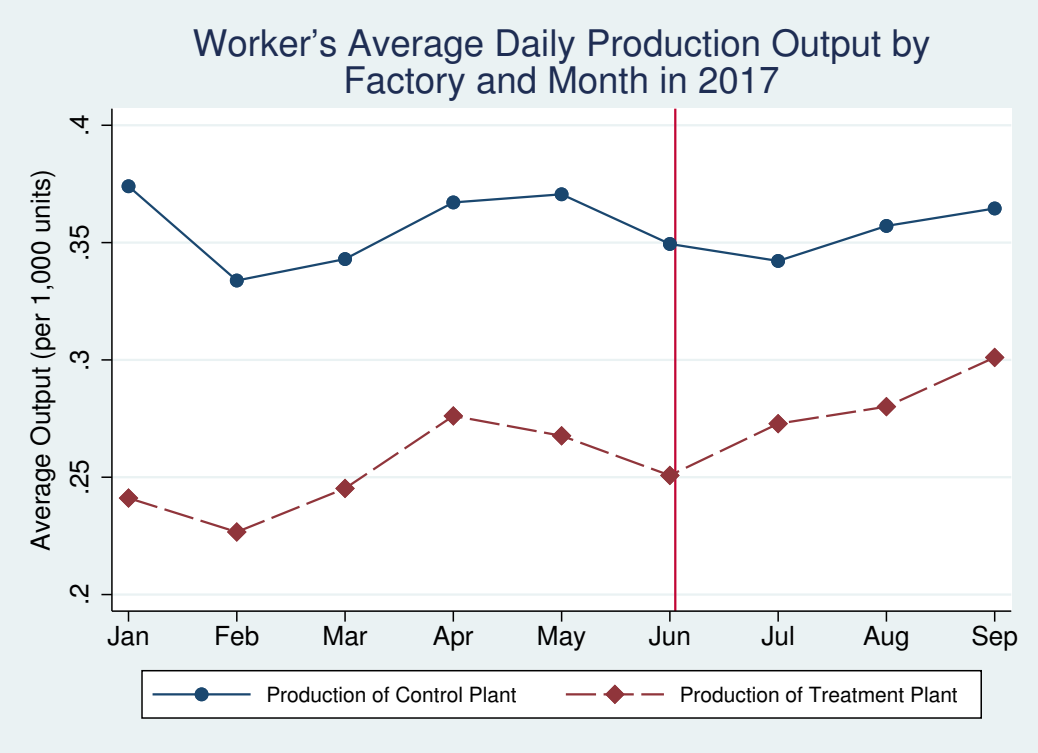
6 Results

6.1 Descriptive evidence

Ranking of Team Leaders Figure A4 in the Appendix presents the rankings of the team leaders (in terms of their aggregate evaluation scores) during the intervention period. It is worth noting that one group of team leaders (Lines A, B, and C) remained in the bottom half of the ranking throughout the intervention (6 is the highest rank), while the other three team leaders remained in the top half. This suggests possible heterogeneity in leadership ability as well as in leadership effort. But team leaders's ranks did vary within their groups, which is consistent with some competition taking place between team leaders.

Productivity Figure 2 shows the average daily production output of workers, including team leaders, for both the treatment and the control factory in each month of 2017 until the end of our experiment.

Figure 2: Production Trend in Both Factories



Notes: The vertical line indicates the beginning of our intervention in the treatment plant.

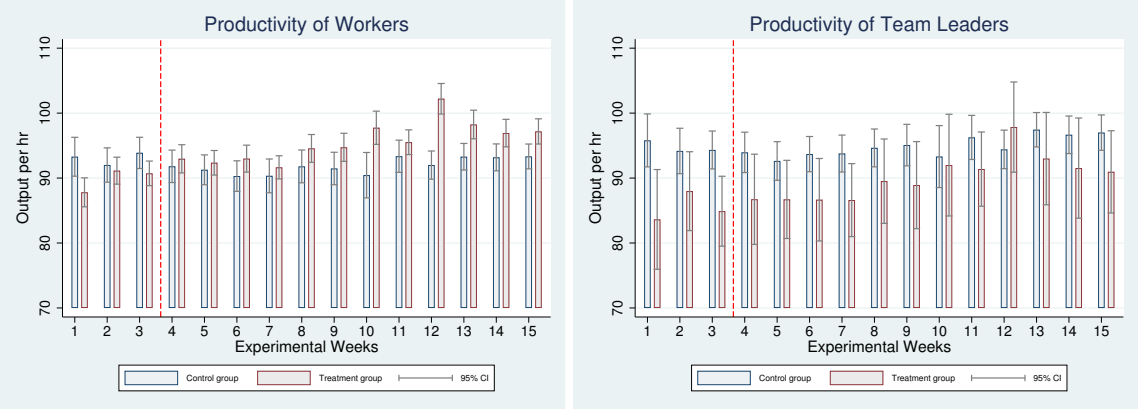
Two interesting observations emerge. First, both factories exhibit very similar time trends in daily average output per worker before our intervention. This is very reassuring and supports the assumption of parallel trends in outcome variables of interest before our intervention, which is needed for our difference-in-difference approach to yield a causal interpretation. Our data start in January 2017, since data from before and after Chinese New Year are not necessarily easily comparable as there is usually large workforce turnover around Chinese New Year, which may be accompanied by changes in team assignments and management practice.

Second, the intervention seems to be associated with a change in average daily output per worker in the treatment plant (dashed line) compared to the control plant (solid line). This change appears to stem from a level effect.

According to the company, output per worker typically drops in July in both factories,

because of the weather.[14] Indeed, a drop in output per worker is observed in the control plant, but not in the treatment plant, which is consistent with our expectations for the intervention. After July both factories display similar time trends.

Figure 3: Mean Productivity of Workers and Team Leaders



Notes: The vertical dashed line indicates the start of the intervention.

Our model would, however, predict that an increase in output stems from an increase in the workers' productivity, namely output per hour worked, while team leaders' productivity would increase by less or indeed fall (Prediction 1). Figure 3 appears to partially confirm this expectation. The figure shows average output per hour worked (solid lines) by factory for each of the 15 weeks of our experiment, accompanied by the 95% confidence intervals (dashed lines). Reassuringly, in the pre-intervention period (weeks 1 to 3) there is little discernible difference in time trends (and even in levels for workers) between the factories, supporting our identification strategy.[15]

From about mid August, after about five weeks of intervention, worker productivity in the treatment plant starts to significantly surpass that in the control factory.[16] Notably

[14]Temperatures peak in July and August (data from WorldWeatherOnline.com show a maximum temperature in both cities above 35 centigrade in July and August, see online appendix). The factories produce medical appliances, requiring workplaces to be sterile and workers to wear impervious gowns. At high temperatures the workplace environment becomes less comfortable and productivity tends to fall.

[15]The difference in levels for team leaders may well be driven by individual-specific heterogeneity as we only observe 6 respectively 7 team leaders per factory. While team leader productivity in the treatment plant appears to catch up with the one in the control plant, this may be due to the fact that we exclude one team leader who was replaced due to sickness from the sample from week 9. Plotting Figure 3 for only the 5 team leaders who we have data for throughout the experiment shows much slower productivity growth in the treatment plant.

[16]Productivity data was only available during our experiment; the factories did not systematically record hours worked as the remuneration was by piece rates. Figure 3 does not change qualitatively if we substitute productivity with daily output per worker. However, measured output contains more noise, as working

worker productivity remains constant in the control factory, but increases in the treatment factory.[17] This is very much consistent with a scenario where team leaders start to invest in leadership activities at the beginning of the intervention, and the investment takes a few weeks to take full effect.

The picture markedly differs for team leaders, for whom no such effect is apparent. Indeed team leader productivity did not appear to change significantly in the treatment factory, neither in absolute terms nor relative to the control factory. In light of the model this would indicate that organisational and productive tasks may not conflict very much, i.e. κ is low.

Table B3 in the Appendix presents the unconditional means of output per worker underlying (part of) Figure 2, as well as productivity data underlying Figure 3 and further descriptive data. The table also distinguishes workers from team leaders.[18]

6.2 Regression Results

Table 3 presents estimated treatment effects from our difference-in-differences specification on our sample of workers. Columns (1) and (5) suggest that the intervention is associated with an increase in worker daily output and, more importantly, worker productivity by 9% and 7% respectively. This improvement is statistically significant at 1% level. The remaining columns show effects over time. As all workers in our sample had more than three months of experience on the job, learning by doing should not play a role. But if team leaders invested in organisational effort and these investments take some time to produce results or if these effects wore off over time, treatment effects may well fluctuate over time. Indeed, we find that the estimates of the treatment effect on worker productivity increase over time, essentially doubling over three months, see columns (6) to (8), and the differences are statistically significant. This is not the case for daily output per worker, see columns (2) to (4), indicating that workers may have adjusted their working hours.[19]

times fluctuated, e.g. when workers were late for work or left work early (e.g. due to sickness).

[17] Indeed, during the intervention average productivity in all teams increased and there is little evidence for sizable heterogeneity of effects across teams. If anything, initially weaker teams caught up, see Figure A9 in the Appendix.

[18] Observations per month vary somewhat. In August one worker in the control plant was absent for the whole month because of illness, and one worker in the treatment plant was assigned to another production unit which is not included in our sample. In September one team leader in the treatment plant was on sick leave for two weeks. She was temporarily replaced by a factory manager, who does not perform manufacturing tasks. Since we have neither data on output nor relative performance we have dropped observations of workers from this line and the replacement team leader is not included in our sample.

[19] It is noteworthy that in September a large number of defective products were returned to the treatment plant. Workers who participated in our experiment were responsible for unpacking these products, but this

Table 3: Treatment Effect on Outcomes for Workers

	Log(Output)				Log(Productivity (output/hour))			
	Jul-Sep	Jul	Aug	Sep	Jul-Sep	Jul	Aug	Sep
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$T_{f,t}$	0.091***	0.096***	0.106***	0.084***	0.068***	0.040***	0.059***	0.103***
	(0.017)	(0.016)	(0.019)	(0.025)	(0.016)	(0.014)	(0.016)	(0.021)
Observations	6,105	3,004	2,875	2,892	6,105	3,004	2,875	2,892
Clusters	62	62	62	62	62	62	62	62
R^2	0.547	0.679	0.513	0.609	0.769	0.798	0.759	0.780
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 present the results for the full sample, which includes observations from June 7^{th} until September 30^{th}, while Columns 2-4 and 6-8 only contain pre-treatment observations and each post-treatment month separately. Productivity is measured as output per hour. $T_{f,t}$ is an indicator variable that takes the value 1 for individuals in the treatment factory during the intervention period (i.e. 1^{st} July or after) and 0 otherwise. Individual, week, and day of the week fixed effects (e.g. Monday), and indicator variables for sitting in another production line, sick leave, and organisational errors 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.

Table 4 presents estimated treatment effects for team leaders, confirming the observation in Figure 3.[20] The intervention led to a statistically and economically significant increase in daily output for team leaders, of similar size as the one observed for workers. However, and in contrast to the case of workers, the intervention did not lead to a statistically significant increase in team leader productivity. While the point estimates indicate a positive effect on productivity, its size is about half that of the increase for workers. Interpreting these results through the lens of our model in Section 3, this suggests a scenario in which the opportunity cost of organisational effort (κ) is about the same as the resulting increase in the team leader's individual productivity (β).

These observations are consistent with our expectations and standard economic reasoning. Incentivising team leaders explicitly to provide inputs to a local public good that increases productivity of all team members should increase workers' productivity, both

task was not incentivised monetarily nor reflected in daily output number. Worker productivity is, however, adjusted for this change. Team leaders recorded the time spent on unpacking and we discounted that time when computing productivity.

[20] 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.

Table 4: Treatment Effect on Outcomes for Team Leaders

	Log(Output)				Log(Productivity (output/hour))			
	Jul-Sep	Jul	Aug	Sep	Jul-Sep	Jul	Aug	Sep
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$T_{f,t}$	0.084**	0.107***	0.106**	0.053	0.035	0.021	0.037	0.052*
	(0.031)	(0.029)	(0.040)	(0.042)	(0.021)	(0.021)	(0.023)	(0.026)
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	0.313	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 team leader i. Dependent variables in Columns 1-4 and Columns 5-8 are the log of individual daily output and the log of individual productivity, respectively. Columns 1 and 5 show the results for the full sample including observations from June 7^{th} until September 30^{th}, while Columns 2-4 and 6-8 compare the observations from the pre-treatment period (June) to each post-treatment month separately. Productivity is measured as output per hour. $T_{f,t}$ is an indicator variable that takes the value 1 for individuals in the treatment factory during the intervention period (i.e. 1^{st} July or after) and 0 otherwise. Individual, week and day of the week fixed effects (e.g. Monday), and indicator variables for sitting in another production line, sick leave and organisational errors 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.

in absolute terms, and relative to that of team leaders.

This reasoning is backed by statements by managers in the treatment factory. They claim that after the introduction of the subjective evaluations and monetary prizes, team leaders indeed engaged more frequently in organisational tasks. This helped them to develop different styles of leadership and further equipped them with a variety of organisational skills. With more organisational experience, team leaders were able to organise the workers more efficiently. An organisational task that took the team leader half an hour in July might only take ten minutes in September. Therefore, the intervention indeed helped transforming the appointed workers into effective leaders.

6.3 Was Performance Evaluation Accurate?

A necessary condition for our intervention to impact the productivity of team members by increasing the leadership skills of team leaders is that the performance evaluation of leadership activities, especially those that foster worker productivity, is sufficiently accurate (the theoretical model assumes an informative signal). To test this condition, we

examine the correlation between worker productivity and subjective performance evaluations of team leaders. Performance measures encompass four dimensions of leadership: organising the work, maintaining high productivity, maintaining high quality, and team building. We would expect a strong positive correlation for the productivity measure, while the correlation for the organisation measure may be positive but weaker, as some of the effects may occur with a lag. Expectations regarding the quality score are unclear, as there may be no trade-off if applying good working techniques reduces interruptions and improves work flow. Regarding team building, we would expect a low or negative correlation as team building is likely to produce long-term effects.

Table 5: Correlation of Workers' Weekly Average Productivity and Team Leaders' Scores During Intervention, Treatment Group

	Log(Workers' Weekly Average Productivity (output/hour))					
	(1)	(2)	(3)	(4)	(5)	(6)
Overall score	0.0010**					
	(0.0003)					
Organisation score		0.0028**				0.0003
		(0.0008)				(0.0006)
Productivity score			0.0037***			0.0035**
			(0.0009)			(0.0010)
Quality score				0.0027**		-0.0005
				(0.0010)		(0.0010)
Relationship score					0.0029*	0.0003
					(0.0012)	(0.0012)
Observations	72	72	72	72	72	72
R^2	0.339	0.245	0.396	0.169	0.281	0.400

Notes: The unit of observation is product line (equivalent to team leader i) in each week in the treatment factory during the intervention period (i.e. 1^{st} July or after). Dependent variables in Columns 1-6 are the log of the average of workers' weekly productivity in a production line. Column 1 presents the results for the correlation between the aggregated score and worker productivity in the same week, Columns 2-5 show the results for the correlation between each performance score and worker productivity in the same week, and Column 6 includes all four dimensions in a single regression. Productivity is measured as output per hour. No control variables are included in any of the regressions. Robust standard errors clustered at the product line (or team leader i) level are reported in brackets below the estimates. *** Significant at 1% level, ** significant at 5% level, * significant at 10% level.

Table 5 presents the simple correlations between performance scores awarded (both aggregate and in each of the four dimensions) and worker productivity in the same week during the intervention. Because evaluations took place weekly, we now use weekly av-

erages. The results align closely with our expectations, providing a high degree of confidence in the subjective performance measures. For instance, approximately 40% of the variation in worker productivity is explained by the variation in the productivity performance measure. The point estimate indicates that a one percent increase in evaluation score is associated with a 0.24% increase in the average productivity of the workers.

In contrast, when conducting the same exercise for the control factory (Table B4), no significant correlations are observed for any measure. The coefficients are very close to zero, and most are indeed negative, and the performance measure is unable to explain more than 2.5% of the variation in worker productivity.

These observations indicate that the performance evaluation was only informative about team productivity during the actual intervention period when workers, team leaders, and managers were all informed of the evaluation process and when payoffs were tied to the scores. Recall that throughout the duration of our experiment, managers in both factories evaluated the performance of team leaders. However, only in the treatment factory, and only during the intervention period (July to September), was information about team leaders' scores shared with workers and team leaders. Absolute scores were not published at all. Moreover, only team leaders in the treatment plant were informed about the dimensions of the performance evaluation, and this information was disclosed at the end of week 3 (see Section 4.4).

That is, it appears that the intervention not only had an effect on productivity but also on the accuracy of subjective performance measurement. A likely explanation for this finding is that the public disclosure of rankings on all dimensions of performance induced a degree of accountability for the evaluators. This increased accountability may have enhanced their incentives to provide more accurate assessments of team leaders' performance, leading to a stronger correlation between the performance scores and team productivity.

Based on the above findings, we conclude that the intervention plausibly induced performance evaluation that was informative about the outcome of interest, namely productivity. This suggests strongly that the treatment effect on worker productivity indeed occurred through the postulated mechanism, namely the increased leadership activities of team leaders.

6.4 Impact on hours of work

One assumption underlying our model in Section 3 is that team leaders face an opportunity cost when engaging in leadership activities, which comes in the form of higher marginal cost of productive effort. This would imply that an increase in daily output and a small, possibly negligible increase in productivity for team leaders associated with the

intervention may be explained by exerting more effort in aggregate, as suggested by the model. While we cannot test for effects on (unobservable) effort, we can assess whether team leaders spent more time working overall during the intervention. Table 6 presents regression results for specification 1, but using minutes worked on the job per day as the dependent variable.

Table 6: Minutes Team Leaders Worked in a Day

	Jul-Sep	Jul	Aug	Sep
	(1)	(2)	(3)	(4)
$T_{f,t}$	28.417	56.578**	35.944	0.580
	(23.911)	(18.725)	(33.185)	(30.614)
Observations	1,312	644	621	621
Clusters	13	13	13	13
R^2	0.317	0.420	0.295	0.372
Controls	YES	YES	YES	YES

Notes: The unit of observation is team leader i. The dependent variable in Columns 1-4 is a team leader's working time per day (in minutes). Columns 1 shows the results for the full sample including observations from June 7^{th} until September 30^{th}, while Columns 2-4 compare the observations from the pre-treatment period (June) to each post-treatment month separately. $T_{f,t}$ is an indicator variable that takes the value 1 for individuals in the treatment factory during the intervention period (i.e. 1^{st} July or after) and 0 otherwise. Individual, week, and day of the week fixed effects (e.g. Monday), and indicator variables for sitting in another production line, sick leave and organisational errors 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.

The intervention led to an increase in the working time of team leaders of about 28 minutes per day, but this increase is not statistically significant. Decomposing results by intervention month reveals a statistically significant increase of about one hour in the first month of the intervention (Column (2)), but not subsequently. This is consistent with a one-off investment by team leaders of their time in optimising their team organisation in the first month. The effect size is consistent with the design of our monthly prizes, which is precisely aimed to motivate the team leaders to spend one hour per day on organising the team instead of packing the products. By comparison, workers also increased their working time during the intervention as shown in Table B5 in the Appendix. But the effect size is much smaller, as point estimates for workers are more than 50% smaller than the ones for the team leaders, although the difference is not statistically significant.

Overall, our findings on the impacts of the intervention on workers' and team leaders' working time, productivity, and output are very much consistent with the idea that multitasking team leaders, when given a high-powered incentive for the organisational task,

increased their total working time to spend some time on the organisational task, which increased all employees productivity on the productive task, just as the model in Section 3 suggests.

7 Conclusions

In this paper, we address a central challenge in organisational economics: how to incentivize team leaders. This is a formidable problem because team leaders are asked to both contribute to team production and to provide leadership and managerial inputs. This constitutes a multitasking problem where the different tasks not only differ in their observability and measurability, but also in the degree to which they generate positive externalities for the team. Standard economic reasoning would predict that tasks that are harder to measure and provide more positive externalities, such as leadership, will be underprovided. We examine the effects of adding a relative subjective evaluation of appointed team leaders' organisational and leadership behavior to a piece-rate reward scheme based on output.

In a field experiment, this intervention yielded the desired effect: we observed a 7% increase of productivity of the workers in the treatment factory compared to the control factory. Interestingly, the productivity of team leaders did not show a comparable increase, although their daily output increased in line with the workers. The data indicate that team leaders increased their working time both in absolute terms and relative to workers. Our findings are consistent with a model of multitasking, where team leaders divide their effort between productive and organisational tasks, with the latter contributing to the overall productivity of the team. Explicitly incentivising organisational inputs through subjective performance evaluation shifts team leaders' effort towards the organisational task.

The intervention was profitable for the firm, as the additional cost of the subjective performance pay scheme amounted to about 50% of a worker's earnings, while the productivity effect was comparable to hiring two additional workers. Indeed, the company decided to roll out a similar, albeit slightly revised, intervention in the control plant in September 2017, while maintaining our intervention in the treatment plant. This decision effectively ended our field experiment. The data collected for the following three months, until January 2018, indicate that worker productivity in the treatment plant fluctuated around the levels observed in September, suggesting that the effect of the intervention lasted for at least six months.

An intriguing finding of our study is that the subjective performance evaluation scores exhibit a strong correlation with worker productivities only during the intervention pe-

riod in the treatment factory. Outside of this period, the correlation is very close to zero or negative. This suggests that the subjective performance scores are only informative about worker productivity when all employees are aware of the evaluation procedure and the scores, and when the rankings have an impact on the payoffs of team leaders. Possible explanations for this observation include implicit incentives for evaluators when their evaluation is made public and subject to scrutiny. Additionally, communicating to team leaders key dimensions of leadership (see e.g., Manthei et al. (2023)) may have facilitated learning as well as enabling to identify and share best practices (see e.g. Song et al. (2018)). Future research could further explore these mechanisms and quantify the impact of publicizing performance rankings.

Our intervention, although effective in increasing worker productivity, is likely not the optimal incentive mechanism for maximising total surplus or profit. As such, we view our intervention more as a proof of principle, providing a lower bound on the possible effects that can be achieved. The design of optimal reward schemes based on subjective performance evaluation is an open question. Similarly, while we employed a simple average of performance scores across four dimensions of leadership developed in collaboration with the company, the optimal design of performance evaluation is another key area of interest for future research (see Adhvaryu et al. (2022) for some important pointers).

References

Adhvaryu, A., Nyshadham, A., and Tamayo, J. (2022). Managerial Quality and Productivity Dynamics. *The Review of Economic Studies*.

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.

Baker, G., Gibbons, R., and Murphy, K. J. (1994). Subjective performance measures in optimal incentive contracts. *The Quarterly Journal of Economics*, 109(4):1125–1156.

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.

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.

Bol, J. C. (2011). The determinants and performance effects of managers' performance evaluation biases. *The Accounting Review*, 86(5):1549–1575.

Bol, J. C. and Smith, S. D. (2011). Spillover effects in subjective performance evaluation: Bias and the asymmetric influence of controllability. *The Accounting Review*, 86(4):1213–1230.

Boudreau, K. J., Lakhani, K. R., and Menietti, M. (2016). Performance responses to competition across skill levels in rank-order tournaments: field evidence and implications for tournament design. *The RAND Journal of Economics*, 47(1):140–165.

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.

de Janvry, A., He, G., Sadoulet, E., Wang, S., and Zhang, Q. (2023). Subjective performance evaluation, influence activities, and bureaucratic work behavior: Evidence from China. *American Economic Review*, 113(3):766–799.

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.

Demeré, B. W., Sedatole, K. L., and Woods, A. (2019). The role of calibration committees in subjective performance evaluation systems. *Management Science*, 65(4):1562–1585.

Dewatripont, M., Jewitt, I., and Tirole, J. (2000). Multitask agency problems: Focus and task clustering. *European Economic Review*, 44(4-6):869–877.

Drago, R. and Garvey, G. T. (1998). Incentives for helping on the job: Theory and evidence. *Journal of Labor Economics*, 16(1):1–25.

Englmaier, F., Grimm, S., Grothe, D., Schindler, D., and Schudy, S. (2021). The value of leadership: Evidence from a large-scale field experiment. Cesifo working paper.

Englmaier, F., Grimm, S., Grothe, D., Schindler, D., and Schudy, S. (2023). The efficacy of tournaments for non-routine team tasks. *Journal of Labor Economics*.

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.

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.

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

Hong, F., Hossain, T., and List, J. A. (2015). Framing manipulations in contests: a natural field experiment. *Journal of Economic Behavior & Organization*, 118:372–382.

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.

Itoh, H. (1991). Incentives to help in multi-agent situations. *Econometrica*, pages 611–636.

Jones, D., Tonin, M., and Vlassopoulos, M. (2018). Paying for what kind of performance? performance pay and multitasking in mission-oriented jobs.

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.

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

Krueger, M. and Friebel, G. (2022). A pay change and its long-term consequences. *Journal of Labor Economics*, 40(3):543–572.

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

List, J. A., Van Soest, D., Stoop, J., and Zhou, H. (2020). On the role of group size in tournaments: Theory and evidence from laboratory and field experiments. *Management Science*, 66(10):4359–4377.

Manthei, K. and Sliwka, D. (2019). Multitasking and subjective performance evaluations: Theory and evidence from a field experiment in a bank. *Management Science*, 65(12):5861–5883.

Manthei, K., Sliwka, D., and Vogelsang, T. (2023). Talking about performance or paying for it? A field experiment on performance reviews and incentives. *Management Science*, 69(4):2198–2216.

Rosaz, J. and Villeval, M. C. (2012). Lies and biased evaluation: A real-effort experiment. *Journal of Economic Behavior & Organization*, 84(2):537–549.

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

Song, H., Tucker, A. L., Murrell, K. L., and Vinson, D. R. (2018). Closing the productivity gap: Improving worker productivity through public relative performance feedback and validation of best practices. *Management Science*, 64(6):2628–2649.

A Appendix: Other Figures

Figure A1: Factory Floor

Note that any identifying information has been obscured in the picture.

Figure A2: Leader Board

Note that any identifying information has been obscured in the picture.

Figure A3: Sliders for Ranking the Team Leaders

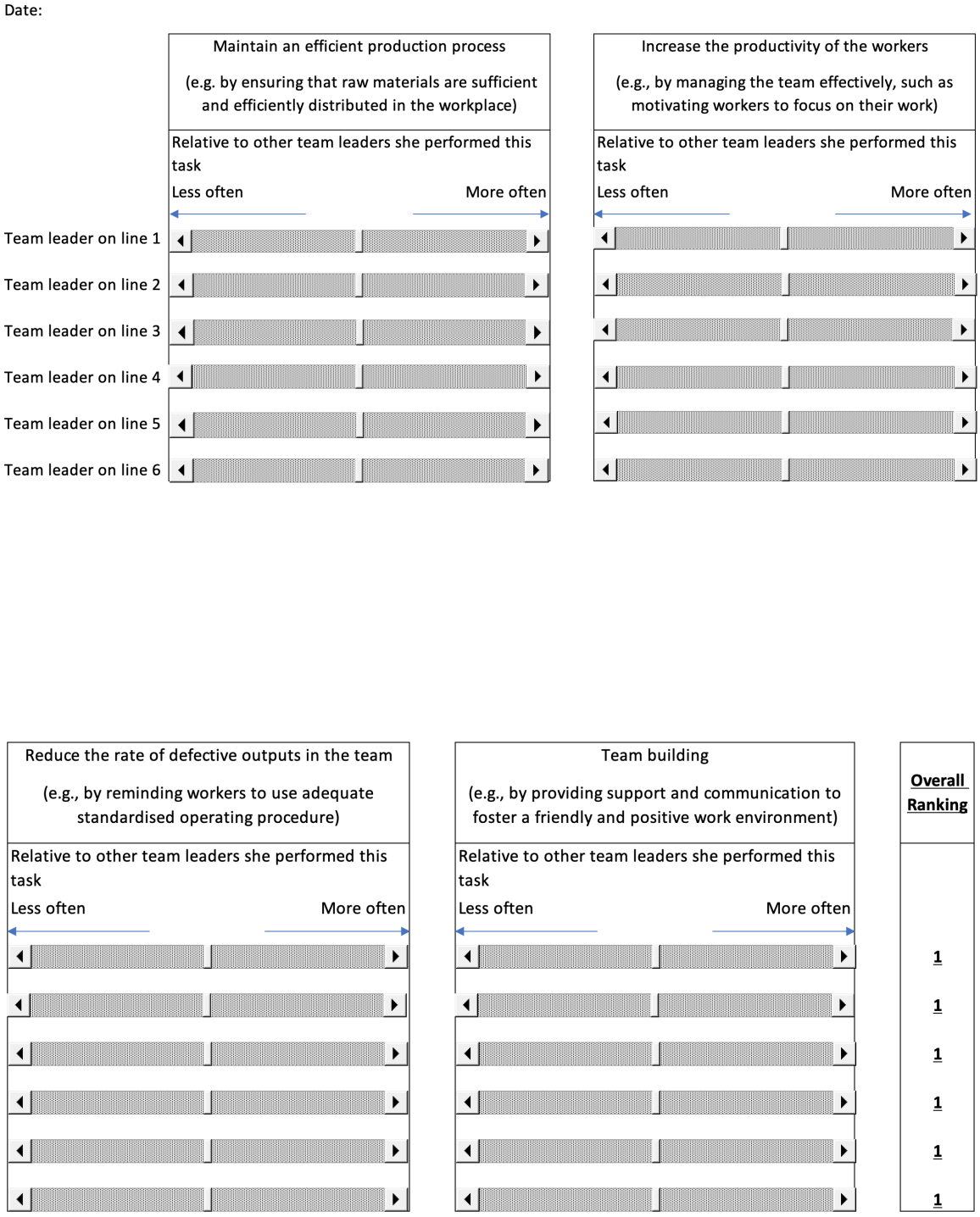
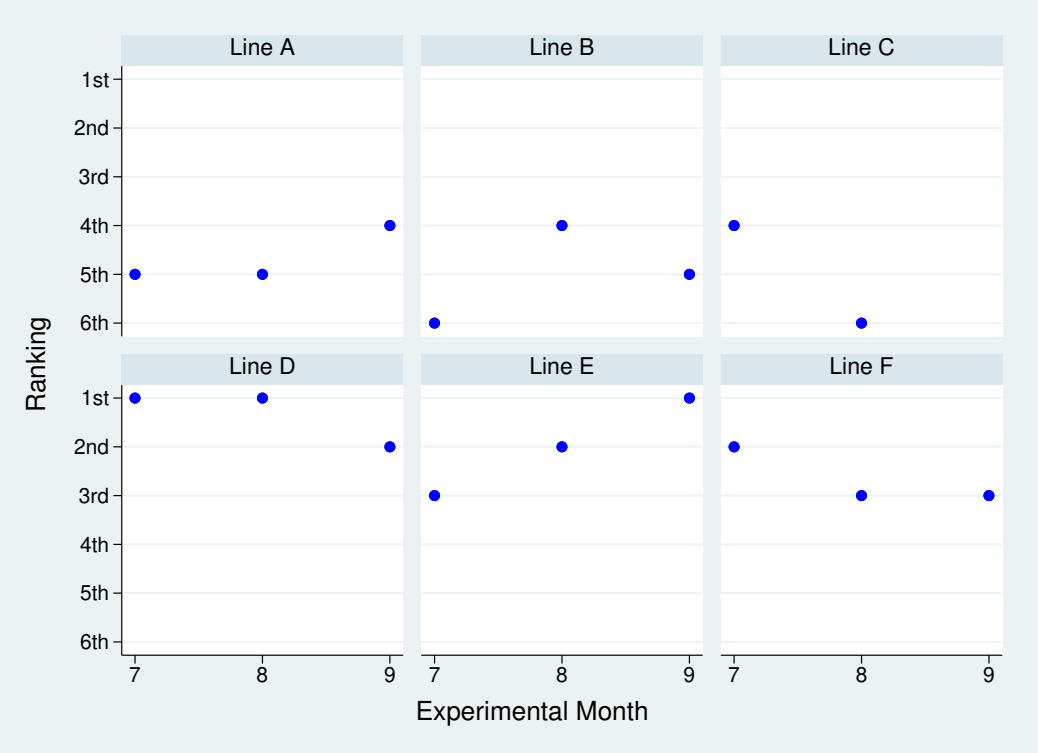
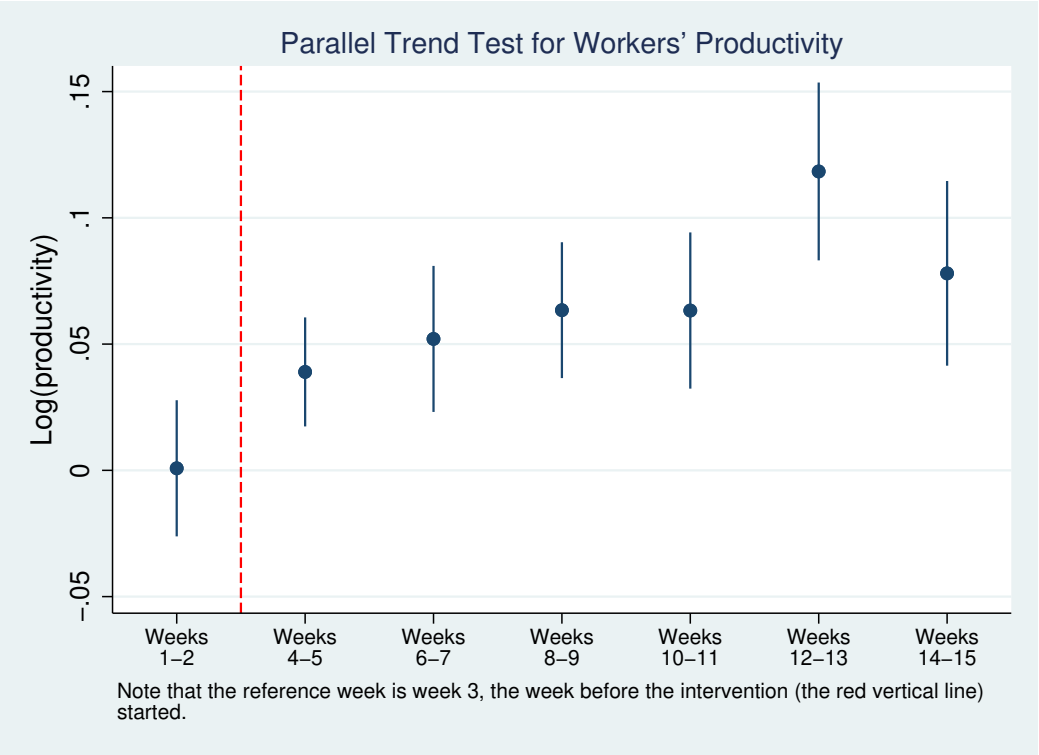


Figure A4: Ranking of Team Leaders in Treatment Plant during the Treatment Period



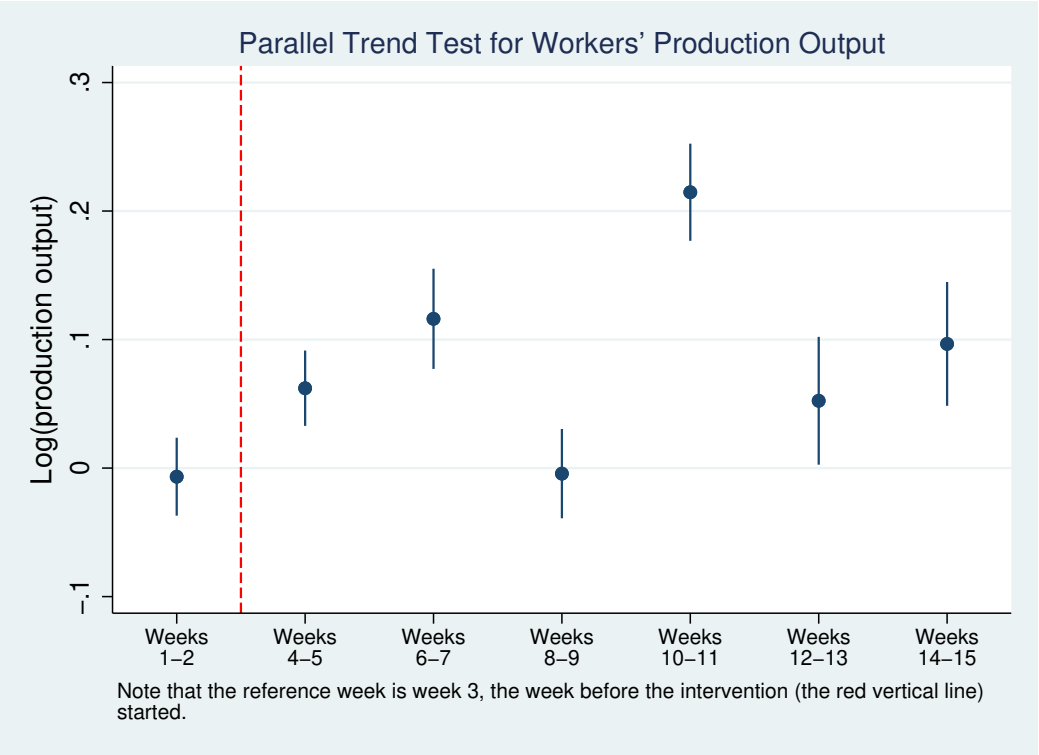
Notes: The ranking indicates the aggregated rank for each team leader in each month, which is used to determine the reward payment of the intervention.

Figure A5: The Parallel Trend Assumption Test for Worker's Productivity



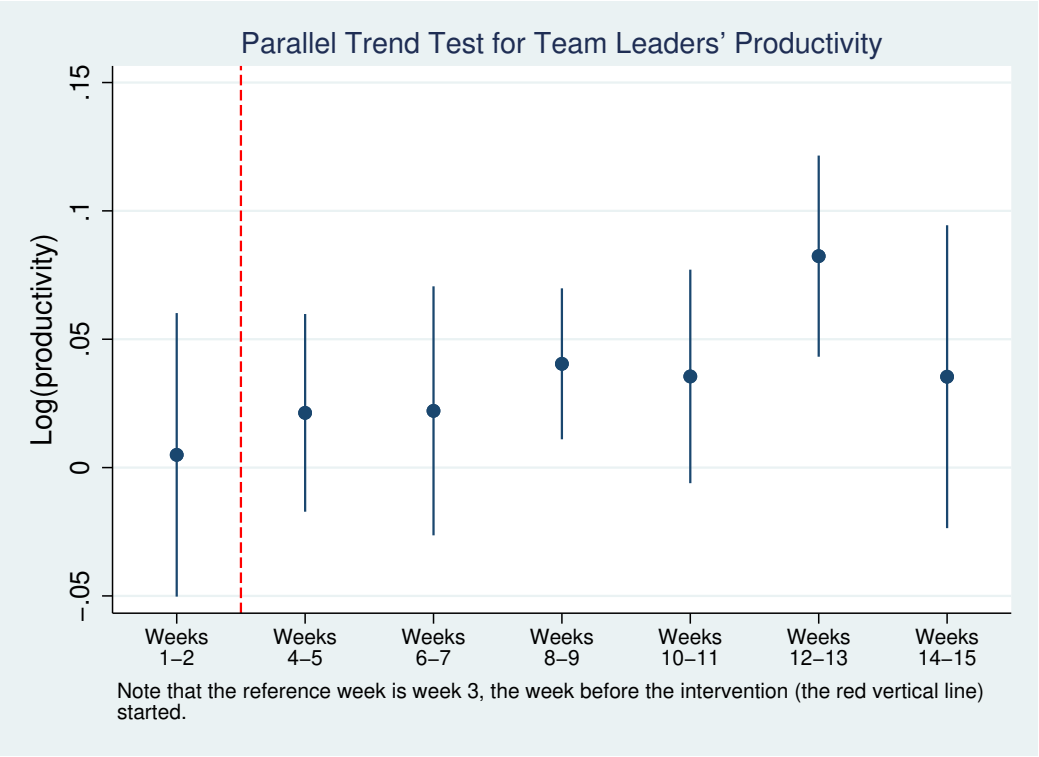
Notes: The unit of observation is worker i. The vertical line indicates the beginning of our intervention in the treatment plant. The dependent variable is the log of worker's productivity. Productivity is measured as output per hour. Individual, week, and day of the week fixed effects (e.g. Monday), and indicator variables for sitting in another production line, sick leave and organisational errors 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.

Figure A6: The Parallel Trend Assumption Test for Worker's Production Output



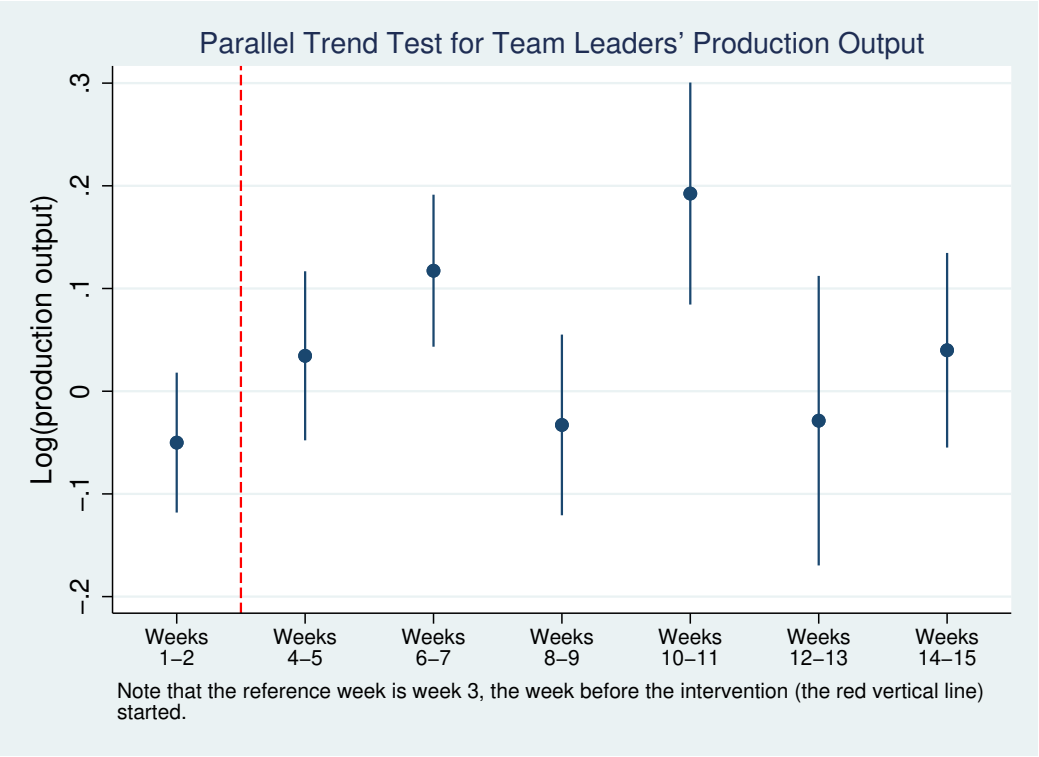
Notes: The unit of observation is worker i. The vertical line indicates the beginning of our intervention in the treatment plant. The dependent variable is the log of worker's production output. Individual, week, and day of the week fixed effects (e.g. Monday), and indicator variables for sitting in another production line, sick leave and organisational errors 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.

Figure A7: The Parallel Trend Assumption Test for Team Leaders' Productivity



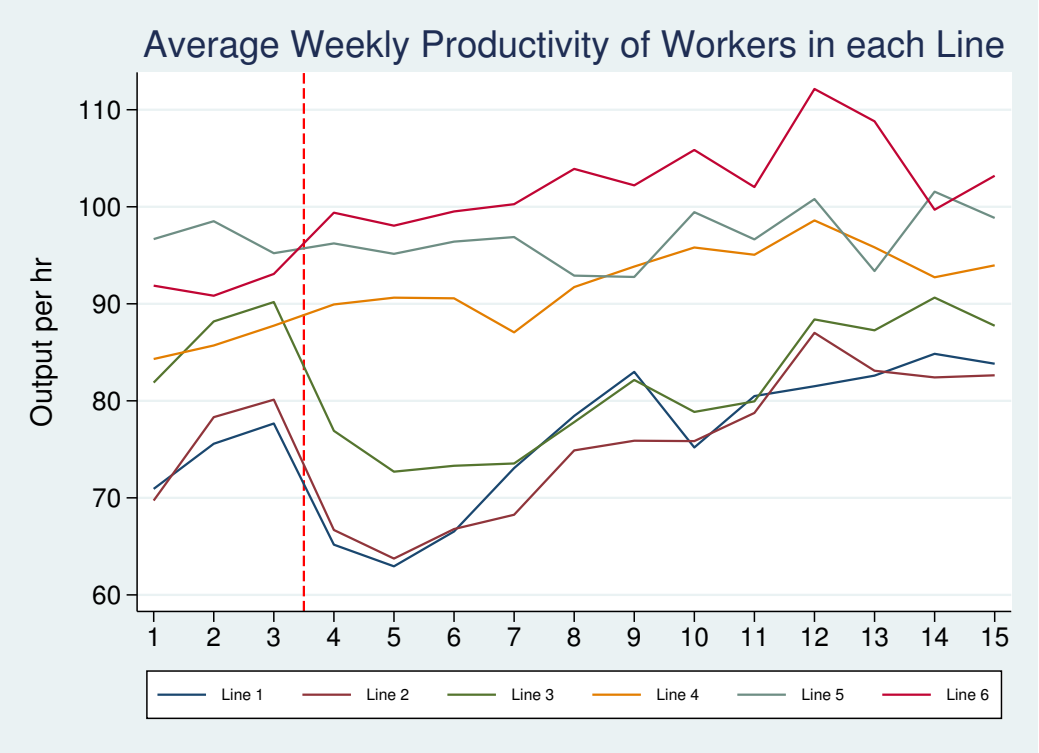
Notes: The unit of observation is team leader i. The vertical line indicates the beginning of our intervention in the treatment plant. The dependent variable is the log of team leader's productivity. Productivity is measured as output per hour. Individual, week, and day of the week fixed effects (e.g. Monday), and indicator variables for sitting in another production line, sick leave and organisational errors 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.

Figure A8: The Parallel Trend Assumption Test for Team Leader's Production Output



Notes: The unit of observation is team leader i. The vertical line indicates the beginning of our intervention in the treatment plant. The dependent variable is the log of team leader's production output. Individual, week, and day of the week fixed effects (e.g. Monday), and indicator variables for sitting in another production line, sick leave and organisational errors 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.

Figure A9: Workers' Weekly Average Productivity in Treatment Plant Over Time



Notes: The vertical line indicates the beginning of our intervention in the treatment plant. The horizontal axis indicates the experimental weeks.

B Appendix: Other Tables

Table B1: Summary Statistics for Other Individual Characteristics (Treatment group)

	N	mean	sd	min	max
Married	48	0.979	0.144	0	1
Live in the factory	48	0.250	0.438	0	1
Commute by factory bus	48	0.729	0.449	0	1
Commute by bike	48	0.0417	0.202	0	1
Commute by motorbike	48	0.125	0.334	0	1
Number of years worked in the factory	48	2.625	2.100	0	7
Number of different types of products worked per day	48	1.946	0.300	1.630	2.435
Number of different products worked per day	48	2.319	0.387	1.917	3.016
Number of temporary coworkers from other lines	48	1.523	1.089	0	3.041
Education level:					
Illiterate	47	0.234	0.428	0	1
Primary school	47	0.426	0.500	0	1
Secondary school	47	0.298	0.462	0	1
High school	47	0.043	0.204	0	1

Table B2: Summary Statistics for Other Individual Characteristics (Control group)

	N	mean	sd	min	max
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	0.0156	1.010	1.049
Number of different products worked per day	27	1.047	0.0280	1.010	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 B3: Summary Statistics

	June	Jul-Sep	Jul	Aug	Sep
	(1)	(2)	(3)	(4)	(5)
Panel A. Treatment group					
Number of Employees	48	45.52	48	47	41
Number of Lines	6	5.695	6	6	5
Worker Daily Output	1,090.1	1,138.7	1,138.5	1,121.8	1,157.7
	(254.7)	(264.9)	(251.6)	(280.3)	(261.1)
Worker Productivity	90.03	95.38	92.44	95.51	98.62
	(17.85)	(17.56)	(16.82)	(18.12)	(17.20)
Leader 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)
Leader Productivity	85.65	89.98	86.67	90.40	93.34
	(20.03)	(20.10)	(19.24)	(20.52)	(20.13)
Panel B. Control group					
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 Productivity	93.04	91.98	90.95	91.94	92.96
	(16.68)	(13.60)	(14.24)	(14.23)	(12.39)
Leader 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)
Leader Productivity	94.72	94.97	93.51	95.00	96.39
	(12.62)	(10.35)	(10.07)	(11.08)	(9.809)

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

Table B4: Correlations between Workers' Weekly Average Productivity and Team Leader's Scores, Control Group

	Log(Workers' Weekly Average Productivity (output per hour))					
	(1)	(2)	(3)	(4)	(5)	(6)
Overall scores	-0.0001					
	(0.0003)					
Organisation scores		-0.0004				-0.0003*
		(0.0007)				(0.0001)
Productivity scores			-0.0004			-0.0003
			(0.0008)			(0.0005)
Quality scores				0.0001		0.0005*
				(0.0006)		(0.0002)
Relationship scores					-0.0003	-0.0001
					(0.0008)	(0.0005)
Observations	98	98	98	98	98	98
R^2	0.013	0.025	0.025	0.001	0.013	0.057

Notes: The unit of observation is product line (equivalent to team leader i) in each week in the control factory during the experimental period (i.e. 7^{th} June or after). Dependent variables in Columns 1-6 are the log of the average of workers' weekly productivity in a production line. Column 1 presents the results for the correlation between the aggregated score and worker productivity in the same week, Columns 2-5 show the results for the correlation between each performance score and worker productivity in the same week, and Column 6 includes all four dimensions in a single regression. Productivity is measured as output per hour. No control variables are included in any of the regressions. Robust standard errors clustered at the product line (or team leader i) level are reported in brackets below the estimates. *** Significant at 1% level, ** significant at 5% level, * significant at 10% level.

Table B5: Number of Minutes Workers Worked in a Day

	Jul-Sep	Jul	Aug	Sep
	(1)	(2)	(3)	(4)
$T_{f,t}$	11.364	37.429***	13.859	-6.929
	(8.942)	(9.948)	(9.522)	(11.857)
Observations	6,105	3,004	2,875	2,892
Clusters	62	62	62	62
R^2	0.437	0.593	0.428	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 7^{th} until September 30^{th} while Columns 2-4 compare the observations from the pre-treatment period (June) to each post-treatment month separately. $T_{f,t}$ is an indicator variable that takes the value 1 for individuals in the treatment factory during the intervention period (i.e. 1^{st} July or after) and 0 otherwise. Individual, week, and day of the week fixed effects (e.g. Monday), and indicator variables for sitting in another production line, sick leave and organisational errors 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.