Delfgaauw, Josse; Dur, Robert; Non, Arjan; Verbeke, Willem

Working Paper

Dynamic Incentive Effects of Relative Performance Pay: A Field Experiment

IZA Discussion Papers, No. 7652

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Delfgaauw, Josse; Dur, Robert; Non, Arjan; Verbeke, Willem (2013) : Dynamic Incentive Effects of Relative Performance Pay: A Field Experiment, IZA Discussion Papers, No. 7652, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
http://hdl.handle.net/10419/89974

DISCUSSION PAPER SERIES

IZA DP No. 7652

Dynamic Incentive Effects of Relative Performance Pay: A Field Experiment

Josse Delfgaauw
Robert Dur
Arjan Non
Willem Verbeke

September 2013

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Dynamic Incentive Effects of Relative Performance Pay: A Field Experiment

Josse Delfgaauw
Erasmus University Rotterdam and Tinbergen Institute

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

Arjan Non
Maastricht University and CESifo

Willem Verbeke
Erasmus University Rotterdam and ERIM

Discussion Paper No. 7652
September 2013

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 7652
September 2013

ABSTRACT

Dynamic Incentive Effects of Relative Performance Pay: A Field Experiment*

We conduct a field experiment among 189 stores of a retail chain to study dynamic incentive effects of relative performance pay. Employees in the randomly selected treatment stores could win a bonus by outperforming three comparable stores from the control group over the course of four weeks. Treatment stores received weekly feedback on relative performance. Control stores were kept unaware of their involvement, so that their performance generates exogenous variation in the relative performance of the treatment stores. As predicted by theory, we find that treatment stores that lag far behind do not respond to the incentives, while the responsiveness of treatment stores close to winning a bonus increases in relative performance. On average, the introduction of the relative performance pay scheme does not lead to higher performance.

JEL Classification: C93, M52

Keywords: dynamic incentives, relative performance pay, field experiment

Corresponding author:

Robert Dur
Department of Economics H 9-15
Erasmus University Rotterdam
P.O. Box 1738
3000 DR Rotterdam
The Netherlands
E-mail: dur@ese.eur.nl

* We thank a co-editor and two referees of this journal, Pablo Casas-Arce, and participants of the 2011 conference on Tournaments, Contests, and Relative Performance Evaluation, North Carolina State University.

1 Introduction

Non-linear pay-for-performance plans have dynamic incentive effects when employees receive intermediate performance information over the course of the incentive period. For instance, consider a salesman who can earn a bonus by attaining a monthly sales target while receiving daily or weekly sales figures. When realised sales during the month are such that it remains challenging but possible to reach the target, the bonus scheme provides strong incentives. The incentive effect is much weaker, however, when realised sales during the month are particularly high or low. High intermediate sales imply that the salesman can hardly miss the bonus, while low intermediate sales imply that the target is practically out of reach.

More generally, workers can use intermediate performance information to determine how much additional performance is necessary to obtain a bonus. This creates dynamic incentive effects, where the incentive effect of the pay-for-performance plan at each point in time depends on realised performance until then. Incentive plans based on relative performance, where prizes are awarded for outperforming sufficiently many competitors, are particularly prone to dynamic incentive effects. Sports leagues are a common example. In the workplace, examples range from employee-of-the-month contests, to beat-the-index bonuses for stock brokers, and to job promotion contests.[1]

Casas-Arce and Martinez-Jerez (2009) show formally that for contests with a large number of participants, the incentive effect of a relative performance incentive scheme is hump-shaped in lagged relative performance. Competitors who find themselves trailing far behind may perceive catching up to be impossible and consequently give up trying. Similarly, competitors who are far ahead may perceive losing as impossible and slack off as well. In contrast, incentives are highly salient for competitors who find themselves almost tied in intermediate performance and are at the margin of winning a prize. Analyzing sales contests among retailers of a commodities company, Casas-Arce and Martinez-Jerez (2009) find indeed that competitors in winning positions reduce performance when their lead increases. However, the performance of trailing competitors does not decrease when they lag further behind. The authors conjecture that this result might be affected by attrition bias. Frank and Obloj (2011) do find the predicted hump-shaped pattern in their analysis of a competition among units of a retail bank. Ludwig and Lünser (2012) run a lab experiment to study the role of intermediate relative performance information in tournaments, and also find that leading competitors slack off, while trailing competitors increase their stated efforts.[2]

[1] As a concrete example, more than half of the remuneration of the executive directors of oil company Shell is based on a ranking of Shell's performance relative to its four main competitors on four publicly available measures. The incentive plan has a three-year horizon, during which the companies regularly release the latest figures with respect to these performance measures (Royal Dutch Shell, 2009).

[2] Relatedly, Fershtman and Gneezy (2011) let kids run side-by-side and find that increasing incentives yield higher performance but also a higher fraction of kids giving up during the race. Following the early literature on tournament theory (Lazear and Rosen, 1981, Green and Stokey, 1983, Nalebuff and Stiglitz, 1983), most of the literature has abstracted from dynamic incentive effects of tournaments. A recent string of theoretical papers studies the cost and benefit to a principal of providing intermediate relative performance feedback during

Testing for the presence and strength of dynamic incentive effects is hampered by two issues. First, in contests with a limited number of participants, a competitor's optimal strategy depends on (its perception of) its competitors' strategies. For example, a trailing competitor may be best off by accepting its loss when the other competitors keep effort high, but not when they would slack off. Second, serial correlation in performance biases the estimates of the effect of intermediate relative performance on subsequent performance. For instance, positive serial correlation would imply that a positive shock to performance in the previous period increases both relative intermediate performance and current performance. Casas-Arce and Martinez-Jerez (2009) and Frank and Obloj (2011) employ a method developed by Arellano and Bond (1991) that relies on taking first differences and using lagged values of independent variables as instruments for the independent variables to correct for this bias.

In this paper, we take a unique approach in tackling both issues by setting up a relative performance pay scheme where only one of the 'competitors' can earn a prize, while the other participants are kept unaware of their involvement. This implies that the strategies of all non-competing participants are exogenous, allowing us to use their performance as an instrument for intermediate relative performance of the competing participant. More specifically, we study the dynamic incentive effects of this relative performance pay scheme by conducting a natural field experiment in a Dutch retail chain. We provide the employees of 93 stores randomly selected from 189 of the company's stores with the opportunity to earn a bonus. The bonus is awarded when a treatment store outperforms three comparable stores from the control condition over the course of a four-week period (February 2010). Each week, treatment stores receive a poster with the intermediate performance of all four stores in their group. Importantly, the employees of the three comparison stores do not receive the poster, cannot earn a bonus, and do not learn that another store can earn a bonus by beating their performance. This way the treatment stores compete against stores that are not competing, and treatment stores were informed about this.

This setup has two advantages. First, we can use the performance of the three comparison stores as an instrument for trailing behind or being ahead: lagged performance of the comparison stores does affect intermediate relative performance, but does not affect current performance of the treatment store other than through lagged relative performance. Hence, using this instrument, our estimates are not biased by serial correlation in stores' own performance. Moreover, we eliminate the possible influence of perceptions of competitors' strategies from the estimations. This makes for a clean test of dynamic incentive effects. A disadvantage of our setup is that the estimated effects of our competition may not generalize to settings where all contestants compete. Thus, while our experiment cleanly tests whether the competing stores respond to intermediate relative performance feedback as predicted by theory, it may not yield a reliable estimate of the overall effect of standard tournaments on performance. In the terminology of Ludwig et al. (2011), we conduct a mechanism experiment.

a contest between his agents (Aoyagi 2010, Ederer 2010, Gershkov and Perry 2009, Goltsman and Mukherjee 2011).

Our results are as follows. First, we find a positive effect of intermediate relative performance on current performance for stores close to the target, particularly in the last two weeks of the experiment. This effect is substantial: a one percentage point increase in intermediate relative performance increases current performance by 0.73 percent. Stores lagging far behind do not respond to intermediate relative performance. This suggests that the employees in these stores gave up trying to win. Hence, as predicted by theory, we find that changes in intermediate relative performance matter more for competitors that perform close to target than for competitors that lag far behind. During the contest, hardly any treatment store managed to get far ahead of all its comparison stores. Hence, we cannot test the hypothesis that high-performers slack off as their lead increases.

Second, we find no average treatment effect of introducing the contest, neither for the four weeks taken together nor for one of the weeks separately. This contrasts with several recent findings on the incentive effects of tournaments. In another retail chain, we do find a substantial positive effect of introducing a standard tournament among shops (Delfgaauw et al. 2013), as do Erev et al. (1993) and Bandiera et al. (2013) among teams of fruit pickers and Casas-Arce and Martinez-Jerez (2009) among retailers of a commodities company. Even more striking, several recent papers suggest that the mere provision of relative performance feedback can be sufficient to trigger higher performance (Azmat and Iriberri 2010, Blanes i Vidal and Nossol 2011, Delfgaauw et al. 2013, and Kosfeld and Neckermann 2011). Bandiera et al. (2013) and Barankay (2012) obtain an opposite result. A possible explanation for the lack of a significant average treatment effect in our experiment is that only a limited number of stores happened to be close to the target for winning, while the majority of stores lagged far behind. As a result, the number of stores positively responding to the contest is just too small to be reflected in a significant average treatment effect. An alternative explanation is that beating unaware contestants, as in our setting, is less exciting than beating competing contestants. Relatedly, participants in our benchmark competitions may anticipate weaker feelings of envy after losing as compared to participants in regular tournaments (Eisenkopf and Teyssier 2013).

Our experiment also relates to a literature on the incentive effects of non-linear payment schemes. Forbes et al. (2012) find that airline personnel who are rewarded for on-time performance reduce taxi-in times only when the expected arrival time is just around the critical threshold for the flight being recorded as 'late'. Schweitzer et al. (2004) and Cadsby et al. (2010) find in lab experiments that non-linear incentive schemes invite substantial lying to meet the target. These findings are well in line with ours in the sense that individuals respond to performance feedback if they are sufficiently close to the target. An important difference with our study is that we aim to identify increases in sales performance rather than artifical improvements in recorded performance or outright lying.

Our experiment involves one incentive period of four weeks. When incentive schemes are repeated over time, as with monthly or year-on-year targets, other types of dynamic incentive effects may arise. For instance, sales may be shifted forward or backward in time around the

incentive commencement date in order to meet the current target or to alleviate the difficulty of meeting the next target; see Asch (1990), Oyer (1998), Courty and Marschke (2004), and Larkin (2013) for empirical evidence. Furthermore, when the targets in repeated incentive schemes are based on historical performance, workers have an incentive to beat the target by only a limited amount even it would be possible to greatly outperform the target. Bouwens and Kroos (2011) find evidence in line with such ratchet effects, using store-level data from a retail chain. Cooper et al. (1999) and Charness et al. (2010) find ratchet effects in the lab. Ratchet effect considerations may be yet another explanation for why we find no average treatment effect, as workers may have feared that a strong response to the introduction of the relative performance pay scheme would result in higher targets in their regular incentive scheme.

We proceed as follows. In the next section, we describe the context and design of our experiment in detail. Then, in section 3, we describe the econometric model and estimation strategy. Section 4 presents the results of the estimations and a number of robustness checks. Finally, section 5 concludes.

2 Experimental context and design

2.1 Experimental set-up

The experiment took place in February 2010 in a retail chain in The Netherlands that sells computer games, music, and movies. At the start of 2010, the retail chain owned 208 geographically dispersed stores, operating under two different brands. Each store employs on average 5 employees, including a store manager. The company's central management decides on the range of products sold, pricing, and advertisement. New products arrive in stores complete with instructions on how to sell them. Store managers are responsible for day-to-day operations.

In this environment, store employees may have limited scope to affect sales. Still, the company's management is convinced that employees can contribute to sales, in particular through cross-selling. The company instructs employees to show interest in potential customers, to help and give advice whenever possible, and to suggest related products. Employees receive rather weak incentive pay on top of their base salary, based on their shop's yearly sales growth and a subjective performance evaluation. The company's management was not fully satisfied with this incentive scheme and wished to learn more about the effects of short-term incentives, in particular of sales contests. The pre-existing incentive scheme remained in place during the experiment. Hence, even though incentives may not have very large effects in this retail chain, the company's management sees sufficient scope for incentives to have a beneficial effect on performance.

We designed a relative performance incentive scheme to be implemented in a randomly selected subset of stores (the treatment condition), while the rest of the stores comprised the control condition. All employees (including the shop manager) of a store in the treatment condition could earn a bonus by outperforming three preselected stores from the control condition

by a sufficient margin. Stores in the control condition could not earn a bonus, and employees in the treatment stores were informed about this. Performance is measured as cumulative sales revenue in percentage deviation of budgeted sales in February 2010 (a period of 4 weeks). Budgeted sales are set in advance by the company's central management and cannot be affected by stores.[3] Let $y_{s,w}$ be sales and $b_{s,w}$ budgeted sales of store s in week w, respectively. Weekly performance $p_{s,w}$ is given by

$$p_{s,w} = \frac{y_{s,w} - b_{s,w}}{b_{s,w}} \cdot 100\% \tag{1}$$

and cumulative performance over February 2010 is given by

$$p_s^{CU} = \frac{\sum_{w=E1}^{E4}(y_{s,w} - b_{s,w})}{\sum_{w=E1}^{E4} b_{s,w}} \cdot 100\% \tag{2}$$

where the summation is over the four experimental weeks $E1 - E4$, namely week 5, 2010 to week 8, 2010. Below, we will refer to cumulative performance during the experimental weeks as performance in the tournament.

All employees of a treatment store received a bonus of gross 150 euro each when their shop's performance in February 2010 was at least 10 percentage points higher than the performance of all three comparison stores. When a treatment store scored between 5 and 10 percentage points above all three comparison stores, all of its employees received 75 euro.[4] Lastly, outperforming all three comparison stores by less than 5 percentage points yielded a cake for the treatment store, but only if the treatment store also performed above budget.[5]

All communication on the experiment towards the shops went through the company's regular channels, so shop managers and employees were not aware of our involvement. Hence, our experiment classifies as a natural field experiment (Harrison and List, 2004). In January 2010, the company informed all store managers and employees that a randomly selected set of stores would get the opportunity to earn a bonus in February 2010, and that all other stores would have a similar opportunity later that year.[6] On January 22, all store managers and employees in the treatment stores were informed about the details of the relative performance incentive scheme, which would start on Monday February 1. At this point, we did not inform treatment stores with whom they were matched. Control stores were not informed about their role in the treatment stores' incentive scheme.

[3] The budgeted sales are forecasts for shops' weekly sales as determined by the company's management in October 2009 (at the start of the financial year) for a year onwards. These budgeted sales boil down to a forecast for total sales of the whole chain, with each store expected to bring in a fixed share of total sales. Hence, a combination of week and store fixed effects explains all variation in the log of budgeted sales in our data. The company gives shop managers weekly feedback on sales relative to budgeted sales, which makes it a well-known and natural measure of performance.

[4] For employees who did not have a full-time contract, the size of the bonus was proportional to the contractual number of hours. Hourly wages are close to the minimum wage, which makes that receiving the high bonus would increase monthly earnings by about 10%.

[5] The latter requirement only applied for the cake, not for any of the two bonuses. This requirement was a last-minute addition by the company's management to the rules.

[6] We organised two-stage elimination tournaments in the Fall of 2010, see Delfgaauw et al. (2011b).

During the experiment, we provided weekly feedback to the treatment stores on their relative performance in the form of a poster. The poster contained the cumulative sales relative to budget figures of the treatment shop and its three comparison shops, ranked in descending order. In order to ensure credibility, we published the identity of the comparison stores along with their performance figures, see Figure 1 for an example. Thus, the identity of the comparison stores was revealed when the first poster arrived, on February 9. Furthermore, in the first week of the experiment, all treatment stores received a large poster, with room to glue on the four posters with weekly rankings to be received in the following weeks. We created the feedback posters and sent them in the company's envelops by regular mail to the stores. Store managers were instructed to put up these posters in the store's canteen.[7] Stores in the control condition did not receive posters.

It is possible that control stores learned about the details of the experiment, or that treatment stores contacted their comparison stores after receiving the first feedback poster. According to the central management staff, normally there is some communication between stores, in particular between store managers within a region. To reduce possibilities of collusion, treatment stores were never matched with another store from their region (the assignment procedure is discussed in detail in the next subsection). For control stores, engaging in collusive actions is not attractive, as it reduces their regular incentive pay. During the experiment, central management staff did not receive questions about the incentive event from control stores, nor did they hear of any treatment store contacting their comparison stores. Hence, we are quite confident that control stores were not aware of the details of the experiment.[8]

Our design has two advantages as compared to a regular competition. First, as treatment stores only receive a bonus when they outperform comparable stores from the control condition, the payout is relatively low when the incentive has little effect on performance. This was seen as a major benefit by the company's management. Second, performance of the comparison stores is exogenous to the incentive scheme, as these stores neither could earn a bonus, nor received relative performance feedback, and were not aware that their performance played a role in the incentive scheme. We exploit differences in comparison stores' performance during the experiment to analyse how treatment stores' intermediate relative performance affects the effect of the incentive scheme in subsequent weeks.

2.2 Assignment procedure

The aim of our assignment procedure is to match stores from the treatment condition to similar stores from the control condition. It is important to create homogeneous groups of stores,

[7] The company's regional managers were instructed to verify that store managers actually put up the posters in the canteen. We have not heard about a single store manager who refused to do so.

[8] Another potential source of contamination of the control group is relocation of employees or store managers from treatment stores to control stores. We have no information on the frequency of such relocations. However, relocations typically take place at the beginning or end of the month, so when stores were not yet informed on the identity of their competitors. Therefore, and because our experiment lasts only four weeks, we don't think this could possibly influence our results.

as theory predicts that differences in ability between contestants weaken incentive effects of tournaments (Lazear and Rosen 1981, O'Keefe et al. 1984, Rosen 1986; see Fonseca 2009 and Höchtl et al. 2011 for empirical evidence). We therefore use weekly sales and budget data for the weeks 40 to 53 in 2009 to match stores on the basis of their historical performance. As the company's management excluded a specific group of 14 stores from participating in the experiment, 194 of the company's 208 stores were included in the matching procedure.[9] Besides historical performance, the company's management argued that store size was an important characteristic to take into account when matching treatment stores with comparison stores, as employees in small treatment stores might perceive it as unfair when matched to large comparison stores or vice versa (e.g. because of differences in local demand conditions, quality of management, free rider effects, and so on). Therefore, we first created four equally large strata based on store size as measured by average weekly sales revenues. Randomly, half of the stores in each stratum was assigned to the treatment condition, while the remaining half of the stores were assigned to the control condition. Subsequently, we matched each treatment store to three control stores from the same stratum. Our randomization procedure ruled out that, by chance, there would be relatively many treatment stores in a particular stratum, which could impede the creation of a level playing field for treatment stores in this stratum (as there would be few control stores with similar past performance and similar store size left). To reduce opportunities for collusion, we imposed that each treatment store was matched to control stores located in other regions, as discussed in the previous subsection. The company distinguishes between 12 geographically-clustered regions, each led by a different regional manager. We used the same region classification in our matching procedure. Apart from this regional separation, treatment stores were matched to the control stores that were most comparable in terms of the performance measure (cumulative sales revenue relative to the budget) for the period of week 40 to week 53 in 2009. Note that a control store can be matched to multiple treatment stores. After this assignment procedure, we excluded one treatment store from the experiment as its budget figures were unavailable during the experimental period, which made it impossible to determine performance in the tournament. Furthermore, 3 treatment stores and 1 control store were shut down in January 2010.[10] This leaves us with 189 stores in the experiment: 93 stores in the treatment condition and 96 stores in the control condition. For each of these stores, we have weekly sales and budget figures for a period of in total 22 weeks, from week 40 in 2009 to week 8 in 2010.

[9] The group that is excluded from participation consists of all stores that are located in railway stations. The company's management considered those stores as special cases. As will be explained below, we had to drop another 5 stores from the experiment, leaving us with a final sample of 189 stores.

[10] The decision to terminate these stores has been made before we conducted the randomization, but was not communicated to us. The closure of these stores is therefore not related to assignment to the treatment condition. Moreover, these stores were already closed before January 22, when we informed the stores about the experiment. Hence, possible relocations of personnel from closed treatment stores to control stores took place before we communicated the details of the experiment to the treatment stores.

2.3 Descriptive statistics

Figures 2 and 3 show weekly sales (indexed) and weekly performance (as described by (1)), respectively, averaged over all stores. Average weekly sales show two spikes in December 2009, related to Sinterklaas and Christmas festivities, respectively.[11] Removing these weeks from the analysis does not affect the results. Average performance varies between plus and minus 20 percent. The spikes in sales in December 2009 are anticipated by the company's management when determining budgeted sales, as the spikes in sales do not carry over to average performance.

The descriptive statistics in Table 1 show that average sales do not differ significantly between treatment stores and control stores, neither for the whole period nor for the first 14 weeks in the data used to stratify the stores. The same holds for budgeted sales and for performance as measured by (1). Note that on average, sales are below budget, but that variation in average performance across stores is large. As noted above, there is substantial variation in sales and performance over time, but sales and performance are relatively stable during the experimental weeks, and common shocks capture a substantial fraction of week-to-week variation in performance. There are no significant differences between treatment and control stores regarding the within-store standard deviation of sales and performance. Further, the number of employees per store does not differ significantly between the treatment and control stores. Lastly, in week 7 of 2010, a total of 29 stores were closed for one or two days in relation to carnival festivities, mainly in the south of The Netherlands. Treatment stores were slightly less often closed than control stores, but not significantly so. In all estimations below, we correct for the effect of carnival.

As a first hint of the overall effect of the relative performance incentive, Table 1 shows that there is no difference in average sales and performance between treatment and control stores for the treatment weeks (week 5, 2010 to week 8, 2010). Figures 4 and 5 provide further insight into the overall treatment effect, by plotting the differences between the treatment and control condition in average sales and in average performance, respectively, by week. The experiment took place in the final four weeks of the period shown. Both figures show no sign of a positive treatment effect, possibly with the exception of the final week. A second hint of the overall treatment effect is given by the fact that the number of treatment stores in leading position during the tournament is smaller than the expected number when determined by chance: in all treatment weeks at most 23 treatment stores were in leading position.[12] In the end, only 13 stores earned a prize: 5 stores earned the high bonus, another 5 stores earned the low bonus, and 3 stores were entitled to cake. Figure 6 shows the combined distribution of the intermediate relative performance feedback the treatment stores received at the start of each of the 3 final experimental weeks. The distribution roughly corresponds to what we would expect when random shocks determine the outcome of the tournament: in 23.7% of all cases, treatment

[11] Sinterklaas takes place the fifth of December and is widely celebrated in the Netherlands. The essence of the festivities is an evening of gift-giving among relatives and friends, much like Christmas in many other countries.

[12] To be exact, 23, 22, 21, and 19 stores were in the leading position at the end of tournament week 1, 2, 3, and 4, respectively. Absent treatment effects, the expected number is 93/4=23.25.

stores were in leading position. Importantly, there is substantial variation in intermediate performance feedback. On average, treatment stores lagged 11% behind their best comparison store. In about 25% of all cases, stores were informed that they lagged more than 20% behind, while 13.6% of the stores were at some point more than 5% ahead of their best competitor. Stores were rarely more than 10% ahead: this happened in only 6% of all cases.

3 Method

In this section, we describe our empirical strategy. We estimate the effects on sales rather than on the performance measure used in the incentive scheme as described by (1). The reason is that budgeted sales are set in advance by the company's central management. Hence, shops can affect their performance only through sales. Estimation results for sales and performance as measured by (1) are therefore similar. Since effects on sales are more easy to interpret, we use sales rather than performance as our main outcome measure. We include 22 weeks in the estimations, where the final four weeks are the experimental weeks (February 2010). We assess the average effect of the relative performance incentive scheme on sales using OLS with week-fixed effects and store-fixed effects. That is, we estimate

$$\ln(y_{s,w}) = \alpha_s + \theta_w + \gamma B_{s,w} + \kappa F_{s,w} + \varepsilon_{s,w} \tag{3}$$

where $\ln(y_{s,w})$ is the log of sales of store s in week w. Store and week-fixed effects are given by α_s and θ_w, respectively. $B_{s,w}$ is a dummy variable that takes the value one from week 5 to week 8 in 2010 for stores in the treatment condition, and is zero otherwise. $F_{s,w}$ measures the number of days shop s is closed for carnival festivities in week w (this variable takes positive values only in week 7, 2010), and $\varepsilon_{s,w}$ is an error term, possibly serially correlated (i.e. $E(\varepsilon_{s,w}\varepsilon_{s,w-1}) \neq 0$).

The main goal of this paper is to analyse the effect of intermediate relative performance on subsequent performance. First, we introduce some additional notation. Let T and C be the sets of stores in the treatment and the control condition, respectively. Further, denote by $c_t \in C$ a control store matched to treatment store $t \in T$. Lastly, let $p^{CU}_{s,w-1}$ denote the *cumulative* performance of store s during the experiment up to but not including week w, as measured by cumulative sales over budget in February 2010 (i.e. weeks 5 to 8 in 2010, which corresponds to weeks 19-22 in our dataset):

$$p^{CU}_{s,w-1} = \begin{cases} \frac{\sum_{w=19}^{w-1}(y_{s,w}-b_{s,w})}{\sum_{w=19}^{w-1} b_{s,w}} \cdot 100\% \text{ if } w \in [20,21,22] \\ 0 \text{ if } w \notin [20,21,22] \end{cases} \tag{4}$$

Hence, $p^{CU}_{s,w-1}$ is the performance figure for store s as depicted on the poster received at the start of week w during the experiment. The effect of intermediate performance of treatment stores relative to their best-performing comparison store on subsequent sales can be estimated

by

$$\ln(y_{s,w}) = \alpha_s + \theta_w + \gamma B_{s,w} + \mu \left(p^{CU}_{t,w-1} - \max_{c_t} \left[p^{CU}_{c_t,w-1} \right] \right) B_{s,w} + \kappa F_{s,w} + \varepsilon_{s,w} \tag{5}$$

where the term $p^{CU}_{t,w-1} - \max_{c_t} \left[p^{CU}_{c_t,w-1} \right]$ gives the difference in cumulative performance during the experiment between treatment store t and its best-performing comparison store c_t up to and including the previous week.[13] Since the experiment lasted four weeks, we have three intermediate relative performance figures per treatment store, corresponding to a total of 279 treatment store-week observations.

The estimation of the effect of intermediate relative performance on subsequent sales, represented by μ in equation (5), is biased in case sales are serially correlated. When $\varepsilon_{s,w}$ is correlated with $\varepsilon_{s,w-1}$, cumulative past performance $p^{CU}_{t,w-1}$ is no longer exogenous to current sales $y_{s,w}$. This can be seen from (4): $p^{CU}_{s,w-1}$ is a function of $y_{s,w-1}$, which is correlated with $y_{s,w}$ via the serial correlation in the error structure.[14] Thus, serial correlation in treatment stores' own performance leads to biased estimates of the parameter of interest μ. By construction, past performance of the best control store matched to treatment store t, $\max_{c_t} \left[p^{CU}_{c_t,w-1} \right]$, is exogenous to the treatment store's sales $y_{t,w}$.[15] Control stores do not earn a bonus and do not receive feedback, implying that their past performance only influences the treatment stores' current performance via the feedback provided to the treatment stores. Therefore, we instrument the difference in intermediate performance $p^{CU}_{t,w-1} - \max_{c_t} \left[p^{CU}_{c_t,w-1} \right]$ by the expected difference

$$D_{t,w-1} = E\left[p^{CU}_{t,w-1} \right] - \max_{c_t} \left[p^{CU}_{c_t,w-1} \right]. \tag{6}$$

The expected cumulative performance of treatment store t in the experiment, $E\left[p^{CU}_{t,w-1} \right]$, is based on the average performance prior to the start of the experiment, excluding the last three weeks to rule out that serial correlation in performance affects $E\left[p^{CU}_{t,w-1} \right]$. That is, we average performance over the period running from week 40, 2009 to week 1, 2010, which amounts to 15 weeks in total. Moreover, we account for week-fixed effects in performance during the experiment:

$$E\left[p^{CU}_{t,w-1} \right] = \begin{cases} \frac{1}{15} \sum_{w=1}^{15} p_{t,w} + \frac{\sum_{w=19}^{w-1} b_{t,w} \theta^p_w}{\sum_{w=19}^{w-1} b_{t,w}} \text{ if } w \in [20, 21, 22] \\ 0 \text{ if } w \notin [20, 21, 22] \end{cases}, \tag{7}$$

where θ^p_w is the week-fixed effect from estimating

$$p_{s,w} = \alpha^p_s + \theta^p_w + \gamma^p B_{s,w} + \kappa^p F_{s,w} + \varepsilon^p_{s,w},$$

[13] This term is set to zero for control stores.

[14] Note that $p^{CU}_{s,w-1}$ is also endogenous when $\varepsilon_{s,w}$ is correlated with $\varepsilon_{s,w-2}$ or $\varepsilon_{s,w-3}$, as $p^{CU}_{s,w-1}$ also includes these longer lags in weeks 21 and 22, respectively.

[15] Exogeneity of control stores' performance is violated when there are region-specific shocks that are correlated both across regions and over time. As a robustness check, we will also estimate (5) including region-specific week-fixed effects, see section 4.2.

with superscript p denoting that the estimates relate to performance as dependent variable.[16] By adjusting $E\left[p^{CU}_{t,w-1}\right]$ to account for week-fixed effects, we ensure that the expected difference in intermediate performance, $D_{t,w-1}$, has the same meaning in each of the treatment weeks, independent of the realisation of the common shock θ_{w-1}. As week-fixed effects by definition show up in the best comparison store's cumulative performance, θ_{w-1} and $D_{t,w-1}$ would be negatively correlated if we would not adjust $E\left[p^{CU}_{t,w-1}\right]$ to account for week-fixed effects. Most importantly, this implies that for each treatment store, variation in $D_{t,w-1}$ across experimental weeks stems solely from variation in $p^{CU}_{c_t,w-1}$, which is unrelated to $\varepsilon_{t,w}$ given the design of our experiment.

In order to obtain precise estimates, it is important that our instrument is a strong predictor of actual intermediate relative performance. The second column in Table 3 reports the estimation results of the first-stage regression. Actual intermediate relative performance increases one-for-one with our instrument. This instrument alone explains about 50 percent of the total variation in intermediate relative performance in the last three weeks of the experiment. Figure 7 shows this relation graphically. It depicts the relation between the actual difference in intermediate cumulative performance between the treatment stores and their best comparison stores, $p^{CU}_{t,w-1} - \max_{c_t}\left[p^{CU}_{c_t,w-1}\right]$, and the expected difference, $D_{t,w-1}$, for the last three weeks of the experiment. Clearly, there exists a strong association between the two (the correlation coefficient is 0.65).

Equation (5) estimates a linear effect of intermediate relative performance. However, the incentive scheme is likely to have the biggest effect when treatment stores learn that they are close to the relative performance targets for winning a bonus (Casas-Arce and Martinez-Jerez, 2009).[17] Treatment stores lagging far behind in the intermediate ranking may give up, and treatment stores far ahead may reduce their efforts when they anticipate that they can hardly miss the bonus. As we have seen in the previous section, in the course of the experiment, we have many treatment stores that face an uphill battle, while there are only few stores that are comfortably ahead. In total, we have only 8 store-week observations where treatment stores' intermediate relative performance is more than 10 percentage points above the target for the high bonus (i.e. with $p^{CU}_{t,w-1} - \max_{c_t}\left[p^{CU}_{c_t,w-1}\right] > 20\%$). This implies that we cannot test whether stores that greatly outperform their comparison stores reduce their efforts.[18] We can test whether the marginal effect of intermediate relative performance on current performance differs between stores that lag far behind and stores that are still in the running, by allowing the estimated effect to differ between both groups.

In determining which stores still have a chance of earning a bonus, we cannot use the actual difference between the lagged cumulative performance of the treatment store and its

[16] We weight the week-fixed effects by budgeted sales $b_{t,w}$ to account for the fact that weeks with a higher absolute budgeted sales volume have a higher weight in cumulative performance, see (4).

[17] The theory developed by Casas-Arce and Martinez-Jerez (2009) predicts that performance is hump-shaped in intermediate relative performance, but does not predict the exact level of relative performance at which the incentive effect peaks.

[18] Excluding these 8 observations from the analysis does not affect any of the results.

best control, as given by (4). Serial correlation in $y_{t,w}$ would bias the estimates. Hence, we again use the estimated difference (6) to determine stores' chances of earning a bonus. Rather arbitrarily, we set the bar for being too far behind at a 5 percentage point lag relative to the best performing comparison store. Note that stores that lag 5 percentage points behind need to improve their relative performance by 5 percentage points in order to win a cake and by at least 10 percentage points to obtain a bonus. We do vary the bar to assess the robustness of the results. Let $I_{t,w-1}$ be a dummy that takes value 1 during experimental weeks for treatment stores whenever $D_{t,w-1} > -5\%$, and zero otherwise. This yields 47 store-week observations where $I_{t,w-1} = 1$, out of a total of 279 treatment store-week observations with intermediate relative performance figures. We estimate

$$\begin{aligned} \ln(y_{s,w}) &= \alpha_s + \theta_w + \gamma B_{s,w} + \mu \left(p^{CU}_{t,w-1} - \max_{c_t} \left[p^{CU}_{c_t,w-1} \right] \right) B_{s,w} + \\ &+ \delta I_{t,w-1} B_{s,w} + \nu \left(p^{CU}_{t,w-1} - \max_{c_t} \left[p^{CU}_{c_t,w-1} \right] \right) I_{t,w-1} B_{s,w} + \kappa F_{s,w} + \varepsilon_{s,w} \end{aligned} \tag{8}$$

again instrumenting the difference in intermediate performance $p^{CU}_{t,w-1} - \max_{c_t} \left[p^{CU}_{c_t,w-1} \right]$ by the expected difference $D_{t,w-1}$ as given by (6).[19] Note that, as compared to equation (5), equation (8) allows for a differential treatment effect for stores close to winning as well as for a different effect of intermediate relative performance for stores close to winning.

Treatment stores' performance may, in addition to the distance to the best control store, also depend on the distance to the second-best control store. Unfortunately, we cannot disentangle the two effects due to problems of multicollinearity. As it turns out, the correlation between distance to second best control store and best control store is extremely high (0.62 excluding the control stores and non-experimental weeks; the correlation between the respective instruments is even higher: 0.86). Therefore, we will concentrate our analysis on the effect of the distance to the best comparison store.

In all of our estimations, we cluster standard errors at the store level to correct for serial correlation within stores, as well as for heteroscedasticity across stores (see Bertrand et al. (2004) for a discussion of the importance of correcting for serial correlation in Difference-in-Difference estimation).

4 Results

4.1 Estimation results

This subsection presents the estimation results. We subsequently present our estimates of the average treatment effect, the average treatment effect for stores close to winning, and how these

[19] Instead of estimating (8), we could estimate a quadratic specification of intermediate relative performance. However, the estimates for the quadratic specification would be heavily affected by the many treatment store-week observations with sizable negative intermediate relative performance (see Figure 6). Hence, we would learn little about the marginal effect of intermediate relative performance for stores close to winning a bonus.

treatment effects depend on interim relative performance. In the next subsection, we present the results of a number of robustness checks we conducted.

The estimates of the average treatment effect are presented in Table 2. The first column in Table 2 gives the results of estimating (3). On average, the relative performance incentive scheme did not significantly affect sales. This result is not due to a lack of statistical power: given the size of the estimated standard errors, we should be able to detect reasonably small effect sizes. However, the point estimate is also very close to zero. The second column of Table 2 shows that there is some variation in the estimated treatment effect by week, but none of the estimates differs significantly from zero. This suggests that the absence of a treatment effect is not due to stores gradually becoming discouraged, which would imply a negative trend in the estimated treatment effects, or learning to improve, which would imply an upward trend.

The main aim of our analysis is to estimate how the treatment effect depends on intermediate relative performance. As OLS estimation is possibly biased by serial correlation in the error term, we use the predicted difference, $D_{t,w-1}$ as defined by (6), as an instrument for the actual difference in cumulative performance between the treatment stores and their best comparison stores, $p^{CU}_{t,w-1} - \max_{c_t}\left[p^{CU}_{c_t,w-1}\right]$. In the third and fourth column of Table 2, we allow the average treatment effect to differ for stores that lag far behind and for stores that can be expected to be close to the winning positions. Specifically, we interact the treatment dummy with $I_{t,w-1}$, the dummy indicating that a treatment store's expected relative performance $D_{t,w-1} > -5\%$. On average, the treatment effect for stores close to winning is estimated to be 3.7 percentage points higher than for trailing stores (p-value = 0.11). As compared to the control group, sales of stores close to winning are estimated to be 2.9% higher, but a Wald test shows that this difference is not statistically significant (p-value = 0.23). In the fourth column of Table 2, we estimate the treatment effects for each week separately. In all weeks, stores competing for the winning positions show a larger response to treatment than stores that lag further behind, but the difference is statistically significant at the 10% level only in the second week. The estimated average treatment effect for stores close to winning is largest in the final week. The point estimate is sizeable, namely 5.5% additional sales as compared to the control stores, but imprecisely estimated (p-value = 0.15).

Next, we run regressions where we allow the treatment effect to depend linearly on expected intermediate relative performance, $D_{t,w-1}$. For comparison purposes, we first estimate the linear effect of the actual intermediate relative performance (as in equation (5)) using OLS. The OLS estimation in the first column of Table 3 shows that intermediate relative performance is significantly positively related to subsequent sales. Its point estimate suggests that a percentage point increase in lagged relative performance increases current sales of treatment stores by 0.26 percent. However, in the IV-2SLS estimation, reported in the third column, the point estimate is more than halved and is no longer significantly different from zero. This underlines the importance of correcting for serial correlation in sales. Figure 8 visualises these results for the relevant subset of observations: treatment stores in the three final weeks of the experiment. It

plots the residuals of the estimation of the average treatment effect (3), as presented in the first column of Table 2, against the predicted difference $p^{CU}_{t,w-1} - \max_{c_t}\left[p^{CU}_{c_t,w-1}\right]$ as estimated by the first-stage regression of the IV-2SLS estimation (second column of Table 3). In line with the estimation results, there is no easily discernible relation between predicted performance and sales, corrected for week and store-fixed effects.

The estimation in the third column of Table 3 assumes that the effect of intermediate relative performance is the same for all experimental weeks. In the fourth column of Table 3, we allow the effect to differ between weeks.[20] The results show that there are no statistically significant effects of intermediate relative performance after the first two experimental weeks, while the effect at the start of the final week of the experiment is positive and significant at the 10% level. Moreover, our estimates show that stores that perform as well as their best-performing comparison store in the first three weeks of the experiment increase sales in the final experimental week by 4.5%. Stores that outperform their best comparison store further increase sales by 0.25% per percentage point distance to their best competitor, while stores that lag behind their best comparison store likewise show a weaker increase in sales. These regression results show that the effects of the treatment and of intermediate relative performance are concentrated in the final week.

The estimations in Table 3 assume that the effect of intermediate relative performance is linear. The first column of Table 4 reports the results of estimating (8), where the effect of intermediate relative performance is allowed to vary between stores that lag far behind and stores that are close to or above the target for winning a bonus. Graphically, we allow the effect of intermediate relative performance to differ between observations to the left and right of the dashed line in Figure 8.[21] We find that intermediate relative performance does not affect current sales for stores that lag far behind. In contrast, current sales of treatment stores that lag less than 5 percentage points behind increases by 0.73 percent per percentage point increase in past relative performance. This is a substantial increase, particularly in the light of the stores' limited opportunities to boost sales. However, as only few treatment stores find themselves substantially ahead of their best comparison stores, this is not reflected in the average treatment effect reported in Table 2. A Wald test shows that the overall treatment effect is significantly different from zero for stores that are at least 6 percentage points ahead of their best comparison store (implying that a further 4 percentage points increase in relative performance results in winning a more valuable prize). These results confirm the intuition that stores' employees respond positively to the incentive scheme only when the bonus is within reach, and that this response is stronger when stores become closer to winning.

[20] We also estimated separate regressions for big stores and small stores. We find no differences in the response to intermediate relative performance. The estimates are reported in Delfgaauw et al. (2011a).

[21] Note that our indicator of stores being close to winning, $I_{t,w-1}$, is defined as $D_{t,w-1} > -5\%$, while the horizontal axis in Figure 8 plots the expected difference $p^{CU}_{t,w-1} - \max_{c_t}\left[p^{CU}_{c_t,w-1}\right]$ from the first-stage regression rather than our instrument $D_{t,w-1}$. As can be seen from the first-stage regression (second column of Table 3), the two move very close together, implying that virtually all stores to the right of the vertical line are considered to be close to winning.

The second column of Table 4 shows that the marginal effect of intermediate relative performance on current sales of relatively good-performing stores is significantly positive in all experimental weeks, with magnitudes ranging from 0.55 to 2.19 percent per percentage point. This relation is particularly strong in the third and fourth week of the experiment. For stores that lag far behind, there is no such effect in any of the weeks. This is not due to a lack of statistical power. Given the size of the standard errors, we should be able to detect at least some statistically significant effects if stores that lag far behind respond in a similar way as stores close to winning.

4.2 Robustness checks

Our results are qualitatively robust to varying the level of intermediate relative performance at which stores are deemed to be close to winning between -2% and -10%. Table 5 shows the estimated effects when the cut-off level is -5%, -2%, -7,5%, and -10%, respectively. In all cases, the effect of intermediate relative performance on current sales for stores deemed to stand a chance is positive and statistically significant. Quantitatively, the estimated effects are larger when the cut-off level is closer to -2%. The effects by week are less robust, and not statistically significant for all weeks. This is perhaps unsurprising, given that an increase in the cut-off level (to -2%) reduces the number of relevant observations and hence statistical power, while lowering the cut-off level reduces stores' perceived winning probabilities, in particular in the last week of the experiment.

Whether store employees perceive themselves as close to winning, however, may depend not only on the distance to the best performing comparison store, but also on the distance to the second-best comparison store. As noted earlier, we cannot disentangle the two effects due to problems of multicollinearity. As an alternative, we ran regressions where we used a modified definition of being close to winning. Specifically, we define stores as being close when, in addition to lagging at most 5% behind their best rival, they are at least 2.5%, at least 5%, or at most 5% ahead of their second-best comparison store. The point estimates are qualitatively similar irrespective of the distance to the second-best rival, but the coefficient loses significance if the number of stores defined as being close becomes too small.[22]

A causal interpretation of the estimations discussed above crucially depends on the exogeneity of our instrument. Our instrument is constructed under the assumption that performance of the comparison stores is unrelated to subsequent performance of the treatment stores, except for common shocks and the effect we aim to identify. More formally, we assume that $\varepsilon_{c_t,w-1}$ is not correlated with $\varepsilon_{t,w}$. One possible channel via which performance of the best comparison stores can influence subsequent performance of treatment stores is the existence of regional shocks that are correlated *across* regions and over time. In that case, regional shocks may influence the

[22] Results are available upon request. Of the 47 store-week observations where the treatment store lags at most 5% behind the best comparison store, 22 are at least 5% ahead of the second-best comparison store, and 36 are at least 2.5% ahead of the second-best comparison store.

predicted difference, $D_{t,w-1}$, as well as $y_{t,w}$. Note that the mere presence of serially-correlated, region-specific shocks is not sufficient to bias our estimates. As our group assignment procedure ensures that treatment and control stores are located in different regions, bias only arises if region-specific shocks in sales are correlated *across* regions, as well as over time.

To address this issue, we include region-specific week fixed effects in equations (5) and (8).[23] That is, we estimate θ_w in (5) and (8) for each region separately. We distinguish between 12 geographically-clustered regions, using the same region classification as the company does. The estimation results are reported in the fifth and sixth column of Table 3 and in the third column of Table 4, respectively. Our main results are by and large robust. The point estimate of the effect of intermediate relative performance for stores close to winning, reported in the third column of Table 4, is in the same order of magnitude (0.63% rather than 0.74%), and still statistically significant (at the 5% level). However, the estimated effects of treatment and intermediate relative performance in the final week of the experiment, reported in the sixth column of Table 3, are much smaller and no longer statistically significant. Possibly, we do not have a sufficiently large number of observations to identify region-specific week fixed effects as well as the effect of intermediate relative performance split out by week.[24]

Next, to assess the exogeneity of our instrument we conducted a placebo-experiment. In particular, we pretend that our experiment would have taken place in weeks 1-4 instead of weeks 5-8 of 2010, following similar estimation procedures as above.[25] If our instrument is free from serial correlation, we expect no significant effects of intermediate relative performance on subsequent performance. Serial correlation can bias our IV-estimates, as treatment stores' expected performance $E\left[p^{CU}_{t,w-1}\right]$ is based on past performance. Although we exclude the 3 weeks prior to the experiment from $E\left[p^{CU}_{t,w-1}\right]$, serial correlation in stores' own performance may still bias our instrument if there are long term trends in stores' performance. The results are reported in Table 6. The first column reports the OLS estimates, the second column the linear IV-estimates. As expected, the OLS estimates suggest a significant positive effect of intermediate relative performance on subsequent performance, but this effect vanishes in the IV-2SLS estimation. The third column of Table 6 estimates the effect of relative intermediate performance for stores that are close to winning. We find a positive treatment effect for stores close to winning, and this effect is decreasing in intermediate relative performance. Both effects are statistically significant at the 10%-level. At first sight, this seems to suggest that our instrument is biased by serial correlation, although the effect of intermediate relative performance for stores close to winning goes in the opposite direction of the main estimations. However, in contrast to the

[23] We do not adjust our instrument to allow for regional-specific shocks. By construction, our instrument, $D_{t,w-1}$, is neutral with respect to week fixed effects, but it is not possible to filter out regional shocks. The reason is that treatment stores and their best comparison store are not located in the same region, and hence are affected by different regional shocks.

[24] For this reason, we do not estimate the effect for stores close to winning separately by week.

[25] We keep the group assignment constant, as the assignment is based on performance in the weeks 40-53 in 2009. We reconstruct our instrument using the relevant time period. Hence, as before, $E\left[p^{CU}_{t,w-1}\right]$ is set equal to the average performance prior to the start of the 'experiment', excluding the last three weeks, i.e. average performance in weeks 40-50. Finally, we exclude the true experimental weeks from the estimations.

findings reported above, this effect becomes insignificant when we include region-specific week fixed effects: the point estimate drops to zero, while the standard errors remain of similar size (see column 4). Thus, the correlation is completely driven by region-specific shocks that are correlated both across regions and over time. By contrast, if serial correlation in stores' own performance would bias our instrument, adding regional-specific shocks should not matter.

4.3 Discussion

Taken together, our results paint the following picture. We have found that stores lagging too far behind do not respond to the incentive scheme, nor to the intermediate relative performance information. However, as stores are closer to winning a bonus, sales increase significantly with lagged relative performance. This effect is strongest in the second half of the experiment. This result contrasts with Casas-Arce and Martinez-Jerez (2009), who do not find that performance decreases when trailing contestants lag further behind. One explanation for their result, as conjectured by the authors, is attrition bias, which is absent in our study. Frank and Obloj (2011) also study a competition without attrition and find, like us, that performance is increasing in intermediate relative performance for contestants that lag behind.

On average, the relative performance incentive scheme that we study had no effect on sales. Possibly, many stores have perceived the relative performance targets as too ambitious from the start, particularly in the light of their limited means to boost sales. Such a perception would be reinforced after receiving the first poster with rankings, as only 23 treatment stores ranked on top of the first poster, and 64 stores lagged more than 5 percent behind their best-performing comparison store on the first poster.[26] An alternative explanation for the weak response is that the prospect of competing against non-competitors did not excite employees in the treatment stores as much as a real tournament would. It is hard to distinguish between these two interpretations. Attainability of the target clearly plays a role, as stores close to the winning positions respond positively to the treatment. This finding, however, does not rule out that a competition against competing competitors would have induced more substantial treatment effects. One specific reason why this would be the case is that feelings of envy play a more prominent role in a conventional tournament than in a comparable benchmark competition. In a conventional tournament, prize money is divided unequally among tournament participants, which gives losers reason to envy the winners. This stimulates effort among competitors, because by exerting effort, competitors can reduce the probability of experiencing envy (see Grund and Sliwka 2005 and Bartling 2011). The same reasoning does not apply to our benchmark competition, as employees in the treatment stores do not have a reason to envy employees in the comparison

[26] There is substantial persistence in the weekly rankings, but stores that do not rank first after the first week still have a reasonable probability of winning. Specifically, about 40% of the eventual winners of a prize did not rank first after the first week. These stores lagged behind by 8 to 9 percentage points, on average. Likewise, stores that rank first after the first week, have 52% probability of finishing in first position. So, there is sufficient persistence in the rankings for intermediate performance feedback to be valuable, but changes in the final rankings are still possible.

stores (since the latter cannot win a prize). In a recent lab experiment, Eisenkopf and Teyssier (2013) examine the role of envy in competitions by comparing stated effort in a conventional two-person tournament with stated effort in a similar benchmark competition. In the benchmark competition, effort is evaluated against the effort of a randomly chosen participant from the conventional tournament treatment. Hence, the two competitions are identical, except for the externality effort imposes on others. They find that the conventional tournament induces individuals to put in significantly more effort than the benchmark competition, suggesting that (anticipated) feelings of envy boost effort in a tournament. This may also explain why we fail to find a statistically significant treatment effect in our setting.[27]

5 Concluding Remarks

We have reported the results of a field experiment on dynamic incentive effects of relative performance pay among stores of a retail chain. We find that intermediate relative performance feedback affects subsequent performance of stores close to the bonus target. These stores show significantly higher performance, particularly near the end of the incentive period. Stores lagging far behind do not respond to the incentive scheme, nor to intermediate relative performance. As many treatment stores happen to trail far behind bonus targets over the course of the experiment, we find no improvement in performance on average.

Our findings underline the importance of dynamic incentive effects. When in the course of a contest the target moves out of reach, people give up, which renders the incentive scheme fruitless. On the other hand, learning that intermediate performance is closer to target encourages people to increase effort. Hence, the incentive effect of competitions is path-dependent.

References

[1] Aoyagi, Masaki (2010), Information Feedback in a Dynamic Tournament, *Games and Economic Behavior*, vol. 70, pp. 242-260.

[2] Arellano, Manuel, and Stephen Bond (1991), Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations, *Review of Economic Studies,* vol. 58(2), pp. 277-297.

[3] Asch, Beth J. (1990), Do Incentives Matter? The Case of Navy Recruiters, *Industrial and Labor Relations Review,* vol. 43, pp. 89S-106S.

[27] Yet another explanation might be the absence of competitors' strategic responses. We are not aware of any study testing this hypothesis. In Eisenkopf and Teyssier (2013), strategic considerations play no role: the tournament is one-shot and participants in the benchmark tournament are informed about their competitors' incentives.

[4] Azmat, Ghazala, and Nagore Iriberri (2010), The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment using High School Students, *Journal of Public Economics,* vol. 94(7-8), pp. 435-452.

[5] Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2013), Team Incentives: Evidence from a Firm Level Experiment, *Journal of the European Economic Association*, vol. 11(5), pp. 1079-1114.

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

[7] Bartling, Björn (2011), Relative Performance or Team Evaluation? Optimal Contracts for Other-Regarding Agents, *Journal of Economic Behavior and Organization*, vol. 79(3), pp. 183-193.

[8] Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004), How Much Should We Trust Differences-in-Differences Estimates?, *Quarterly Journal of Economics*, vol. 119(1), pp. 249-275.

[9] Blanes i Vidal, Jordi, and Mareike Nossol (2011), Tournaments without Prizes: Evidence from Personnel Records, *Management Science,* Vol. 57, pp. 1721-1736.

[10] Bouwens, Jan, and Peter Kroos (2011), Target Ratcheting and Effort Reduction, *Journal of Accounting and Economics,* vol. 51(1-2), pp. 171-185.

[11] Cadsby, C. Bram, Fei Song, and Francis Tapon (2010), Are You Paying Your Employees to Cheat? An Experimental Investigation, *B.E. Journal of Economic Analysis & Policy*, vol. 10(1), article 35.

[12] Casas-Arce, Pablo, and F. Asis Martinez-Jerez (2009), Relative Performance Compensation, Contests, and Dynamic Incentives, *Management Science,* vol. 55(8), pp. 1306-1320.

[13] Charness, Gary, Peter Kuhn, and Marie-Claire Villeval (2010), Competition and the Ratchet Effect, *Journal of Labor Economics,* vol. 29(3), pp. 513 - 547.

[14] Cooper, David, John Kagel, Wei Lo, and Qing Liang Gu (1999), Gaming Against Managers in Incentive Systems: Experimental Results with Chinese Students and Chinese Managers, *American Economic Review,* vol. 89(4), pp. 781-804.

[15] Courty, Pascal, and Gerald R. Marschke (2004), An Empirical Investigation of Gaming Responses to Explicit Performance Incentives, *Journal of Labor Economics* vol. 22(1), pp. 23-56.

[16] Delfgaauw, Josse, Robert Dur, Arjan Non, and Willem Verbeke (2011a), Dynamic Incentive Effects of Relative Performance Pay: A Field Experiment, Tinbergen Institute Discussion Paper 2010-124/1

[17] Delfgaauw, Josse, Robert Dur, Arjan Non, and Willem Verbeke (2011b), The Effects of Prize Spread and Noise in Elimination Tournaments: A Natural Field Experiment, Tinbergen Institute Discussion Paper 2011-120/1.

[18] Delfgaauw, Josse, Robert Dur, Joeri Sol, and Willem Verbeke (2013), Tournament Incentives in The Field: Gender Differences in The Workplace, *Journal of Labor Economics*, vol. 32(2), pp. 305-326.

[19] Ederer, Florian (2010), Feedback and Motivation in Dynamic Tournaments, *Journal of Economics & Management Strategy*, vol. 19(3), pp. 733-769.

[20] Eisenkopf, Gerald, and Sabrina Teyssier (2013), Envy and Loss Aversion in Tournaments, *Journal of Economic Psychology*, vol. 34, pp. 240–255.

[21] Erev, I., G. Bornstein, and G. Rachely (1993), Constructive Intergroup Competition as a Solution to the Free Rider Problem: A Field Experiment, *Journal of Experimental Social Psychology,* vol. 29(6), pp. 463-478.

[22] Fershtman, Chaim, and Uri Gneezy (2011), The Trade-off between Performance and Quitting in High-Power Tournaments, *Journal of the European Economic Association,* vol. 9(2), pp. 318-336.

[23] Fonseca, Miguel A. (2009), An experimental investigation of asymmetric contests, *International Journal of Industrial Organization*, vol. 27(5), pp. 582-591.

[24] Forbes, Silke J., Mara Lederman, and Trevor Tombe (2012). Quality Disclosure Programs with Thresholds: Misreporting, Gaming and Employee Incentives, mimeo, University of California, San Diego.

[25] Frank, Douglas, and Tomasz Obloj (2011), Reference Points and Organizational Performance: Evidence from Retail Banking, mimeo, INSEAD.

[26] Gershkov, Alex, and Motty Perry (2009), Tournaments with Midterm Reviews, *Games and Economic Behavior,* vol. 66, pp. 162-190.

[27] Goltsman, Maria and Arijit Mukherjee (2011), Interim Performance Feedback in Multistage Tournaments: The Optimality of Partial Disclosure, *Journal of Labor Economics,* vol. 29(2), pp. 229-265.

[28] Grund, Christian and Dirk Sliwka (2005), Envy and Compassion in Tournaments, *Journal of Economics and Management Strategy*, vol. 14(1), pp. 187-207.

[29] Harrison, Glenn, and John A. List (2004), Field Experiments, *Journal of Economic Literature*, vol. 42(4), pp. 1009-1055.

[30] Höchtl, Wolfgang, Rudolf Kerschbamer, Rudi Stracke, and Uwe Sunde (2011), Incentives vs. Selection in Promotion Tournaments: Can a Designer Kill Two Birds with One Stone?, IZA discussion paper no 5755.

[31] Kosfeld, Michael, and Susanne Neckermann (2011), Getting More Work for Nothing? Symbolic Awards and Worker Performance, *American Economic Journal: Microeconomics,* vol. 3(3), pp. 86-99.

[32] Larkin, Ian (2013), The Cost of High-Powered Incentives: Employee Gaming in Enterprise Software Sales. *Journal of Labor Economics,* forthcoming.

[33] Lazear, Edward P., and Sherwin Rosen (1981), Rank-Order Tournaments as Optimum Labor Contracts, *Journal of Political Economy*, vol. 89(5), pp. 841-864.

[34] Ludwig, Jens, Jeffrey R. Kling, and Sendhil Mullainathan (2011), Mechanism Experiments and Policy Evaluations, *Journal of Economic Perspectives*, 25(3), pp. 17-38.

[35] Ludwig, Sandra, and Gabriele Lünser (2012), Observing your competitor - The role of effort information in two-stage tournaments. *Journal of Economic Psychology,* vol. 33(1), pp. 166-182.

[36] O'Keeffe, Mary, W. Kip Viscusi, Richard J. Zeckhauser (1984), Economic Contests: Comparative Reward Schemes, *Journal of Labor Economics*, vol. 2(1), pp. 27-56.

[37] Oyer, Paul (1998), Fiscal Year Ends and Nonlinear Incentive Contracts: The Effect on Business Seasonality, *Quarterly Journal of Economics,* vol. 113(1), pp. 149-185.

[38] Rosen, Sherwin (1986), Prizes and Incentives in Elimination Tournaments, *American Economic Review,* vol. 76(4), pp. 701-715.

[39] Royal Dutch Shell (2009), Annual Report and Form 20-F for the Year ended December 31, 2009.

[40] Schweitzer, Maurice, Lisa Ordóñez, and Bambi Douma (2004), Goal-setting as a Motivator of Unethical Behavior. *Academy of Management Journal*, vol. 47(3): 422-433.

Figure 1: Example of the weekly feedback posters (translated into English)

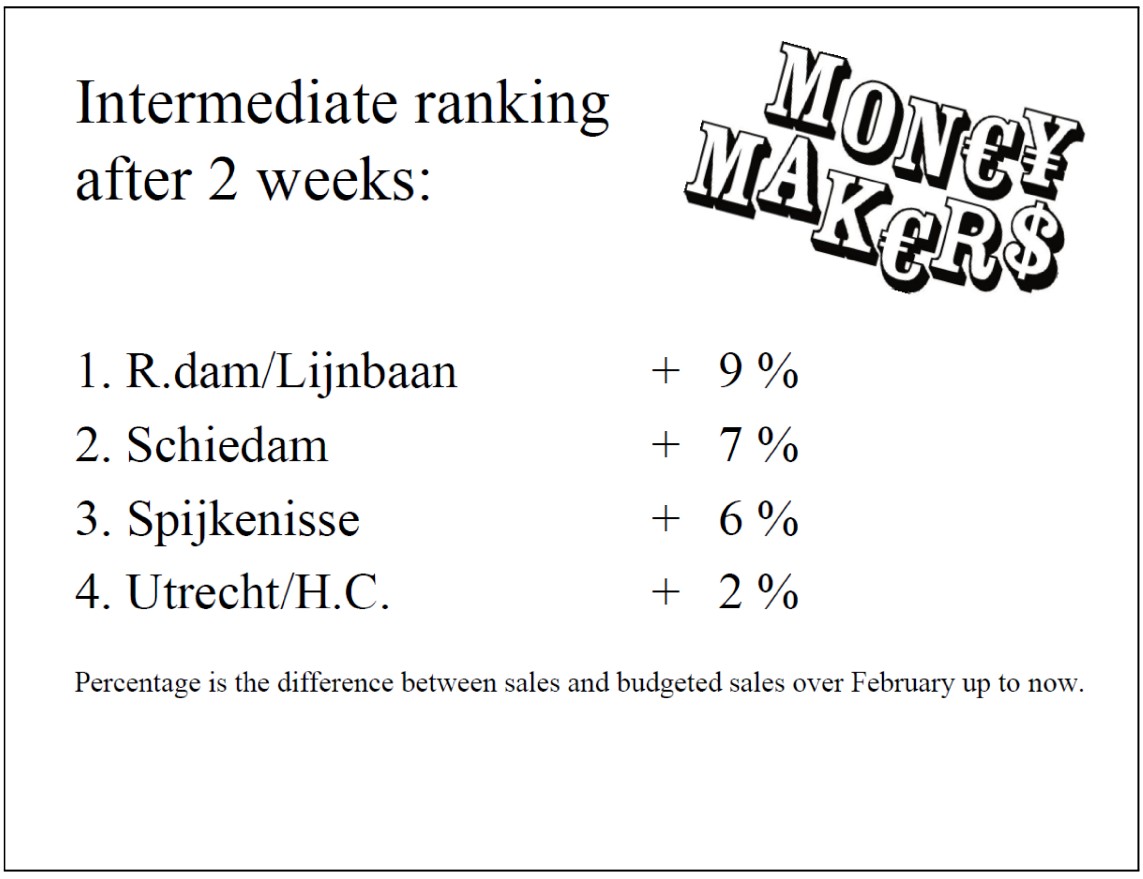


Figure 2: Average sales per store

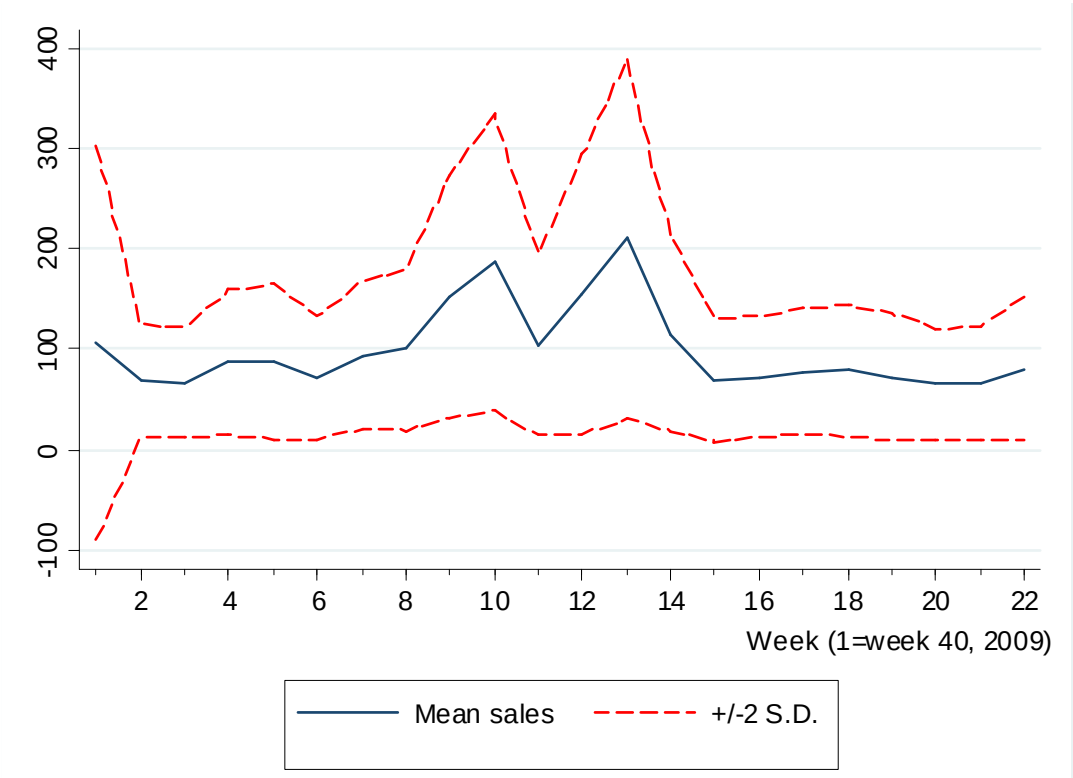


Figure 3: Average weekly performance

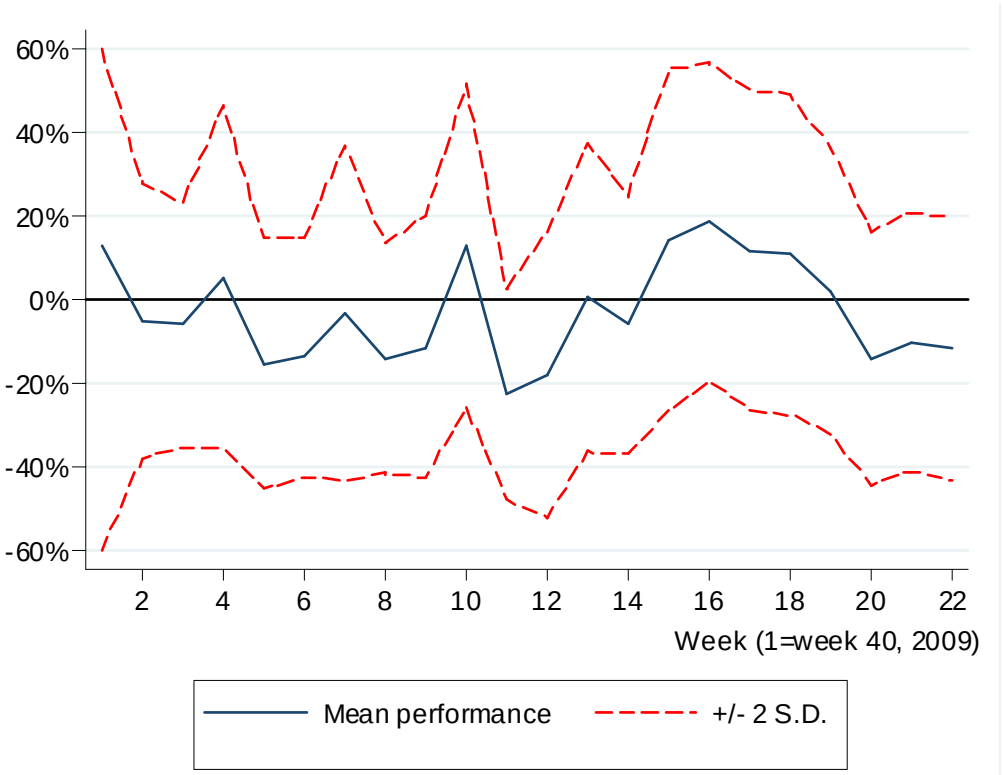


Figure 4: Average sales of treatment stores divided by average sales of control stores

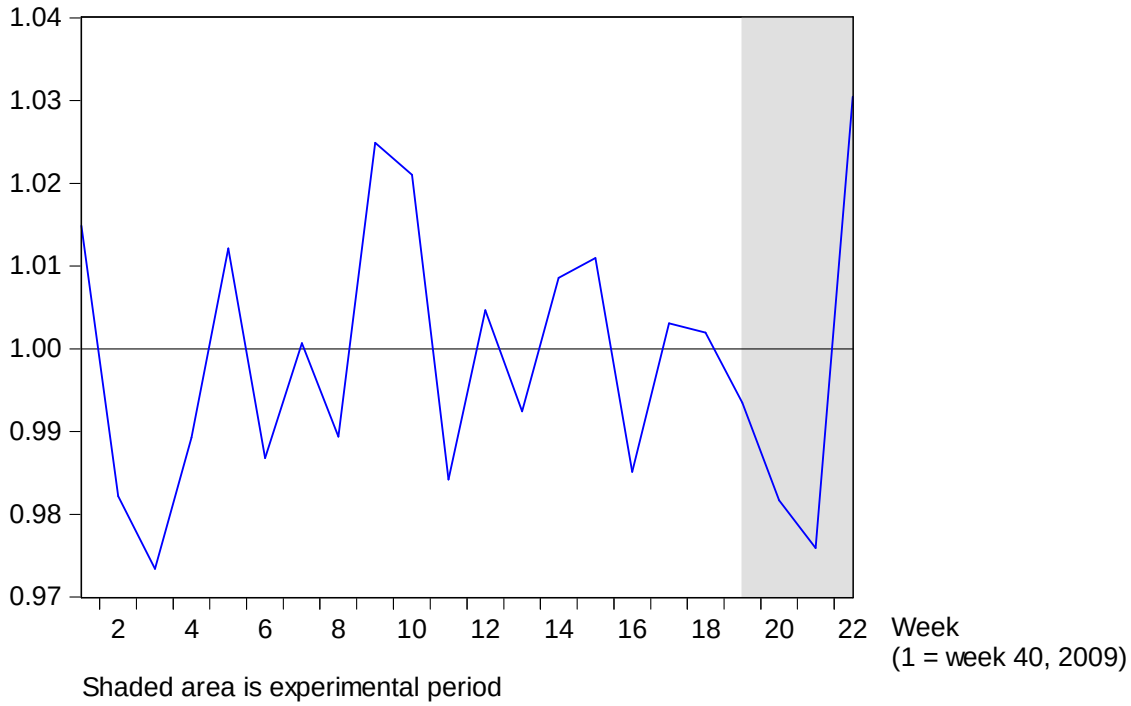
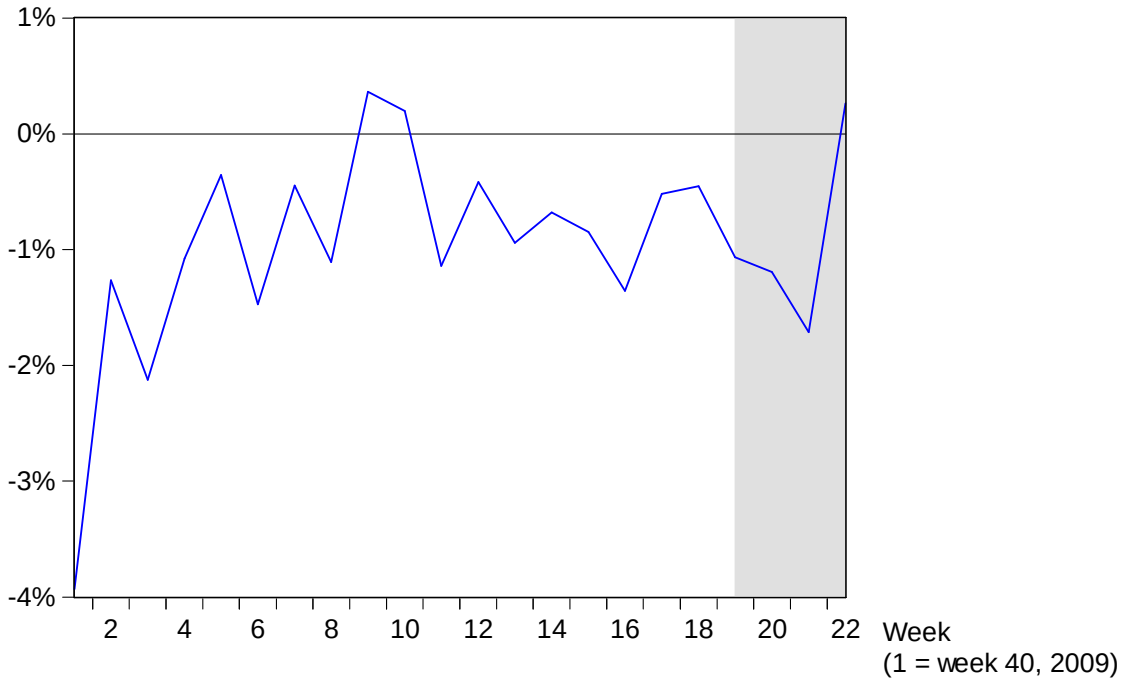
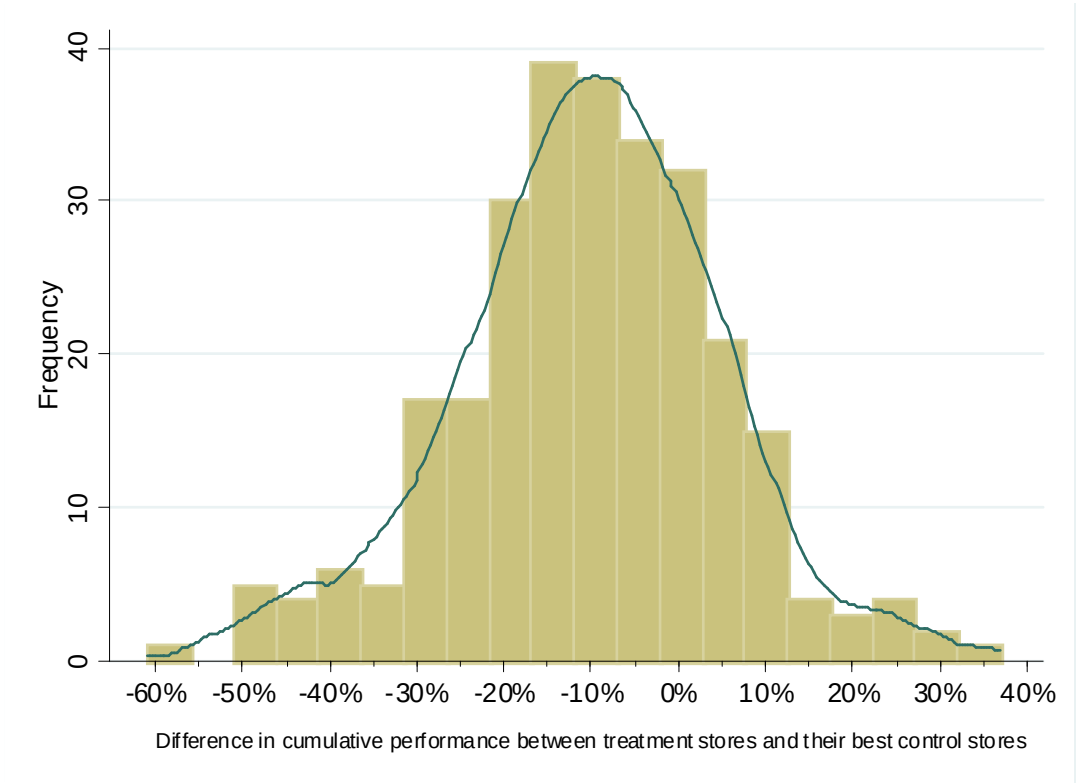


Figure 5: Difference in average performance between treatment stores and control stores



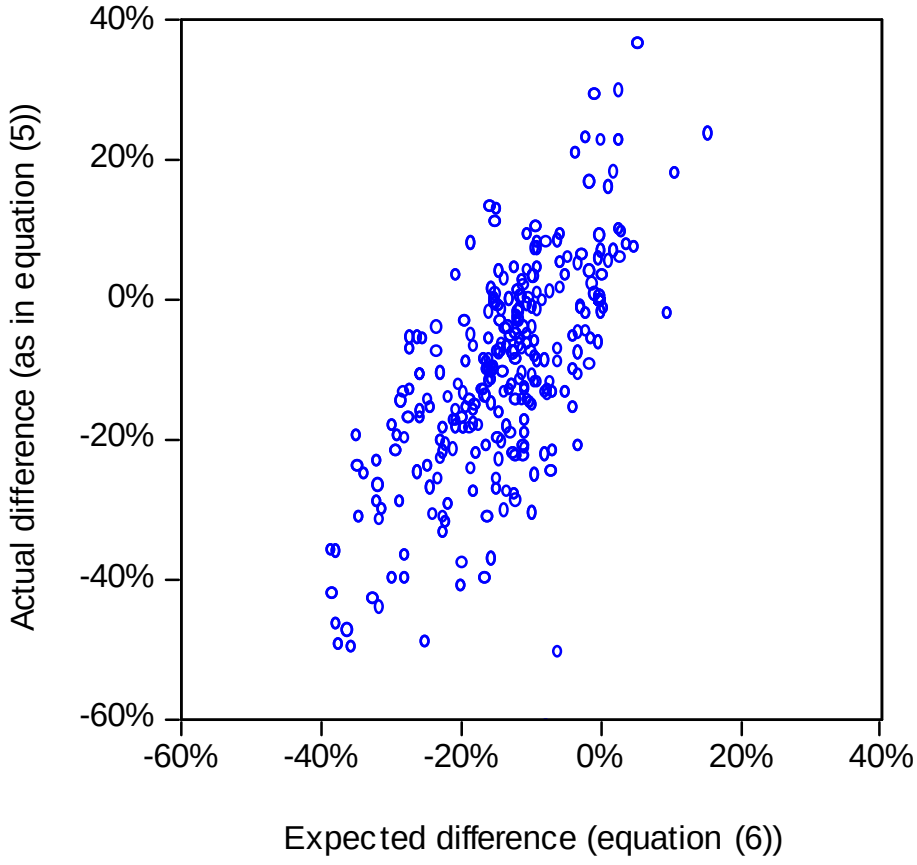
Performance is measured as sales in % deviation from budgeted sales
Shaded area is experimental period

Figure 6: Distribution of actual differences in intermediate cumulative performance between the treatment store and its best comparison store



(One outlier not shown (-101%))

Figure 7: Actual and expected difference in intermediate cumulative performance between the treatment store and its best comparison store ($p^{CU}_{t,w-1} - \max_{c_t}\left[p^{CU}_{c_t,w-1}\right]$)



One outlier not shown (actual difference -101%; expected difference -24%)

Figure 8: The relation between the expected difference $E(p^{CU}_{t,w-1}) - \max_{c_t}\left[p^{CU}_{c_t,w-1}\right]$ and the residuals from estimating (3)

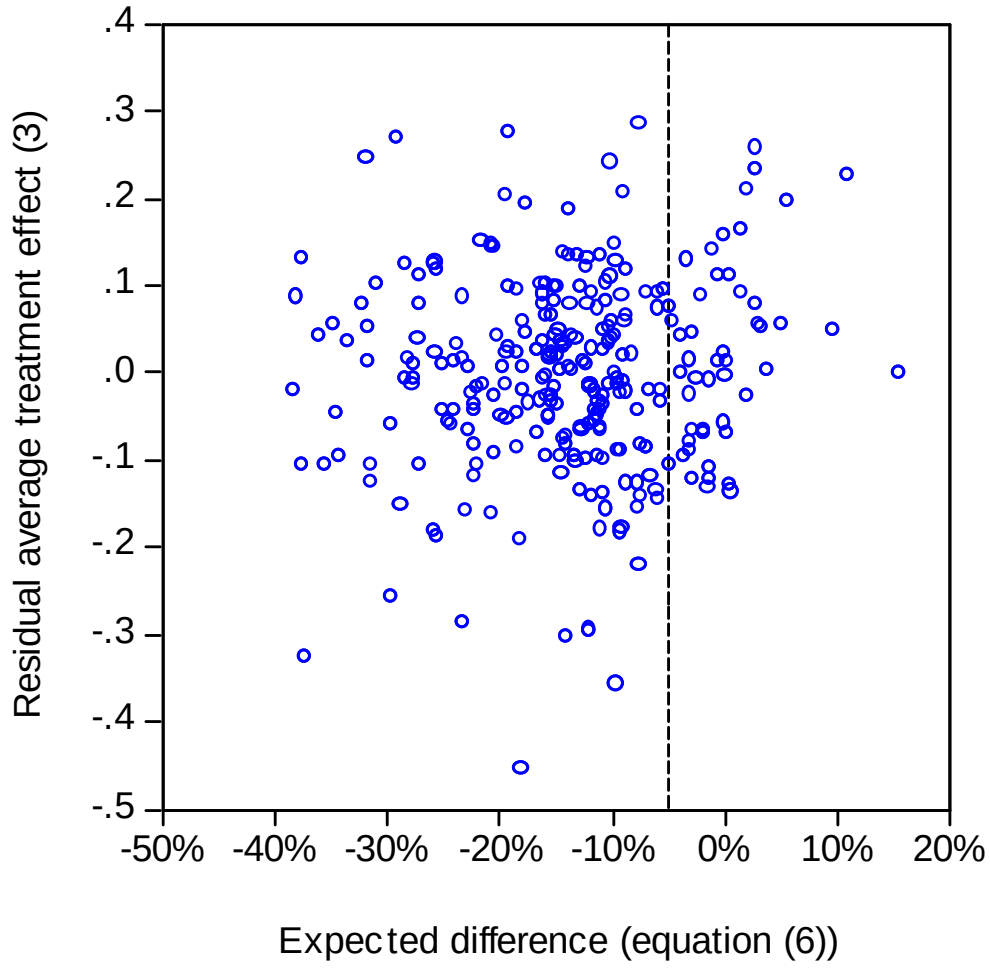


Table 1: Descriptive statistics

	All stores		Treatment stores		Control stores	
	mean	Std	Mean	Std	Mean	Std
Sales	100.00	40.31	100.04	43.77	99.96	36.89
Sales weeks 40/2009 - 53/2009	115.18	46.62	115.27	50.81	115.08	42.43
Sales weeks 5/2010 - 8/2010	71.67	29.90	71.56	31.85	71.78	28.05
Budgeted sales	104.56	41.54	105.33	45.42	103.81	37.63
Budgeted sales weeks 40/2009 - 53/2009	122.90	48.82	123.80	53.38	122.02	44.23
Budgeted sales weeks 5/2010 - 8/2010	78.52	31.19	79.10	34.10	77.96	28.86
Performance	-0.03	0.12	-0.04	0.11	-0.02	0.12
Performance weeks 40/2009 - 53/2009	-0.06	0.11	-0.07	0.10	-0.05	0.12
Performance weeks 5/2010 - 8/2010	-0.08	0.14	-0.09	0.13	-0.08	0.14
Average within-store standard deviation sales	44.27	21.06	44.40	22.90	44.15	19.11
Av. within-store st. dev. sales weeks 40-53 /2009	49.18	24.16	49.33	26.16	49.03	22.06
Av. within-store st. dev. sales weeks 5-8 /2010	9.55	5.97	9.88	6.65	9.22	5.22
Average within-store st. dev. performance	0.19	0.12	0.18	0.10	0.20	0.14
Av. within-store st. dev. performance weeks 40-53	0.18	0.16	0.17	0.13	0.19	0.19
Av. within-store st. dev. performance weeks 5-8	0.11	0.05	0.10	0.04	0.11	0.06
Average within-store st. dev. performance corrected for common shocks	0.14	0.13	0.13	0.10	0.15	0.15
Number of employees	5.45	1.99	5.24	1.69	5.66	2.22
Number of days closed for carnival (week 7/2010)	0.25	0.63	0.19	0.56	0.31	0.69
Number of stores	189		93		96	

Performance is defined as (sales-budgeted sales)/budgeted sales
For confidentiality reasons, sales and budgeted sales figures are indexed to the average sales per store per week over the whole sample.
None of the differences between treatment stores and control stores are significant at the 10%-level.

Table 2: Average treatment effect

	Dependent variable: ln(sales)			
	(1)	(2)	(3)	(4)
Treatment	-0.004		-0.008	
	(0.013)		(0.013)	
Treatment*close			0.037	
			(0.023)	
Treatment week 1		-0.003		-0.003
		(0.019)		(0.019)
Treatment week 2		-0.013		-0.025
		(0.020)		(0.022)
Treatment week 3		-0.020		-0.022
		(0.017)		(0.018)
Treatment week 4		0.021		0.015
		(0.019)		(0.019)
Treatment week 2*close				0.056*
				(0.030)
Treatment week 3*close				0.017
				(0.032)
Treatment week 4*close				0.039
				(0.037)
Carnival	-0.026*	-0.028*	-0.026*	-0.028*
	(0.015)	(0.015)	(0.015)	(0.015)
Store-fixed effects	yes	yes	yes	yes
Week-fixed effects	yes	yes	yes	yes
Store-week observations	4158	4158	4158	4158
Stores	189	189	189	189
R^2	0.9281	0.9281	0.9281	0.9281

Standard errors clustered at the store level in parentheses.
***, **, * denote statistically significant effects at the 1%, 5%, and 10% level, respectively.

Table 3: Dynamic incentives

	Dependent variable: ln(sales)					
					Region-Week-fixed effects	
	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	IV-2SLS	IV-2SLS	IV-2SLS	IV-2SLS	IV-2SLS
Stage IV-2SLS		First	Second	Second	Second	Second
Treatment	0.017	0.024**	0.003		-0.002	
	(0.012)	(0.011)	(0.014)		(0.012)	
Relative intermediate performance	0.0026***		0.001		0.0011	
	(0.0005)		(0.0008)		(0.0007)	
Expected intermediate relative performance, $D_{t, w-1}$ (equation (6))		1.003***				
		(0.077)				
Treatment week 1				-0.003		-0.004
				(0.019)		(0.018)
Treatment week 2				-0.0003		-0.0118
				(0.022)		(0.023)
Treatment week 3				-0.023		-0.012
				(0.023)		(0.024)
Treatment week 4				0.045**		0.0128
				(0.023)		(0.026)
Relative performance after week 1				0.001		0.0014
				(0.001)		(0.001)
Relative performance after week 2				-0.0003		0.0006
				(0.0014)		(0.0015)
Relative performance after week 3				0.0025*		0.0008
				(0.0013)		(0.0017)
Carnival	-0.020	-0.013*	-0.024	-0.029*	-0.075***	-0.076***
	(0.014)	(0.007)	(0.018)	(0.016)	(0.021)	(0.021)
Store-fixed effects	yes	yes	yes	yes	yes	yes
Week-fixed effects	yes	yes	yes	yes	no	no
Region-Week-fixed effects	no	no	no	no	yes	yes
Store-week observations	4158	4158	4158	4158	4158	4158
Stores	189	189	189	189	189	189
R^2	0.9284	0.6358	0.9283	0.9283	0.9542	0.9542

Standard errors clustered at the store level in parentheses.
The dependent variable in the first-stage regression shown in column (2) is relative intermediate performance.
***, **, * denote statistically significant effects at the 1%, 5%, and 10% level, respectively.

Table 4: Dynamic incentives separate for stores close to winning a bonus

	Dependent variable: ln(sales)		
	(1)	(2)	Region-week-fe (3)
	IV-2SLS	IV-2SLS	IV-2SLS
	Second stage	Second stage	Second stage
Treatment	-0.009		-0.008
	(0.017)		(0.014)
Relative intermediate performance	-0.0001		0.0007
	(0.001)		(0.0009)
Treatment*close	-0.0049		-0.021
	(0.025)		(0.030)
Relative intermediate performance*close	0.0074***		0.0063**
	(0.0017)		(0.0032)
Treatment week 1		-0.003	
		(0.019)	
Treatment week 2		-0.046	
		(0.039)	
Treatment week 3		-0.051	
		(0.034)	
Treatment week 4		0.039	
		(0.034)	
Relative performance after week 1		-0.0012	
		(0.0019)	
Relative performance after week 2		-0.0022	
		(0.0021)	
Relative performance after week 3		0.0019	
		(0.0021)	
Treatment week 2*close		0.057	
		(0.044)	
Treatment week 3*close		-0.112	
		(0.128)	
Treatment week 4*close		-0.037	
		(0.045)	
Relative performance after week 1*close		0.0055*	
		(0.0030)	
Relative performance after week 2*close		0.0219**	
		(0.011)	
Relative performance after week 3*close		0.0087***	
		(0.003)	
Carnival	-0.024	-0.027	-0.074***
	(0.015)	(0.018)	(0.021)
Store-fixed effects	yes	yes	yes
Week-fixed effects	yes	yes	no
Region-week-fixed effects	no	no	yes
Store-week observations	4158	4158	4158
Stores	189	189	189
R^2	0.9282	0.9279	0.9542

Standard errors clustered at the store level in parentheses.
"Close" is a dummy variable that takes value one when the store's expected intermediate performance is at most 5 percentage points below its best comparison store, i.e. when $D_{t,\,w-1} > -5\%$ (see equation (6)).
***, **, * denote statistically significant effects at the 1%, 5%, and 10% level, respectively.

Table 5: Varying the cut-off for being close

	Dependent variable: ln(sales)							
Cut-off level	-5%	-5%	-2%	-2%	-7.5%	-7.5%	-10%	-10%
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Relative intermediate performance*close	0.0074***		0.0098**		0.0056***		0.0049***	
	(0.0017)		(0.0044)		(0.0013)		(0.0015)	
Relative performance after week 1		0.0055*		0.0046		0.0051**		0.0062**
		(0.0030)		(0.0045)		(0.0026)		(0.0024)
Relative performance after week 2		0.0219**		0.0609		0.0099*		0.0068*
		(0.0111)		(0.0839)		(0.0055)		(0.0039)
Relative performance after week 3		0.0087***		0.0102**		0.0040		0.001308
		(0.0030)		(0.0045)		(0.0027)		(0.0038)
Store-fixed effects	yes	yes	yes	yes	yes	yes	yes	yes
Week-fixed effects	yes	yes	yes	yes	yes	yes	yes	yes
Region-Week-fixed effects	no	no	no	no	no	no	no	no
Store-week observations	4158	4158	4158	4158	4158	4158	4158	4158
Number of stores that are close	47		31		61		90	
Number of stores that are close, week20		20		15		22		26
Number of stores that are close, week21		14		9		20		32
Number of stores that are close, week22		13		7		19		32
Stores	189	189	189	189	189	189	189	189
R^2	0.9282	0.9279	0.9282	0.9265	0.9283	0.9281	0.9285	0.9285

Standard errors clustered at the store level in parentheses.
The regression specification is described in equation (8), and is identical to the regressions shown in table 4.
The variables treatment, relative intermediate performance, treatment*close, and carnival are included, but not reported.
***, **, * denote statistically significant effects at the 1%, 5%, and 10% level, respectively.

Table 6: Results of a placebo-experiment in weeks 1-4, 2010

	Dependent variable: ln(sales)			
	(1) OLS	(2) IV-2SLS Second stage	(3) IV-2SLS Second stage	(4) Region-week-fe IV-2SLS Second stage
Placebo treatment	0.015	0.0011	-0.0052	-0.022
	(0.013)	(0.014)	(0.017)	(0.014)
Placebo relative int. performance	0.0017**	-0.0003	-0.0004	-0.0009
	(0.0007)	(0.0007)	(0.001)	(0.0008)
Placebo treatment*close			0.063**	0.044
			(0.031)	(0.025)
Placebo relative int. performance*close			-0.0072*	-0.0011
			(0.0038)	(0.0032)
Store-fixed effects	yes	yes	yes	yes
Week-fixed effects	yes	yes	yes	no
Region-week-fixed effects	no	no	no	yes
Store-week observations	3402	3402	3402	3402
Stores	189	189	189	189
R^2	0.9246	0.9243	0.9233	0.9535

Standard errors clustered at the store level in parentheses.
"Close" is a dummy variable that takes value one when the store's expected intermediate performance is at most 5 percentage points below its best comparison store, akin to $D_{t,\,w-1} > -.05$ (see equation (6)), but adjusted to the placebo period.
***, **, * denote statistically significant effects at the 1%, 5%, and 10% level, respectively.