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

DISCUSSION PAPER SERIES

IZA DP No. 12446

Talking about Performance or Paying for it? Evidence from a Field Experiment

Kathrin Manthei
Dirk Sliwka
Timo Vogelsang

JUNE 2019

I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12446

Talking about Performance or Paying for it? Evidence from a Field Experiment

Kathrin Manthei
RFH Koeln

Dirk Sliwka
University of Cologne and IZA

Timo Vogelsang
University of Cologne

JUNE 2019

ISSN: 2365-9793

IZA – Institute of Labor Economics

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

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

www.iza.org

ABSTRACT

Talking about Performance or Paying for it? Evidence from a Field Experiment*

We investigate the causal effect of conversations about performance and performance pay implementing a 2x2 field experiment in a retail chain. In the performance pay treatments, managers receive a bonus for profit increases. In the performance review treatments, managers have regular meetings with their supervisors discussing their activities to increase profits. We find that review conversations raise profits by 7%-8%. However, when additionally receiving performance pay this effect vanishes. Analyzing an extension of Bénabou and Tirole (2006), we rationalize this effect formally and provide empirical evidence that the use of performance pay changes the nature of conversations undermining their value.

JEL Classification: J3, L2, M5, C93

Keywords: performance pay, performance reviews, monitoring, feedback, field experiment, management practices

Corresponding author:
Dirk Sliwka
University of Cologne
Faculty of Management, Economics and Social Sciences
Albertus-Magnus-Platz
50923 Köln
Germany

E-mail: sliwka@wiso.uni-koeln.de

* The field experiments were registered with the ID AEARCTR-0002128.

1. Introduction

Influencing employees to act in the interest of employers has been a key focus of the literature in organizational economics (see, e.g., Prendergast 1999, Gibbons and Roberts 2013, or Lazear 2018 for surveys). Economists have traditionally stressed the importance of performance pay to align employees' behavior with the objectives of employers. However, organizations also adopt non-monetary practices to guide the behavior of employees.[1] In particular, in recent years many larger companies have revised their practices to manage employee performance – often reducing the role of individual rewards and focusing more on establishing regular conversations about performance between supervisors and subordinates.[2] In his Nobel lecture, Holmström (2017), for instance, argues that such practices affect performance as they trigger reputational concerns.[3] The key aim of this paper is to study whether introducing regular conversations about a specific outcome variable can indeed raise performance, and how the effect of these conversations compares to and interacts with effects of performance pay tied to the same outcome variable.

From an economic perspective, performance review conversations may be viewed as monitoring devices. Employees anticipate that they will be observed and have to explain their actions, which increases their costs of shirking and results in higher performance. In this case, monitoring and performance pay can be viewed as alternative solutions to reduce moral hazard problems.[4] Lazear and Oyer (2013, p. 486), for instance, state "*An alternative to financial incentives is to simply monitor workers. If a supervisor can keep close watch over employees, she can ensure that the employee takes the best actions.*" Therefore, both can be substitutes: both management practices raise performance, but the gain from using performance pay may be smaller when performance reviews are in place and vice versa.

[1] For recent surveys on the economics of non-monetary incentives see e.g. Ellingsen and Johannesson (2007), Rebitzer and Taylor (2011) or Cassar and Maier (2018).

[2] Cappelli and Tavis (2016) survey the development of performance management practices in larger firms. Examples for larger firms that recently entirely stopped or strongly revised standard annual performance ratings tied to rewards and instead established regular feedback conversations are for instance Adobe, Deloitte, Lear, Microsoft, IBM, Bosch, or SAP.

[3] Holmström, for instance, states that "*The craving for appreciation and the desire to impress superiors explains why a mere change in the accounting system can have a big impact on the behavior of employees*" (Holmström 2017, p.1753).

[4] When monitoring generates more precise information that can be used for performance pay, then monitoring and performance pay can also be complements (compare, for instance Milgrom and Roberts (1992, pp. 226). However, in this study we consider a setting where performance pay uses an objective key figure that is not generated through monitoring. See also Holmström (1979), Jensen and Meckling (1976), Eisenhardt (1989) for early discussions on monitoring from an economic perspective.

We investigate the causal effects of introducing performance review conversations and performance pay as well as their interaction in a field experiment in a retail chain with 224 store managers.[5] Store managers were randomly assigned to one of four treatments in a 2x2 design: performance pay, performance reviews, receiving both, or none of the two (the control group). Store managers in the performance pay conditions received bonus payments that were simple linear functions of the profits achieved above a threshold value. Store managers in the performance review conditions had meetings with their supervisors every two weeks in which they had to report their activities to raise profits as well as their planned next steps. Store managers in all treatments and the control group also received an "information package" consisting of an online training and information about profit margins in order to refresh knowledge on the stores' profit production function and to exclude the possibility that results are driven by mere attention or "experimenter demand" effects.

We hypothesize that the performance review conversations generate an additional (psychological or economic) cost for an agent who is not able to demonstrate efforts to increase profits to his supervisor in the review meetings. As illustrated in a simple formal model, performance reviews and performance pay should both increase performance. However, the instruments are substitutes as the additional effect of performance pay should be weaker when performance reviews are conducted and vice versa.

Our key result is that the introduction of performance reviews indeed increases profits by approximately 7%-8% in the treated stores. Yet, this positive effect of the performance reviews vanishes when it is accompanied by performance pay. Hence, while we hypothesized that performance pay reduces the *marginal* effect of introducing performance reviews, we find that – in contrast to our expectation – it even reduces the *absolute* effect of this practice. Moreover, performance pay alone does not raise profits significantly above the effects of the information on the respective outcome variable provided to all stores. These results are robust to a variety of specifications and different estimation techniques.

We provide an economic rationale for this finding extending the Bénabou and Tirole (2006) model. In the model, agents benefit from a higher reputation about their willingness or ability to exert effort. The use of monetary rewards can reduce the power of the reputational incentive mechanism as such rewards (in the words of Bénabou and Tirole (2006, p. 1652))

[5] See Harrison and List (2004), Bandiera et al. (2011) or Floyd and List (2016) for recent surveys on field experiments.

"*create doubt about the true motive*" for which an action is taken. In turn, bonus payments can reduce performance when reputational incentives are strong. The key idea of our extension is that performance reviews generate more transparency about the agent's activities to raise profits, facilitating the signaling of motivation which leads to higher powered incentives.[6] We show that in this framework any detrimental effect of bonus payments is naturally stronger when performance reviews are in place. When supervisors receive more precise direct information about efforts exerted, the reputational incentive mechanism is strengthened. But, by the same token, the detrimental effect of bonuses on the reputational incentive mechanism is larger when reviews are in place. In other words, performance pay may undermine the reputational incentives triggered through the review meetings and, in turn, can affect the quality of the interaction between supervisor and subordinate.

We then explore the character of the review conversations empirically, finding evidence that the nature of the conversations indeed changes when performance pay is used. Supervisors conducting the reviews in our field experiment had been asked to write down short protocols noting activities undertaken by the respective store managers, occurring problems, and next steps. Analyzing these protocols, we find that a significantly smaller number of problems are stated by store managers in the performance pay condition. In fact, in about 60% of the cases not a single problem is mentioned in any of the meetings when the bonus is in place, but this fraction drops to 22% when no bonus is paid, which indicates a significant change in the nature of the conversations. Moreover, when we consider only stores in which conversations included an open discussion of problems, the negative effect of performance pay tends to vanish.

While it was not our initial hypothesis, our results are also closely related to the literature in behavioral economics on potential detrimental effects of incentives (e.g. Gneezy and Rustichini 2000a, 2000b, Bénabou and Tirole 2003, 2006, Fehr and Rockenbach 2003, Fehr and List 2004, Falk and Kosfeld 2006, Sliwka 2007, Ellingsen and Johannesson 2008, Ariely, Bracha, Meier 2009) which has so far mostly shown the existence of such detrimental effects in laboratory experiments rather than within firms. In our study, performance pay is not detrimental in itself, but it undermines the beneficial effect that structured conversations between supervisors and subordinated have on performance. Moreover, our study complements arguments put forward in the literature on biases in subjective performance evaluations of

[6] This is essentially a classical Holmström (1999)-type career concerns effect: reducing noise in performance signals increases efforts.

employees. Prendergast (1999), for instance, discusses the argument that bonus payments can change the nature of these biases and that "[…] *many firms now explicitly separate pay setting from subjective evaluations*" (Prendergast 1999, p. 30).[7]

We also contribute to the literature on performance feedback and monitoring. The effect of pure quantitative feedback on performance has been studied extensively in recent years with rather mixed results. While some studies find positive effects on performance (e.g. Blanes i Vidal and Nossol 2011, Tran and Zeckhauser 2012) other find negative effects (Barankay 2012, Ashraf et al. 2014, Bradler et al. 2016) or no effect (Lourenço 2016). However, to the best of our knowledge the interplay between qualitative supervisor feedback and performance pay has not been previously studied in a field experiment. There are only few studies on the causal effects of monitoring in firms. Nagin et al. (2002) find a heterogeneous effect of monitoring intensity with a negative effect among those workers who perceived the monitoring as unfair and no effect among those who do not. Banker et al. (2018) investigate the interaction of performance pay and an already existing monitoring scheme and find a decreasing marginal impact of performance pay the higher the level of supervisor monitoring.[8]

Finally, the paper also contributes to the growing economic literature on the use of management practices within firms (Ichniowski et al. 1997, Bartel et al. 2004, Bloom and Van Reenen 2007, Merchant and Van der Steede 2017) and their causal effects on firm performance (see, e.g., Bandiera et al. 2011, Bloom et al. 2013, 2015, Delfgaauw et al. 2013, Friebel et al. 2017, Manthei et al. 2018). While most of the existing field experiments have varied the use of a single practice, we study the interplay of two management practices in a 2x2 experimental design.

The paper proceeds as follows. Section 2 describes the organization we study. Section 3 describes the details of the experimental design and its implementation. Section 4 describes the key hypothesis we had at the outset. Section 5 presents the experimental results and section 6 concludes.

[7] His argument, however, is different from the mechanism we suggest. He conjectures that poor feedback may be harder to communicate for supervisors once a monetary bonus is attached to the rating. In our setting, the performance reviews do not determine the bonus payment, as the bonus is based on objective performance measures and the supervisors have no influence on the size of the bonus.

[8] See also Boly (2011) and Belot and Schröder (2016) for lab and field experiments on monitoring in which sanctions are tied to the result of the monitoring outcome such that the effects of monitoring are not disentangled from the effects of performance pay. Campbell et al. (2011) use data from a casino chain in which each casino could decide about the intensity of the monitoring. They find that tight monitoring leads to strong implicit incentives which leads to less experimentation and learning.

2. The Organization

The company is a nationwide retailer, operating discount supermarkets in Germany consisting of several larger geographical regions. Each region has a regional top manager and sales area managers. The sales area managers supervise about 4-6 district managers. District managers are responsible for 5-8 store managers. The main duty of district managers is supervision of store managers, whom they visit approximately twice per week. A store consists of approximately 5-8 FTE and the store manager is responsible for the daily routines within the store.

Store managers have limited leeway within their operational tasks as discount retailing is generally characterized by highly standardized tasks and processes (for instance, concerning the placement and ordering of products). While a computer system generates recommendations for order quantities of products, the store manager can overwrite these suggestions. Moreover, they can decide on special placements of goods within a limited area in the store. Store managers' main duty lies in the execution of daily operational tasks such as keeping the store clean, the presentation of products, the availability of products in the shelves and an efficiently working cash desk (see Table A1 in the Appendix for an overview of possible tasks).

3. The Experiment

We introduced performance pay and performance reviews for store managers, implementing a 2x2 factorial experimental design over three months (April – June 2017). Prior to the intervention, store managers were not systematically trained to work with store profits and were mostly concerned with managing sales.[9] One aim of the intervention was to get store managers to focus more on profits, thus broadly taking into account the effects of their actions on both sales and the respective costs. The key performance metric for this intervention is a simplified form of the store's profits

Profit = Net Sales – Cost of Goods Sold – Staff Costs –Inventory Losses

which covers all key elements of company performance a store manager can influence.[10]

[9] See Manthei et al. (2018) for a performance pay field experiment within the same organization using a sales-base performance measure.

[10] It excludes, for example, rent payments or costs of renovations on which store managers have no impact.

We randomly assigned each district within the region to one of the following treatments: *BONUS*, *REVIEW*, *BONUS&REVIEW*, or the *CONTROL* group.[11] Store and district managers did not know that this was part of an experiment and we thus maintained a natural environment. Table 1 summarized the treatments.[12]

Table 1: Treatments

	Review	No Review
Bonus	N=63	N=51
No Bonus	N=50	N=60

Importantly, prior to the intervention all store managers (including the control group) received an information package about the profit metric as we wanted to avoid any attention or demand effects (i.e. that treatment effects are merely driven by creating attention for the new performance metric). All store managers participated in an online training session about possible ways to increase stores profits. From the end of March onwards, they had access to the online training consisting of a video and a quiz.[13] Additionally, they received novel information about the relative profit margin (e.g. *(sales price - procurement price)/sales price*) of each product. For this, all products are ranked according to their margin and then divided into five equal sized groups and named SP1 (highest margin) to SP5 (lowest margin). All store managers further received a monthly report about the development of profits (and its components) apart from the possible monthly bonus notifications. Hence, all treatment effects are effects over and above the effects of the information provision. In Manthei et al. (2019) we show in an experiment conducted in a different region of the same firm that the information provision itself already raises profits by about 3%.[14]

[11] The experiment was preregistered under AEARCTR-0002128. Note that we initially registered two regions. However, in one region (North-West Germany) the regional manager told us already in the early weeks of the interventions that higher-level management did not back the project due to many refurbishments in the region and difficult external market influences during that time. Due to this, district managers did not regularly hold the conversations with the store managers (average number of 3.5 conversations per store in the focus region, 2.4 conversations in the other region, MWU p<0.001). Moreover, they did significantly fewer first conversations in the first two weeks (90% conversations in the focus region, 52.48% conversations in the other region, Signed Rank Test p<0.001) Moreover, bonus payments were delayed by the regional manager and corrected ex-post.

[12] Differences in the sample size per treatment occur due to our randomization on the district level.

[13] Both, the video and the quiz were designed by us but provided with a company label. One of the authors was the presenter in the training video to further ensure full control about its content.

[14] Moreover, we find that a combination of bonus and information provision does not raise profits significantly above the effect of the information provision alone.

3.1. Treatment BONUS

In each of the three months from April to June 2017, store managers in this group were eligible to receive a bonus according to the following formula:

$$\text{Bonus (in €)} = [\text{Profit} - (0.8 \cdot \text{Planned Profit})] \cdot €0.05$$

Store managers received €0.05 for every €1 profit above a threshold of 80% of the planned value.[15] Accumulated bonuses were paid out after the three months of the experiment with the store managers' salary. Store managers were informed with personalized letters each month from April to June 2017.[16] The letter reported the achieved profit of the store and all of its components of the previous month. Moreover, the initially planned values (as determined by the accounting department in the beginning of the year based on previous performance as well as other expected influence factors such as renovations, opening or closing of stores by competitors etc.) were also provided. Additionally, store managers received feedback about the bonus for the respective month.[17]

3.2. Treatment REVIEW

In this treatment, store managers had systematic biweekly conversations with their supervisors (district managers) about actions taken during the past two weeks to influence the profits, occurring problems and their strategies for the upcoming weeks. A "Conversation Guide" with specific questions to be discussed during the conversation was provided to district managers:

(a) "What did the store manager do to increase profits?"

(b) "What problems occurred?"

(c) "What would the store manager like to do before the next meeting?"

[15] This is substantial compared to, for instance, a usual CEO compensation of $3.25 for $1000 change in shareholder wealth (Jensen and Murphy 1990).

[16] More precisely, due to a delay in calculating staff costs, the data was always delayed by one month. Hence, for instance, by the end of May letters were sent out with the calculations for April.

[17] Note that all store managers in this region also received an annual bonus for sales, inventory and a mystery shopping score which accumulated to on average €233 per store manager in 2017. The design of this existing bonus scheme was such that payments were fixed within brackets so that bonus payments hardly varied over time.

District managers were asked by the Human Resource (HR)-department to write a protocol documenting the responses to these questions and send it back to the HR office. The HR office then sent them further to us. District managers received emails every two weeks to remind them to have the conversations. They were also not aware that they were part of an experimental study.[18]

3.3. Treatment BONUS&REVIEW

This treatment is a combination of individual monetary performance pay and the biweekly conversations with the district managers.

3.4. Implementation

We use a stratified randomization depending on a prediction of the district profits for the first treatment month (see, e.g., Athey and Imbens 2017). To construct the stratification groups, we predict profits for the district in April 2017 using one year of past data through January 2017 with a simple time-series model.[19] We then randomly assigned the treatments within groups of four with similar predicted values. This aims to reduce the standard error in our main variable of interest. Randomization was conducted at the district level in order to avoid possible spillover effects between different stores.[20] We provide a balancing table in Appendix A2.[21]

Personalized letters were sent to the store managers' home addresses in the last week of March to inform them about the changes. The letters were signed by the regional manager and regional HR manager, and sent from the company's post office.[22] We also ran two online surveys with store and district managers before and after the experiments. Again, personalized letters were sent to the store managers' home addresses in February 2017 as well as in the last week of June 2017.

[18] We use the following phrase in the introductory letter "We would like you to have an intensive, personal conversation with your store manager each and every second week". Store managers were informed at the beginning of the treatment phase that their supervisors will meet them every second week for the review conversations.

[19] We had to randomize three months in advance as the data on profits, as explained above, were accessible with a delay of one month.

[20] Store managers within a district may interact from time to time. Interactions are substantially less likely between store managers belonging to different districts.

[21] The small imbalances that are visible in the balancing table are idiosyncratic. Moreover, they are time invariant and should be controlled for in fixed effects regression. However, we further control for the imbalanced variables in Table A3 in the Appendix to show robustness of our results.

[22] Exemplary letters are provided in the Appendix 9.3.

4. Key Hypotheses

To illustrate our pre-registered key hypotheses, consider the following simple extension of a standard linear principal agent model. An agent can exert an effort e to raise store profits π at personal costs $c(e)$ where $c''(e) > 0$ and $c'(\overline{e}) = 0$ for some $\overline{e} > 0$. Profits are given by

$$\pi = e + \varepsilon$$

where ε is a noise term with mean m and variance σ^2. The agent receives a wage $\alpha + \beta \cdot \pi$. For simplicity assume that the agent is risk neutral and maximizes her utility

$$\alpha + \beta \cdot \pi - c(e).$$

Suppose now that the principal can also introduce a monitoring activity (performance review) $r \in \{0,1\}$ carried out by the agent's respective supervisor. The agent anticipates that she will incur psychological or economic costs from "underperforming" some effort level $\hat{e} > \overline{e}$ when performance reviews are in place (where ê may, for instance, be the first-best effort level).[23] If reviews are in place (i.e. $r = 1$) her utility is reduced by $g(\hat{e} - e)$ where $g(\varDelta) = 0$ for $\varDelta \leq 0$ and $g' > 0$, $g'' \geq 0$ for $\varDelta > 0$. The agent then maximizes

$$\max_e \beta(e + m) - c(e) - r \cdot g(\hat{e} - e).$$

This leads to the following result:

Proposition 1: *Performance pay and performance reviews raise performance. Both instruments are substitutes: The introduction of performance reviews has a weaker additional effect on performance if performance pay is in place.*

Proof: See Appendix.

Hence, our key hypotheses for the field experiments are that (i) performance reviews raise performance, (ii) bonus payments raise performance and (iii) both instruments are (partial) substitutes.

[23] This is a simple way to formally express Lazear and Oyer's (2013, p. 486) claim cited in the introduction that "*An alternative to financial incentives is to simply monitor workers. If a supervisor can keep close watch over employees, she can ensure that the employee takes the best actions*". In section 6.1 we develop a model that more explicitly captures how performance review affect reputational incentive concerns.

5. Results

First, it is instructive to consider the number of conversations conducted and the actual bonuses paid out. On average, store managers had 3.5 review conversations with their district managers within the three months period in the respective treatment groups (median=4, SD=1.63) and 91.52% of the first conversations took place within the first two weeks after the start of the project. The average total bonus payment was €535.19 (median=421.65, SD=506.73) with only 17 of the 117 store managers receiving no bonus at all.

Table 2 shows the estimated average treatment effects from fixed effects regressions of store profits on the treatment dummies. Column 1 shows results of the (store) fixed models controlling for refurbishments of the stores as well as plan values of profits as predicted by the accounting department in the beginning of the year. [24] Column 2 additionally controls for district and store manager fixed effects. Column 3 and 4 use the same specifications with the log of profits as the dependent variable.

[24] Some of the stores were refurbished before the intervention. Refurbishment controls include a dummy indicating whether a refurbishment took place in the given month and a dummy indicating whether the store has been refurbished.

Table 2: Main Treatment Effects on Profits

	(1)	(2)	(3)	(4)
	FE	FE	Log FE	Log FE
Treatment Effect	-51.85	156.2	-0.00441	0.0141
BONUS	(607.3)	(710.5)	(0.0417)	(0.0569)
Treatment Effect	1370.2**	1492.3**	0.0732***	0.0858**
REVIEW	(559.0)	(666.2)	(0.0238)	(0.0411)
Treatment Effect	-376.3	-397.7	-0.00485	-0.00390
BONUS&REVIEW	(605.1)	(564.3)	(0.0351)	(0.0501)
Wald test	p=0.0162	p=0.0090	p=0.0218	p=0.0330
REVIEW=BONUS&REVIEW				
Time FE	Yes	Yes	Yes	Yes
Store FE	Yes	Yes	Yes	Yes
District Manager FE	No	Yes	No	Yes
Store Manager FE	No	Yes	No	Yes
Refurbishments	Yes	Yes	Yes	Yes
Planned Profits	Yes	Yes	Yes	Yes
N of Observations	3975	3777	3966	3768
N of Stores	224	224	224	224
Cluster	31	31	31	31
Within R^2	0.2370	0.2722	0.1621	0.1875
Overall R^2	0.7577	0.5955	0.6158	0.4316

Note: The table reports results from a fixed effects regression with the profits on the store level as the dependent variable. The regression accounts for time and store fixed effects and adds fixed effects for district manager and store managers in column 2&4. The regressions compare pre-treatment observations (January 2016 - March 2017) with the observations during the experiment (April 2017 – June 2017). *Treatment Effect* thus refers to the difference-in-difference estimator. All regressions control for possible refurbishments of a store and the companies planned value of profits. Observations are excluded when a store manager switched the store during the treatment period. Robust standard errors are clustered on the district level of the treatment start and displayed in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

As column 1 shows, performance reviews (*REVIEW*) significantly increases monthly profits by, on average, €1370. The result remains robust when including store manager and district manager fixed effects. Column 2 displays an estimated treatment effect of €1,492. According to the log specification in columns (3) and (4) performance reviews increase profits by about 7%. However, the *BONUS* and *BONUS&REVIEW* treatments have no significant effect on the outcome variable relative to the control group in all specifications and thus do not raise performance above any effects of the information provision (that has been introduced for all).[25] Table A3 in the Appendix provides robustness checks with simple OLS regressions using only the treatment time.[26]

[25] To exclude that the results are driven by outliers in specific districts we also ran this regression repeatedly excluding each single district in one regression and the result remains stable.

[26] Note that the results remain qualitatively robust when estimating wild bootstrap standard errors.

We had hypothesized that performance pay and performance reviews both have a positive effect on performance and that both instruments are substitutes. Thus, we expected that performance pay may reduce the *marginal* effect of performance reviews. But in fact the results show it reduces the *absolute* effect of performance reviews which we did not expect. In all specifications *BONUS&REVIEW* is significantly smaller than *REVIEW* (Wald test, $p<0.05$).

Table A4 in the appendix further shows treatment effects by the respective month of the treatment period including two months after the end of the treatment.[27] The most interesting insight here is that treatment effects in the *REVIEW* treatment vanish after the end of the treatment indicating that the reviews do not have the character of coaching. Thus, the reviews do not create persistent human capital, which is in line with the idea that the reviews trigger temporary incentives to perform. Figure A1 in the Appendix shows the treatment effects depending on whether the number of review conversations conducted per store manager is below or above the median number of conversations (4). While we caution that the number of conversations is not exogenously assigned and thus the figure has no clean causal interpretation, it indicates that the treatment effect is slightly higher when more conversations are conducted but this difference is not significant.

Of course, the performance reviews also come at a cost which is essentially the time invested by district and store managers in the conversations. Approximating the duration of the conversation with a maximum of 30 minutes for each and using hourly wages of store and district managers, the opportunity cost of a meeting is less than €40. Hence, the total opportunity cost for the pure *REVIEW* intervention are substantially smaller than the estimated treatment effects.[28] The bonus, however, did not provide a return above its costs neither alone nor in combination with the performance reviews.

[27] In the three months starting with September 2017, the control group and *REVIEW* stores received a bonus for three months (in order to alleviate fairness concerns as all store managers in the end of the year had a bonus for three months). Hence, only July and August are informative to study post intervention effects.

[28] Recall that district managers visit the stores anyway during the week so that typically no additional travel costs occurred.

6. Bonuses and the Quality of Performance Reviews: Theory and Further Evidence

While we initially hypothesized that performance reviews should increase performance, we did not expect that the introduction of bonus payments would reduce the impact of performance reviews.[29] It is of course important to understand the reason for this effect. Apparently, the fact that store managers received a bonus undermined the quality of the review conversations. We now provide a theoretical framework to organize detrimental effects of bonuses and their interplay with performance reviews and then analyze further data we obtained from surveys, and review protocols in more detail.

6.1. Theoretical Framework

In our initial model, we treated the effect of performance reviews as a “black box”, assuming that reviews generate (psychological or economic) costs for underperforming agents. Moreover, in this prior model we assumed that these costs are additively separable from the utility from income generated through performance pay – which is clearly refuted by the data. In a next step we thus develop a theoretical framework that aims at opening this black box, explicitly modeling the role of performance reviews as a device to create more transparency about the agent’s efforts which allows to study the interaction between “reputational” and “material” incentives in a more detailed manner.

In order to study this, we use and extend a standard career or image concerns model. In essence our model is an extension of Bénabou and Tirole (2006), who have explored the interplay between reputational concerns and the provision of incentives. Assume that agents differ in their ability/motivation to exert effort which is determined by η_e and by their preferences for money η_m. Following Bénabou and Tirole (2006), an agent receives image utility which is a function of the supervisor’s posterior expectation about their intrinsic willingness η_e to raise profits. Or, as an alternative interpretation, suppose that the agent’s career success will depend on the supervisor’s beliefs about the agent’s ability η_e as in a Holmström (1999)-type career concerns model. The agent’s utility function is

[29] The absence of a bonus effect is in line with the result obtained in another region of the company. In the companion paper Manthei et al. (2019), we evaluated the effect of an information provision intervention conducting a 2x2 design in a different region of the firm (treatments: information, bonus, and bonus&information). We find that introducing the same bonus as in this study and the information provision as explained in section 3 alone both raises performance. However, bonus and information together do not significantly outperform the information without bonus.

$$\eta_m \beta e - \frac{c}{2}(e - \eta_e)^2 + \zeta E[\eta_e | I_S]$$

where I_S is the supervisor's posterior information she can use to form inferences about η_e. Hence, η_e is the agent's "bliss point" for effort, i.e. the effort level she would choose without extrinsic incentives. Higher levels of η_e thus can reflect both, higher levels of ability as well as a higher intrinsic motivation to perform.[30] Assume that the parameters η_m and η_e are independently normally distributed with

$$\begin{pmatrix} \eta_e \\ \eta_m \end{pmatrix} \sim N\left(\begin{pmatrix} m_{\eta_e} \\ m_{\eta_m} \end{pmatrix}, \begin{pmatrix} \sigma_{\eta_e}^2 & 0 \\ 0 & \sigma_{\eta_m}^2 \end{pmatrix}\right).$$

We model the introduction of performance reviews as a change in the observability of effort. Without performance reviews the supervisor only observes profits π and thus the agent maximizes

$$E\left[\eta_m \beta e - \frac{c}{2}(e - \eta_e)^2 + \zeta E[\eta_e | \pi]\right]$$

and when reviews are introduced the supervisor observes the effort level e such that the agent's objective function becomes

$$E\left[\eta_m \beta e - \frac{c}{2}(e - \eta_e)^2 + \zeta E[\eta_e | e]\right].$$

Hence, the key difference is that performance reviews generate a more precise signal of the agent's efforts.[31] Characterizing Perfect Bayesian Equilibria of this setting we obtain the following result:

Proposition 2: *There exists a Perfect Bayesian Equilibrium in which expected gross profits are*

$$\Pi^N(\beta, r) = m_{\eta_e} + \frac{m_{\eta_m}\beta}{c} + \zeta \frac{\sigma_{\eta_e}^2}{c\sigma_{\eta_e}^2 + \sigma_{\eta_m}^2 \frac{\beta^2}{c} + (1 - r) \cdot c\sigma_\varepsilon^2}.$$

(i) Performance reviews raise performance (i.e. $\Pi^N(\beta, 1) > \Pi^N(\beta, 0)\ \forall \beta \geq 0$).

(ii) The benefit of introducing performance reviews $\Pi^N(\beta, 1) - \Pi^N(\beta, 0)$ is decreasing in β.

(iii) The introduction of a bonus β>0 will reduce performance (i.e. $\Pi^N(\beta, r) < \Pi^N(0, r)$) if and only if reputational concerns ζ are sufficiently strong.

[30] Note that this is equivalent to maximizing $\eta_m \beta e + \eta_e e - \frac{c}{2}e^2 + \zeta E[\eta_e | I_S]$ which is the utility function used in the linear normal setting in Bénabou and Tirole (2006).

[31] The qualitative results do not hinge on the assumption that reviews make efforts perfectly observable, but that reviews lead to less noisy signals of effort.

(iv) Such a detrimental effect of bonus payments will always be stronger when performance reviews are in place.

Proof: See Appendix.

Performance reviews raise performance as reviews make it easier for agents to signal their underlying motivation or ability. Without performance reviews supervisors infer this motivation from (noisy) profits which yield a less precise signal of effort. When performance reviews are in place supervisors observe a direct signal of effort. In turn, efforts have a stronger effect on reputation and marginal returns to increasing effort are higher. In other words, performance reviews strengthen the reputational incentive mechanism.[32]

The result in claim (iii) that bonuses can reduce performance is the Bénabou and Tirole (2006) result: When bonus payments are used, it becomes harder to signal motivation as a supervisor cannot perfectly disentangle whether an agent chose a higher effort level because she is more able and motivated (i.e. has a higher η_e) or because of having a stronger preference for money (i.e. a higher η_m). Hence, performance pay can reduce profits if reputational concerns are sufficiently strong.[33]

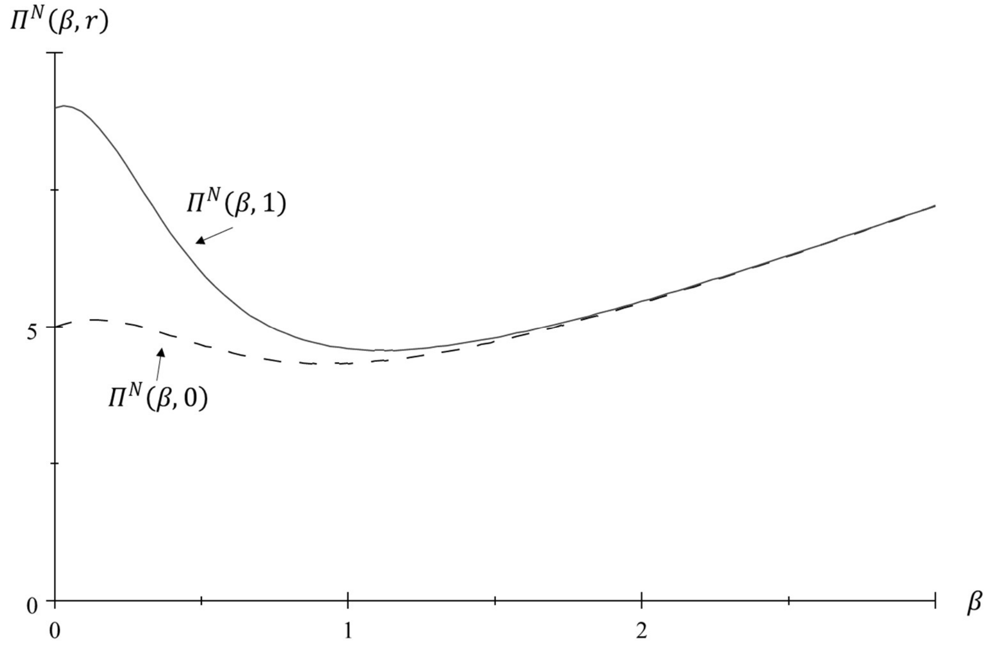
The novel result here is that this framework naturally explains why performance pay can reduce the benefits of performance reviews when both are introduced together. The reason is due to the interaction of the results explained above: On the one hand, performance reviews create more precise information on performance which strengthens the reputational incentive mechanism. But on the other hand, performance pay undermines this reputational mechanism as observed efforts are a less useful signal of the agent's true underlying motivation when bonus payments are used. In turn, the reputational mechanism that is triggered by performance reviews naturally becomes less effective when performance pay is used.

Figure 1 plots the profit function $\Pi^N(\beta, r)$ for a specific parameter constellation. The solid line depicts expected profits when performance reviews are in place, the dashed line plots profits without performance reviews.

[32] Note that this is analogous to a standard comparative statics results in Holmström (1999) type career concerns models: increasing the observability of effort by reducing noise increases career concerns incentives.

[33] Ariely, Bracha, and Meier (2009) provide experimental evidence for the importance of this mechanism in real effort experiments.

Figure 1: Profits as a function of bonus and reviews



As the figure illustrates, performance reviews – by reducing noise – strengthen the reputational incentive mechanism and therefore the solid line always exceeds the dashed line. As in Bénabou and Tirole (2006), introducing a bonus can decrease performance – and this effect is stronger when performance reviews are in place. Note that the model of course does not necessarily imply that the bonus has initially a negative effect on performance (when reputational concerns are weak both profit functions will be strictly increasing in β). But when reputational concerns are sufficiently strong (i.e. ζ is sufficiently large), there will always be an interval in which the function is downward sloping in β. Moreover, as Proposition 2 shows, this performance loss will always be larger when reviews are in place. The model thus yields an explanation for both key observations in the field experiment where (i) the bonus itself had no significant effect on performance and (ii) the bonus significantly reduced the benefits of the performance reviews. And this explanation rests on standard assumptions in organizational economics that agents care for their reputation and that supervisors make inferences about an agent's type from their accessible information.

It is instructive to also consider a slight reinterpretation of the model. Suppose that we just consider the *interaction within the meeting* and that e now captures the agent's intensity of communication to exchange ideas with the supervisor. Suppose that an agent with a higher η_e

will be more able/willing to use the review meetings more intensively in this respect. Think of the model now as solely capturing the value of interaction within the meetings which is given by

$$m_{\eta_e} + \frac{m_{\eta_m}\beta}{c} + \zeta \frac{\sigma_{\eta_e}^2}{c\sigma_{\eta_e}^2 + \sigma_{\eta_m}^2 \frac{\beta^2}{c}}. \qquad (1)$$

There are two countervailing effects analogous to the effects described above: On the one hand, the bonus should lead the agent to use the meetings more intensively as he benefits more from an increase in profits achieved through the interaction. But on the other hand, there is again the reputational incentive effect which will be undermined by performance pay: If the agent uses the meetings more intensively (i.e. proposes more suggestions or tries to acquire more support from the supervisor etc.), a supervisor will be less sure whether this is due to a higher motivation/ability or a stronger preference for money when the bonus is in place. Bonus payments thus can naturally reduce the agent's reputational incentives to exert effort in the meetings.

This effect could even be exacerbated if supervisors themselves can exert effort to increase the quality of the conversations and invest more in the conversations, the more they believe that the agent's motivation is genuine. When reputational concerns are sufficiently strong, performance pay may thus undermine the quality of conversation within the review meetings. This mechanism thus may also reflect some claims made in a recent practitioner debate on traditional performance appraisals that performance pay changes the nature of feedback interactions between supervisors and subordinates as they become more "politicized".[34]

6.2. Further Evidence

Following an "insider econometrics" approach (Ichniowski and Shaw 2003, Bartel et al. 2004), we explore further details on the specific job and collect evidence from surveys and protocols to build a better understanding of what managers actually did in general to raise profits and how the review meetings were used.

[34] See e.g. DiDonato (2014).

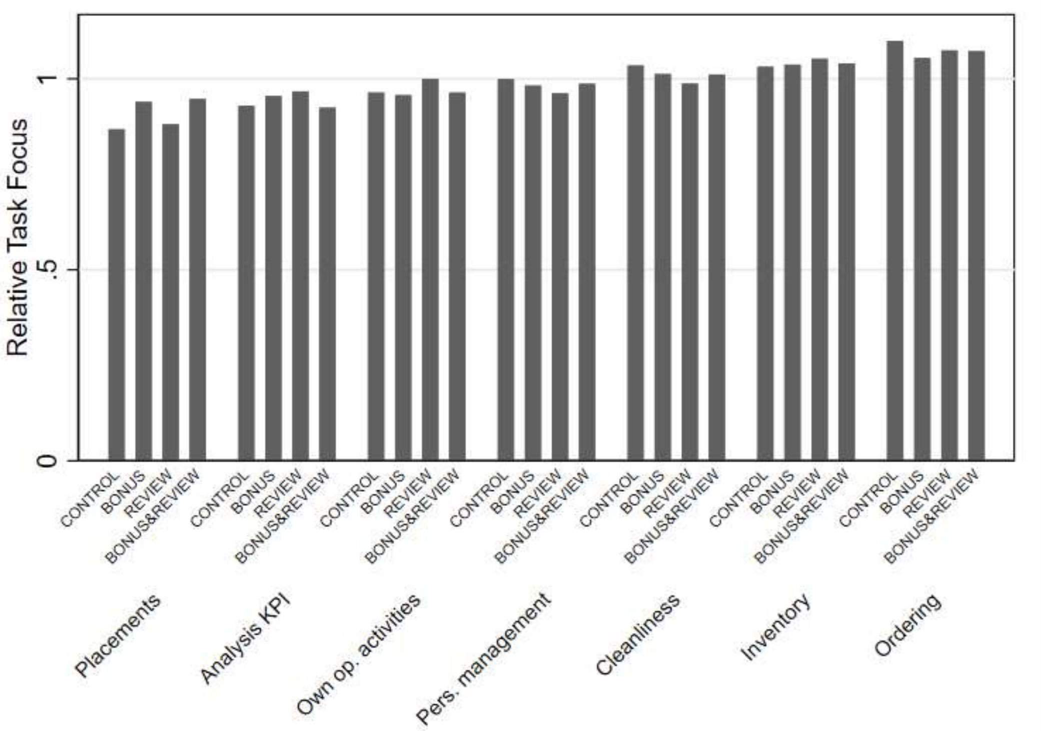
6.2.1. What did the Store Managers do?

We have different sources of information to assess what store managers actually did in their daily jobs and what they did to increase profits. First, we analyzed structured job description documents of store managers provided by the firm and had additional meetings with store, district, and sales territory managers. We distilled a list of 29 different tasks that should capture nearly all key activities of a store manager. We clustered these 29 tasks into 7 task dimensions: *personnel management, ordering, cleanliness, inventory management, placements*, *analysis of key performance indicators*, and *own operational activities* (cash desk, own customer interaction). This classification is visible in Table A1 in the Appendix.

To assess the relative importance of these tasks, we included 29 items in our post-experimental survey listing the different tasks and asked store managers to state to what extent they had focused in the previous months on the respective task. Figure 2 displays average ratings for the respective task dimensions normalized by dividing the focus rating for a task by the average focus rating across all tasks.

As Figure 2 shows, store managers in all groups put a particular focus on ordering (e.g. ordering of meat, vegetables, fruit and bakery products) and inventory management (e.g. analysis of shrinkage, checking of incoming goods). In general, they put a relatively low focus on the placements of goods (which, for instance, includes secondary placements of profitable products, decisions on placements on aisle ends) and the analysis of Key Performance Indicators. Here we hardly see sizeable treatment differences and the overall allocation of attention seems to be very similar across treatments.

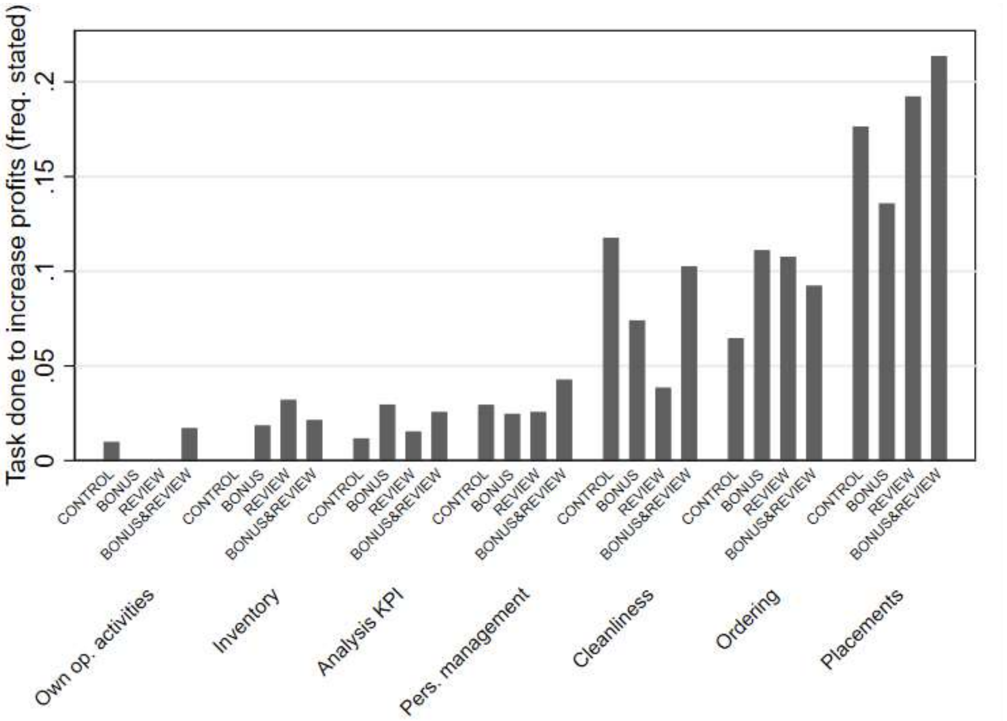
Figure 2: Relative Task Focus (Post-experimental survey)



Note: The figure displays the average rating of focus on specific tasks (1=low focus,6=high focus) obtained from an online questionnaire. Tasks were clustered into 7 dimensions. The average focus of a dimension was then divided by the average focus of all dimensions.

While the previous items asked about the general focus in the store manager's work our survey also included an open-ended question in the survey asking store managers explicitly what they did to increase profits. Recall that store managers in the control group also had received the online training on the profit metric allowing for this open question to be asked in all four groups. Research assistants assigned the responses to the 29 task categories. Figure 3 shows the respective frequencies for the different task dimensions.

Figure 3: Self-Reported Tasks Done to Increase Profits



Note: The figure displays the share of stated tasks dimensions to increase profits obtained from open questions of an ex-post questionnaire.

Most notably, by their own assessment, store managers tried to increase profits through improved placements (many for instance stated that they explicitly tried to place articles with high margins in prominent positions), ordering, and improved cleanliness. But again, we do not see sizeable systematic treatment differences that stand out and between-task differences in the frequency of the tasks mentioned are much more substantial than within-task treatment differences. This indicates that treatment differences may not be driven so much by what store managers did but rather how they did it.[35]

6.2.2. Frequency and Content of Review Conversations

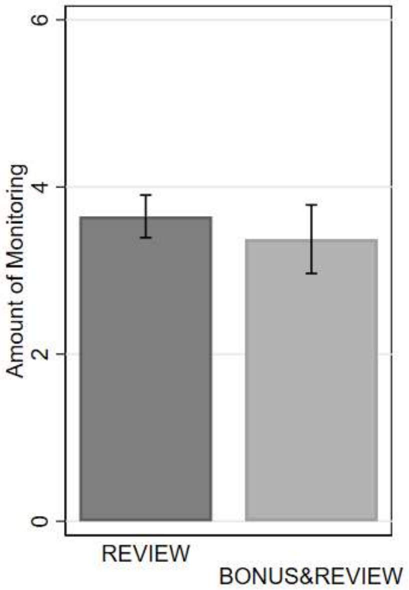
In order to explore this further, we now investigate the content of the meetings. At our request the company had asked district managers conducting the review meetings to fill out a short form documenting the contents of the conversation in a concise manner after each

[35] In fact, the only task dimension for which we see a statistically significant difference in the responses between store managers in *REVIEW* as compared to *BONUS&REVIEW* is own operational effort, which is significantly more often mentioned in *BONUS&REVIEW* (as it is never mentioned in *REVIEW*, *p*=0.042 in an OLS regression with clustered standard errors).

meeting. We can use the content of these protocols to assess potential differences in the way in which the reviews were conducted.

As a first step, we counted the number of meetings and measured the lengths of the protocols (number of notes/sentences stated in each protocol). First of all, the number of meetings is not statistically significantly different between the treatments with or without bonus payments. District managers in the *REVIEW* treatment conducted only slightly more conversations (3.66) than in *BONUS&REVIEW* (3.38, see also Figure 4) and this difference is not significant (MWU, *p*=0.4282; OLS regression with clustered standard errors, *p*=0.705). Hence, bonus payments to store managers apparently did not affect the district managers' general willingness to conduct the meetings. It thus seems unlikely that performance reviews didn't work as well when bonuses were used because district managers were less willing to spend time on these meetings.[36]

Figure 4: Average Conversations Conducted

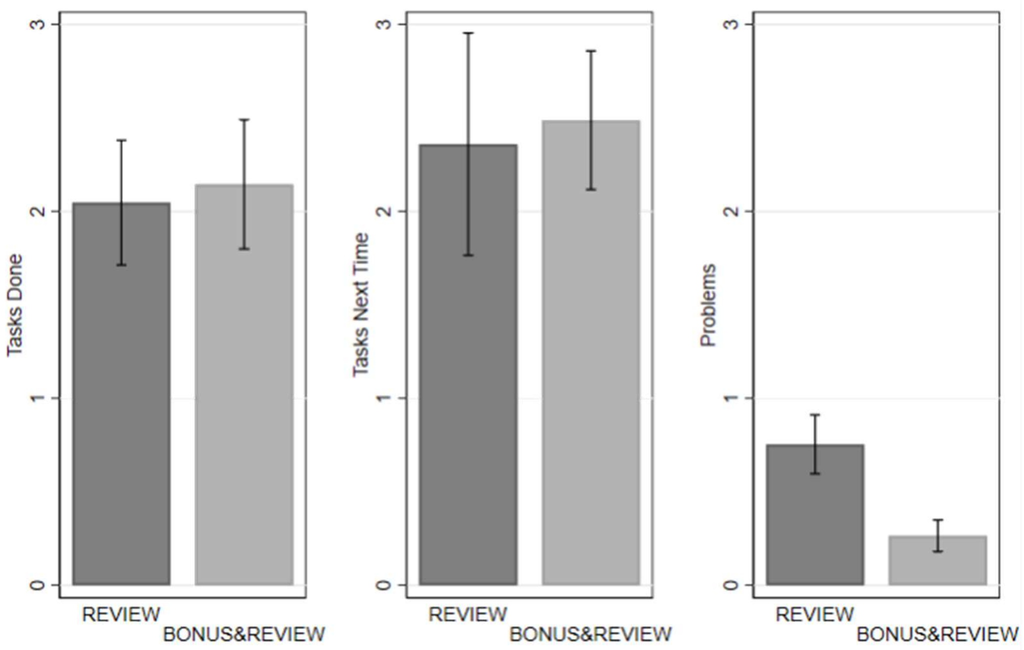


Note: The figure displays the average number of conversations per store manager. 95% confidence bars are displayed.

[36] This also makes it unlikely that district managers invested less into the conversation as they thought that performance pay would raise store manager performance anyway (potential reasons for such behavior could be that district managers felt that their input would be needed to a lesser extent or because of envy towards store managers who now received a bonus). In fact, district managers would have harmed themselves by lowering their effort as higher profits also lead to higher bonuses for themselves (district managers received annual bonuses based on regional performance). Moreover, salaries of district managers are more than 80% higher than salaries of store managers which seems to make a feeling of envy relative to store managers rather unlikely.

To explore whether the quality of the review conversations changed, as we conjectured in the above, we now explore the contents of the review meetings in more detail. Recall that district managers were asked to go through three sections in the meetings. For each of these sections they were asked to protocol what the store managers reported. Figure 5 displays information about the intensity of the protocolled parts of the conversations. It shows the average number of notes in the sections (i) "What did the store manager do to increase profits?" (ii) "What problems occurred?", and (iii) "What would the store manager like to do before the next meeting?".

Figure 5: Average Number of Notes per Conversation



Note: The figure displays the per session average number of notes/sentences per store manager in the respective category. 95% confidence bars are displayed.

There are no significant difference in sections (i) and (iii) regarding reports on the tasks done and tasks planned. However, the use of performance pay substantially reduced the amount of problems stated by store managers during the conversations. In *BONUS&REVIEW* store managers only state on average 0.27 problems per conversations, in *REVIEW* store managers state nearly three times as many (0.75) problems per conversation (MWU, *p*-value<0.001; OLS regression with clustered standard errors, p=0.003).[37] In fact, for 60.3% of the stores not a single problem was mentioned in any of the meetings in *BONUS&REVIEW* while this fraction is only 22.2% in the *REVIEW* treatment.

An interpretation of this finding in the light of the above reasoning is that the incentives to state a problem are changed as such an action is perceived in a different manner when performance pay is in place. For instance, without bonus payments stating a problem a store manager encountered when trying to raise profits should be a rather genuine signal of her motivation (or ability) to identify a problem. If, however, performance pay is in place, it is less clear whether the problem was raised simply to receive a higher bonus. In a set of laboratory experiments, Vohs et al. (2006) find that priming subjects with the concept of money leads to fewer requests for help to solve a problem and argue that money creates a state of self-sufficiency. The reasoning based on the Bénabou and Tirole (2006) model yields an economic rationale for such behavior: when money is involved the signaling value of specific actions is altered, which in turn will affect the incentives to provide these actions.

A key question is of course whether this finding indeed captures a core mechanism that explains our main result. To explore this, we now consider only the subset of stores from the *REVIEW* and *BONUS&REVIEW* treatments in which at least one problem has been mentioned in a review conversation. In other words, we include only review meetings that entailed sufficiently open conversations. We then replicate the results from Table 2 on the reduced data set and report the results in Table 3. The coefficients of *BONUS&REVIEW* are now substantially larger than in the full sample and the treatment effect of 5,2% (as for instance estimated in the log specification in column (3)) now comes much closer to those of the *REVIEW* treatment. This indicates that the fact that problems are mentioned openly is indeed an indicator for the quality of review conversations and that the bonus payments undermined this quality. In none of the specifications the coefficient of *BONUS&REVIEW* is now significantly smaller than that of *REVIEW*.

[37] As Figure A2 in the appendix shows, the effect is not driven by single district managers, but the pattern is very similar across district managers of the respective groups (MWU with one observation per district manager averaged across all stores, p=0.0421). Figure A3 in the appendix shows the timing of stated problems within conversations. Table A5 in the appendix reports regression estimates for this treatment differences also controlling for tenure and performance evaluations.

Table 3: Main Treatment Effects on Profits (only reviews where problems mentioned)

	(1) FE	(2) FE	(3) Log FE	(4) Log FE
Treatment Effect BONUS	-0.0126 (0.620)	0.207 (0.713)	-0.00484 (0.0421)	0.0147 (0.0572)
Treatment Effect REVIEW	1.261** (0.531)	1.402** (0.596)	0.0719*** (0.0213)	0.0827** (0.0373)
Treatment Effect BONUS&REVIEW	0.643 (0.706)	1.126 (0.745)	0.0515* (0.0268)	0.0687 (0.0415)
Wald test REVIEW=BONUS&REVIEW	*p*=0.4070	*p*=0.7168	*p*=0.3436	*p*=0.5929
Time FE	Yes	Yes	Yes	Yes
Store FE	Yes	Yes	Yes	Yes
District Manager FE	No	Yes	No	Yes
Store Manager FE	No	Yes	No	Yes
Refurbishments	Yes	Yes	Yes	Yes
Planned Profits	Yes	Yes	Yes	Yes
N of Observations	3046	2917	3040	2911
N of Stores	172	172	172	172
Cluster	30	30	30	30
Within R^2	0.301	0.354	0.158	0.183
Overall R^2	0.822	0.748	0.606	0.522

Note: The table reports results from a fixed effects regression with the profits on the store level as the dependent variable. The regression accounts for time and store fixed effects and adds fixed effects for district manager and store managers in column 2&4. The regressions compare pre-treatment observations (January 2016 - March 2017) with the observations during the experiment (April 2017 – June 2017). *Treatment Effect* thus refers to the difference-in-difference estimator. All regressions control for possible refurbishments of a store and the companies planned value of profits. Observations are excluded when a store manager switched the store during the treatment period. Observations are further excluded if no problem was mentioned in any performance review. Robust standard errors are clustered on the district level of the treatment start and displayed in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Finally, we can also explore the topics covered in the review conversations. In order to do so, protocol items were classified by research assistants again into the 29 different tasks for each of the up to 6 meetings held in each store. Figure A4 in the Appendix displays the respective frequencies of mentioning a topic in the seven tasks dimensions (summed up across all three sections of the protocol). The ranking of the task dimensions is well in line with the ranking of the relative importance from the open-ended question to store managers (compare Figure A4 in the Appendix): placements, ordering and personnel management were the key focus areas. Moreover, when bonuses are in place, review meetings tend to be more concerned with placements, personnel management and own operational activities and less concerned with ordering behavior and cleanliness (see Table A6 in the Appendix for regression results).

6.2.3. Post-Experimental Questionnaire: Satisfaction and Perceptions

We also conducted a post-experimental online survey asking store managers about their activities and perceptions. In the first part of the survey we asked store managers about their overall satisfaction with their job as well as specific job domains such as their compensation and their workload. As columns (1)-(3) in Table 4, which display results from regressions of the respective survey items on treatment dummies, show, the treatments did not affect employee satisfaction in a detectable manner.

The survey then includes an item about their own perceived aim to raise profits ("*I have tried to increase profits in the last few months*") as well as items eliciting store managers' perceptions on the interaction with their respective district manager ("*My district manager gave me regular feedback*", "*My district manager motivated me regularly to do better*"). Columns (4)-(6) of Table 4 report the respective regression results.

Table 4: Survey Results Perceptions on Activities

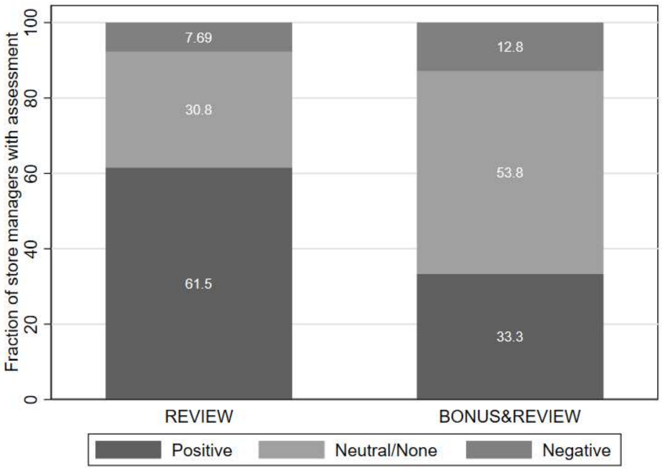
	(1)	(2)	(3)	(4)	(5)	(6)
	Satisfaction Job	Satisfaction Compens.	Satisfaction Workload	Profit Aim	Feedback	Motivate
Treatment Effect	-0.313	0.341	-0.257	-0.0657	0.138	0.289
BONUS	(0.304)	(0.262)	(0.550)	(0.268)	(0.269)	(0.345)
Treatment Effect	0.114	-0.031	-0.314	0.128	0.931***	0.831*
REVIEW	(0.254)	0(0.358)	(0.554)	(0.249)	(0.304)	(0.445)
Treatment Effect	-0.133	0.0138	-0.551	0.538**	0.385	0.00922
BONUS&REVIEW	(0.311)	(0.228)	(0.445)	(0.236)	(0.248)	(0.343)
Wald test REVIEW=BONUS&REV.	*p*=0.3986	*p*=0.8949	*p*=0.5904	*p*=0.1304	*p*=0.0428	*p*=0.0372
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N of Observations	97	97	97	95	96	96
Cluster	28	28	28	28	28	28
Overall R^2	0.140	0.303	0.093	0.177	0.189	0.174

Note: The table reports results from OLS regressions with the respective survey response as the dependent variable (scale from 1-6). "Job" is general job satisfaction, "Compens." is satisfaction with the compensation and "Workload" is satisfaction with the workload. Further controls are store size, number of employees, store manager's age and prior performance evaluation, as well as randomization group. Standard errors are clustered on the district level of the treatment start and displayed in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

In contrast to the results on satisfaction, here we observe specific patterns in the store managers' perceptions. Store managers state the strongest aim to increase profits when they receive a bonus and have review meetings (column (4) of Table 4).[38] As columns (5) and (6) show, however, the store manager's perception that they receive regular feedback from and feel motivated by their district managers is highest in the *REVIEW* treatment. And these positive effects vanish when they also receive a bonus: The coefficient of the *REVIEW* treatment is significantly larger than that of the *BONUS&REVIEW* treatment (*p*=0.0428 and *p*=0.0372 for the feedback and motivation item respectively). This lends further support to the idea that bonus payments changed the nature of the feedback conversations and reduced the quality of interaction.

Moreover, we asked store managers an open question about their opinion on the project ("*How did you perceive the regular conversations with your district manager?*") and categorized the answers into positive, neutral/none and negative. The results are displayed in Figure 6. Of all store managers responding to the survey, 61.5% stated a positive opinion in the *REVIEW* treatment against only 33.3% in the *BONUS&REVIEW* treatment (MWU, *p*=0.0262; OLS Regression with clustered standard errors, *p*=0.025). The fraction of negative assessments is higher in the *BONUS&REVIEW* treatment, but this difference is not statistically significant (MWU, *p*= 0.5168; OLS Regression with clustered standard errors, *p*= 0.420).

Figure 6: Project Assessment by Store Managers



Note: The figure displays the fraction of store managers with a positive, neutral/none, or negative assessment about the project stated in an open question in the post-experimental survey. N=65.

[38] It is conceivable that this due to demand/social desirability effects: this group had the most intensive set of practices to increase profits (and the most costly investment by the firm): hence, store managers may have felt the obligation to state the strongest consent when they receive both a bonus and the review meetings.

6.2.4. Benefits of Reputation

An important part of this explanation is that store managers have career or image concerns, i.e. they benefit from having a better reputation in the eyes of their district managers. One channel through which this is the case is due to the district manager's role in affecting annual performance evaluations and future salaries. To assess the importance of this mechanism empirically, we study the predictive power of performance evaluations for the wages of store managers on which we have information from 2018. District managers evaluate the store managers' performance on a scale from 1 (low performer) to 4 (high performer). We can access this data for 5 regions in Germany. In Table 5 we report regressions of wages in 2018 on performance evaluations in 2016.

Table 5: Monthly Wages and Performance Evaluations

	(1) Log Wage	(2) Log Wage
Perf. Eval.=1 (low performer)	*Reference Group*	
Perf. Eval.=2	0.0259**	0.0280**
	(0.0121)	(0.0115)
Perf. Eval.=3	0.0511***	0.0515***
	(0.0113)	(0.0108)
Perf. Eval.=4 (high performer)	0.0808***	0.0778***
	(0.0151)	(0.0149)
Tenure		0.0014***
		(0.0003)
Store Space		0.0001
		(0.0001)
N of Observations	764	764
Overall R^2	0.0468	0.0755

Note: The table reports results from OLS regressions with log monthly wages of store managers in 2018 as the dependent variable. *Perf. Eval.* is a set of dummy variables and refers to the store managers' annually made subjective performance evaluation of supervisors (district managers) with 1=low performer and 4=high performer. *Perf. Eval.*=0 (the lowest group) is the reference group and thus omitted. Robust standard errors are displayed in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

In Column 1 of Table 5 we regress the log monthly wages only on a set of dummy variables referring to the store managers' evaluated performance. Column 2 controls for store managers' tenure and the store space. The estimates show a monotonically increasing relationship between the performance evaluation and monthly wage. For instance, a store manager who is evaluated as a high performer in 2016 has a 7.8% higher wage in 2018 than somebody with the lowest performance rating (p<0.001). While we caution that this is of course no causal statement, it indicates that district manager's judgements may have a substantial impact on future wages and the career of store managers, lending support to the assumption that reputational concerns matter for store manager's actions.

7. Conclusion

Implementing appropriate management practices to align the behavior of employees with the interests of the employer is one of the biggest challenges in the design of organizations. We provide evidence that simply implementing regular conversations between supervisors and subordinates on a specific performance objective can lead to greater performance improvements (profit increases) than bonus payments based on this key figure. But more importantly, the use of performance pay in our setting substantially reduced the benefits of the review conversations.

We provide a potential explanation for this finding studying an extension of the Bénabou and Tirole (2006) model. When performance reviews increase the scope to signal intrinsic motivation and ability, the existence of bonus payments will naturally undermine reputational incentives. In other words, the existence of bonus payments can divert attention towards the instrumental value of receiving a bonus, affecting the usefulness of the review conversations as an open and honest dialogue between supervisor and subordinate. Exploring data from surveys and protocols we find that the use of bonuses indeed changed the nature and quality of review conversations in our field experiment. More precisely, we find that store managers state substantially less problems during review conversations when they receive additional performance pay. Moreover, when considering only stores in which conversations included an open discussion of problems, the negative effects of performance pay tend to vanish.

Following Roth's (2002) call for economists to develop an "engineering" approach or Duflo's (2017) related postulation that economists should adopt the mindset of a plumber to help decision makers in their design choices in practice, our results also shed some light on a recent debate on the use of performance reviews and performance pay in firms. Traditionally, performance reviews have often been used to assess performance and allocate bonuses (see, e.g. Cappelli and Conyon 2018). But in recent years many firms have intentionally shifted the focus in performance reviews away from the allocation of rewards. Several larger companies have entirely stopped or strongly revised standard annual performance ratings and instead established regular feedback conversations (see, for instance, Buckingham and Goodall 2015, Cappelli and Tavis 2016). Frequently, this change has been triggered by a feeling that a continuous dialogue between supervisor and subordinate is a key driver for performance and may be more important than incentives set through evaluation and compensation. Moreover, it has even been claimed that bonuses may undermine open communication and in turn harm performance.[39] Our results indicate that such claims are not lacking substance.

On a broader level, our results show that different organizational practices may interact in non-trivial ways. As has been stressed in the literature on complementarity in organizations[40], the performance effect of introducing a specific management practice may be contingent on the use of other practices. Whether and how specific practices interact depends on the interplay of different economic motives and behavioral mechanisms. Brynjolfsson and Milgrom (2013) describe challenges in the empirical assessment of interdependencies between organizational practices, stating that the opportunities to run designed experiments in firms are "underexploited" in this respect. RCTs that simultaneously vary the use of two practices are still rare, but can advance our understanding of the role of such interdependencies for firm performance and at the same time allow to study the relevance of different behavioral mechanisms in field settings.

[39] Tom DiDonato, Chief Human Resource Officer of Lear Corporation for instance claimes that "*Performance reviews that are tied to compensation [...], discourage straight talk, and too easily become politicized.*" (DiDonato 2014). Uwe Schirmer, Head of HR Policies at Bosch, world's largest auto parts supplier, for instance, claims that "*feedback discussions have become less tactical*" since Bosch has abolished individual performance bonuses in 2015 (Handelsblatt, Nov, 11 2018).

[40] See, e.g. Milgrom and Roberts (1990, 1995), Ichniowski et a. (1997), or Brynjolfsson and Milgrom (2013) for a recent survey.

8. References

Ariely, Dan, Anat Bracha, and Stephan Meier. "Doing good or doing well? Image motivation and monetary incentives in behaving prosocially." *American Economic Review* 99.1 (2009): 544-55.

Ashraf, Nava, Oriana Bandiera, and Kelsey Jack. "No margin, no mission? A field experiment on incentives for public service delivery." *Journal of Public Economics* 120 (2014): 1-17.

Athey, Susan and Guido Imbens. "The Econometrics of Randomized Experiments.", in A. Banerjee and E. Duflo, *Handbook of Field Experiments* 1 (2017): 73-140.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. "Field Experiments with Firms." *Journal of Economic Perspectives*, 25.3 (2011), 63-82.

Banker, Rajiv D., Shunlan Fang, and Seok-Young Lee. "Conflict between Supervisory Monitoring and Monetary Incentives – Evidence from a High-End Retail Store." *Working Paper* (2018).

Barankay, Iwan. "Rank Incentives: Evidence from a Randomized Workplace Experiment." *Working Paper* (2012).

Bartel, Ann, Casey Ichniowski, and Kathryn Shaw. "Using 'insider econometrics' to study productivity." *American Economic Review* Papers and Proceedings 94. 2 (2004): 217-223.

Belot, Michèle, and Marina Schröder. "The spillover effects of monitoring: A field experiment." *Management Science* 62.1 (2016): 37-45.

Bénabou, Roland, and Jean Tirole. "Intrinsic and extrinsic motivation." *The Review of Economic Studies* 70.3 (2003): 489-520.

Bénabou, Roland and Jean Tirole. "Incentives and Prosocial Behavior." *American Economic Review* 96.5. (2006): 1652-1678.

Buckingham, Marcus, and Ashley Goodall. "Reinventing performance management." *Harvard Business Review* 93, no. 4 (2015): 40-50.

Blanes i Vidal, Jordi and Mareike Nossol. "Tournament Without Prizes: Evidence from Personnel Records." *Management Science* 57.10 (2011): 1721-1736.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts. "Does Management Matter? Evidence from India." *Quarterly Journal of Economics* 128.1 (2013): 1-51.

Bloom, Nicholas, James Liang, John Roberts, and Zhichung Jenny Ying. "Does Working from Home Work? Evidence from a Chinese Experiment." *Quarterly Journal of Economics* 130.1 (2015): 165-217.

Bloom, Nicholas, and John Van Reenen. "Measuring and Explaining Management Practices across Firms and Countries." *Quarterly Journal of Economics* 122.4 (2007): 1351-1408.

Boly, Amadou. “On the incentive effects of monitoring: evidence from the lab and the field.” *Experimental Economics* 14.2 (2011): 241-253.

Bradler, Christiane, Robert Dur, Susanne Neckermann, Arjan Non. “Employee Recognition and Performance: A Field Experiment.” *Management Science* 62.11 (2016): 3085-3391.

Brynjolfsson, Erik, and Paul Milgrom. “Complementarity in Organizations.” *Handbook of Organizational Economics* (2013): 11-55.

Campbell, Dennis, Marc J. Epstein, and F. Asis Martinez-Jerez. “The Learning Effects of Monitoring.” *The Accounting Review* 86.6 (2011):1909-1934.

Cappelli, Peter, and Martin J. Conyon. “What Do Performance Appraisals Do?” *ILR Review* 71.1 (2018): 88-116.

Cappelli, Peter, and Anna Tavis. “The performance management revolution.” *Harvard Business Review* 94.10 (2016): 58-67.

Cassar, Lea, and Stephan Meier. “Nonmonetary Incentives and the Implications of Work as a Source of Meaning.” *Journal of Economic Perspectives* 32.3 (2018): 215-38.

Delfgaauw, Josse, Robert Dur, Joeri Sol, and Willem Verbeke. “Tournament incentives in the field: Gender differences in the workplace.” *Journal of Labor Economics* 31.2 (2013): 305-326.

Didonato, Tom. “Stop Basing Pay on Performance Reviews.” *Harvard Business Review Digital Articles* (2014).

Duflo, Esther. “Richard T. Ely Lecture: The Economist as Plumber.” *American Economic Review* 197.5 (2017): 1-26.

Eisenhardt, Kathleen M. “Agency theory: An assessment and review”. *Academy of Management Review* 14.1 (1989): 57-74.

Ellingsen, Tore, and Magnus Johannesson. “Paying respect.” *Journal of Economic Perspectives* 21.4 (2007): 135-150.

Ellingsen, Tore, and Magnus Johannesson. “Pride and prejudice: The human side of incentive theory.” *American Economic Review* 98.3 (2008): 990-1008.

Falk, Armin, and Michael Kosfeld. “The Hidden Costs of Control.“ *American Economic Review* 96.5 (2006):1611-1630.

Fehr Ernst, and John List. “The Hidden Costs and Returns of Incentives-Trust and Trustworthiness Among CEOs.” *Journal of the European Economic Association* 2.5 (2004): 743-771.

Fehr, Ernst, and Bettina Rockenbach. “Detrimental effects of sanctions on human altruism.” *Nature* 422.6928 (2003): 137.

Floyd, Eric and John A. List. “Using Field Experiments in Accounting and Finance.” *Journal of Accounting Research* 54.2 (2016): 437-475.

Friebel, Guido, Matthias Heinz, Miriam Krüger, and Nikolay Zubanov. “Team incentives and performance: Evidence from a retail chain.” *American Economic Review* 107.8 (2017): 2168-2203.

Gibbons, Robert and John Roberts. “Economic Theories of Incentives in Organizations,” *Handbook of Organizational Economics* (2013): 56-99.

Gneezy, Uri, and Aldo Rustichini. “Pay enough or don't pay at all.” *The Quarterly Journal of Economics* 115.3 (2000a): 791-810.

Gneezy, Uri, and Aldo Rustichini. “A fine is a price.” *The Journal of Legal Studies* 29.1 (2000b): 1-17.

Harrison, Glenn W. and John List. “Field Experiments.” *Journal of Economic Literature* 42.4. (2004): 1009-1055.

Holmström, Bengt. “Moral Hazard and observability.” *The Bell Journal of Economics*. 10.1 (1979): 74-91.

Holmström, Bengt. “Managerial incentive problems: A dynamic perspective.” *The Review of Economic Studies* 66.1 (1999): 169-182.

Holmström, Bengt. “Pay for Performance and Beyond.” *American Economic Review* 107(7) (2017): 1753-1777.

Ichniowski, Casey, Kathryn Shaw, and Giovanna Prennushi. “The effects of human resource management practices on productivity: A study of steel finishing lines.” *American Economic Review* (1997): 291-313.

Ichniowki, Casey, and Kathryn Shaw. “Beyond Incentive Pay: Insiders’ Estimates of the Value of Complementary Human Resource Management Practices.” *Journal of Economic Perspectives* 17.1 (2003):155-180.

Jensen, Michael C., and William H. Meckling. “Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure.” *Journal of Financial Economics* 3.4 (1976): 305-360.

Jensen, Michael C., and Murphy, Kevin J. “Performance pay and top-management incentives.” *Journal of Political Economy*, 98.2 (1990): 225-264.

Lazear, Edward P., and Paul Oyer, “Personnel Economics" in in: R. Gibbons and J. Roberts (eds.), *Handbook of Organizational Economics*. (2013): 479-519.

Lazear, Edward P. “Compensation and Incentives in the Workplace” *Journal of Economic Perspectives* 32.3 (2018): 195-214.

Lourenço, Sofia M. “Monetary Incentives, Feedback, and Recognition – Complements or Substitutes? Evidence from a Field Experiment in a Retail Service Company.” *The Accounting Review* 91.1 (2016): 279-297.

Manthei, Kathrin, Dirk Sliwka and Timo Vogelsang. “Performance Pay and Prior Learning: Evidence from a Retail Chain.” *IZA Discussion Paper No. 11859* (2018).

Manthei, Kathrin, Dirk Sliwka, and Timo Vogelsang. “Information Provision and Performance Pay.” *mimeo* (2019).

Merchant, Kenneth A., and Wim A. Van der Stede. “Management Control Systems – Performance Measurement, Evaluation and Incentives.“ Harlow, UK: Pearson (2017).

Milgrom, Paul, and John Roberts. “The economics of modern manufacturing: Technology, strategy, and organization.” *American Economic Review* 80.3 (1990): 511-528.

Milgrom, Paul, and John Roberts. “Economics, Organization & Management.” Prentice Hall (1992).

Milgrom, Paul, and John Roberts. “Complementarities and fit strategy, structure, and organizational change in manufacturing.” *Journal of Accounting and Economics* 19.2-3 (1995): 179-208.

Nagin, Daniel S., James B. Rebitzer, Seth Sanders, and Lowell J. Taylor. “Monitoring, Motivation, and Management: The Determinants of the Opportunistic Behavior in a Field Experiment.” *American Economic Review* 92.4 (2002): 850-873.

Prendergast, Canice. “The Provision of Incentives in Firms.” *Journal of Economic Literature* 37.1. (1999): 7-63.

Rebitzer, James B., and Lowell J. Taylor. “Extrinsic rewards and intrinsic motives: standard and behavioral approaches to agency and labor markets.” *Handbook of Labor Economics*, Vol. 4 (2011): 701-772.

Roth, Alvin E. “The economist as engineer: Game theory, experimentation, and computation as tools for design economics.” *Econometrica* 70.4 (2002): 1341-1378.

Sliwka, Dirk. “Trust as a signal of a social norm and the hidden costs of incentive schemes.” *American Economic Review* 97.3 (2007): 999-1012.

Tran, Anh, Richard Zeckhauser. “Rank as an inherent incentive: Evidence from a field experiment.” *Journal of Public Economics* 96. 6 (2012): 645-650.

Vohs, Kathleen D., Nicole L. Mead, Miranda R. Goode. “The Psychological Consequences of Money.” *Science* 314.5802 (2006): 1154.1156.

9. Appendix

9.1. Proof of Proposition 1

The first derivative of the agent's objective function is

$$\begin{cases} \beta - c'(e) + rg'(\hat{e} - e) & if \quad e < \overline{e} \\ \beta - c'(e) & if \quad e \geq \overline{e}. \end{cases}$$

Without performance reviews and performance pay the agent chooses $e = \overline{e}$. If $c'^{-1}(\beta) \geq \hat{e}$ (i.e. if β is sufficiently large) then the agent always chooses $e = c'^{-1}(\beta)$ irrespective of the monitoring activity. If this is not the case, optimal efforts are characterized by

$$\beta - c'(e) + rg'(\hat{e} \ - e) = 0.$$

Hence, efforts are in this case increasing in r: when performance reviews are in place, marginal returns to efforts are higher, as higher efforts (below ê) then additionally reduce the psychological costs of underperformance. In this case by the implicit function theorem we have that

$$\frac{\partial e}{\partial \beta} = \frac{1}{c''(e) + rg''(\hat{e} \ - e)}$$

which implies that

$$\left.\frac{\partial e}{\partial \beta}\right|_{r=0} > \left.\frac{\partial e}{\partial \beta}\right|_{r=1} > 0.$$

■

9.2. Proof of Proposition 2

Consider first the case without performance reviews. Suppose for the moment that conditional expectations are linear in profits, i.e.

$$E[\eta|\pi] = \tau_0 + \tau_1 \pi.$$

We will show that there indeed exists a PBE in which expectations are linear.[41] If this is the case the first order condition of the agent's objective function is

[41] As Bénabou and Tirole (2006) show, this equilibrium is unique in the class of equilibria with differentiable strategies.

$$\eta_m \beta - c(e - \eta_e) + \zeta E[\eta_e|\pi]\eta_m \beta + c\eta_e - ce + \zeta\tau_1 = 0$$

and optimal efforts become $e = \frac{\eta_m \beta + \zeta\tau_1}{c} + \eta_e$. Using that for normally distributed random variables $E[Y|X] = E[Y] + Cov[X,Y]/V[X]\,(X - E[X])$ we obtain that

$$\begin{aligned} E[\eta_e|\pi] &= m_{\eta_e} + \frac{Cov[\eta_e, \eta_e + \eta_m\beta + \zeta\tau_1/c + \varepsilon]}{V[\eta_e + \eta_m\beta + \zeta\tau_1/c + \varepsilon]}\left(\pi - E\left[\eta_e + \frac{\eta_m\beta + \zeta\tau_1}{c} + \varepsilon\right]\right) \\ &= m_{\eta_e} + \frac{\sigma_{\eta_e}^2}{\sigma_{\eta_e}^2 + \sigma_{\eta_m}^2 \frac{\beta^2}{c^2} + \sigma_\varepsilon^2}\left(\pi - m_{\eta_e} - \frac{m_{\eta_m}\beta + \zeta\tau_1}{c}\right). \end{aligned}$$

Hence, expectations are then indeed linear in π and thus $\tau_1 = \sigma_{\eta_e}^2 / (\sigma_{\eta_e}^2 + \sigma_{\eta_m}^2 \frac{\beta^2}{c^2} + \sigma_\varepsilon^2)$. In turn,

$$e_{r=0} = \eta_e + \frac{\eta_m \beta}{c} + \zeta \frac{\sigma_{\eta_e}^2}{c\left(\sigma_{\eta_e}^2 + \sigma_\varepsilon^2\right) + \sigma_{\eta_m}^2 \frac{\beta^2}{c}}.$$

when performance reviews are in place then $\sigma_\varepsilon^2 = 0$ and thus

$$e_{r=1} = \eta_e + \frac{\eta_m \beta}{c} + \zeta \frac{\sigma_{\eta_e}^2}{c\sigma_{\eta_e}^2 + \sigma_{\eta_m}^2 \frac{\beta^2}{c}}$$

Taking expectations gives the profit function $\Pi^N(\beta, r)$. The profit effect of introducing performance reviews is

$$\begin{aligned} \Delta\Pi_r^N &= \Pi^N(\beta, 1) - \Pi^N(\beta, 0) \\ &= \zeta\left(\frac{\sigma_{\eta_e}^2}{c\sigma_{\eta_e}^2 + \sigma_{\eta_m}^2 \frac{\beta^2}{c}} - \frac{\sigma_{\eta_e}^2}{c(\sigma_{\eta_e}^2 + \sigma_\varepsilon^2) + \sigma_{\eta_m}^2 \frac{\beta^2}{c}}\right) > 0 \end{aligned}$$

which establishes claim (i) and claim (ii) follows as

$$\frac{\partial \Delta\Pi_r^N}{\partial \beta} = \zeta\left(-\frac{\sigma_{\eta_e}^2 \sigma_{\eta_m}^2 \frac{2\beta}{c}}{\left(c\sigma_{\eta_e}^2 + \sigma_{\eta_m}^2 \frac{\beta^2}{c}\right)^2} + \frac{\sigma_{\eta_e}^2 \sigma_{\eta_m}^2 \frac{2\beta}{c}}{\left(c(\sigma_{\eta_e}^2 + \sigma_\varepsilon^2) + \sigma_{\eta_m}^2 \frac{\beta^2}{c}\right)^2}\right) < 0.$$

To establish claim (iii) note that the profit effect of introducing a bonus is

$$\begin{aligned} \Delta\Pi_{\beta}^{N} &= \Pi^{N}(\beta,r) - \Pi^{N}(0,r) \\ &= \frac{m_{\eta m}\beta}{c} - \zeta\underbrace{\left(\frac{\sigma_{\eta e}^{2}}{c\sigma_{\eta e}^{2}+(1-r)\cdot c\sigma_{\varepsilon}^{2}} - \frac{\sigma_{\eta e}^{2}}{c\sigma_{\eta e}^{2}+\sigma_{\eta m}^{2}\frac{\beta^{2}}{c}+(1-r)\cdot c\sigma_{\varepsilon}^{2}}\right)}_{>0} \end{aligned} \tag{A1}$$

which will be negative if ζ is sufficiently large. To establish claim (iv) note that the expression in brackets in (A1) is larger for $r = 1$ if

$$c\sigma_{\eta_e}^{2}\left(c\sigma_{\eta_e}^{2} + \sigma_{\eta_m}^{2}\frac{\beta^{2}}{c}\right) < \left(c\sigma_{\eta_e}^{2} + c\sigma_{\varepsilon}^{2}\right)\left(c\sigma_{\eta_e}^{2} + \sigma_{\eta_m}^{2}\frac{\beta^{2}}{c} + c\sigma_{\varepsilon}^{2}\right)$$

which always holds. ■

9.3. Tables and Figures

Table A1: Classification of Store Manager Tasks

Task	Classification
Ordering of fruits and vegetables, plants	Ordering
Ordering of baked goods	
Ordering of meat	
Additional Ordering	
Baking of bakery articles	
Preparation of secondary placements	Placements
Presentation and maintenance of special-offer tables (Non-Food/ Food/ end of aisle)	
Maintaining product positioning plans	
Quality checks fruits, vegetables and plants	Cleanliness
Cleanliness of the baked goods stations	
Preservation and maintenance of the condition of the furnishings and the inventory (e.g., shelves, bumpers, freezers, cash desks)	
Guaranteeing the cleanliness and orderliness inside and outside the store	
Analysis of Spoilage	Analysis KPI
Analysis of Sales	
Analysis of Personnel Costs	
Analysis of Hourly Output	
Analysis of Inventory	
Checking minimum durability date (meat, dairy, convenience)	Inventory
Process left overs	
Stocking of goods and maintenance of shelves (colonial goods, frozen goods, load)	
Incoming goods inspection	
Security of goods	
Working on gap listing and inventory care	
Training of cashier employees	Personnel Management
Appraisal interviews / leadership	
Staff planning	
Communication with customers, processing of customer requests	Own Operational Activities
Own cashier work	
(Temporary price reductions)	

Table A2: Balancing Table

	(1) Descriptives Overall	(2) Descriptives Control	(3) Descriptives Bonus	(4) Descriptives Review	(5) Descriptives Bonus&Review
Profits Jan-Mar '17	26511.69 (10963.98)	27776.48 (11949.38)	25549.85 (11373.97)	26138.93 (8450.81)	26381.56 (11544.57)
Planned Profits Jan-Mar '17	28166.79 (10229.54)	28827.52 (11253.13)	28221.86 (10796.1)	27397.36 (8226.94)	28103.59 (10367.23)
Female Store Manager (Y/N)	0.55 (0.50)	0.67 (0.48)	0.68 (0.47)	0.4*** (0.50)	0.44** (0.50)
Walking Customers (Y/N)	0.13 (0.34)	0.12 (0.03)	0.14 (0.35)	0.16 (0.37)	0.13 (0.34)
FTE	6.39 (1.35)	6.54 (1.10)	6.23 (1.49)	6.22 (1.27)	6.52 (1.51)
Age of Store	14.90 (8.79)	13.35 (8.18)	14.09 (9.39)	17.65** (9.19)	14.85 (10.05)
Age Store Manager	41.42 (9.61)	41.93 (9.78)	42.49 (9.75)	39.5 (9.02)	41.60 (10.05)
Tenure Store Manager	15.33 (8.86)	15.80 (8.84)	16.90 (8.09)	13.55 (8.61)	14.98 (9.59)
Tenure District Manager	13.12 (11.05)	14.82 (10.31)	10.10 (10.11)	15.04 (11.00)	12.47 (12.23)
Store Space	710.22 (145.53)	744.57 (134.88)	714.29 (143.39)	689.08 (179.76)	691 (121.86)
Max. Observations	224	60	51	50	63

Note: The table reports means of the respective variables for the different treatment groups and their standard deviations in parentheses. Asterisks display significance levels from t-tests (fisher exact test for binary variables) of the respective treatment group against the control group. * p<0.1, ** p<0.05, *** p<0.001.

Table A3: Regression including only Treatment Months

	(1) OLS	(2) log OLS	(3) OLS	(4) log OLS
Treatment Effect BONUS	-366.59 (580.88)	-0.0502 (0.0365)	-302.60 (627.39)	-0.0357 (0.0404)
Treatment Effect REVIEW	1101.66** (514.86)	0.0650** (0.0296)	1390.47** (534.66)	0.0649** (0.0262)
Treatment Effect BONUS&REVIEW	-733.38 (492.00)	-0.0202 (0.0297)	-638.23 (523.46)	-0.0215 (0.0286)
Wald test REVIEW=BONUS&REVIEW	p=0.0002	p=0.0065	p=0.0001	p=0.0041
Time FE	Yes	Yes	Yes	Yes
Store FE	No	Yes	No	Yes
District Manager FE	No	Yes	No	Yes
Store Manager FE	No	Yes	No	Yes
Refurbishments	Yes	Yes	Yes	Yes
Planned Profits	Yes	Yes	Yes	Yes
Further Controls	No	No	Yes	Yes
N Observations	669	669	669	669
N Stores	224	224	224	224
N Cluster	31	31	31	31
Overall R^2	0.8696	0.6491	0.8726	0.6622

Note: The table reports results from ordinary least squares regressions using only data from the treatment period further controlling for the mean of profits from January 2016-March 2017. All regressions control for possible refurbishments of a store, the randomization pair, and the companies planed profits. Columns 3&4 further control for variables with slight imbalance between treatments (gender, age of the store). Observations are excluded once a store manager switched the store during the treatment period. Robust standard errors are clustered on the district level of the treatment start and displayed in parentheses. * p<0.1, ** p<0.05, *** p<0.001.

Table A4: Monthly Treatment Effects

	(1) FE	(2) FE	(3) ln FE	(4) ln FE
Treatment Effect BONUS	-74.22	293.0	-0.0181	0.00471
1st Month	(637.2)	(684.7)	(0.0325)	(0.0429)
Treatment Effect BONUS	572.8	912.4	0.0327	0.0536*
2nd Month	(726.4)	(820.0)	(0.0314)	(0.0314)
Treatment Effect BONUS	-585.1	-202.9	-0.0237	-0.000987
3rd Month	(928.8)	(1053.1)	(0.0941)	(0.110)
Treatment Effect BONUS	-1379.8	-1014.3	-0.0554	-0.0381
4th Month (after treatment)	(1032.5)	(1074.5)	(0.0449)	(0.0514)
Treatment Effect BONUS	854.5	1225.3	-0.0196	-0.00470
5th Month (after treatment)	(1436.0)	(1626.6)	(0.0271)	(0.0402)
Treatment Effect REVIEW	1417.1*	1465.3	0.0645*	0.0751*
1st Month	(783.4)	(867.4)	(0.0332)	(0.0442)
Treatment Effect REVIEW	2451.7***	2490.4***	0.0957***	0.104***
2nd Month	(618.8)	(692.7)	(0.0291)	(0.0295)
Treatment Effect REVIEW	461.9	966.1	0.0680	0.0922
3rd Month	(782.6)	(889.9)	(0.0551)	(0.0723)
Treatment Effect REVIEW	-1038.2	-461.8	-0.0599	-0.0332
4th Month (after treatment)	(1149.4)	(1255.2)	(0.0493)	(0.0592)
Treatment Effect REVIEW	746.8	1086.4	0.0205	0.0342
5th Month (after treatment)	(685.7)	(1044.9)	(0.0255)	(0.0433)
Treatment Effect BONUS&REVIEW	-590.7	-474.2	-0.0274	-0.0184
1st Month	(590.5)	(511.4)	(0.0294)	(0.0398)
Treatment Effect BONUS&REVIEW	801.1	886.8	0.0267	0.0306
2nd Month	(686.1)	(664.7)	(0.0364)	(0.0381)
Treatment Effect BONUS&REVIEW	-1074.2	-958.3	-0.00156	0.000577
3rd Month	(1030.9)	(1165.6)	(0.0751)	(0.0970)
Treatment Effect BONUS&REVIEW	-656.0	-456.7	-0.0536	-0.0513
4th Month (after treatment)	(1234.0)	(1410.9)	(0.0589)	(0.0768)
Treatment Effect BONUS&REVIEW	-121.7	-30.55	-0.0260	-0.0297
5th Month (after treatment)	(709.6)	(940.2)	(0.0349)	(0.0507)
Fixed Effects (Time, Store)	Yes	Yes	Yes	Yes
Fixed Effects (District & Store Manager)	No	Yes	No	Yes
Refurbishments	Yes	Yes	Yes	Yes
Planned Profits	Yes	Yes	Yes	Yes
Observations	4421	4203	4412	4194
N Store	224	224	224	224
N Cluster	31	31	31	31
Within R^2	0.2407	0.2726	0.1703	0.1938
Overall R^2	0.7484	0.5379	0.6152	0.4063

Note: The table reports results from fixed effects regressions with the profits on the store level as the dependent variable. The regression accounts for time and store fixed effects (column 1-4) and adds fixed effects for district and store managers in columns 2&4. The fixed effects regressions compare pre-treatment observations (January 2016-March 2017) with the observations during the experiment (April 2017 – June 2017). All regressions control for possible refurbishments of a store and the companies planned value. Observations are excluded once a store manager switched the store during the treatment period. *Treatment Effect* thus refers to the difference-in-difference estimator. Robust standard errors are clustered on the district level of the treatment start and displayed in parentheses.* p<0.1, ** p<0.05, *** p<0.001.

Table A5: Treatment Effects on Review Conversation Notes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Reference Group:	Overall	Overall	Tasks	Tasks	Problems	Problems	Tasks for	Tasks for
Treatment REVIEW			Done	Done	with	with	Next	Next
					Tasks	Tasks	Time	Time
Treatment Effect	0.166	0.129	0.0854	0.175	-0.482***	-0.684***	0.136	-0.026
BONUS&REVIEW	(0.765)	(0.557)	(0.734)	(0.634)	(0.140)	(0.140)	(0.628)	(0.641)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
N of Observations	118	89	118	89	118	89	118	89
Cluster	18	17	18	17	18	17	18	17
Overall R^2	0.0018	0.5471	0.008	0.3282	0.1623	0.3706	0.0010	0.1780

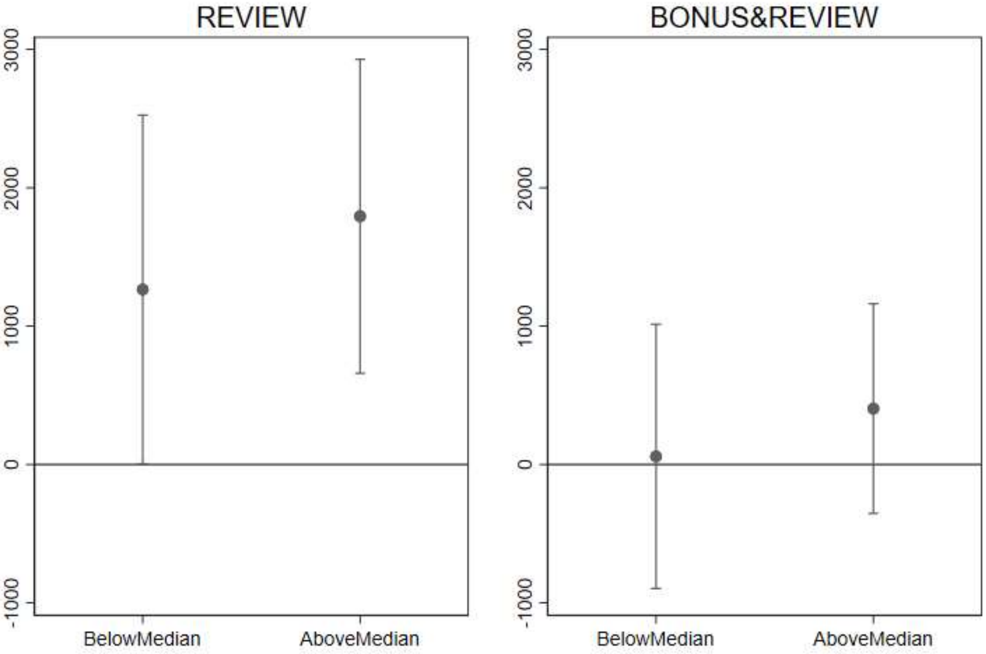
Note: The table reports results from ordinary least squares (OLS) regressions with the different subsections of the review conversations as depending variable. Columns 2,4,6&8 control further for store size, number of employees, store manager's age and prior performance evaluation, as well as randomization group. The Treatment REVIEW serves as the reference group. Robust standard errors are clustered on the district level of the treatment start and displayed in parentheses.* p<0.1, ** p<0.05, *** p<0.001.

Table A6: Content of Review Meetings

Reference Group: *Treatment REVIEW*	(1)	(2)
Ordering	0.0898***	0.0924***
	(0.0182)	(0.0146)
Placements	0.0983***	0.0994***
	(0.0131)	(0.0141)
Cleanliness	0.0386*	0.0332
	(0.0220)	(0.0250)
Analysis KPI	0.00368	-0.00450
	(0.0167)	(0.0174)
Inventory	0.0507***	0.0514***
	(0.0169)	(0.0170)
BONUS x Ordering	-0.0353*	-0.0308*
	(0.0198)	(0.0173)
BONUS x Placements	0.0210	0.0305*
	(0.0136)	(0.0159)
BONUS x Cleanliness	-0.0432**	-0.0367**
	(0.0191)	(0.0167)
BONUS x Analysis KPI	0.0140	0.0267
	(0.0229)	(0.0257)
BONUS x Inventory	-0.0127	-0.0121
	(0.0179)	(0.0170)
Planned task (section 2)	-0.00420	-0.00500
	(0.0128)	(0.0129)
Problems encountered (section 3)	-0.0685***	-0.0670***
	(0.0102)	(0.00860)
Meeting slot 2	0.00438	0.00376
	(0.00562)	(0.00539)
Meeting slot 3	0.00746	0.000295
	(0.00815)	(0.00809)
Meeting slot 4	-0.000128	0.00165
	(0.00590)	(0.00717)
Meeting slot 5	-0.0107	-0.00810
	(0.0135)	(0.0152)
Meeting slot 6	-0.000548	-0.00765
	(0.00976)	(0.00909)
Controls	No	Yes
Observations	35931	27318
Cluster	18	17
Pseudo R^2	0.1377	0.1492

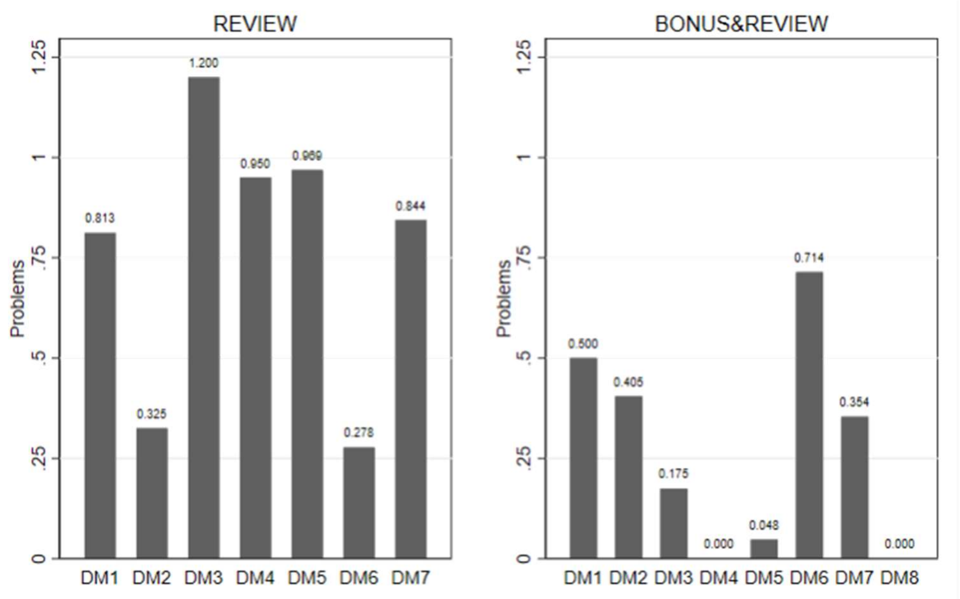
Note: The table reports results from Probit regressions. Dependent variable y_{kst} is a dummy variable indicating whether a task k was mentioned in section s of a review meeting conducted in time slot t. Further controls in columns (4)-(6) are store size, number of employees, store manager's age and prior performance evaluation, as well as randomization group. The Treatment REVIEW serves as the reference group. Standard errors are clustered on the district level at treatment start and displayed in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Figure A1: Average Treatment Effects Depending on the amount of Review Conversations



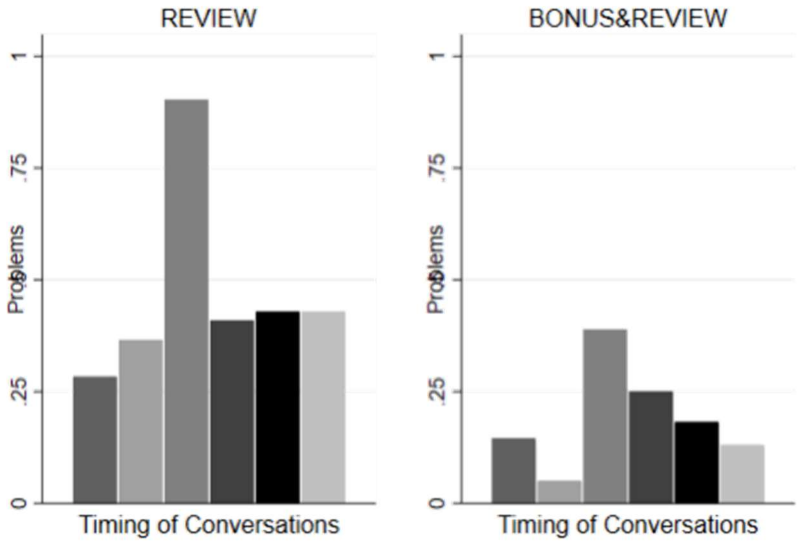
Note: The figure displays separately estimated treatment effects from our standard fixed effects regression specification depending on whether the number of performance reviews conducted is below or above/equal to the median (4). 95% confidence bars are displayed.

Figure A2: Average Number of Notes in Subsection “Problems” per Conversation By District Manager



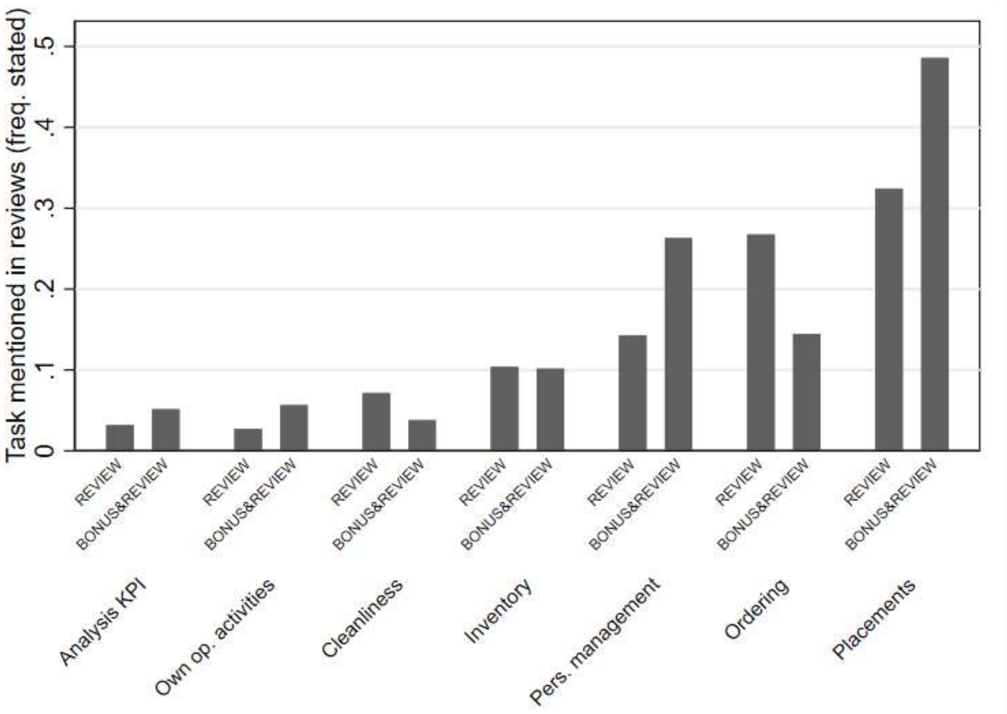
Note: The figure displays the average number of problems (in notes/sentences) per session displayed for each district manager separately.

Figure A3: Average Number of Notes in Subsection “Problems” per Conversation Depending on Time of Conversation



Note: The figure displays the average number of problems (in notes/sentences) per session. The average number of problems is displayed for each time point separately (1^{st} bar= 1^{st} two weeks, 2^{nd} bar= 2^{nd} two weeks).

Figure A4: Tasks covered in review conversations



Note: The figure displays the frequency of mentioned tasks per review conversation (all categories –tasks done, problems, tasks next time- pooled).

9.4. Instructions (Online Appendix)

9.4.1. Store Manager - CONTROL Group (sent to their home address, originally in German)

Project DB1[42]

Dear Mr./Mrs. XXX,

a positive DB1 profit measure is important for the economic success of [*the company*]. For this reason, the DB1 project will be implemented in your region during the next few months. Within the scope of the DB1 project, you will have the opportunity to earn an additional bonus, receive a learning unit and have a regular DB1-Conversation with your district manager in the near future.

You will now have access to the information package.

Learning Unit[43]:

In order to renew and deepen your knowledge about the DB1, we have put together an online learning unit for you. This consists of a short learning video and a quiz afterwards. In order for [*the company*] to remain economically strong, you should finish this learning unit by 08.04.2017!

The learning unit is provided by the University of Cologne. You can complete the learning unit using the access data listed below in the EDP (Home left> Section "Other"), with your private computer or your smartphone. Please see the access data listed below.

Access data learning unit:
Please visit the following website for the learning unit:

Your password is:

Alternatively, you can also use the following QR code directly:

In order for you to keep track of the explained figures, you will receive a separate DB1-report in the Store Data Warehouse at the end of the following month.

We would like to thank you sincerely in advance for your participation and support.

If you have any questions, please contact your district management / personnel management.

[42] The company uses „DB1" (short for Deckungsbeitrag 1/ contribution margin) as an internal title for the simplified profit measure explained above in our study: *Profit = Net Sales – Cost of Goods Sold – Staff Costs –Inventory Losses*

[43] Due to previous company wording, the company uses "learning unit" as an internal description for the learning video, the quiz, the margin information and the monthly feedback. We refer to this as "information package" in the above.

Yours sincerely

9.4.2. Store Manager – BONUS&REVIEW Group (send to their home address, originally in German)

Project DB1

Dear Mr./Mrs. XXX,

a positive DB1 profit measure is important for the economic success of [*the company*]. For this reason, the DB1 project will be implemented in your region during the next few months. Within the scope of the DB1 project, you will have the opportunity to earn an additional bonus, receive a learning unit and have a regular DB1-Conversation with your district manager in the near future.

Your bonus period starts on 01.04.2017 for 3 months. You will now have access to the learning unit. Your district manager will contact you regarding the DB1-Conversation.

Bonus:

Within this project, you will be able to earn an additional bonus in your store over the next three months (April, May, June) for increasing the DB1 profit measure.

Therefore, the DB1 profit measure of your store will be compared monthly with the plan DB1 of the respective month. If your DB1 profit measure is more than 80% of the plan DB1, you will receive a bonus. From the difference between the DB1 profit measure and 80% of the plan DB1, you are paid-out 5% as a premium in euros.

Calculation: DB1-Bonus (in €) = (DB1 – 80% of the Plan DB1) * 0,05

The DB1-Bonus is always calculated at the end of the month. The sum of the bonuses from the three months will be paid out to you in September 2017 with your payroll. This means that the bonus amount can be negative in a single month (if the plan achievement is under 80%). Should you still have a negative amount after the end of the three months, you will be paid € 0. Please see the attached info sheet for the bonus calculation.

Information about your bonus amount will always be send by post to your home at the end of the following month.

Learning Unit:

In order to renew and deepen your knowledge about the DB1, we have put together an online learning unit for you. This consists of a short learning video and a quiz afterwards. In order for [*the company*] to remain economically strong, you should finish this learning unit until 08.04.2017!

The learning unit is provided by the University of Cologne. You can complete the learning unit using the access data listed below in the EDP (Home left> Section "Other"), with your private computer or your smartphone. Please see the access data listed below.

Access data learning unit:

Please visit the following website for the learning unit:

Your password is:

Alternatively, you can also use the following QR code directly:

District manager DB1-Conversation:

Your district manager will also have an in-depth DB1-Conversation with you every two weeks. Within this conversation, he will ask you about actions you have already taken to increase the DB1 profit measure. In addition, you can discuss possible problems with him.

In order for you to keep track of the explained figures, you will receive a separate DB1-report in the Store Data Warehouse at the end of the following month.

We would like to thank you sincerely in advance for your participation and support.

If you have any questions, please contact your district management / personnel management.

Yours sincerely

Information about the DB1-Bonus (*added to both BONUS treatments*)

The DB1 profit measure represents the economic success of [*the company*]. The more positive it is, the stronger [*the company*] is positioned. The DB1 profit measure is the net sales minus influenceable costs such as inventory and personnel costs.

Please find attached the details for the calculation as well as a fictitious example.

Calculation DB1-Bonus

From 01.04.2017 up to and including 30.06.2017, you will be informed monthly about the increase of your DB1 profit measure compared to your plan of the DB1.

If your DB1 profit measure is at least 80% of the plan DB1, you will receive a bonus. From the difference between your actual DB1 profit measure and 80% of the plan DB1, you are paid-out 5% as a bonus in euros.

Amount in euros = (DB1 – 80% plan DB1) * 0,05

This amount in euros is added up for the months of April, May and June and then paid out to you with your payroll in September.

Fictious Example

Month April: The DB1 in April was 30.000 with a plan DB1 of 28.000.
This results in a euro amount of (30000 – 0.8 * 28000) * 0.05 = 380 Euro.

Month May: The DB1 in April was 24.000 with a plan DB1 of 29.000.
This results in a euro amount of (22000 – 0.8 * 29000) * 0.05 = - 60 Euro.

Month June: The DB1 in April was 28.000 with a plan DB1 of 29.000.
This results in a euro amount of (28000 – 0.8 * 29000) * 0.05 = 240 Euro.

Total bonus paid: 380 (April) – 60 (May) + 240 (June) = 560€

Thus, in September 560 € would be paid as a bonus.

9.4.3. Monthly Communication to Store Manager (sent to their home address, originally in German)

Project DB1

Dear Mr./Mrs. XXX,

Please find below a summary of your key figures in the first month of the project.

<u>Summary of your DB1 profit measure in April 2017:</u>[44]
(Amounts are not rounded until the end)

Sales:
Cost of good sold:
Personnel costs:
Inventory:

This results in a DB1 April/2017:
For a plan DB1 April/2017:

The resulting bonus amount for the month of April is:

(DB1 – 0.8 * plan DB1) * 0.05 € =

<u>Summary of your bonus amounts since April 2017:</u>
(Amounts are not rounded until the end)

Bonus amount April 2017: € (gross)

The sum of the bonus amounts (if greater than 0) will be paid-out at the end of the three-month period in September 2017 with your payroll. Please note that positive bonus amounts are offset against negative ones. There will only be one bonus payment of the grand total in September.

For further questions, please contact your district manager / personnel management.

[44] For accounting reasons, the letter in May came with additional information: “In April, adjusting entries through accounting were posted to the region only and not distributed to the branches. Their profit margin is therefore too well represented. These bookings will be made up with the May-finalization. Therefore, you will find the margin correction in your May letter with a reversed sign. In sum of April and May, the correction value will be € 0.00. We ask for your understanding.”

9.4.4. District Manager – Review Group (sent to their e-mail address, originally in German)

Project DB1

A positive DB1 profit measure is important for the economic success of [*the company*]. For this reason, the DB1 project will be implemented in your region during the next few months.

Within the scope of the DB1 project, all store managers will participate in a learning unit about the DB1 profit measure in the near future. In addition, stores in randomly selected districts receive an additional DB1-Conversation. Moreover, an additional bonus for store managers is introduced in all stores of the region. For administrative reasons, the bonus will be introduced in the districts at different times. The assignment happens randomly according to a statistical procedure

From 27.03.2017, store managers will have access to a learning unit regarding the DB1 profit measure. Please make sure that the learning unit is completed by the store managers in your district.

From 01.04.2017, an additional DB1-Conversation will be introduced in your district.
In your district, store managers will receive the DB1-Bonus at a later date. You will be informed in sufficient time about the exact time frame.

Your store managers will be informed elaborately and separately by mail.

<u>Your store manager DB1-Conversation:</u>

We would like to ask you to hold an in-depth personal conversation with the store managers in your district every two weeks about the development of the DB1 profit measure (DB1-Conversation).
For this DB1-Conversation, we have attached a guideline for you which we would like you to fill out in note form with every conversation and send it back to your personnel management. During your conversation, you should not only inquire and examine what the store manager did, but also communicate what they should do differently until the next meeting. The DB1-Conversation should happen every two weeks on the key dates 18.04.2017, 02.05.2017, 16.05.2017, 30.05.2017, 13.06.2017, 27.06.2017. Store managers will be informed individually in a separate letter.

<u>Store manager learning unit:</u>

In order to renew and deepen the knowledge of store managers regarding the DB1 profit measure, we have put together an online learning unit for your store managers. This consists of a learning video and a quiz afterwards. If you are interested, you can also watch the learning video (provided by the University of Cologne) with the following link:
Your personal password is: XXXXX

Communication upon inquiries of store managers:

If your store managers ask why they are not getting a bonus for the increased DB1 profit measure, we also ask you to communicate that this is a random selection and that the store managers in your district will in any case receive a bonus at a later date.

For a neat evaluation, it is important that all district managers strictly follow this language regulation. Please do not pass any further information on to store managers and only discuss the bonus if a store manager explicitly asks for it.

The findings of this project are of great importance to [the company].

For inquiries your personnel management is at your disposal at any time.

Yours sincerely

Conversation guideline

Key date: □18.04.2017 □ 2.05.2017 □16.5.2017 □ 30.05.2017 □13.06.2017 □ 27.06.2017

Store Manager:

What has the store manager done to increase the DB1?

What problems have occurred?

Which measures / which next steps does the store manager want to carry out until the next meeting?