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

DISCUSSION PAPER SERIES

IZA DP No. 11228

Team Incentives, Task Assignment, and Performance: A Field Experiment

Josse Delfgaauw
Robert Dur
Michiel Souverijn

DECEMBER 2017

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11228

Team Incentives, Task Assignment, and Performance: A Field Experiment

Josse Delfgaauw
Erasmus University Rotterdam and Tinbergen Institute

Robert Dur
Erasmus University Rotterdam, Tinbergen Institute, CESifo and IZA

Michiel Souverijn
Erasmus University Rotterdam and Tinbergen Institute

DECEMBER 2017

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

Team Incentives, Task Assignment, and Performance: A Field Experiment*

The performance of a work team commonly depends on the effort exerted by the team members as well as on the division of tasks among them. However, when leaders assign tasks to team members, performance is usually not the only consideration. Favouritism, employees' seniority, employees' preferences over tasks, and fairness considerations often play a role as well. Team incentives have the potential to curtail the role of these factors in favor of performance – in particular when the incentive plan includes both the leader and the team members. This paper presents the results of a field experiment designed to study the effects of such team incentives on task assignment and performance. We introduce team incentives in a random subsets of 108 stores of a Dutch retail chain. We find no effect of the incentive, neither on task assignment nor on performance.

JEL Classification: C93, M12, M52

Keywords: team incentives, task assignment, field experiment

Corresponding author:
Robert Dur
Erasmus University Rotterdam
Department of Economics H9-15
P.O. Box 1738
3000 DR Rotterdam
The Netherlands
E-mail: dur@ese.eur.nl

* We gratefully acknowledge comments and suggestions by the Editors and two anonymous reviewers of this journal, Sacha Kapoor, Adriaan Soetevent, Otto Swank, seminar participants at Erasmus University Rotterdam, and participants of the 2015 Colloquium on Personnel Economics at the University of Vienna, the 2017 Conference on Economics and Leadership at the University of Groningen, and the Field Days 2017 at the University of Copenhagen.

1 Introduction

In many organisations, employees work in teams that perform a variety of tasks. A team's ultimate performance depends on how well employees perform their tasks as well as on the division of tasks among the employees. For instance, allocating more important tasks to more talented employees will often improve a team's performance. However, in practice, performance considerations are rarely the only determinant of task allocation, in particular when some tasks are more interesting or pleasant than others. In such cases, preferences of team members for tasks may affect the allocation, as well as fairness concerns or seniority. When team managers decide on task allocation, favouritism may play a role as well. The importance of these factors can be diminished by introducing or strengthening incentive pay based on team performance for teams and their managers. In addition to inducing workers to perform better on their tasks, team incentives may induce teams or their managers to reallocate tasks in a performance-enhancing way.

In this paper, we present the results of a field experiment designed to study the effects of team incentive pay on team performance and task allocation in teams. The experiment took place in a retail chain in The Netherlands comprising 108 geographically dispersed stores. We randomly selected 60 stores to participate in a short-term sales tournament. Each participating sales team competed with two comparable sales teams for a period of six weeks on the basis of sales relative to a pre-determined sales target. Employees and the manager of the best-performing store earned a bonus of 50 euro each, which amounts to more than 3% of monthly employee earnings. During the tournament, participating stores received weekly feedback informing them about the current ranking in their group.

We use administrative sales data to analyse the effects of the team incentive on performance. Furthermore, we conducted surveys among employees and managers of all stores before and after the tournament period to learn about task allocation in stores. The surveys ask store employees and managers about the importance of several aspects for task allocation in their team, including employee ability, employee preferences over tasks, fairness concerns, seniority, and managerial favouritism. In addition, we collected data on employees' job satisfaction. By conducting identical surveys before and after the tournament in both the treatment and control group, we are able to estimate the effect of the team incentive on task allocation within teams (as perceived by employees) and on job satisfaction.[1]

The theoretical predictions are as follows. The team incentive increases the importance of team performance to employees and managers. Consequently, they should have a stronger incentive to exert effort, leading to better performance. Furthermore, team performance should play a more important role in the allocation of tasks among employees. Hence, we predict that employee ability becomes more important in task

[1] An alternative measure of task allocation would be to ask employees about the tasks they have actually performed. Any observed changes would be hard to interpret, though, because we lack reliable information about the relation between task allocation and performance.

assignment, while the other considerations (employee preferences, fairness concerns, seniority, and favouritism) become less important. This revised task allocation should also result in better team performance. Job satisfaction may suffer as workers' individual task preferences play a smaller role in task assignment. On the other hand, some workers' job satisfaction may actually increase as considerations like favoritism and seniority are muted. The overall implications for job satisfaction are thus theoretically unclear.

Our experiment follows the tradition in organizational economics to focus on incentives and contract design. However, by hypothesizing about and collecting detailed data on how leaders assign tasks to employees, we move a step in the direction of the literature on leadership in management and psychology, which allows for a much richer role of leadership than the economics literature (Zehnder et al. 2017). In particular, we stress the role of leaders in coordinating employees' complementary actions that jointly determine the team's success. We share this feature with Burgess et al. (2010), who analyse the introduction of team pay-for-performance at the UK tax authorities. The incentive scheme covered only a part of the tasks that teams are responsible for. The findings indicate that team incentives increased performance, part of which can be attributed to a change in task assignment within teams, as managers disproportionately reallocated efficient workers to the incentivised tasks. Two other papers in economics studying task assignment by managers are Bandiera et al. (2007 and 2009). They introduce incentive pay for managers in a UK fruit farm, making their pay dependent of their subordinates' performance. This induced managers to assign more productive employees to the incentivised task (picking fruit), leading to a substantial increase in productivity. By contrast, when paid a flat wage, managers were more likely to assign this task to employees who were socially connected to them. Hence, providing performance-based incentives to managers reduced favouritism in task assignment. Contrary to these earlier studies, we examine the effects of an incentive scheme that rewards overall team performance, not performance on a subset of tasks. We lack administrative data on individual task assignment and productivity. Instead, we use surveys to assess whether the incentives affect how tasks are allocated, as perceived by the employees.[2]

Our results are as follows. We find no effect of the team incentive on sales performance. This average treatment effect is fairly precisely estimated. Furthermore, we find no evidence supporting the hypothesis that task allocation has changed to enhance performance in treated stores. In particular, we find no effect on the importance of employee ability in the allocation of tasks. Nor do employees in treated stores report more often that "the division of tasks is such that the best possible sales performance is achieved". The average treatment effect on job satisfaction is not statistically significant either.

[2] Another recent paper examines task assignment in teams in the lab. Cooper and Sutter (2017) compare teams with random task assignment to teams with endogenous task assignment, finding that the latter do not outperform the former, despite a positive selection effect. In contrast to our setting, their teams have no leader. Muehlheusser et al. (2016) study task allocation by managers in the context of professional soccer.

Together, these results suggest that business continued as usual despite the treatment. We provide an extensive discussion of possible interpretations of the null results in the final section of our paper.

The estimated average treatment effect in this study is at the lower end of the range of estimates in comparable studies in the retail sector. In Delfgaauw et al. (2013, 2014, 2015), we implement comparable team-based tournament incentives schemes across shops of different retail chains and obtain average treatments effects on performance ranging from 0% to 5% increase in sales. Friebel et al. (2017) implement a (non-competitive) team incentive based on performance targets in a German bakery chain and find an average treatment effect of 3%. Casas-Arce and Martinez-Jerez (2009) find a more substantial effect of a tournament incentive implemented among independent retailers of a commodities firm in which there was substantially more money at stake than in the current study. We deviate from these previous studies in our focus on task allocation within teams. By contrast, earlier studies have looked at how treatment effects relate to the gender composition of the team and gender of the manager (Delfgaauw et al. 2013), the prospect of participation in further tournament rounds and volatility in performance (Delfgaauw et al. 2015), the extent to which employees are able to influence waiting times (Friebel et al. 2017), and intermediate rankings during the tournament (Casas-Arce and Martinez-Jerez 2009, Delfgaauw et al. 2014). The emerging evidence suggests that even within a sector, comparable incentive schemes can lead to different responses across organisations.

Experiments on the use of team incentives have also been conducted in other settings. Erev et al. (1993) recruit students to pick oranges and find that participants grouped into teams of 4 under team-based pay are 30% less productive than participants under individual incentive pay. This negative effect of team-based pay is mitigated when the teams competed under a tournament incentive scheme. Bandiera et al. (2013) study endogenous team formation and find that providing relative performance information reduces performance while relative performance pay increases performance, partially due to changes in team composition. Studies in education have found mixed effects of introducing team incentive pay for teachers (Lavy 2002, Glewwe et al. 2010, Springer et al. 2010, Muralidharan and Sundararaman 2011, Goodman and Turner 2012, Fryer 2013).[3]

The relationship between performance-related pay and job satisfaction is the subject of a small literature. With a single exception (Friebel et al. 2017), the existing evidence is correlational. Heywood and Wei (2006) and Green and Heywood (2008) document a positive relation between job satisfaction and performance pay, including profit-sharing,

[3]In a lab experiment, Chen and Lim (2013) find that contests between teams of participants yield higher productivity than contests between individuals, but only when (potential) teammates could meet before the contest. Babcock et al. (2015) offer students a reward for meeting a target regarding study room visits and find that team-based pay can outperform individual-based pay, but only when the teammates know each other.

in US and British panel survey data, respectively. Studying the introduction of incentive pay in two firms, Welbournel and Cable (1995) document a similar relation. Using cross-sectional British data, Petrescu and Simmons (2008) find no relation between team performance pay and job satisfaction. For individual incentive pay, they find a positive relation with satisfaction with pay, but no relation with overall job satisfaction. McCausland et al. (2005) find a positive relation between incentive pay and job satisfaction for highly paid workers but the reverse for lower paid workers. We add to this literature by examining the effect of introducing a short-term team incentive on job satisfaction using an experimental design, allowing for a causal interpretation of our findings.

The rest of this paper is organized as follows. The next section discusses the experimental setting and design. In Section 3 we discuss the estimation procedure. Descriptive statistics are presented in Section 4 and the estimation results in Section 5. Section 6 discusses our findings and concludes.

2 Experimental Context and Design

2.1 Experimental context

The experiment took place from October 2013 to January 2014 among 108 stores of a retail chain selling lingerie and swimwear in the Netherlands. All stores are company-owned, there are no franchisers. All managers and employees are female. A store has a manager and on average 7 employees. The majority of employees works part-time or on-call.

Employees earn an hourly wage slightly above the legal minimum hourly wage. They can occasionally earn team bonuses during incentive periods that are usually timed to coincide with marketing efforts. These bonuses are generally earmarked for team outings.[4] The company's management was interested in conducting this field experiment as it wished to explore a more extensive use of incentive pay.

Decisions regarding the product range, pricing, and marketing are made by the retail chain's management. The primary tasks of store staff include advising customers, attending the register, administration, and keeping the displays stocked and tidy. According to the company's management, employees have substantial influence on store performance, especially through an assertive and commercial attitude when advising customers. Furthermore, within the company it is widely acknowledged that store employees differ substantially in their ability to generate sales when advising customers. The chain's management and the store managers we spoke to considered task allocation to be an important channel through which store managers affect store performance. This anecdotal evidence suggests that, in this retail chain, task allocation matters for team performance, providing a good setting for an experiment on whether team incentives affect task assignment.

[4] No such incentive period ran concurrently with our experiment.

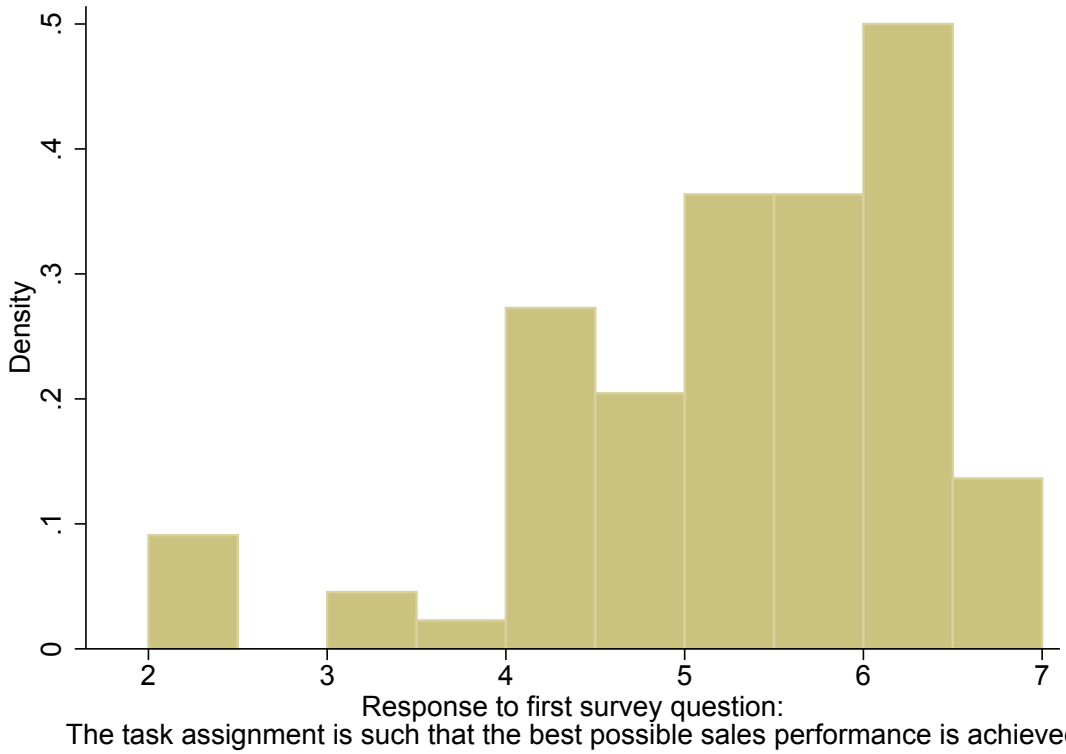
The store manager is responsible for staffing of the store and has the authority to assign tasks to employees. In practice, employees are consulted and can express their preferences. Furthermore, employees typically perform multiple tasks, prioritizing advising customers and attending the register when there are many customers and cleaning and stocking during quiet moments. In the survey we conducted before the experiment, the average response of store managers to the question "I determine the allocation of tasks" equals 6.3 on a 7-point scale ranging from completely disagree (1) to neutral (4) to completely agree (7). The average response to the question "I monitor whether everyone performs their tasks well" equals 6.4 on the same 7-point scale. Hence, store managers feel that they are in charge of task allocation. The mean response among employees to the statement 'I decide myself which tasks to perform' was 4.3 and their mean response to the statement 'The store manager decides about the allocation of tasks in the store' was 5.2, both on a 7-point Likert scale. This suggests that the store manager coordinates the allocation of tasks, but that employees do have some leeway in deciding which tasks to perform at a given time. Importantly, we also asked employees whether the task allocation in their store achieved the best possible sales performance. The mean response to this statement was 5.2. Taking averages across employees at the store level, Figure 1 gives the distribution of responses across stores. This indicates that while task allocation is catered towards enhancing performance, employees in many stores do see room for improving sales performance through changes in task allocation. We discuss the surveys in more detail in Subsection 2.3.

2.2 Experimental Design

We implemented tournaments between subsets of stores over a period of six weeks. A randomly selected subset of stores was assigned to groups of three stores each; we discuss the assignment procedure at length in subsection 2.4 below. Within these groups, stores competed for a prize. The performance measure in the tournament was a store's cumulative realized sales over the period of six weeks as a percentage of cumulative sales targets. These weekly targets for each store's sales are set by the company's management at the start of the financial year, long before we set up the experiment.[5] These targets take into account variation in sales due to e.g. seasonal effects, holidays, and planned marketing efforts, as well as store-specific factors. In our data, time and store-fixed effects account for 92.7% of the variation in sales targets. Realized sales relative to sales targets is a familiar performance metric to employees. Store managers receive this performance measure on a weekly basis. Let $r_{s,w}$ denote sales revenues realized by store s in week w, and $b_{s,w}$ the sales target for store s in week w. Cumulative performance of

[5] Our experiment ran at the end of the financial year, and the sales targets had been determined well before the experiment was planned. Hence, the targets are not affected by the experiment.

Figure 1: Task assignment and efficiency, store averages



store s after w weeks is given by

$$P_{s,w} = \frac{\sum_1^w r_{s,w}}{\sum_1^w b_{s,w}} \cdot 100\% \tag{1}$$

All employees and the manager of the store with the best performance in the group after six weeks received a bonus. Full-time employees received €50, part-time employees received a bonus proportional to their contract size. The bonus amounts to approximately 2.5% of employees' earnings in a six-week period. Depending on the size of the store and experience, store managers earn about 20 to 50 percent more than employees. Hence, the bonus amounts to 1.7% to 2.1% of store managers' pay in a six-week period. This is comparable to Burgess et al. (2010), where employees could earn a bonus of about 3% of pay and managers a bonus of either 2% or 4%, depending on the treatment. In Bandiera et al. (2007, 2009), the bonus was considerably higher. It could reach up to 25% of total compensation, and actual bonus payout was about 7% of total pay.

Out of fairness considerations, the company's management insisted on allowing each store the opportunity to compete. Therefore, we implement tournaments in two periods. The second tournament period started three weeks after the end of the first tournament period. In the first period, 60 randomly chosen stores participate while the remaining

48 stores comprised the control group. The relatively large number of stores in the treatment group serves two purposes. First, as we expect the variance of performance to be larger among treatment stores than among control stores during the tournament, a larger treatment group serves to increase the power of our analysis (List et al. 2011). Second, a larger treatment group enables us to have a set of stores compete in both periods, to analyse potential spill-over effects. In the second tournament period, all 48 stores that did not participate in the first period were assigned to participate. In addition, we randomly selected 15 stores that did participate in the first tournament period to also participate in the second period.

All communication regarding the tournaments was sent through regular company channels using company material. Store personnel was unaware of our involvement in the incentive scheme. Prior to the first tournament period, the company's management announced that several incentive events would be held in the near future. Stores were informed that – while each store would participate at least once – they would not necessarily participate in all events. Stores were not informed about an upcoming tournament if they would not be participating.[6] In the communication towards the stores, there was no mentioning of a specific interest in task allocation.

The tournaments were announced and explained to participating stores in the week prior to the start of each tournament. The company decided to run the tournaments under the name 'Sexy Super Cup'.[7] Participating stores received a large poster specifically designed for this event, which contained the rules of the contest and was supposed to be glued to a wall or door in the backoffice. During the tournament, each week the stores received a small poster with the ranking in their group (see Figure 2 for an example; the original poster was in Dutch), which could be glued to dedicated spaces on the large poster. This allowed stores to track their (relative) performance during the tournament.

2.3 Surveys

We have asked all store employees twice to complete an online survey. The first survey was administered prior to the first tournament period and the second survey after the first tournament period. The two surveys were identical. The goal of the surveys was to measure the importance of different considerations driving the allocation of tasks in the store, as well as employees' job satisfaction. Invitations to participate in the survey were sent on behalf of Erasmus University Rotterdam to all employees and managers of all 108 stores. The invitation for the first survey was sent three weeks before the start of the first tournament period, and the invitation for the second survey was sent one week after the end of the first tournament period. The invitations included a personal code that allowed us to link survey responses. Employees were given two weeks to complete each

[6] However, personnel may in some instances have learned of an ongoing tournament due to contacts with other stores.

[7] Pun intended.

survey. In the second week we called stores as a reminder, using a call script. Neither the surveys nor the call script mentioned the tournaments. As an incentive to complete the surveys, one randomly chosen respondent in each wave was awarded with a tablet with a retail value of 150 euro.

Figure 2: Example of intermediate ranking provided to stores

Table 1: Survey questions on task assignment

Variable	Survey question
	Please indicate which answer best describes the situation in the past two months.
	The following statements are concerned with the task assignment in your store.
Task assignment	
Ability	The division of tasks is such that everyone does that which she does best.
Preference	The division of tasks is such that everyone does that which she enjoys most.
Seniority	Employees who have been with the store longer carry out more pleasant tasks.
Fairness	The divisions of tasks is fair.
Favouritism	Friends of the manager carry out more pleasant tasks.
Efficiency	The division of tasks is such that the best possible sales performance is achieved.
Job satisfaction	How satisfied are you with your job at ...

The statements on task assignment and efficiency came with a 7-point Likert scale ranging from 'Completely disagree' to 'Compeletely agree'. The question on job satisfaction came with a 7-point Likert scale ranging from 'Very dissatisfied' to 'Very satisfied'.

In the surveys, respondents had to evaluate statements using a 7-point Likert scale. In particular, we asked employees about the importance of employee ability, employee preferences, fairness concerns, favouritism, and seniority in the determination of their store's task allocation in the past two months. Furthermore, we asked whether the store's task assignment is such that the best possible sales performance is achieved. These questions are presented in Table 1.

Conducting the survey before and after the first tournament period among both first-period treatment and first-period control stores allows us to analyse the effects of the tournament treatment on task allocation (as perceived by employees) and job satisfaction. We did not conduct a third wave of surveys after the second tournament period, as at that point all stores would have been treated at least once. This makes it impossible to distinguish between time-specific determinants of task allocation and changes in task allocation as a result of the treatment.

Store managers and store employees received similar surveys, where the wording in the survey for managers was slightly altered to reflect their specific role in allocating tasks.[8] Managers may have stronger strategic or image concerns in answering survey questions regarding task allocation than employees. Therefore, in the analysis we only use the survey responses of employees.[9] Response rates among employees were 34.5% (258 out of 747) on the first survey and 18.5% (140 out of 739) on the second survey. At the store level, we have at least one respondent for 82.4% of stores (89 out of 108) in the first survey and for 61.1% (66 out of 108) in the second survey. We will discuss possible selection bias in Section 4.

2.4 Assignment procedure

For the first tournament period, we used a stratified random assignment procedure to create balanced treatment and control groups in terms of task allocation and prior performance, as follows. First, for each of the 89 stores with at least one respondent in the first survey, we calculated the average response to the statement 'the division of tasks is such that everyone does that which she does best', measuring the importance of employee ability in allocating tasks. We ranked these stores based on this score. We added one randomly chosen store from the set of 19 stores with no survey response, at a randomly chosen rank. From this ranking of 90 stores, we created 5 strata of 18 stores where the top 18 stores constituted one stratum, as well as stores ranked 19 to 36, and so on. The remaining 18 stores with no survey response constituted a sixth stratum. Within each of these six strata, we ranked stores based on average weekly performance (sales relative to sales target) over the 36 weeks prior to the experiment. We divided each stratum into two substrata constituting the top 9 and bottom 9 stores in terms of

[8] In the survey for managers, we left out one item, the one on favouritism.

[9] Indeed, in the first survey, managers' average response to the question on the efficiency of task allocation is equal to 6.2 out of 7. This is significantly higher than employees' average response of 5.2 to this question.

prior performance. Finally, we randomly selected five stores out of the 9 stores in each of the 12 substrata to participate in the first tournament period. Hence, the treatment group comprises 60 stores and the control group contains 48 stores.

In the second tournament period, all stores in the first-period control group participated in the tournament because the companies' management wanted all stores to participate in at least one tournament. In addition, we selected 15 stores from the first-period treatment group. We randomly selected one first-period treatment store from each of the 12 substrata created in the first-period assignment procedure and added 3 more stores randomly chosen out of the remaining first-period treatment stores. The remaining first-period treatment stores did not participate in the second period competition.

In assigning the participating stores to groups of three stores in the tournament, we aimed at maximising the level of competition by grouping the stores together on the basis of past performance. For the first tournament period, we ranked the participating stores by cumulative sales over 36 weeks prior to the experiment. Next, we grouped the three best performing stores together, as well numbers 4 to 6, and so on. To reduce possible sabotage opportunities, we made in total 4 adjustments to prevent that stores from the same regional area (Dutch province) were grouped together. For the second tournament period, we followed a similar procedure. We ranked the participating stores based on cumulative past performance over 45 weeks prior to the tournament period and divided them into groups of three similarly performing stores. This time, we had to make 8 adjustments to prevent stores within the same regional area from competing with one another. Upon informing the companies' management of the assignment, we learned that two stores, which had not participated in the first tournament period, would be closed for refurbishment during the second tournament period. We drop these stores' observations during and after their refurbishment from the analysis. We did not replace these two stores in the tournament, so that in two second-period tournament groups only two stores competed.[10] Hence, 61 stores participated in the second tournament period.

3 Estimation

We estimate the effect of participating in the experiment on weekly performance, task allocation, and job satisfaction using OLS with period-fixed effects and either store- or individual-fixed effects. The average treatment effect on stores' performance is estimated by

$$Y_{s,w} = \alpha_s + \gamma_w + \beta T_{s,w} + \varepsilon_{s,w}, \tag{2}$$

where $Y_{s,w}$ is weekly performance of store s in week w, measured as actual sales revenue over targeted sales: $Y_{s,w} = \dfrac{r_{s,w}}{b_{s,w}}$. Store and week-fixed effects are given by α_s and γ_w,

[10] We keep these four stores in the analysis below. None of the results change if we drop these stores after the first tournament period.

respectively. $T_{s,w}$ is a dummy denoting whether store s is participating in a tournament in week w, so that β measures the average treatment effect. The error term $\varepsilon_{s,w}$ is clustered at the store level to account for possible serial correlation.[11] We will also estimate the effects separately for the first and second tournament period.

For the first tournament period, stores have been randomly allocated to either treatment or control, implying that in expectation the control group provides a reliable counterfactual for the performance of the treatment stores in the tournament. In the second tournament period, however, all non-participating stores have participated in the first period. If participation leads to carry-over effects, for instance due to learning or fatigue, these stores do not constitute a proper control group for the stores that compete for the first time in the second period. We use the 15 stores that participate in both waves to analyse possible carry-over effects. To prevent short-term carry-over effects from affecting the estimates, the three weeks in between the two tournament periods are excluded from the analysis.[12]

Estimation of the incentive effect on job satisfaction and task allocation is based on the first tournament period, as the surveys were administered before and after this period. We estimate the effect of the team incentive both at the individual employee level and at the store level. For the analysis at the store level, we use the average response on the survey items across all respondents employed in a given store. The average treatment effect is estimated by

$$R_{i,t} = \alpha_i + \gamma_t + \beta T_{i,t} + \varepsilon_{i,t}, \tag{3}$$

where $R_{i,t}$ is the survey response of unit i (individual or store) in survey t (before or after the first tournament period). Observation unit-fixed effects and survey-fixed effects are given by α_i and γ_t, respectively. $T_{i,t}$ is a dummy that takes value 1 for responses on the survey after the first-period tournament if the store (of individual) i was part of the first-period treatment group. Hence, β measures the average treatment effect. Finally $\varepsilon_{i,t}$ is the error term, clustered at the store level.

The store level is the natural level of analysis, as we randomized at the store level. Furthermore, task assignment affects the team as a whole. However, there are three caveats in analyzing the results at the store level. First, for a given store the respondents to the first survey may differ from the respondents to the second survey. Insofar as selection into and out of the survey is correlated with assignment to the first-period treatment group, this may bias the results. Below, we analyse the self-selection of employees into answering the first and second survey and find no indication of selection related to first-period assignment. Second, stores are given the same weight in the

[11] For ease of interpretation, we estimate the incentive effect on weekly performance as opposed to cumulative performance (as given by (1)). This is inconsequential for the estimation results since (1) determines performance by comparing total sales to total targeted sales over the tournament period. Thus sales staff cannot strategically focus efforts on apparantly "easy" weeks as a sale counts equally towards performance regardless of the week in which it occurs.

[12] Including these weeks in the analysis does not influence the results.

analysis independent of the number or fraction of employees that answered a given survey. Weighing stores by the number of respondents (in either the first or the second survey) does not affect the estimates. Third, store-level averages mask within-store differences in responses. Among the questions on task allocation in the first survey, the fraction of total variation explained by store-fixed effects ranges from 0.34 to 0.44. Hence, there is sizable heterogeneity across stores, but also considerable differences within stores. Therefore, we also present the average treatment effects estimated at the individual employee level, accounting for individual-fixed effects.

4 Descriptive statistics

We have weekly data on store performance covering a year starting in February 2013. In addition, prior to the first tournament period we received personnel data of all stores. This includes information on employees' age, tenure, contractual hours (measure in full-time equivalent, fte), and position. Figure 3 depicts weekly targeted and actual sales averaged across stores for the 52 weeks in our dataset. Sales is highly volatile, but most of it is predicted by the company's management as sales and targeted sales follow by and large the same pattern.

Figure 3: Average sales and average targeted sales

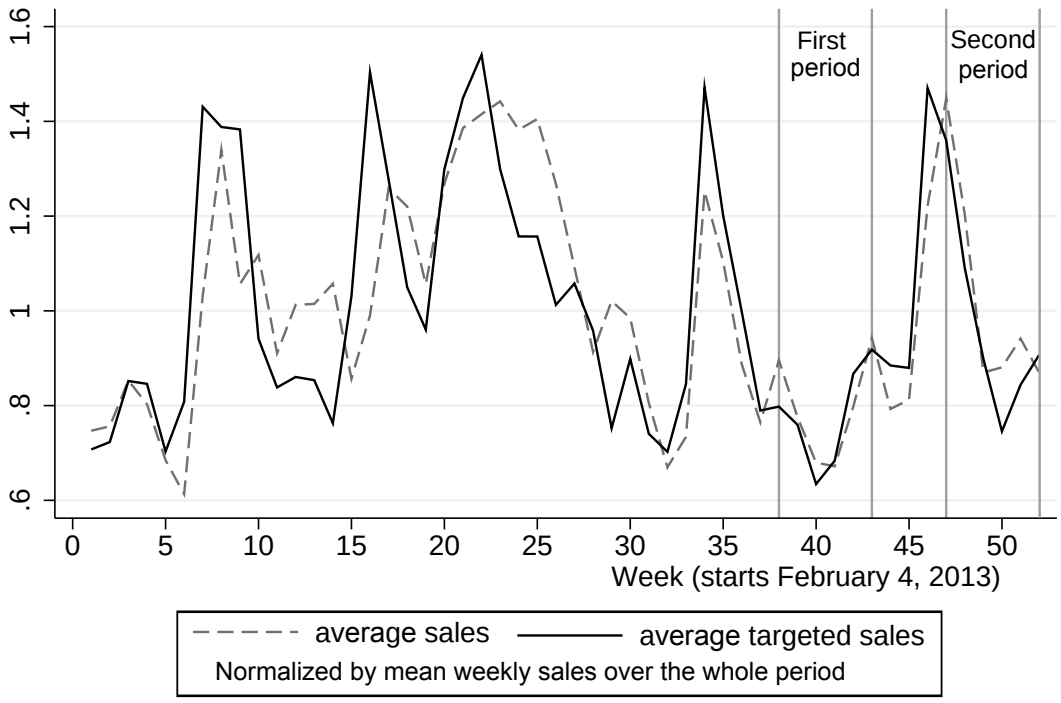


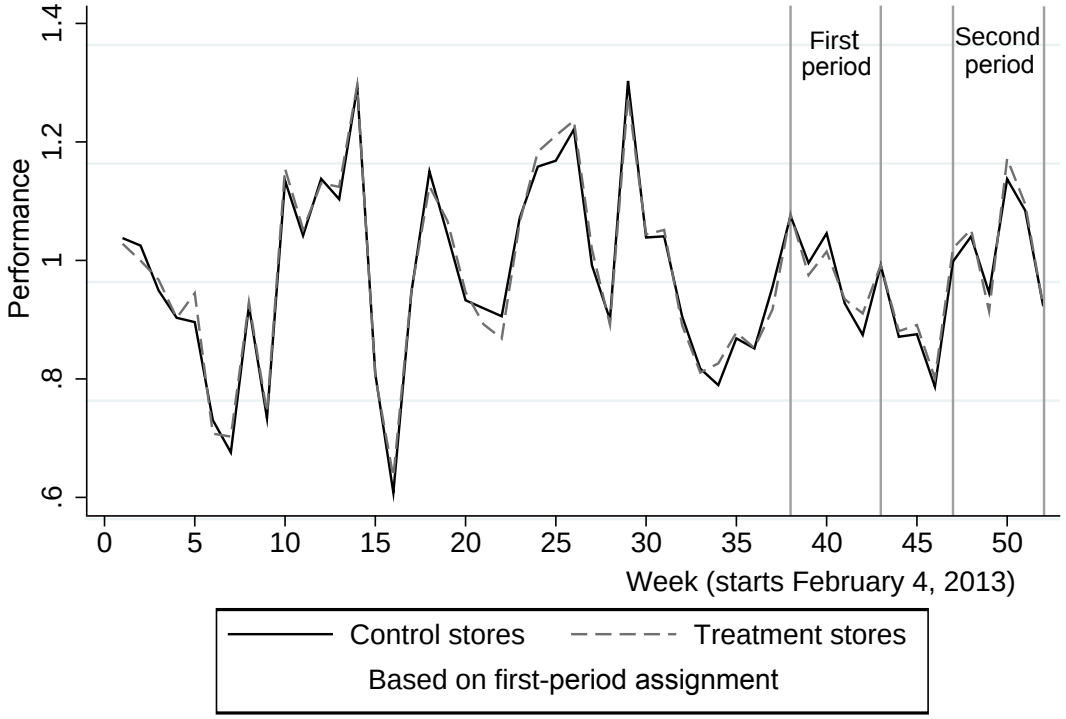
Table 2: Store characteristics, by first-period assignment

	Total		Control		Treatment	
	mean	sd	mean	sd	mean	sd
Average prior performance	0.95	0.08	0.95	0.10	0.96	0.07
Team size	7.91	3.30	8.19	3.61	7.68	3.04
Aides	1.91	1.56	1.75	1.62	2.03	1.51
Age manager	43.79	11.06	43.53	11.70	44.00	10.62
Tenure manager	15.43	13.88	13.86	8.94	16.67	16.80
Fte manager	0.89	0.08	0.89	0.07	0.88	0.09
Average age staff	38.40	7.06	38.33	6.77	38.46	7.34
Average tenure staff	8.70	4.94	9.05	5.34	8.42	4.61
Average fte staff**	0.48	0.13	0.51	0.13	0.45	0.12
Observations	108		48		60	

*, **, and *** indicate that the difference between treatment stores and control stores is statistically significant at the $p < .1$ level, $p < .05$ level, and $p < .01$ level, respectively.

Descriptive statistics are given in Table 2. The first column shows that, on average, stores' sales fell short of sales targets by about 5 percent. Stores employ on average 8 individuals (including the manager), of whom two work on-call. Most regular employees work part-time. All employees and managers are female. Table 2 also reports these statistics separately for the treatment group and the control group in the first tournament period. This shows that the two groups are similar in terms of past performance and personnel characteristics. The only exception is average contract size ('fte') which is slightly but significantly larger in the control stores. Figure 4 depicts average performance separated by first-period assignment. This shows that both prior to and during the tournament periods performance in both groups was similar.

Figure 4: Performance of treatment and control stores.



Besides the performance and personnel data, we also have survey data. The response rates of the two surveys were 34.5% (258 out of 747) and 18.5% (140 out of 739) for employees on the first and second survey respectively. 92 employees completed both surveys. We averaged employees' survey responses by store. This results in 89 and 66 stores with staff survey data for the first and second survey, respectively, and 60 stores for which we have respondents for both surveys. Due to some item non-response, the exact number of stores with at least one response varies a bit across questionnaire items.

The left-hand side of Table 3 reports descriptive statistics of the first employee survey, conducted before the first tournament period. On average, employees indicate that fairness considerations and employee ability are the most important drivers of task assignment in their store. Employee preferences matter to a smaller extent while favouritism and seniority are not perceived as important drivers of task assignment. We find limited differences across first-period treatment and control stores. Only favouritism is perceived as slightly more important by employees in treatment stores as compared to employees in control stores (p-value 0.098). The average level of job satisfaction is also similar across treatment and control.[13]

[13] Across the different determinants of task allocation as listed in Table 1, between 34% and 44% of variation in responses to the first survey across employees is explained by store-fixed effects. Furthermore, restricting attention to stores that were in the control group during the first experimental period, correlations of average responses at the store level between the first and second survey range between 0.27 and 0.60. Hence, our measures of task allocation probably contain some measurement error through individuals' subjective assessment, but do seem to capture the underlying actual task allocation.

Table 3: Results of first and second survey, store averages

	Survey 1				Survey 2			
	Control		Treatment		Control		Treatment	
	mean	sd	mean	sd	mean	sd	mean	sd
Job satisfaction	5.17	1.08	5.35	1.00	5.51	0.94	5.63	1.08
Task allocation:								
Ability	4.09	1.04	4.05	1.23	4.42	1.17	4.11	1.33
Preference	3.24	1.06	3.35	1.07	3.71*	0.94	3.29	0.99
Seniority	2.08	0.81	2.27	0.85	2.36	0.91	2.42	1.09
Fairness	5.28	1.37	5.55	0.92	5.60	0.75	5.33	1.20
Favouritism	2.09*	0.85	2.43	1.04	2.30	0.93	2.38	0.96
Efficiency	5.14	1.23	5.17	1.00	5.57	0.76	5.24	1.12
Observations	34-39		47-50		27-28		33-38	

*, **, and *** indicate that the difference between treatment stores and control stores in a given survey is statistically significant at the $p < .1$ level, $p < .05$ level, and $p < .01$ level, respectively.
Table 1 gives the exact wording of the questionnaire items. The exact number of stores varies across questionnaire items due to item non-response.

Table 4 presents correlations between the store averages on the main items in the first survey. There is a rather strong positive correlation between the perceived importance of employee ability and the perceived importance of employee preferences in task allocation as well as between the perceived importance of seniority and favouritism. Not surprisingly, the perceived importance of favouritism and seniority both correlate negatively with fairness considerations. Interestingly, the extent to which task allocation is geared towards sales performance ('efficiency') is most strongly related to fairness considerations, and relates positively (negatively) to the importance of employee ability and employee preferences (favouritism and seniority) in task allocation. This could reflect that people are more productive under a fair task allocation. Another interpretation is that an efficient allocation is considered to be fair. Job satisfaction is most strongly related to the perceived importance of fairness and is also positively correlated with the importance of employee ability and employee preferences.

Table 4: Correlations between drivers of task allocation in first survey, store-level averages

Task Allocation	Ability	Preference	Seniority	Fairness	Favouritism	Efficiency
Preference	0.560***	1				
Seniority	0.118	0.073	1			
Fairness	0.222**	0.310***	-0.406***	1		
Favouritism	0.008	0.020	0.746***	-0.401***	1	
Efficiency	0.372***	0.423***	-0.333***	0.773***	-0.283***	1
Job Satisfaction	0.290***	0.300***	0.029	0.352***	-0.090	0.318***

*, **, and *** indicate that the coefficient is statistically significantly different from zero at the $p < .1$ level, $p < .05$ level, and $p < .01$ level, respectively.

The right-hand side of Table 3 gives the outcomes of the second survey, conducted after the first tournament period. Compared to the first survey, job satisfaction is somewhat higher but not specifically among the treatment stores. The difference in the perceived importance of favouritism in task allocation between first-period treatment and control stores is no longer present in the second survey. We do find that employee preferences are considered less important for task assignment among stores that did participate in a first-period tournament than among stores that did not participate.

A concern in evaluating the effects of the team incentive on job satisfaction and task allocation is that only part of the employees completed the surveys, which may lead to selection effects. When this selection is related to stores' assignment to the treatment or control group, it yields biased estimates of the treatment effects. For the first survey, self-selection is unlikely to be problematic. At that point, employees were not yet aware of the upcoming experiment, so that self-selection cannot be based on assignment to treatment or control. Panel A in Table A.1 in the Appendix shows that first-survey respondents are older than non-respondents, have a longer history with the company, and work more hours, suggesting that employees with a stronger connection to the firm were more likely to complete the survey. Separating this by first-period assignment, we find indeed that across the treatment and control groups a comparable set of employees completed the survey. A similar pattern arises in the second survey as shown in Panel B. Furthermore, survey attrition is not significantly related to first-period treatment assignment. Conditional on answering the first survey, 38% of the employees in the control group answered the second survey against 34% in the treatment group.

Aggregating the survey data to store level, Table A.2 shows that stores with and without respondents to the first survey are comparable in terms of observable characteristics. For the second survey, we do find some differences between stores with and without survey respondents. Stores that performed relatively well before the experiment, that are headed by an older and more experienced manager, and that have a larger team are more likely to have at least one employee responding to the survey. Comparing the differences between stores with and without survey across the treatment and the control group, we find that these patterns arise in both groups (not reported for brevity). Hence, both the individual-level and the store-level data show that self-selection into the first and second survey appears to be unrelated to treatment assignment.

Self-selection can also be related to stores' task allocation. Again, this type of self-selection is most problematic when it differs between stores in the treatment and the control group. As assignment to treatment and control was stratified by the response to the first survey, self-selection into the first survey is unlikely to affect our estimated treatment effects. Furthermore, we can analyse whether self-selection into the second survey conditional on first-survey responses differs between treatment and control stores. Panel A in Table A.3 reports the average response to the key questions in the first survey, comparing employees who only responded to the first survey with employees

who responded to both surveys, separated by first-period assignment. This shows no substantial differences between employees who only responded to the first survey with employees who responded to both surveys. None of these differences differs significantly between the treatment group and the control group. Panel B in Table A.3 reports average responses to the second survey comparing employees who answered only to the second survey with employees who answered both surveys. In the control group, employees who completed both surveys indicate significantly higher importance of seniority and favouritism in their stores' task allocation compared to employees who only respond to the second survey. A similar, but less pronounced pattern arises in the treatment group. Comparing these differences across the treatment and control group, only for ability we find a statistically significant difference (p-value 0.06). Employees in the control group who answered only to the second survey indicate lower importance of ability than employees who answered both surveys, while the reverse holds in the treatment group. Assuming that the non-participation of respondents in the first survey is unrelated to treatment assignment, this implies that we may underestimate the treatment effect on the importance of ability in task allocation in the estimations at the employee level.

In Table A.4, we report similar figures aggregated at store level, comparing stores where at least one employee answered to each of the surveys with stores where none of the employees responded to one of the surveys. At the store level, we find that none of the differences between these types of stores differs significantly between the treatment and the control group. All together, the available evidence suggests that self-selection into the surveys does not affect our estimates.

5 Results

The first column of Table 5 gives the results of estimating (2). The estimated treatment effect of participating in a tournament is a reduction in performance by 0.6 percentage points. This effect is precisely estimated, with a standard error of 0.8 percentage points. Hence, the 95 percent confidence interval of the average treatment effect lies between -2.1 and 1.0 percentage point. In Column 2 of Table 5, we separate the average treatment effect by tournament period. In both periods, the estimated effect is small and statistically insignificant, and the difference between the estimated treatment effects for the two periods is small as well. In the third column, we estimate the effect of participating in the second tournament period separately for the 15 stores who had also participated in the first period, to establish whether there are carry-over effects of participating in the first tournament to performance during the second tournament period. We find no statistically significant carry-over effect, suggesting that having participated in the first tournament period does not affect stores' response during the second tournament

period.[14]

Table 5: Estimation performance effect (intermediate weeks discarded)

	Dependent variable: Performance			
Treatment	-0.006			
	(0.008)			
Treatment in period 1		-0.004	-0.004	
		(0.013)	(0.013)	
Treatment in period 2		-0.008	-0.004	
		(0.014)	(0.016)	
Carry-over effects			-0.013	
			(0.022)	
Placebo-treatment				0.002
				(0.011)
Store-fixed effects	yes	yes	yes	yes
Week-fixed effects	yes	yes	yes	yes
Observations	5274	5274	5274	3880
Stores	108	108	108	108
within R^2	0.565	0.565	0.565	0.615

Standard errors clustered at the store level in parentheses.
* and ** indicate that the coefficient is statistically significantly different from zero at the $p < .1$ level and $p < .05$ level, respectively.

The absence of a positive treatment effect is not due to low power. Given our difference-in-difference design, the power of our analysis depends on serial correlation across observations within stores (Bertrand et al. 2004). Using the period before the experiment took place (weeks 1 – 36), we regress performance on store- and week-fixed effects (as we also do in the main analysis). The residuals of this regression have a standard deviation of 0.12, as well as an estimated first-order auto-correlation of 0.19 and effectively no higher-order auto-correlation. We are not aware of an exact way to determine the power of an analysis in the presence of serial correlation. It is possible to give upper and lower bounds, though. In the absence of correlation across observations within stores, our design would allow us to detect an effect of 1.9 percentage point with a power of 0.8. On the other hand, if all observations within stores would have a 0.19 correlation, we could detect an effect of 3.3 percentage points with 0.8 power. As only a small subset of observations within stores are correlated, the power of our analysis will be closer to the former than to the latter.[15]

[14] Within a tournament period, there is limited variation in the estimated treatment effect across weeks. In the first (second) half of the first tournament period, the estimated treatment effect is -1.8% (1.5%). In the first (second) half of the second tournament period, the estimated treatment effect is -0.2% (-1.3%). None of these coefficients differs significantly from zero.

[15] As shown by Bertrand et al. (2004), clustering standard errors at store level corrects for any serial correlation within stores. Consistent with the presence of weak serial correlation in the data, clustering increases the standard errors of the estimated treatment effects to a limited extent. An alternative way to handle serial correlation is to remove the time dimension from the data. Hence, we collapse our data into three periods, by taking average store performance before the experiment, during the first

Despite the random assignment, it could be that relatively many of the stores that participated in the first tournament period experienced a positive shock to performance just before that period. During the tournament period, performance may have moved back to normal levels, giving rise to a downward bias in the estimated treatment effect. To assess this possibility, we pretend a tournament took place in the period before the first real tournament, i.e. in weeks 31 to 36 in our data-set (see Figure 4).[16] Thereto, we construct a dummy that takes value 1 in these six weeks for all stores that take part in the first tournament period. Column 4 of Table 5 gives the results of estimating the effect of this 'placebo-treatment', dropping all weeks afterwards. The estimated effect is very close to zero. Hence, the stores that participated in the first tournament period did not experience a positive shock to performance in the weeks before the tournament.

The absence of a positive average treatment effect on performance does not necessarily imply that the team incentive did not affect behaviour of managers and employees. The team incentive may have induced them to try out new ways of improving performance, for instance by making changes to task assignment in stores. Such attempts could be successful in some stores but fail in others. If so, we would observe an average treatment effect close to zero accompanied by a relatively large standard error. Comparing the standard errors on the actual treatment in Column 2 with the standard error on the placebo-treatment in Column 4 (which are all based on six-week periods), we find that the standard errors on the actual treatment are only slightly higher. Hence, the zero average treatment effect does not mask a large increase in the heterogeneity of store performance.[17]

Next, we use the survey data to assess whether employees perceived a change in task allocation in the first tournament period in response to the treatment. Table 6 gives the estimated effects of participating in a tournament in the first period on task allocation within teams, estimated at the store level. In contrast to our predictions, we do not find that our treatment increased the importance of employee ability in allocating tasks. The point estimate is negative, but not significantly different from zero. For the other considerations, we also find negative point estimates, all of them insignificant except for the importance of employee preferences in task allocation. Estimating these effects at the individual employee level yields a similar picture, as presented in Table 7. Summarising, while the reduced emphasis on fairness, favouritism, and in particular employee preferences in allocating tasks is in line with our predictions, this should have been accompanied by an increased emphasis on employee ability. Instead we find an insignificant negative effect. Overall, our findings suggest that the introduction of team

experimental period, and during the second experimental period. Running the difference-in-difference estimation using only these three periods, we obtain very similar results: the estimated effect of the treatment equals -0.004, with a standard error of 0.008.

[16] We leave out the week before the first tournament period (week 37), as that may pick up an effect of the announcement of the tournament.

[17] All results discussed carry over when we focus only on the effects in the first week or the first three weeks of each treatment period.

incentives had little effect on the allocation of tasks within the teams.[18]

Table 6: Task allocation estimates, store level

Task allocation:	Ability	Preferences	Seniority	Fairness	Favouritism
Treatment	-0.243	-0.562*	-0.136	-0.450	-0.300
	(0.324)	(0.287)	(0.267)	(0.321)	(0.292)
Store-fixed effects	yes	yes	yes	yes	yes
Period-fixed effects	yes	yes	yes	yes	yes
Stores	60	60	60	60	60
within R^2	.009	0.101	0.015	.039	.028

Standard errors clustered at the store level in parentheses.
Dependent variables measured on a 7-point Likert scale, see Table 1 for the exact wording of the survey questions.
* and ** indicate that the coefficient is statistically significantly different from zero at the $p < .1$ level and $p < .05$ level, respectively.

Table 7: Task allocation estimates, worker level

Task allocation:	Ability	Preferences	Seniority	Fairness	Favouristism
Treatment	-0.352	-0.750**	0.114	-0.412	-0.006
	(0.286)	(0.325)	(0.223)	(0.325)	(0.386)
Worker-fixed effects	yes	yes	yes	yes	yes
Period-fixed effects	yes	yes	yes	yes	yes
Workers	91	91	92	90	92
within R^2	0.013	0.067	0.091	.027	.009

Standard errors clustered at the store level in parentheses.
Dependent variables measured on a 7-point Likert scale, see Table 1 for the exact wording of the survey questions.
* and ** indicate that the coefficient is statistically significantly different from zero at the $p < .1$ level and $p < .05$ level, respectively.

Table 8 reports the estimated effects of the treatment on employees' job satisfaction at the store level and at the worker level. The first and third column show that the average treatment effect is small and statistically insignificant. This average treatment effect may mask heterogeneity across treated stores, in particular between stores that won their tournament and stores that did not win. However, we cannot estimate the effect of winning a tournament by simply including a dummy for stores that won their tournament, as performing relatively well may also affect job satisfaction in the absence of an tournament incentive. Hence, in order to differentiate between the effects of winning the tournament and the effect of attaining relatively high performance, we determine 'winners' and 'losers' of a pseudo-competition among stores that were part of the control group in the first tournament period. The pseudo-competition was conducted as follows.

[18] In line with this interpretation, we also find a small and insignificant treatment effect on employees' perception regarding the 'efficiency' of their store's task allocation (point estimate of -0.108; standard error of 0.239).

We assigned the control stores to groups of three in the same manner as we did with the treatment stores for the actual tournament. Next, we determined for each group of control stores the 'winning' store, based on stores' cumulative performance during the first tournament period. The second and fourth column of Table 8 give the treatment effect for winning and non-winning stores separately. The first coefficient gives the treatment effect on non-winning stores, which is statistically insignificant both in the store-level estimation and in the worker-level estimation. Hence, participating in a tournament without winning it does not affect job satisfaction significantly compared to stores that did not participate and performed relatively poor as well. The second coefficient shows that job satisfaction goes down in control stores that outperformed two similar control stores during the tournament period, and significantly so in the store-level estimation. One explanation is that the higher performance is due to higher employee effort. Winning an actual tournament mitigates this effect, as seen by the third coefficient, although the effects are not statistically significant.

Table 8: Job satisfaction

	Dependent variable: Job satisfaction			
	Store level		Worker level	
Treatment	0.212	0.154	0.328	-0.356
	(0.333)	(0.366)	(0.307)	(0.374)
Best in group		-1.241**		-0.599
		(0.608)		(0.588)
Best in treatment		0.575		0.281
		(0.707)		(0.648)
Store/employee-fixed effects	yes	yes	yes	yes
Period-fixed effects	yes	yes	yes	yes
Stores / workers	53	53	59	59
within R^2	0.055	0.189	0.030	0.066

Standard errors clustered at the store level in parentheses.

Dependent variable measured on a 7-point Likert scale, see Table 1 for the exact wording of the survey question.

* and ** indicate that the coefficient is statistically significantly different from zero at the $p < .1$ level and $p < .05$ level, respectively.

6 Discussion

Overall, our results show that the introduction of the team incentive neither affected team performance nor task assignment within teams. These results are in contrast to the results of earlier studies in the literature. An important question is: why? One way to address this question is to compare studies along important dimensions, and we will do so in the remainder of this section. However, before we delve into this, we would like to stress that the number of existing empirical studies providing causal estimates on the

effects of team incentives is still quite small. In building up a body of evidence, it should be no surprise to sometimes find contrasting evidence (Antonakis 2017).

Our result on the effect of team incentives on task assignment is in contrast to Bandiera et al. (2007) and Burgess et al. (2010), who find that supervisors directed more competent workers towards the incentivized tasks. In contrast to these earlier studies, we do not observe the tasks actually performed by workers. Instead, we use subjectively reported drivers of store's task assignment. This indirect method may underestimate changes in actual task allocation. Alternatively, the bonus offered may have been too low to induce changes in task assignment, which is in line with the absence of an overall treatment effect. For managers, the monetary incentive offered was comparable to the incentives offered in Burgess et al. (2010), but considerably lower than the incentive offered in Bandiera et al. (2007), as discussed in detail in subsection 2.2.

In our earlier field experiments with comparable designs and rewards conducted in other retail chains, we found average treatment effects on performance varying from 0%, 1.5%, to 5%, the latter two statistically significantly different from zero (Delfgaauw et al. 2013, 2014, 2015). Friebel et al. (2017) report a significant increase in sales of about 3% after the introduction of a team bonus for meeting sales targets in a retail setting. Hence, in earlier work, similar incentive schemes did induce higher performance.[19] Furthermore, part of our treatment entailed the provision of relative performance feedback. Studies in various settings find that the provision of relative performance feedback alone, absent relative performance pay, induces higher performance (Azmat and Iriberri 2010, 2016, Blanes-i-Vidal and Nossol 2011, Bradler et al. 2016, Delfgaauw et al. 2013, Kosfeld and Neckermann 2011, Kuhnen and Tymula 2012). Barankay (2012) and Bandiera et al. (2013), in contrast, find negative effects of relative performance feedback on performance at work. This suggests that the bonus level alone cannot explain the absence of a positive average treatment effect.[20]

[19] While the bonus in our field experiment is equal to about 2.5% of earnings over a 6-week period and paid to a third of the participating employees (i.e. those of the winning store in a group of three stores), in Delfgaauw et al. (2013) employees could earn a bonus of 3.8% of earnings over a 6-week period, but this bonus was paid only to a fifth of the participating employees as stores competed in groups of five stores. However, the bonus scheme in that field experiment also included a bonus of 1.9% for the employees of the runner-up. Note, however, that Delfgaauw et al. (2013) also studied a treatment where stores compete in groups of five without any monetary prizes, yielding effects on performance of similar size. In Delfgaauw et al. (2014), employees could earn a bonus of 5% of monthly pay when the store would outperform a benchmark by a small margin and a bonus of 10% when outperforming the benchmark by a large margin. It turned out to be hard to qualify for the bonus: less than 11% of the stores earned a bonus, with half of these winning the low bonus. In Delfgaauw et al. (2015), teams entered elimination tournaments lasting a maximum of two times four weeks, with expected earnings of about 2% of monthly earnings, with prize money ranging from 1.2% to 6% of monthly earnings. Lastly, in Friebel et al. (2017), about 40% of the employees received a bonus of on average 4% at least once, implying an increase in total wage compensation of about 2%. We conclude that, by and large, the strength of the monetary incentives in these closely related studies is comparable to those in our study. Moreover, there is no clear relation between the strength of the incentive and the performance effect, which ranges from a null finding in Delfgaauw et al. (2014), to an increase in sales growth of about 5 to 7 percentage points in Delfgaauw et al. (2013), a 1,5% increase in the average number of products per customer in Delfgaauw et al. (2015), to a 3% increase in sales in Friebel et al. (2017).

[20] Several studies show a non-monotonic relation between the level of incentives and performance,

Another possible explanation for the absence of an effect of our treatment is that the competitive element in our incentive design did not strike a chord due to the all-female composition of the teams. Gneezy et al. (2003) find in a lab experiment that females respond less to competitive incentives than males. Subsequent studies show that this gender difference in the response to competition depends on the specific task and the environment (Niederle and Vesterlund 2011). Studying a competitive business game played in groups of three students, Apesteguia et al. (2012) finds that all-female teams perform worse than any other team in terms of gender composition. However, Delfgaauw et al. (2013) implemented similar tournaments among stores of another retail chain and found that the treatment effect increases significantly in the fraction of female employees, provided that the store manager is female (which holds for all stores in the current study). Hence, this suggests that the all-female team composition likely does not drive the lack of response.[21]

According to the company's management, the employees were actively engaged in the tournament. In some stores, the weekly rankings were eagerly awaited. However, the company's management perceived that many employees faced difficulties translating their engagement into higher sales, possibly due to a lack of skills to recognize and act on sales opportunities. Arguably, if employees are uncertain about how to increase performance, incentives may have little effect, at least in the short-run. After the current experiment, the company decided to invest in commercial training of its employees. Furthermore, it adopted an incentive scheme based on individual performance, which (unfortunately for us) was implemented in all stores at once.

References

[1] Antonakis, J. (2017), On doing better science: From thrill of discovery to policy implications, *Leadership Quarterly*, 28(1), 5-21.

[2] Apesteguia, J., Azmat, G., & Iriberri, N. (2012), The impact of gender composition on team performance and decision-making: Evidence from the field, *Management Science*, 58(1), 78–93.

where performance is lower for weak incentives than in the absence of incentives (Gneezy and Rustichini 2000, Gneezy and Rey-Biel 2014). Hence, it could be that the average treatment effect would have been higher if we had only provided relative performance information. Still, our earlier experiments suggest that the current level of bonus pay can induce positive treatment effects.

[21] Ceiling effects could be another explanation. If many stores perform as good as it gets, there is no room for improvement. Figure 1 suggests otherwise, however. As a further check we ran regression (2) adding the interaction between the first-period treatment dummy and the stores' average response to the question on efficiency in the first survey. This yields a negative coefficient for the interaction, suggesting that stores with more room for improvement respond relatively stronger to the team incentive. However, the effect is not statistically significant.

[3] Azmat, G., & Iriberri, N. (2010), The importance of relative performance feedback information: Evidence from a natural experiment using high school students, *Journal of Public Economics*, 94(7), 435-452.

[4] Azmat, G., & Iriberri, N. (2016), The provision of relative performance feedback: An analysis of performance and satisfaction, *Journal of Economics and Management Strategy*, 25(1), 77-110.

[5] Babcock, P., Bedard, K., Charness, G., Hartman, J., & Royer, H. (2015), Letting down the team? Evidence of social effects of team incentives, *Journal of the European Economic Association*, 13(5), 841-870.

[6] Bandiera, O., Barankay, I., & Rasul, I. (2007), Incentives for managers and inequality among workers: evidence from a firm-level experiment, *Quarterly Journal of Economics*, 122(2), 729-773.

[7] Bandiera, O., Barankay, I., & Rasul, I. (2009), Social connections and incentives in the workplace: Evidence from personnel data, *Econometrica*, 77(4), 1047-1094.

[8] Bandiera, O., Barankay, I., & Rasul, I. (2013), Team incentives: evidence from a firm-level experiment, *Journal of the European Economic Association*, 11(5), 1079-1114.

[9] Barankay, I. (2012), Rank Incentives: Evidence from a Randomized Workplace Experiment, *mimeo*, University of Pennsylvania.

[10] Bertrand, M., Duflo, E., & Mullainathan, S. (2004), How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, 119(1), 249-275.

[11] Blanes-i-Vidal, J., & Nossol, M. (2011), Tournaments without prizes: evidence from personnel records, *Management Science*, 57(10), 1721-1736.

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

[13] Burgess, S., Propper, C., Ratto, M., Scholder, K., von Hinke, S., & Tominey, E. (2010), Smarter Task Assignment or Greater Effort: The Impact of Incentives on Team Performance, *Economic Journal*, 120(547), 968-989.

[14] Casas-Arce, P., & Martinez-Jerez, F. A. (2009), Relative performance compensation, contests, and dynamic incentives, *Management Science*, 55(8), 1306-1320.

[15] Chen, H., and Lim, N. (2013), Should managers use team-based contests?, *Management Science*, 59(12), 2823-2836.

[16] Cooper, D.J., & Sutter, M. (2017), Endogenous role assignment and team performance, *International Economic Review*, forthcoming.

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

[18] Delfgaauw, J., Dur, R., Non, A., & Verbeke, W. (2014), Dynamic incentive effects of relative performance pay: a field experiment, *Labour Economics*, 28, 1-13.

[19] Delfgaauw, J., Dur, R., Non, A., & 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.

[20] Erev, I., Bornstein, G., & 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.

[21] Friebel, G., Heinz, M., Krüger, M., & Zubanov, N. (2017), Team incentives and performance: Evidence from a retail chain, *American Economic Review*, 107(8), 2168-2203.

[22] Fryer, R. (2013), Teacher incentives and student achievement: Evidence from New York City public schools, *Journal of Labor Economics,* 31(2), 373-407.

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

[24] Gneezy, U., Niederle, M., & Rustichini, A. (2003), Performance in competitive environments: Gender differences, *Quarterly Journal of Economics*, 118(3), 1049-1074.

[25] Gneezy, U. & Rey-Biel, P. (2014), On the relative efficiency of performance pay and noncontingent incentives, *Journal of the European Economic Association*, 12(1), 62-72.

[26] Gneezy, U., & Rustichini, A. (2000), Pay enough or don't pay at all, *Quarterly Journal of Economics*, 115(3), 791-810.

[27] Goodman S.F., & Turner, L.J. (2013), The design of teacher incentive pay and educational outcomes: Evidence from the New York City bonus program, *Journal of Labor Economics,* 31(2), 409-420.

[28] Green, C., & Heywood, J. S. (2008), Does performance pay increase job satisfaction?, *Economica*, 75(300), 710-728.

[29] Heywood, J. S., & Wei, X. (2006), Performance pay and job satisfaction, *Journal of Industrial Relations*, 48(4), 523-540.

[30] Kosfeld, M., & Neckermann, S. (2011), Getting more work for nothing? Symbolic awards and worker performance, *American Economic Journal: Microeconomics*, 3(3), 86-99.

[31] Kuhnen, C.M., & Tymula, A. (2012), Feedback, self-esteem, and performance in organizations, *Management Science,* 58(1), 94-113.

[32] Lavy, V. (2002), Evaluating the effect of teachers' group performance incentives on pupil achievement, *Journal of Political Economy*, 110(6), 1286-1317.

[33] McCausland, W. D., Pouliakas, K., & Theodossiou, I. (2005), Some are punished and some are rewarded: a study of the impact of performance pay on job satisfaction, *International Journal of Manpower*, 26(7/8), 636-659.

[34] List, J., Sadoff, S., & Wagner, M. (2011), So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design, *Experimental Economics*, 14, 439-457.

[35] Muehlheusser, G., Schneemann, S., & Sliwka, D. (2016), The impact of managerial change on performance: The role of team heterogeneity, *Economic Inquiry*, 54(2), 1128-1149.

[36] Muralidharan, K., & Sundararaman, V. (2011), Teacher performance pay: Experimental evidence from India, *Journal of Political Economy*, 119(1), 39-77.

[37] Niederle M., & Versterlund, L. (2011) Gender and competition, *Annual Review of Economics*, 3, 601–630.

[38] Petrescu, A. I., & Simmons, R. (2008), Human resource management practices and workers' job satisfaction, *International Journal of Manpower*, 29(7), 651-667.

[39] Springer, M.G., Ballou, D., Hamilton, L., Le, V.-N., Lockwood, J.R., McCaffrey, D.F., Pepper, M., & Stecher, B.M. (2010), Teacher pay for performance: Experimental evidence from the project on incentives in teaching, Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

[40] Welbournel, T. M., & Cable, D. M. (1995), Group incentives and pay satisfaction: Understanding the relationship through an identity theory perspective, *Human Relations*, 48(6), 711-726.

[41] Zehnder, C., Herz, H., & Bonardi, J.-P. (2017), A productive clash of cultures: Injecting economics into leadership research, *Leadership Quarterly*, 28(1), 65-85.

A Appendix

Table A.1: Characteristics of first and second survey (non-)respondents, by treatment

A: Survey 1						
	Overall		Control		Treatment	
	Non-Resp.	Resp	Non-Resp.	Resp	Non-Resp.	Resp
Tenure	7.28***	9.34	7.31**	9.60	7.24**	9.12
	(7.71)	(8.48)	(7.82)	(8.78)	(7.64)	(8.24)
FTE	0.44***	0.54	0.47***	0.58	0.41***	0.51
	(0.33)	(0.30)	(0.33)	(0.28)	(0.33)	(0.32)
Age	35.56***	39.49	35.00***	39.87	36.07**	39.17
	(13.78)	(13.03)	(13.34)	(13.02)	(14.17)	(13.09)
Observations	473-483	254-257	224-227	117-118	249-256	137-139
B: Survey 2						
	Overall		Control		Treatment	
	Non-Resp.	Resp	Non-Resp.	Resp	Non-Resp.	Resp
Tenure	7.83	8.71	7.96	8.68	7.72	8.73
	(8.23)	(7.20)	(8.46)	(7.12)	(8.03)	(7.32)
FTE	0.46**	0.54	0.48***	0.60	0.44	0.48
	(0.33)	(0.30)	(0.32)	(0.29)	(0.33)	(0.31)
Age	36.40**	39.23	36.25	38.49	36.52*	39.87
	(13.80)	(12.76)	(13.62)	(12.68)	(13.97	(13.13)
Observations	589-601	138-139	277-281	64	312-320	74-75

Mean of each variable with standard deviation in parentheses.

The number of observations varies due to partial missing data.

*, **, and *** indicate that the difference in means between non-respondents and respondents within a group is statistically significant at the $p < .1$ level, $p < .05$ level, and $p < .01$ level, respectively. None of the differences between non-respondents and respondents differs significantly across the treatment and control group at the 0.1 level.

Table A.2: Store characterics and survey response

	Survey 1				Survey 2			
	No response		A response		No response		A response	
	mean	sd	mean	sd	mean	sd	mean	sd
Prior performance	0.96	0.14	0.95	0.06	0.93*	0.08	0.96	0.08
Team size	7.11	3.38	8.08	3.28	6.88***	2.51	8.56	3.58
Aides	1.74	1.69	1.94	1.53	1.57*	1.09	2.12	1.77
Age (manager)	41.23	14.06	44.34	10.33	40.80**	11.46	45.69	10.44
Tenure (manager)	14.07	12.51	15.71	14.21	11.05***	9.36	18.21	15.55
FTE (manager)	0.90	0.08	0.88	0.09	0.88	0.08	0.89	0.09
Age (staff)	39.55	9.43	38.16	6.48	38.31	6.68	38.46	7.33
Tenure (staff)	8.99	5.62	8.64	4.81	8.04	5.41	9.13	4.60
FTE (staff)	0.48	0.15	0.48	0.12	0.47	0.14	0.48	0.12
Observations	19		89		42		66	

Comparing stores with and without responses on a given survey, *, **, and *** indicate that the difference in means is statistically significantly different from zero at the $p < .1$ level, $p < .05$ level, and $p < .01$ level, respectively.

Table A.3: Average worker responses by survey participation and treatment

	A: Survey 1				B: Survey 2			
	Control		Treatment		Control		Treatment	
Participants in survey:	1	1 & 2	1	1 & 2	2	1 & 2	2	1 & 2
Job satisfaction	5.52	5.11	5.33	5.38	5.47	5.53	5.95	5.40
	(1.21)	(1.63)	(1.43)	(1.46)	(1.36)	(1.07)	(1.00)	(1.33)
Task allocation:								
Ability	4.01	4.42	4.01	4.30	3.89*	4.59	4.46	4.13
	(1.67)	(1.53)	(1.46)	(1.83)	(1.78)	(1.26)	(1.56)	(1.44)
Preference	3.15	3.27	3.16*	3.65	3.22*	3.76	3.23	3.43
	(1.39)	(1.34)	(1.35)	(1.74)	(1.59)	(1.21)	(1.39)	(1.23)
Seniority	2.11	2.36	2.35	2.13	1.72**	2.67	2.12	2.55
	(1.30)	(1.38)	(1.34)	(1.21)	(0.89)	(1.43)	(0.99)	(1.33)
Fairness	5.52	5.36	5.47	5.53	5.89	5.40	5.38	5.21
	(1.50)	(1.19)	(1.34)	(1.54)	(1.02)	(1.09)	(1.36)	(1.56)
Favouritism	2.15	2.51	2.54	2.36	1.67**	2.64	2.23	2.49
	(1.34)	(1.65)	(1.56)	(1.50)	(0.84)	(1.58)	(1.18)	(1.41)
Observations	56-73	37-45	81-93	40-47	15-18	30-45	20-26	35-47

Means of each variable with standard deviation in parentheses.

Comparing employees with and without responses on a given survey, *, **, and *** indicate that the difference in means is statistically significantly different from zero at the $p < .1$ level, $p < .05$ level, and $p < .01$ level, respectively.

Table A.4: Store average responses by survey participation and treatment

	A: Survey 1				B: Survey 2			
	Control		Treatment		Control		Treatment	
Respondents in survey:	1	1 & 2	1	1 & 2	2	1 & 2	2	1 & 2
Job sat.	5.17	5.18	5.49	5.28	7.00	5.46	5.00	5.69
	(0.97)	(1.13)	(1.13)	(0.95)	(.)	(0.91)	(1.73)	(1.02)
Task allocation:								
Ability	3.81	4.21	3.97	4.09	6.00	4.36	4.83	4.00
	(1.32)	(0.90)	(1.22)	(1.25)	(.)	(1.15)	(1.62)	(1.27)
Preference	3.30	3.21	3.31	3.38	2.00	3.77	2.73	3.37
	(1.39)	(0.91)	(1.07)	(1.09)	(.)	(0.90)	(0.83)	(0.99)
Seniority	1.83	2.19	1.97*	2.43	2.00	2.38	2.03	2.48
	(0.55)	(0.89)	(0.74)	(0.88)	(.)	(0.92)	(0.30)	(1.15)
Fairness	5.46	5.21	5.91**	5.36	6.00	5.58	5.57	5.29
	(1.95)	(1.06)	(0.67)	(0.98)	(.)	(0.76)	(0.43)	(1.28)
Favouritism	1.66**	2.27	1.94**	2.68	2.00	2.31	2.13	2.42
	(0.45)	(0.92)	(0.95)	(1.01)	(.)	(0.95)	(0.51)	(1.01)
Observations	9-12	25-27	15-17	32-33	1	26-27	3-5	30-33

Means of each variable with standard deviation in parentheses.

Comparing stores with and without responses on a given survey, *, **, and *** indicate that the difference in means is statistically significantly different from zero at the $p < .1$ level, $p < .05$ level, and $p < .01$ level, respectively.