# Multitasking and Subjective Performance Evaluations - Theory and Evidence from a Field Experiment in a Bank

Kathrin Manthei*
RFH Cologne

Dirk Sliwka
University of Cologne, CesIfo, IZA

September 6, 2018

**Abstract**

We study the incentive effects of granting supervisors access to objective performance information when agents work on multiple tasks. We first analyze a formal model showing that incentives are lower powered when supervisors have no access to objective measures but assess performance subjectively by gathering information. This incentive loss is more pronounced when the span of control is larger and incentives are distorted towards more profitable tasks. We then investigate a field experiment conducted in a bank. In the treatment group managers obtained access to objective performance measures, which raised efforts and profits. We find that the effects are driven by larger branches and lower margin products.

**Key Words:** Incentives, Subjective Performance Evaluation, Multitasking, Field Experiment, Bank
**JEL Classification:** M52, J33, D23

*University of Cologne, Albertus Magnus Platz, 50923 Köln, Germany, tel: +49 221 470-5887, fax: +49 221 470-5078, e-mail: kathrin.manthei@rfh-neuss.eu, dirk.sliwka@uni-koeln.de.

# 1 Introduction

A key assumption in most principal agent models is that objective and verifiable performance measures are available and can be used to reward an agent's performance. In practice, however, the performance of individual employees is often assessed subjectively by a supervisor as not all aspects of employees' performance can be measured with objective key figures. There is substantial evidence that supervisors are frequently not able or not willing to accurately evaluate performance. Hence, subjective evaluations tend to be biased (see e.g., Bretz et al. (1992), Murphy and Cleveland (1995), Prendergast (1999), Ockenfels et al. (2014)), which may reduce the effectiveness of incentive schemes.

In this paper we study the interplay between subjective evaluations and multitasking incentives. We start by analyzing a formal model illustrating how a lack of objective performance information affects the allocation of efforts across tasks when supervisors endogenously gather information to assess the performance of employees. We then analyze data from a field experiment conducted by a bank and interpret our findings in light of the formal model. To the best of our knowledge, this paper provides the first study that allows a clean evaluation of causal performance effects of the introduction of an extensive set of objective measures brought about by an exogenous intervention implemented in a field setting in a firm.

The experiment was conducted by a retail bank in Germany. In a representatively selected subgroup of its branches, the bank made a comprehensive set of objective performance measures available for supervisors that enabled them to measure the exact profit contributions of the agents in the branch. Prior to the intervention and in the remaining branches, supervisors had no access to this information and assessed performance only subjectively. This exogenous change allows us to study the impact of access to objective performance measures on the overall profitability, as well as on profits in different product categories. By comparing the effects of the intervention on sales in the different product categories and across branches with different spans of control, we thus can infer patterns in the nature of

distortions under subjective evaluations.

Our formal model extends a framework developed by Prendergast and Topel (1996) or Prendergast (2002) to a Holmström and Milgrom (1991)-type multitask environment and endogenizes the attention spent on assessing performance for different tasks. The key idea of the model is (i) to illustrate an economic rationale for "biases" in subjective assessments and (ii) to study their effects on the provision of incentives under multitasking. In the model, a supervisor who is interested in accurately assessing agents' overall profit contributions has to allocate (limited) attention on different tasks to collect information on performance. The more attention is spent on a task, the more precise the supervisor's information on the agent's performance on a given task – and more precise information leads to higher powered incentives. A first implication is that due to limited attention incentives are lower powered under purely subjective assessments as compared to a situation when objective performance indicators are available. This effect is stronger when there is a larger span of control as supervisors of larger teams are not able to devote as much attention on each individual agent.

A further key insight of the model is that without objective performance information, supervisors devote more attention to monitoring more profitable tasks. The reason is that one unit of time spent on monitoring a more profitable task improves the accuracy of the overall assessment to a stronger extent than one unit spent on a less profitable task. This generates a rational "halo" or focus effect according to which subjective assessments are dominated by the more profitable tasks.[1] In turn, under subjective evaluations agents efforts are distorted towards the more profitable tasks: agents work harder on more profitable tasks not only because this generates higher profits but also because these tasks receive more attention from the supervisor. As a direct implication, incentive gains when performance can be measured objectively (and thus incentives are undistorted) are larger for less profitable tasks. Hence, the accessibility of comprehensive objective

[1] Note that here focus or salience effects occur not due to a cognitive bias (as recently analyzed in Bordalo et al. (2012), Kőszegi and Szeidl (2013)), but as a rational reaction to a limited time budget.

performance information should not only increase overall incentives but also lead to a reallocation of effort towards the less profitable tasks that received less attention without objective measures. The model therefore illustrates an economic rationale for two of the "biases" in subjective assessments commonly stressed in the psychological literature on performance evaluation (see, e.g. Murphy and Cleveland (1995), pp. 268), namely that subjective evaluations are too compressed ("central tendency"/"centrality bias") and that these evaluations tend to be dominated by very visible subdimensions ("halo effect"). Furthermore, the model shows the effect of these biases on incentives to exert effort for different tasks.

We then analyze the data from the field experiment and interpret our findings using the arguments developed in the formal model. First, we find that the introduction of objective measurement indeed significantly increases performance. In line with the mechanisms illustrated in the formal model the increase in profits is driven by (i) larger branches, where profits increase by more than 5%, and (ii) higher sales of products that have previously had a lower share of the overall sales volume. In small branches, where there is less division of labor, as essentially all employees sell the same product portfolio, we detect a shift of sales performance from core products (where sales volumes decrease significantly) to more fringe products (where there are significant performance increases).

We also find evidence for a strategic timing of sales in the treatment group. Customer appointments initiated by branch employees (a leading indicator of sales) had already significantly increased after the announcement of the treatment before it was actually implemented. However, profits increased significantly only after the date of the implementation. Moreover, self-initiated customer appointments have a stronger correlation with profits in the subsequent month in the treatment group directly before the intervention. Sales agents in the treatment group apparently contacted more customers prior to the intervention but then delayed actual sales in order to push profits into the time period when individual performance was tracked objectively.

While problems of subjective evaluations have received some attention

by economists in recent years, not much work has been done on the interplay between subjective evaluations and multitasking, and the amount of empirical evidence on the incentive consequences of subjective evaluations is still rather limited.[2] The typical discussion here has centered on the trade-off between greater precision of objective measures when measuring specific aspects of a job versus the advantage of subjective evaluation allowing for a more balanced assessment of different facets of the job and hence avoid the classical multitasking distortions (see, for instance, Gibbs et al. (2003), Zabojnik (2014), or Delfgaauw and Souverijn (2016)). Our paper has a different focus: We look at a situation where a comprehensive set of undistorted objective performance indicators is made available. We use this intervention to study the existence of the prior distortions under subjective assessment, thus aiming at a better understanding of the incentive consequences of these distortions in a multitasking environment.

Several papers have explored the role of subjective performance evaluations in relational contracts (Baker et al. (1994), Bentley MacLeod (2003), Levin (2003)). Importantly, this literature has mostly focused on subjective evaluations carried out by a principal who is a residual claimant on a firm's profits and is concerned about her reputation for honoring a pledge to pay a bonus. Here we study a setting in which subjective evaluations are carried out by supervisors who do not pay the bonus out of their own pockets (as is the case for most employment relationships). Distortions in subjective assessments made by a supervisor who is not the residual claimant have been studied in single-task models (Prendergast and Topel (1996), Prendergast (2002), Golman and Bhatia (2012), Giebe and Gürtler (2012), Kampkötter and Sliwka (2018)). Whereas in these papers the information available to supervisors is exogenously given, we endogenize the information collection, study how this affects the evaluation of performance across multiple tasks and then investigate these patterns empirically.

Our study also contributes to the recent literature using field experi-

[2] Exceptions are Engellandt and Riphahn (2011), Bol (2011), Berger et al. (2013), Takahashi et al. (2014), or Kampkötter and Sliwka (2018), who either use observational data or lab experiments.

ments within firms to evaluate the causal effects of incentive schemes (see, for instance, List and Rasul (2011), Bandiera et al. (2011), Levitt and Neckermann (2014) for overviews). Multitasking incentives have been analyzed in field experiments, for instance by Al-Ubaydli et al. (2015), who argue that multitasking distortions are weaker when the choice of an incentive scheme can signal information on the principal's ability to monitor agents. Hong et al. (2013) show that the introduction of a piece rate in Chinese factories increased quantity but reduced quality significantly. Barankay (2012) studies the effect of rank feedback in a randomized field experiment among furniture salespeople who sell products from different firms. A move by one firm to abandon rank feedback increased sales at this firm, as the feedback had shifted attention away from products that generated negative feedback. In another study with a large European agricultural producer, Englmaier et al. (2016) investigate how the salience of quantity incentives influences performance of harvesting teams. A higher salience of quantity indeed raises quantity, but quality is negatively affected. All of these studies thus show that agents indeed trade-off incentives for different tasks. Our paper builds on these insights, and our formal model illustrates that subjective performance evaluations create a further trade-off as not only agents allocate their efforts but supervisors have to decide how to allocate their limited attention for evaluating the different tasks. This creates specific patterns in the agent's reaction when performance is measured subjectively rather than objectively, which we investigate empirically using the data from the natural field experiment.

The paper proceeds as follows. We first present the setting of the field experiment in section 2. An illustrative formal model of subjective evaluations is analyzed in section 3. Section 4 presents more details on the field experiment and the results of the econometric analysis, and section 5 concludes.

## 2 The Firm Setting

The experiment was conducted by a retail bank with over 250 branches in Germany.[3] Staff at each branch consisted of a branch manager and a team of employees.[4] The branch employees are the bank's sales representatives. Their main jobs were to serve clients by performing administrative tasks at the counter and to sell the bank's products to private customers. In principle, every branch employee was trained and capable of selling products from each category to the customers[5]. Potential new customers were brought into the branches in several ways. First, a central marketing department initiated sales campaigns (e.g., direct mailings, a company website, promotion campaigns). In addition, a central call center organized sales appointments in the different branches.[6] Furthermore, the bank's branch employees themselves could on their own initiative call current customers to make appointments in the branches. At the time of the experiment the bank sold products in the following key categories: loans, investment products, saving plans with building societies, and credit cards.[7] Given the strategy of the bank, loans to private customers were the most important product category in terms of sales revenues, profits per transaction, and overall profits.

The compensation system prior to the intervention consisted of a fixed monthly salary and an additional bonus, which was based on financial targets and paid out quarterly. Branch performance was assessed using a profit measure called "customer net revenue" (henceforth CNR), which tracked

[3] We were not part of the project team ourselves and the bank conducted the experiment at its own inititiative as will be explained below.

[4] In addition to the branch employees, the bank contracts independent sales representatives (mobile sales force), who we exclude from the analysis. Although they are associated with a specific branch they are self-employed and face different compensation conditions.

[5] Thus employees are allrounders rather than specialists. This is due to the fact that business with private customers has a limited scope and the clients' needs are relatively homogenous. Being able to sell all kinds of products to the customers is even more necessary in smaller branches where there is less scope for specialization.

[6] E.g., the call center covered calls generated by direct mailing campaigns. The resulting sales appointments in the branches initiated by the call center were therefore completely independent of efforts in the branches.

[7] The bank also sold insurance products. However, these are not part of this analysis as the available data on insurance sales is incomplete.

the profit contribution of each product in each branch. When the respective target was met, a bonus pool was paid out to the branch, and it was the branch manager's responsibility to allocate the bonus to the employees in the branch.[8] The size of the bonus pool depended on the number of employees in the branch. Above the CNR target the bonus pool was linear in the achieved CNR. Each branch had a single branch manager. Hence, branch size also directly measures the span of control of this manager.

The bank carried out the experiment because it wanted to introduce objective performance meaures and was in negotiations with its works council[9] about the consequences for employees. Prior to the treatment intervention and during the experiment in the control group, the branch managers did not have access to objective performance indicators on sales by indivdiual employees and, hence, the allocation of the bonus was based only on their subjective assessment of the employees' performance.[10] During the intervention in 23 branches, branch managers in the treatment group had access to the CNR measures for all individual employees in their branch. They were told to use this information to set sales targets to individual employees and allocate the bonus pool accordingly based on realized profits. Branch managers had to fill out individual forms (on paper) noting target profits as well as realized profits of each employee and employees were also informed

[8] The average bonus was about 3.5% of a branch employee's quarterly salary. However, when targets were exceeded bonuses could raise significantly above this average level.

[9] In Germany, employees have a right to set up employee-elected works councils in establishments with more than 5 employees. Firms need the consent of works councils when implementing policies to evaluate employees' performance (This is due to § 87 (1) No 6, 10 and 11 of the Works Constitution Act (BetrVG), which specifies, for instance, that "*The works council shall have a right of co-determination in [..] the introduction and use of technical devices designed to monitor the behavior or performance of the employees.*"). In the negotiations, the firm and works council agreed to first run a "pilot" experiment to analyze effects of a change in the way performance is assessed.

[10] In principle, managers could try to collect pieces of objective information also before the intervention (as is probably the case whenever there is subjective performance evaluation). For instance, they could look up written contracts, but could not access the computer software. A member of the project team reported, that some branch managers took notes and used these for the assessment. However, it was a very time-consuming process and typically not done systematically. Note that the formal framework developed below could also be reinterpreted as modeling a supervisor who subjectively decides to collect noisy pieces of objective performance information.

about both.

# 3 A Conceptual Framework

## 3.1 The Model

To investigate the implications of introducing access to objective performance measures in an environment where supervisors a priori had to assess performance subjectively, we analyze a formal model which builds on the framework introduced by Prendergast and Topel (1996) or Prendergast (2002) and extends it to a multitasking setting endogenizing the quality of signals a supervisor gets when allocating attention on multiple tasks.

There is a group of $n$ agents $i = 1, .., n$ whose performance is to be evaluated by one supervisor, such that $n$ measures the supervsior's span of control. As in Holmström and Milgrom (1991) agents can be risk averse with constant absolute risk aversion $r$. They work on a set $J$ of tasks $j = 1, ..., m$. Each agent $i$ exerts effort $e_{ij}$ on task $j$ with a cost function $c(e_{i1}, e_{i2}, ...e_{im})$. For each agent and each task, there is a performance outcome $\pi_{ij} = e_{ij} + a_{ij}$ where the $a_{ij} \sim N\left(\mu_{ij}, \sigma_a^2\right)$ are independently distributed random variables.[11] The performance outcomes of all tasks of an agent $i$ generate a profit for the firm, which is equal to

$$\Pi_i = \sum_{j=1}^{m} b_j \cdot \pi_{ij}$$

such that $b_j$ describes the importance of task $j$ for the firm. Without loss of generality, the tasks are ranked according to profitability such that $b_1 > b_2 > ... > b_m$.

We compare two appraisal regimes, one in which objective performance

[11] The assumption that the $a_{ij}$ are independent is made for analytical tractability. However, one way of reinterpreting the model is to assume that $a_{ij} = \kappa_i + \lambda_j + \zeta_{ij}$ where $\kappa_i$ is the ability of the agent, $\lambda_j$ is a business cycle effect for the product category and $\zeta_{ij}$ is an ideosyncratic shock. If the supervisor can now observe $k_i$ (because she knows the subordinate $i$) and $\lambda_j$ (because she can observe product sales for $j$ across the nation), we can define $\mu_{ij} = k_i + \lambda_j$.

measures are available and one in which performance is assessed subjectively by a supervisor who endogenously collects information on the agents' profit contributions. In both cases, the agents receive a wage that is linear in their (estimated) profit contributions, such that $w_i = \alpha + \beta \cdot \tilde{\Pi}_i$ where $\tilde{\Pi}_i$ is the supervisor's assessment of the agent's profit contribution. When objective performance measures are available, the supervisor directly observes $\Pi_i$ before evaluating the employee.

We can apply the standard result on linear agency models with normally distributed noise and constant absolute risk aversion (see Holmström and Milgrom (1991), or, for instance (Wolfstetter, 2002, p. 347) for a proof) to show that an agent's certainty equivalent is

$$\alpha + \beta \cdot \tilde{\Pi}_i - c\left(e_{i1}, e_{i2}, ...e_{im}\right) - \frac{1}{2} r \beta^2 V\left[\tilde{\Pi}_i\right].$$

Agents maximize this certainty equivalent taking the supervisor's evaluation strategy into account.

When there are no objective performance measures, a supervisor $S$ has to evaluate the performance by collecting pieces of information and aggregating these pieces (which we together define as "subjective performance evaluation"). After the agents have exerted their efforts, the supervisor monitors them by collecting signals on their performance $\pi_{ij}$ for the different tasks.[12] The quality of each signal depends on the time the supervisor spends on monitoring each agent and task. Let $t_{ij}$ be the time spent on the performance of agent $i$ for task $j$. The supervisor has an overall time budget $T$ that she can allocate to the different tasks and agents, such that

$$\sum_{i=1}^{n} \sum_{j=1}^{m} t_{ij} = T.$$

By spending time on monitoring a task, the supervisor collects increasingly

[12] We here assume that supervisors decide on the allocation of attention ex-post or equivalently that agents cannot observe the attention spent on monitoring but infer the supervisor's equilibrium choices. However, when costs are quadratic and additively separable, the results in Propositions 1 and 2 also hold when supervisors first allocate attention and employees then choose their efforts knowing the attention spent on each task.

precise information on the true performance outcome $\pi_{ij}$ for this task.

In each unit of time $\tau$ the supervisor observes a signal $\eta_{ij\tau} = \pi_{ij} + \varepsilon_{ij\tau}$ where the $\varepsilon_{ij\tau}$ are iid and $\varepsilon_{ij\tau} \sim N\left(0, \sigma_{\varepsilon}^2\right)$. Hence, when investing time $t_{ij}$ the supervisor observes a vector $\eta_{ij} \in \mathbb{R}^{t_{ij}}$ of signals. Note that the mean of the observed signals $s_{ij} = \frac{1}{t_{ij}} \sum_{\tau=1}^{t_i} \eta_{ij\tau}$ is a sufficient statistic for $\pi_{ij}$ and – given equilibrium efforts $e_{ij}^*$ – this "observed performance" $s_{ij}$ is normally distributed with

$$E\left[s_{ij}\right] = \mu_{ij} + e_{ij}^* \text{ and } V\left[s_{ij}\right] = \sigma_a^2 + \frac{1}{t_{ij}} \cdot \sigma_{\varepsilon}^2. \tag{1}$$

Hence, we can reinterpret the model as one where a supervisor "subjectively" decides about the collection and aggregation of noisy pieces of objective information.[13] We solve the supervisor's decision problem by treating the $t_{ij}$ as continuous variables.[14]

The supervisor's task is to assess the agents' profit contributions. Following an approach used, for instance, by Prendergast and Topel (1996) or Prendergast (2002) to model subjective evaluations, we assume that the supervisor cares for profits and the accuracy of ratings, and her expected utility is

$$\sum_{i=1}^{n} E\left[\kappa \cdot \Pi_i - \left(\tilde{\Pi}_i - \Pi_i\right)^2 \middle| s_{i1}, s_{i2}, ...s_{im}\right]$$

where $\tilde{\Pi}_i$ is the supervisor's assessment of $i$'s profit and $\Pi_i$ is $i$'s actual profit.[15] This assumption about the supervisor's preferences implies that she optimally reports her own conditional expectation about $\Pi_i$ given the observed signals.[16]

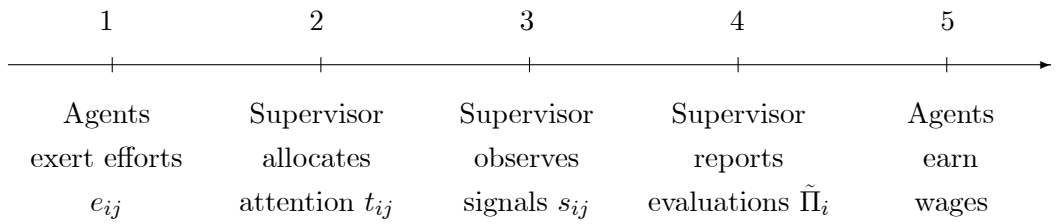
[13] Note that in nearly all cases where subjective evaluations are carried out, supervisors can probably collect some objective pieces of information and decide how to use them in their evaluations.

[14] As the supervisor's objective function will be strictly concave, the optimal discrete choice must be one of the nearest neighbors of the continuous optimum in the discrete grid. In the Appendix we also develop a continuous time interpretation of $s_{ij}$.

[15] Another interpretation is that the principal can verify the report with a certain probability and then imposes a fine $\left(\tilde{\Pi}_i - \Pi_i\right)^2$, where the marginal fine increases with the size of the deviation.

[16] For a proof see, for instance, Theorem 3.1.2 in (Angrist and Pischke, 2008, p. 33).

When the supervisor has access to objective performance information she will report $\tilde{\Pi}_i = \Pi_i$. Hence, the model is then a standard multitasking principal agent model. Without access to objective performance information, however, the timing is as follows:




The supervisor then faces the decision problem at the intermin stage of how to allocate the time budget $T$ to the different tasks and agents, in order to obtain the best estimate of the agents' profit contribution and minimize expected posterior deviations between reported and actual profit contributions. The agents, in turn, will anticipate this allocation of the monitoring intensity and choose their effort levels in order to maximize expected payoffs.

### 3.2 Performance Evaluation and the Allocation of Attention

When objective performance information is available the supervisor will accurately report profits $\Pi_i$. Under subjective evaluations, the supervisor ex-post reports her conditional expectation about $\Pi_i$, which is equal to the least squares estimator of $\Pi_i$ based on the signals $s_i$, or

$$\tilde{\Pi}_i = E\left[\sum_{j=1}^{m} b_j \cdot \pi_{ij} \,\middle|\, s_{i1}, s_{i2}, ...s_{im}\right] = \sum_{j=1}^{m} b_j \cdot \frac{\sigma_\varepsilon^2\left(\mu_{ij} + e_{ij}^*\right) + t_{ij}\sigma_a^2 s_{ij}}{t_{ij}\sigma_a^2 + \sigma_\varepsilon^2} \quad (2)$$

where the latter follows from applying a standard result on the conditional expectation of normally distributed random variables (see, for instance, DeGroot (1970), pp. 169; details given in the Appendix), and the $e_{ij}^*$ are equilibrium effort levels. It is instructive to consider the "subjective" performance report $\tilde{\Pi}_i$ as given by (2). Note that, in particular, when there is only imprecise information available (i.e. high $\sigma_\varepsilon^2$ and low $t_{ij}$) subjective

assessments become "compressed" as $\tilde{\Pi}_i$ varies to a lesser extent with the performance information $s_{ij}$. This occurs because the higher the uncertainty about the true performance, the closer the optimal estimate of performance is to its prior expectation: If a supervisor knows that her assessment is noisy, she will rationally, and to a larger extent, attribute a deviation from prior expectations to errors of perception. As argued already in Prendergast and Topel (1996), this effect yields an economic rationale for the rating compression often observed in subjective assessments (sometimes called the "centrality bias"). In our model, rating compression can be avoided when the monitoring intensity $t_{ij}$ for the task is very large. But limited attention will typically preclude this.

We can now study the optimal allocation of attention. The supervisor's ex-ante expected disutility of misreporting is equal to

$$\sum_{i=1}^{n} E\left[\left(\sum_{j=1}^{m} b_j \cdot \left(\frac{\sigma_\varepsilon^2\left(\mu_{ij}+e_{ij}^*\right)+\sigma_a^2 t_{ij}\left(\pi_{ij}+\frac{1}{t_{ij}}\sum_{\tau=1}^{t_i}\varepsilon_{ij\tau}\right)}{\sigma_\varepsilon^2+t_{ij}\sigma_a^2}-\pi_{ij}\right)\right)^2\right]$$

which (after some rerrangement – see Appendix) simplifies to

$$\sum_{i=1}^{n}\sum_{j=1}^{m} b_j^2 \frac{\sigma_\varepsilon^2\sigma_a^2}{\sigma_\varepsilon^2+t_{ij}\sigma_a^2}.$$

Note that for each task $j$ of each agent $i$ this is a decreasing and convex function of $t_{ij}$. Hence, there are decreasing marginal returns on allocated attention for each task – the more time a manager has spent on collecting information about the performance in a task, the less informative additional signals are. The manager thus optimally allocates attention by balancing the marginal returns under the time constraint.

Using this expression we can characterize the optimal allocation of time spent on assessing the different tasks and obtain the following result:

**Proposition 1** *The supervisor allocates his attention on the $\bar{m}$ most productive tasks, i.e., $t_j > 0$ for $j \leq \bar{m}$. The degree of attention spent on a task*

*j for each agent i is equal to*

$$t_j = \begin{cases} \frac{b_j}{\sum_{j' \leq \bar{m}} b'_j} \left( \frac{T}{n} + \bar{m} \cdot \frac{\sigma_\varepsilon^2}{\sigma_a^2} \right) - \frac{\sigma_\varepsilon^2}{\sigma_a^2} & if \quad j \leq \bar{m} \\ 0 & if \quad j > \bar{m} \end{cases} . \tag{3}$$

*The least productive task that is still monitored $\bar{m}$ is the smallest j for which*

$$\frac{\sigma_\varepsilon^2}{\frac{T}{n}\sigma_a^2 + j \cdot \sigma_\varepsilon^2} \sum_{j'=1}^{j} b_j > b_{j+1}. \tag{4}$$

Hence, the supervisor monitors the $\bar{m}$ most important tasks such that a number of low productivity tasks may get no attention at all. If a task gets some attention, then the time spent on monitoring a task is a function of the "relative productivity share" of this task $b_j / \sum_{j' \leq \bar{m}} b'_j$. To understand this result, recall that the supervisor wants to accurately assess the overall profit contribution. As more productive tasks contribute more to overall profits, the supervisor will invest the most attention assessing these tasks. If productivity differences are sufficiently large, the marginal gains from the last unit of monitoring a productive task may well exceed the marginal gains from starting to monitor a less productive task.

Based on equation (2) this implies that ratings will be more accurate (i.e., closer to the true performance outcome) when a task is more profitable. This may be interpreted as an economic rationale for what psychologists have called the "halo" effect in subjective performance evaluations, according to which there is a tendency for evaluators to excessively focus on very salient characteristics.[17]

Finally, note that all tasks will receive some attention when $T$ is sufficiently large. But, even in this case, more attention is spent on the most profitable tasks.

[17] Note that there are various definitions of "halo effects", and its importance is controversially discussed (see Murphy et al. (1993)).

### 3.3 Incentives

Proposition 1 has a number of implications regarding the potential benefits of objective performance measurement. It is instructive to start by comparing marginal returns to effort under both evaluation regimes. Under objective performance measurement, the principal observes profits. Additional signals therefore do not affect reported profits and the supervisor reports $\Pi_i$. Each agent $i$'s expected wage is $E[\alpha + \beta \cdot \Pi_i]$ such that he maximizes

$$\alpha + \beta \cdot \left( \sum_{j=1}^{m} b_j \left( e_{ij} + \mu_{ij} \right) \right) - c\left( e_{i1}, e_{i2}, ...e_{im} \right) - \frac{1}{2} r \beta^2 V\left[ \Pi_i \right].$$

and, as $V[\Pi_i]$ does not depend on effort choices, the marginal returns to effort for a task $j$ are equal to

$$\beta b_j \tag{5}$$

When the cost function is additively separable, this directly yields the optimal effort levels when objective performance measures are available

$$e_{ij} = c_j'^{-1}\left( \beta b_j \right).$$

Under subjective evaluation we can substitute the supervisor's optimal report $\tilde{\Pi}_i$ given by (2), and an agent $i$'s objective function is

$$\alpha + \beta \cdot E\left[ \sum_{j=1}^{m} b_j \frac{\sigma_\varepsilon^2 \left( \mu_{ij} + e_{ij}^* \right) + t_{ij} \sigma_a^2 s_{ij}}{t_{ij} \sigma_a^2 + \sigma_\varepsilon^2} \right] - c\left( e_{i1}, e_{i2}, ...e_{im} \right) - \frac{1}{2} r \beta^2 V\left[ \tilde{\Pi}_i \right].$$

Again, efforts do not affect $V\left[\tilde{\Pi}_i\right]$ such that we can determine the optimal effort choices by maximizing the difference between expected income and costs of effort. As $E[s_{ij}] = e_{ij} + \mu_{ij}$ marginal returns to effort for task $j$ are equal to

$$\beta b_j \frac{t_j \sigma_a^2}{t_j \sigma_a^2 + \sigma_\varepsilon^2}.$$

By inserting the optimal monitoring choices, rearranging terms and com-

paring to marginal returns of efforts under objective measurement (5), we obtain the following result:

**Proposition 2** *When no objective performance measures are available, the agents' marginal returns from each monitored task $j$ are equal to*

$$\beta b_j \left( 1 - \frac{\sigma_\varepsilon^2}{\frac{b_j}{\sum_{j' \leq \bar{m}} b_{j'}} \left( \frac{T}{n} \sigma_a^2 + \bar{m} \cdot \sigma_\varepsilon^2 \right)} \right) \quad \textit{for } \forall \; j \leq \bar{m}.$$

*When the agents' cost functions are additively separable, then efforts for all tasks are strictly lower than under objective performance evaluation and are decreasing with the span of control $n$. The relative loss in incentives is then larger the lower the relative profitability $b_j / \sum_{j' \leq m} b_{j'}$ of a task.*

As already noted in the above, limited attention under subjective performance evaluations here leads to less differentiated assessments that depend to a weaker extent on observed signals. In turn, marginal returns of effort decrease.

Moreover, the result shows that the strength of this distortion is affected by the span of control $n$. A supervisor who has to monitor a larger number of agents can spend less time on each agent and thus has less precise information on individual performance. In turn, reported ratings will be less differentiated and thus incentives will be lower-powered.[18]

Proposition 2 also shows that the degree of the incentive distortion will depend upon the importance of the tasks. Ratings will put a higher weight on the outcomes of more profitable tasks (interpreted as "halo effect" in the above). The economic consequences in our model are seen in the bias of efforts towards the most important tasks.[19] Agents do not only work less

[18] Note that it can be worthwhile to use several supervisors if the team is large. In the bank studied in the field experiment, however, there was also always only one branch manager in the larger branches.

[19] To make this statement more precise, it is useful to refer to the literature on multitasking incentives: Under objective performance measurement an owner receives a performance signal that is perfectly *congruent* in the sense defined by Feltham and Xie (1994). (Baker (2002) calls such a measure *undistorted* and Schnedler (2008) *aligned*): Under risk neutrality the first-best can be implemented with a simple linear contract based on the

for less profitable tasks because of their lower profitability (this is also the case when performance is measured objectively), but they also work less on these tasks because less profitable tasks are monitored less intensively by supervisors with limited capacities for attention.

If the effort costs are additively separable and equal to $\sum_{j=1}^{m} c_{ij}(e_{ij})$ with $c'_{ij}(e_{ij}) > 0$ and $c''_{ij}(e_{ij}) > 0$, the effects on marginal returns directly translate into effects on equilibrium efforts as efforts are simply a monotonic function $c'^{-1}_{ij}(.)$ of the marginal returns:

**Corollary 1** *If effort costs are additively separable, the use of objective performance measures leads to higher efforts for each task.*

In a slight reinterpretation of the model, this result also applies to a situation in which there is a complete division of labor such that each agent works on exactly one task.[20] In this case there is no interdependence in the costs between the different tasks and, again, efforts for all tasks indeed should be lower when performance is measured subjectively.

However, if there are interdependencies between the tasks with respect to the agents' disutility of effort and if there is no divison of labor, this is no longer clear. In our field experiment we also consider smaller branches, where there is no division of labor. Hence, it is important to also investigate non-separable cost functions. Consider, for instance, the two task case. Under objective performance measurement an internal solution is characterized by

$$\begin{aligned} \beta b_1 - \partial c / \partial e_{i1}(e_{i1}, e_{i2}) &= 0 \\ \beta b_2 - \partial c / \partial e_{i2}(e_{i1}, e_{i2}) &= 0. \end{aligned} \tag{6}$$

When $\frac{\partial^2 c}{\partial e_{i1} \partial e_{i2}} > 0$, higher incentives for task 2 (i.e., a higher $b_2$) will reduce efforts for task 1, reflecting the well known interdependence result from the multitasking literature (see Holmström and Milgrom (1991)).[21] Under

signal. Under subjective evaluations the owner does not receive such a signal from the supervisor: Any contract based on the subjective report that implements a first-best effort for task 1 necessesarily implements less than the first-best for all other tasks.

[20] To see that set $i = 1$ and replace worker 1 with a set of $m$ independent workers each with a separate cost function $c_{1j}(e_{1j})$.

[21] See Fehr and Schmidt (2004) for evidence from a lab experiment on this distortion.

subjective performance evaluation, in any internal solution efforts are characterized by

$$\begin{aligned} \beta b_1 \frac{\frac{T}{n}\sigma_a^2+\left(2-\frac{b_1+b_2}{b_1}\right)\sigma_\varepsilon^2}{\frac{T}{n}\sigma_a^2+2\sigma_\varepsilon^2} - \partial c/\partial e_{i1}\left(e_{i1},e_{i2}\right) &= 0 \\ \beta b_2 \frac{\frac{T}{n}\sigma_a^2+\left(2-\frac{b_1+b_2}{b_2}\right)\sigma_\varepsilon^2}{\frac{T}{n}\sigma_a^2+2\sigma_\varepsilon^2} - \partial c/\partial e_{i2}\left(e_{i1},e_{i2}\right) &= 0. \end{aligned} \tag{7}$$

As already laid out above, marginal returns for both tasks are smaller under subjective performance evaluation. Conversely, a move to objective performance evaluation increases incentives to a stronger extent for task 2, as here the downward distortion under subjective evaluation is larger. This, in turn, can affect the optimal efforts for task 1. The size of this effect depends upon the cross derivative of the cost function (or in economic terms, the degree to which working more on one of the tasks affects the marginal effort costs of the other task). To study this, consider the following simple cost function with a constant cross derivative

$$C\left(e_1,e_2\right)=\frac{c_1}{2}e_1^2+\frac{c_2}{2}e_2^2+c_{12}e_1e_2$$

with $c_1,c_2,c_{12}>0$ and $c_1c_2-c_{12}^2>0$, such that the function is strictly convex. Here $c_{12}$ captures the degree of cost substitution between the tasks, such that increasing effort for task 1 increases the marginal costs of effort for task 2 by $c_{12}$ (and vice versa). By comparing efforts under objective and subjective performance evaluation we now obtain:

**Corollary 2** *When there is no separation of labor and $c_{12}>c_2$, then efforts for task* 1 *decrease, when switching from subjective to objective performance measurement.*

Hence, even though marginal incentives for the most important task 1 increase, efforts for this task can decrease when objective performance measures become available. The gain in marginal incentives is larger for the less important task 2 and, if the substitutability ($c_{12}$) is large, this leads to

a shift in effort away from task 1 towards task 2.[22]

While we caution that the field experiment cannot test all predictions of the model directly such as, for instance, the specific biases in the subjective performance evaluation, the developed framework yields implications for patterns associated with the availability of objective performance measures that we can study in the field setting:

(i) When interdependencies between tasks are weak, the availability of objective performance measures should increase output.

(ii) Output increases to a stronger extent when there are larger spans of control.

(iii) Outputs increase to a weaker extent for the more important tasks or may even decrease when interdependencies are strong and there is no division of labor.

### 3.4 Discussion of Key Assumptions

The assumptions on the supervisor's preferences and her choice problem need some discussion as the real world decision problem considered in the experiment is of course richer in nature. A first important assumption is that supervisors care for the accuracy of ratings as in Prendergast and Topel (1996) or Prendergast (2002). A more general interpretation of (and justification for) using the expected squared deviations between reported and true performance is that it models a supervisor who is motivated to report her best estimate, i.e., her conditional expectation about profits. This has several reasons: First, supervisors are not residual claimants (i.e. owners of the firm). They themselves receive a bonus based on the financial performance of the branch but this bonus constitutes a small share of profits. Second, a key aim of the evaluation procedures – even before the intervention in the firm we study, but also in many other firms – is to give employees feedback

[22] To be more precise, if $c_{12} > c_2$, then increasing effort for task 1 by one unit has a stronger impact on the marginal costs of effort for an additional unit of task 2 than increasing effort for task 2, i.e., task 1 has high opportunity costs due to a strong externality on task 2 (Note that $c_{12} > c_2$ implies that $c_1 > c_2$ as $c_1 c_2 > c_{12}^2$.). An increase in "relative incentives" in favor of task 2 is then accompanied by a shift of effort away from task 1 in order to reduce the marginal costs of effort for the now more attractive task 2.

on their contribution to the firm's success. The preference for accuracy can thus be viewed as a preference to give appropriate feedback on an agent's contribution. While it thus seems reasonable that accuracy of ratings is indeed a key element of supervisors' objectives, other motives such as social preferences will most certainly also play a role in the real world setting. We here have abstracted away from social preferences to illustrate the mere informational aspects in the allocation of attention. But social prefereces can be incorporated in the setting without affecting the key trade-offs we have illustrated. As shown by Prendergast and Topel (1996), in this framework simple linear altruism will lead to an upward shift in ratings and as shown by Kampkötter and Sliwka (2018), inequity aversion will lead to a further compression of ratings.

Moreover, note that in the field setting, supervisors allocate a bonus pool that depends on profits while in the model the supervisor assesses the agents' profit contributions and agents' pay is a linear function of these assessments. Note here that the model can be transformed to an equivalent one in which the supervisor decides on payments $P_i$ to the agents and agents earn wages $\alpha + P_i$. Such a model is equivalent to our setting when substituting $\tilde{\Pi}_i = P_i/\beta$. The underlying equivalent preference assumption is then that supervisors would like to minimize $(P_i/\beta - \Pi_i)^2$ or equivalently $(P_i - \beta\Pi_i)^2$. If now $\beta$ is the share of profits that is paid into the pool by the firm, this assumption implies that supervisors aim at allocating payments such that they proportionaly represent contributions of individual agents to the creation of the pool. Hence, we can think of the model as illustrating preferences of supervisors who follow a contribution-based fairness norm and thus strive for apportioning bonus payments proportional to (estimated) contributions of the individual employees.

Moreover, in the field setting the payout scheme for the branch is piecewise linear, i.e. the size of the bonus pool is increasing in branch profits only if these profits exceed a certain cut-off value. Our model abstracts away from such a threshold. For reasons of analytical tractability we also did not impose the restriction that reported profits have to be equal to realized profits (this will hold in our model only in expected terms). Kampkötter

and Sliwka (2018) study a single task variant of the model where such an assumption is imposed. This induces a related effect that adds to the benefits of objective performance information in larger teams: Namely, when the size of the overall bonus pool depends on the joint performance of the group, then subjective assessments lead to more free-riding which will be more severe in larger teams.[23]

A further effect that can be of importance in the field setting is that the observability of performance in itself can affect employees' well being and, in turn, their performance. If, for instance, an agent's reputation is affected by her performance either because of career concerns (as in Holmström (1999)) or image concerns (as in Bénabou and Tirole (2006)) the observability of objective performance information may produce an incentive effect even without a material benefit. However, note that such reputational concerns can lead to the same payoff-structure as in our model if an agent's reputation is a linear function of her (expected) profit contribution.[24]

# 4 The Experiment

## 4.1 Data and Procedures

Recall that in both the treatment and control groups, employees could receive a bonus in addition to their fixed salaries, which was based on quarterly financial targets for each branch measured in Euro CNR. This key indicator was used throughout the bank for evaluating performance. It is important to stress that CNR is a profit measure, i.e., it tracks the net sur-

[23] See, for instance, Proposition 3 in Kampkötter and Sliwka (2018). Intuitively, even without differentiation according to individual performance, agents then have an interest in maximizing the size of the bonus pool. But when the span of control is larger this effect is smaller due to free-riding.

[24] If we assume that agents' reputational utility is a linear function of the supervisor's conditional expectation of her profit contribution the model is completely identical to the one we analyze (except that $\beta$ has a different interpretation). If, however, reputational utility is a function of the supervisor's conditional expectation of some underlying ability parameter, this would add an additional benefit of objective performance measurement as actual profits would be a more precise signal of ability than estimated profits under limited attention.

plus that the bank's sales organization generates from selling the products. We use monthly data from the bank's branches (>250) for the year 2003 (Jan-Dec).[25]

The bank picked 23 branches, which were assigned to the treatment group using a stratified sampling method to ensure that they were representative in terms of size, performance, and geographical distribution. The procedure was such that members of the project team formed groups of branches along three criteria namely profit per FTE (full time eqivalents)[26], growth over three years and number of FTE. Within these groups the treatment branches were drawn randomly. As a means of control the allocation to treatment and control group was approved by the Works Council that had the right to object when suspecting any irregularities.[27] The authors were not part of the project team and therefore did not witness the process personally. However, we have no reason whatsoever to doubt the declarations given by the bank.

Table 1: Summary Statistics (Average Months 1-3)

| | Control | | Treatment | | | |
|---|---|---|---|---|---|---|
| | mean | sd | mean | sd | Difference | p-Value |
| Employees (in FTE) | 7.88 | 2.75 | 7.93 | 2.33 | -0.05 | 0.933 |
| Appointments | 248.52 | 96.36 | 207.48 | 81.31 | 41.04 | 0.049 |
| Appointments Call Center | 123.47 | 88.82 | 151.45 | 110.92 | -27.98 | 0.158 |
| CNR loans (norm.) | 91.90 | 34.67 | 94.09 | 37.29 | -2.39 | 0.772 |
| CNR investment (norm.) | 5.65 | 3.55 | 4.90 | 3.05 | 13.67 | 0.306 |
| CNR savings (norm.) | .78 | .47 | .66 | 59.80 | .47 | 0.231 |
| CNR crd. cards (norm.) | 1.67 | 1.01 | 1.78 | .92 | -6.57 | 0.617 |
| CNR total (norm.) | 100.00 | 37.65 | 101.43 | 40.21 | -1.43 | 0.862 |

For reasons of confidentiality profit measures are normalized at the mean total CNR of the control group before the intervention.

The number of branches in the treatment group was around 10% of all German branches to limit the workload and economic risks associated with

[25] We cannot reveal the exact number of branches for reasons of confidentiality.

[26] Part time working employees are counted by the fraction of their contractual working time relative to the working time of a full time employee.

[27] Note that works councils in Germany have quite powerful inspection rights.

Table 2: Structure of the field experiment

| Year 2003 | month 1-6 | month 7-12 |
|---|---|---|
| Treatment (23 branches) | subjective assessment | **objective assessment** |
| Control (>250 branches) | subjective assessment | subjective assessment |

the experiment. Descriptive statistics and p-values for comparisons between the treatment and the control group prior to the intervention (i.e., in month 1-3) are reported in Table 1.[28]

We do not find any significant differences for any of the profit variables or the number of employees per branch between treatment and control. The number of self-initiated appointments, however, is significantly smaller in the treatment group prior to the intervention. But as depicted in Figure 1 this difference in levels does not come along with a difference in time trends as time trends are very similar between treatment and control group.

Branches in the treatment group were informed about the new system two months prior to the intervention (i.e., in month 5 and 6). Workshops were conducted with the branch managers in the treatment group to inform them about the mechanics of the new system, the way in which the objective key figures would be made available and how they could be handled. Employees in the other branches were also informed about the fact that a new assessment system was tested in a subset of branches.

Table 2 shows the structure of the experiment. From January to June 2003, purely subjective assessments were used in all branches. The intervention ran from July to December 2003 in 23 branches. After the experiment had expired, the tested system was implemented throughout the bank and is still applied in a similar way today.

Finally, we acknowledge that we do not have access to bonus payments made to indivdiual employees but all information is measured at the level of an individual branch. Hence, we can study the effect of the intervention on customer appointments and profits in the different product categories but cannot analyze the appraisal strategies of individual branch managers.

[28] For reasons of confidentiality, we normalized Euro values by dividing profits in each product category by the average total profits in the control group prior to the intervention.

### 4.2 Employee Effort: Customer Appointments

In the following we analyze the effects of the treatment intervention on the different available outcome variables. Recall that we have more than 250 branches altogether, with 23 branches in the treatment group. We start with the most direct available key figure for employee effort – the number of self-initiated sales appointments.[29] In each month a database tracked how many appointments were arranged by the employees in a branch by actively calling up customers and inviting them to visit the branch. Calling customers is the most direct way in which an employee can try to raise his sales and hence his financial performance.

Figure 1 shows the development of the number of customer appointments per full time equivalent employee (FTE) over time for the treatment and control group, normalized at the average of the four months prior to the intervention.[30]

We analyze the causal effect of the intervention by estimating fixed effects models, with the treatment intervention as the key independent variable and report robust standard errors clustered on branch level. That is, we estimate models of the form

$$Appoint_{bm} = \beta \cdot Treatment_{bm} + \delta_b + \gamma_m + \varepsilon_{bm}$$

where the dummy $Treatment_{bm} = 1$ when a branch $b$ belongs to the treatment group, and the observation is from a month $m$ where the treatment is in place, $\delta_b$ are branch and $\gamma_m$ month fixed effects. The results are reported in Table 3.

There are two points in time after which the treatment group is af-

[29] We will later show (see subsection 4.5) that appointments are indeed strongly associated with sales - but their impact is different for different products.

[30] As mentioned above the branches in the treatment group on average made a lower number of appointments prior to the intervention (see Table 1). This could lead to the concern that the observed effect is a catching up/mean reversion phenomenon. However, mean reversion should rather lead to a continuous increase in the treatment group over time, and it seems unlikely that it explains the jump in May.

Note that the peak in July is a seasonal effect as due to the summer holidays July tends to be a profitable month.

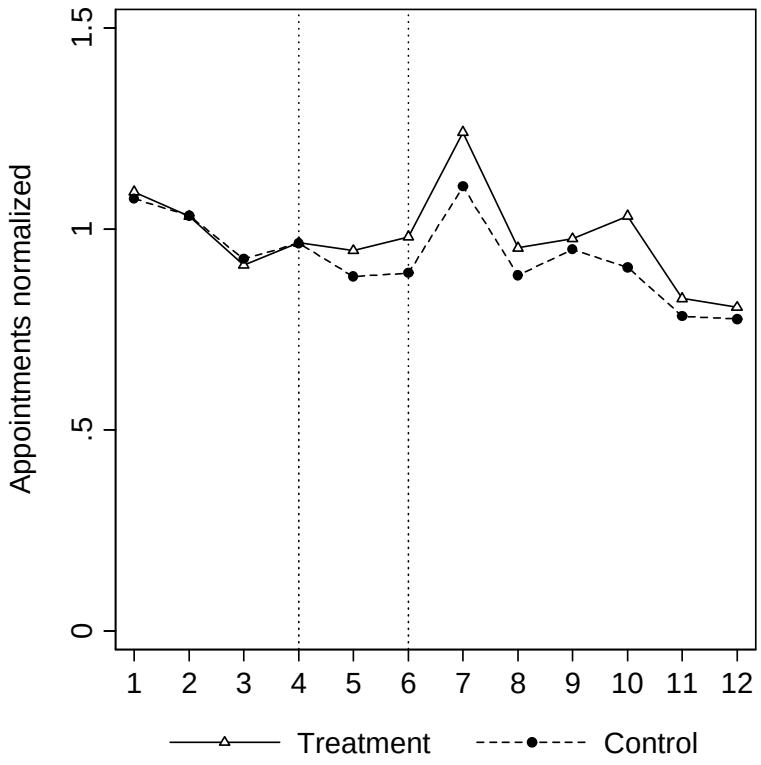



Figure 1: Appointments over time

fected differently by the treatment as compared to the control group. Before month 5, employees in the treatment group were informed that objective performance measures would be used starting with month 7. Hence, it is conceivable that the mere announcement of individual measurement affects performance. As customer appointments typically predate actual product sales, employees may already at an earlier date have an incentive to increase the number of appointments. In additon, we therefore include in column (2) a dummy "Information" for the months 5 and 6, in which employees in the treatment group were already informed about the new system, but it was not yet in place.

Table 3: The impact of objective measurement on self-initiated appointments

| | *Appointments* | |
|---|---|---|
| | (1) | (2) |
| Treatment (month 7-12) | 15.07* | 24.80*** |
| | (7.958) | (8.931) |
| Information (month 5-6) | | 29.19*** |
| | | (7.659) |
| Constant | 258.6*** | 258.6*** |
| | (2.935) | (2.925) |
| $R^2$ | 0.323 | 0.326 |

The dummy "Treatment" takes value 1 in the treatment branches in the months 7-12, and the dummy "Information" takes value 1 in the treatment branches in months 5 and 6, i.e. after the announcement of the intervention but before it is in place. Branch fixed effects and month dummies included. Robust standard errors clusted on branch level, *** p<0.01, ** p<0.05, * p<0.1

Considering the estimates in column (1), the number of self-initiated sales appointments increases by about 6% relative to month 1 through 6. But as column (2) reveals, the treatment effect already starts in month 5. The mere announcement of the treatment intervention increases the number of monthly appointments by about 11%. When the treatment is in place the effect remains stable at roughly 10% relative to the months prior to the announcement. Hence, the introduction of objective performance measures indeed leads to an increase in the number of self-initiated customer appointments, which was already observable as soon as the information on the intervention was provided.

### 4.3 Financial Performance

But how does the treatment affect financial performance? To study this we analyze the profit measure CNR separately by single product categories and aggregated over all categories. Figure 2 shows the development of profits by product category over time, normalized at the mean of the respective

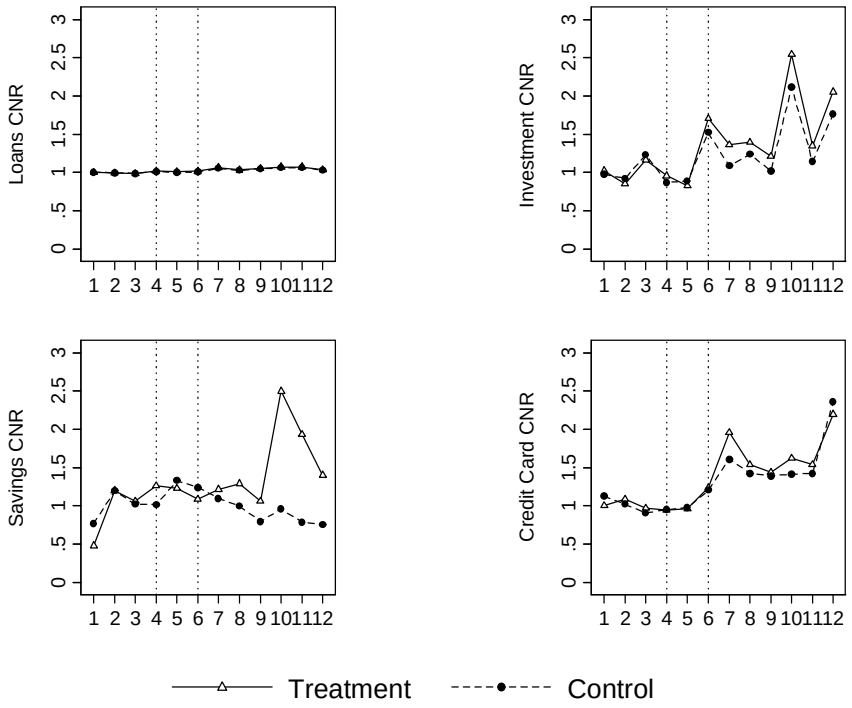



Figure 2: Normalized profits ("Customer Net Revenue") by product category

product category in the four months before the intervention.[31]

Table 4 (upper panel) reports fixed effects regressions, with the (log) profits of the different product categories in columns (1) to (4) and overall profits in column (5) as dependent variables. As can be seen in column (1) of the upper panel of the table, the treatment has no significant impact on consumer loans, the bank's key product, but it does have substantial effects on the other products, which were of lower importance before the intervention. For instance, profits from investment products increased by roughly 18%, those from building society savings by even[32] 42% and of

[31] For reasons of confidentiality all profit measures are normalized as pecentages of the mean total profit of the control group prior to the intervention.

[32] As can be seen in Figure 2, the very high effect for building society savings is

credit cards (albeit weakly significantly) by 10%. As loans were still the predominant product after the intervention, the effect of the treatment on overall profits is at about 2%. Note that, in particular as the intervention did not increase costs, this effect is sizeable when comparing it to outcomes from other recent field experiments in retailing (for instance Friebel et al. (2017) find that the introduction of a (costly) team bonus raised sales in a bakery chain by about 3%).[33]

As Table A1 in the Appendix shows, the results are also robust when we run a simple diff-in-diff regression without month and branch fixed effects and introduce these fixed effects step by step.

Recall that the core product is consumer loans, and the bank estimates that profits per transaction are on average roughly five times larger for a loan transaction as compared to, for instance, the sale of an investment product. The model laid out in section 3 suggests the interpretation that supervisors kept track of this core product even under subjective evaluation ("halo effect" in subjective assessments). Since the non-core products had only a weak share in the overall profitability, these products were not the key focus of their attention. Objective measurement now provided precise information about these minor products at no cost. Hence, employees had

to a large part driven by a boost in sales in the treatment group starting in October. In September 2003 the German Federal Government published a draft law (http://dip21.bundestag.de/dip21/btd/15/015/1501502.pdf) that intended to abolish a subsidy for new savings plans ("Wohnungsbauprämie") and was supposed to enter into effect on January 1, 2004. Later, the subsidy was only reduced and not abolished, but the discussion led to a boost in sales in 2003. Apparently, employees in the treatment branches made much stronger use of this event to raise sales of these savings plans.

[33] From the perspective of statistical power, having 23 treated branches is of course a limiting factor of the data as power would be maximized with a more balanced size of treatment and control group (see, for instance, List et al. (2011)). Hence, it is worthwhile to consider also confidence intervals of the estimates which we display in Figure A2 in the Appendix. The Figure shows the 95%, 90%, 80% and 70% confidence bands respectively for the treatment effects estimated by the regressions reported in Table 4. The confidence bands are indeed wide (in particular for credit card sales as well as savings plans (for which we have expanded the x-axis)). But also note that for investment products the lower boundary of the 95% confidence band is at 9.9%, or for savings plans it at 6% which still would constitute economically meaningful effects.

Table 4: The impact of objective measurement on profits

| | (1) *log CNR loans* | (2) *log CNR investment* | (3) *log CNR savings* | (4) *log CNR credit cards* | (5) *log CNR total* |
|---|---|---|---|---|---|
| Treatment | 0.00652 | 0.182*** | 0.416** | 0.0961* | 0.0200** |
| | (0.00906) | (0.0419) | (0.177) | (0.0552) | (0.00894) |
| $R^2$ | 0.308 | 0.382 | 0.086 | 0.381 | 0.396 |
| Treatment | 0.00936 | 0.194*** | 0.414** | 0.103 | 0.0240** |
| | (0.0109) | (0.0482) | (0.178) | (0.0660) | (0.0108) |
| Information | 0.00850 | 0.0374 | -0.00725 | 0.0210 | 0.0121 |
| | (0.00717) | (0.0571) | (0.193) | (0.0622) | (0.00779) |
| $R^2$ | 0.309 | 0.382 | 0.086 | 0.381 | 0.397 |

The upper panel shows regressions with profits per product category as dependent variables and the treatment dummy as the key independent variable. The lower panel reports regressions that additionally include the dummy "Information" that takes value 1 in the treatment branches in months 5 and 6, i.e., after the announcement of the treatment but before it is in place. Branch fixed-effects, month dummies and call center initiated customer appointments included. Robust standard errors clustered on branches; *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

an incentive to exert substantially more effort on these product categories.[34]

As we have seen above, employees increased their efforts directly following the announcement of the treatment. Hence, it is interesting to see whether financial performance already increases at that point. However, when we again include a dummy for the months after the announcement but before the treatment (see Table 4 lower panel), we find no significant effect of the announcement for any of the product categories. The coefficients of the treatment remain virtually unchanged (the point estimate for the effect on overall profits increases to 2.4%, and the effect on credit card sales, while increasing somewhat in size, is no longer significant). Hence, employees apparently start to prepare for the new system by increasing appointments after the announcement and even before the system is in place, but product sales increase significantly only after they are measured objectively. We explore this issue in more detail when we investigate potential strategic timing of sales in section (4.5).

Note that this observation is also useful for discussing the potential concern that the announcement of the treatment changed behavior through a different channel, namely by sending a signal to the employees that the bank expects them to exert more effort on other (non-loan) products. However, we observe the increase in profits for these products not after the announcement but only when the actual treatment is in place – indicating that it is the measurability of output that affects the incentives to generate profits in the different product categories.

### 4.4 Heterogeneous Treatment Effects: Branch Size

To study further heterogeneous treatment effects, we look more closely at the role of branch size. This is measured by the bank in Full Time Equivalent Employees (FTE), i.e., part time working employees are counted by the fraction of their contractual working time relative to the working time of a

[34] An alternative interpretation of the results is that the sale of loans is less elastic in terms of sales agents' efforts. However, as we show in section 4.4, the intervention leads to a decrease in loan sales in relatively small branches (see Table 6) which would be inconsistent with such an interpretations. Moreover, in section 4.5 we show that branch-initiated appointments predict loan sales (see Table 7).

Table 5: The role of branch size

| | *log CNR total* | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Treatment | 0.0191** | 0.0187* | 0.0368** |
| | (0.00871) | (0.0105) | (0.0146) |
| Treatment x branch size (centered) | 0.00456* | | |
| | (0.00262) | | |
| Treatment x 20% smallest branches | | -0.0343** | |
| | | (0.0144) | |
| Treatment x 20% largest branches | | 0.0329** | |
| | | (0.0140) | |
| Treatment x 30% smallest branches | | | -0.0459*** |
| | | | (0.0163) |
| Treatment x 30% largest branches | | | -0.0175 |
| | | | (0.0182) |
| $R^2$ | 0.397 | 0.399 | 0.398 |

Branch size is the number of (full time equivalent) employees in a branch, centered at the mean. Branch fixed-effects, month dummies and call center initiated appointments included. Robust standard errors clustered on branches, *** p<0.01, ** p<0.05, * p<0.1.

full time employee. Descriptive statistics for treatment and control groups are given in Table 1 and the distributions of branch sizes are shown in Figure A3 in the Appendix.[35] To study whether and how the treatment effect depends on branch size, we interact the treatment variable with the number of full time equivalent employees. Table 5 reports the respective regression results. Model (1) just includes an interaction term with the branch size (centered at the mean) to our baseline specification. Column (2) instead includes dummies for the 20% smallest ($\leq$ 5.5 FTE) and for the 20% largest ($>$ 10 FTE) branches. Analogously, we consider the interaction effect for the 30% smallest ($\leq$ 6 FTE) and the 30% largest ($>$ 8.9 FTE) branches, which are displayed in column (3).

[35] The number of employees in the branches varies slightly over time. As branch size may in principle be endogenously affected by the intervention, we use the branch size from the last month prior to the announcement of the intervention (i.e., month 3) in all regressions.

As the results show, the treatment effect depends on the size of the branch. In the largest 20% of the branches, overall profits increase by more than 5%, whereas there is virtually no effect in the small branches.[36] This is confirmed in model 3, which shows that the intervention does not lead to significant net effects in branches with fewer than 6 employees ($30th$ percentile). The number of observations with fewer than 5.5 or 6 FTE within the treatment group is naturally rather small. In order to rule out that the size effect is driven by a particular branch, we therefore also performed "leave-one-out" robustness checks where we ran the regressions each time leaving out one of the small branches in the treatment group. The results remain robust in all of these specifications.

A straightforward interpretation of this size effect, in light of the model presented above, posits that the larger the branch, the harder it is for a supervisor to keep track of the performance of the employees. Larger branches thus benefit substantially more from the use of objective performance measurement. But beyond the mechanisms illustrated in our model, this effect could also be due to more free-riding in larger branches when no individual objective performance measures are available.[37] Finally, another conceivable explanation is that the intervention freed up time branch managers formerly used to assess performance and thus have more time to do other things such as own sales activities – an effect that should be stronger in larger branches.

While we cannot rule out these alternative explanations entirely, a more detailed analysis reveals a slightly more complex pattern explaining these size effects that is in line with multitasking distortions illustrated in our formal model. To see this pattern consider Table 6, in which we split the sample and report separate regressions for the smallest branches, with 6 or fewer full time equivalent employees (upper panel), and those with more than 6 (lower panel)[38]. The columns (1) through (4) again show the results

[36] Testing the net effect of *Treatment* and *Treatment x 20% smallest branches* shows that it is not significantly different from zero.

[37] Note that as we discuss in the above, the effect is not captured in our model, as we do not impose the assumption that the bonus pool corresponds to a function of true profits (it does so only in expected terms in our model). But in the field setting this could clearly be relevant.

[38] As Figure A3 in the Appendix shows, the distribution of FTE is not symmetric and

Table 6: Treatment effects in small and larger branches

| | (1) *log CNR loans* | (2) *log CNR investment* | (3) *log CNR savings* | (4) *log CNR credit cards* | (5) *log CNR total* |
|---|---|---|---|---|---|
| *Small branches ($\leq$ 6 full time equivalent employees)* | | | | | |
| Treatment | -0.0275** | 0.279*** | 0.0188 | 0.0302 | -0.00469 |
| | (0.0111) | (0.0603) | (0.270) | (0.109) | (0.0124) |
| $R^2$ | 0.174 | 0.351 | 0.090 | 0.309 | 0.234 |
| *Large branches ($>$ 6 full time equivalent employees)* | | | | | |
| Treatment | 0.0159 | 0.152*** | 0.517** | 0.112* | 0.0262*** |
| | (0.00976) | (0.0498) | (0.206) | (0.0627) | (0.00984) |
| $R^2$ | 0.452 | 0.403 | 0.098 | 0.438 | 0.534 |

Separate regressions for smaller (upper panel) and larger (lower panel) branches. Branch fixed-effects, month dummies and call center initiated appointments included. Robust standard errors clustered on branches, *** p<0.01, ** p<0.05, * p<0.1.

for the different product categories and column (5) the effects on overall profits. While the results for the large branches are well in line with the previous observations, results for the small branches reveal an interesting pattern. Even though profits from investment products increased by almost 28% in these branches, here the intervention had a pronounced *negative* effect on loans, the sales of which significantly decrease by about 2.75%.

Hence, in the small branches the intervention caused a shift from the bank's core product loans to the former fringe investment products. This effect is absent in the larger branches. While this negative effect of having more information may seems surprising at first glance, the formal model suggests an explanation. The model predicts that performance should increase for all products only if there is a weak interdependency between the tasks

the majority of branches have more than 6 FTE. In the treatment group only 5 branches belong to the group of smaller branches.

or if there is a division of labor. Indeed, in larger branches different employees tend to specialize in different product categories. In smaller branches specialization is harder to achieve because sales agents have to serve all customers. As the model illustrates, this directly leads to incentive conflicts. A stronger monitoring of "fringe" tasks can indeed lead to a reallocation of efforts, as time spent convincing customers to buy investment products cannot be used to sell loans.[39] If, however, different agents are responsible for the different tasks, this effect should not occur since a change in the incentive structure for one agent does not affect the behavior of other agents.

The result is also informative in a further respect: above we suggested a lower effort elasticity for loans as one possible interpretation of the absence of a positive average performance effect on loans. However, if this were the case, we should not observe a reduction of loan sales in the small branches. Hence, it seems more likely that the stability of loan sales is indeed driven by the fact that loans (which generated more than 90% of profits and also a substantially higher profit per unit of transaction) were already monitored closely before the intervention.[40]

### 4.5 The Role of Appointments

We now take a closer look at the role of appointments initiated by the branch employees. The regression results in section 4.1 have shown that appointments already increase in the two months prior to the intervention. We argued that the mere announcement might have led to this increase because sales deals are concluded during or subsequent to appointments and therefore appointments should precede future profits. To better understand the linkage between appointments and profits, we now study the correlation between self-initiated and call-center initiated appointments,[41]

[39] Note that the model does not predict that total profits should decrease, but rather small profit increases in smaller branches are in line with the model.

[40] Indeed, as described by company representatives, sales agents could actively approach customers to sell loans. For instance, they had access to a software that could predict the likelihood of taking a loan based on customer records. This way they could identify promising customers belonging to this target group and contact them directly.

[41] Recall that a central call center organized sales appointments on behalf of the branches and that this call center was external; consequently these appointments were not influenced

and later profits across the different product categories. We estimate log-log regression models for all variables of interest in order to be able to interpret the results as elasticities (see Table 7). Core independent variables are the number of appointments in the same month and the month before the profits are realized. Columns (1)-(4) show the outcomes for the separate product categories, and column (5) displays the results for total profits. We use these regressions not to make causal statements but rather to uncover patterns in the data.

First, note that appointments are mostly associated with immediate sales in the same month: with the exception of savings for building societies, the highest elasticities are observed in the same month. Hence, generally there does not seem to be a strong time lag between appointments and sales. But, the estimates illustrate that there are pronounced differences in the way self-initiated appointments predict profits, depending on the product in focus. There is a significant but fairly small correlation for loans: a 10% increase in appointments corresponds to a 0.03% growth in loan sales in the same month. This indicates that loans are more rarely sold via self-initiated appointments. However, the association is rather strong for investment products: 10% more self-initiated appointments in the current month are associated with 1.94% higher profits in the same month and 1.3% in the next month. The short-term correlation is even stronger for credit cards, with an elasticity of 2.4%. Savings with building societes are inelastic in appointments in the same month, but the elasticity is rather large for lagged appointments. Hence, products differ in (i) the extent to which appointments matter for profits and (ii) the time lag between the appointment and the profit realization.

The fact that self-initiated appointments increased even before the intervention was in place, but profits only afterwards, suggests that employees used some personal leeway in timing the realization of profits. Indeed, a company representative stated that influencing the timing of sales is possible in particular for investment products. Consider for instance the following situation: A sales agent notices that a customer has a large amount of money

by the branch employees.

Table 7: The association between appointments and profits

| | (1) *log CNR loans* | (2) *log CNR investment* | (3) *log CNR savings* | (4) *log CNR credit cards* | (5) *log CNR total* |
|---|---|---|---|---|---|
| log Appoint.$_t$ | 0.0292*** | 0.194*** | -0.0139 | 0.240*** | 0.0416*** |
| | (0.00813) | (0.0508) | (0.0762) | (0.0399) | (0.00863) |
| log Appoint.$_{t-1}$ | 0.00797 | 0.126*** | 0.254*** | 0.0321 | 0.0196** |
| | (0.00687) | (0.0481) | (0.0890) | (0.0437) | (0.00896) |
| $R^2$ | 0.396 | 0.410 | 0.090 | 0.415 | 0.479 |

log Appoint$_t$ is the logarithm of the number of self-initiated appointments in month $t$. Branch fixed-effects, month dummies and call center appointments in $t$ and $t-1$ included. Robust standard errors clustered on branches, *** p<0.01, ** p<0.05, * p<0.1.

in a checking account. He may then decide to contact this customer, invite him to an appointment in the branch and to advise him on how to improve his financial portfolio to yield higher returns. The outcome of such a meeting could be that the customer buys an investment product, e.g., shares in mutual fund. The sales agent may now have some leeway in determining the date when the actual purchase takes place. Knowing that sales are tracked in July, he may well have incentives to make appointments for June but carry out the sales in July. Hence, we might observe that elasticities of lagged sales are different in July for the treatment group.[42] We explore this conjecture in the following analysis.

In order to investigate this conjecture, we study whether there are differences in the predictive power of lagged appointments for profits in the different product categories across months. To do this, we run OLS regressions of the following form

$$\begin{aligned} x_{bm} &= \alpha + \beta_m \cdot Appoint_{bm} + \gamma_m \cdot Appoint_{bm-1} \\ &\quad + \delta_m \cdot Appoint_{bm-1} \times TreatmentBranch_b \\ &\quad + \eta_m \cdot TreatmentBranch_b + \varepsilon_{bm} \end{aligned}$$

[42] See, for instance, Oyer (1998) or Larkin (2014) for previous empirical studies on the effect of incentive schemes on gaming in the timing of realized performance measures.

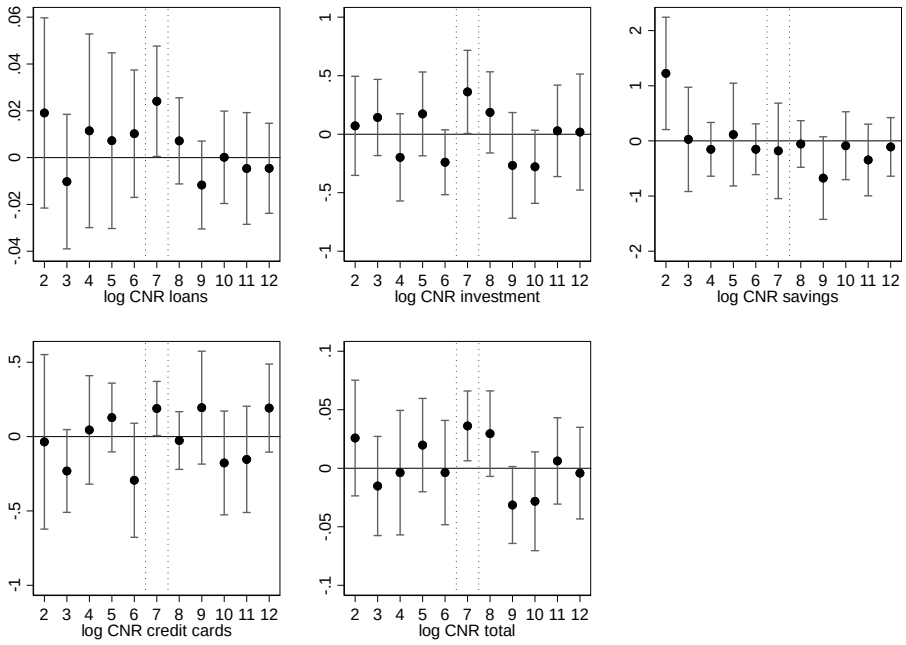



Figure 3: Treatment differences by month in predictive power of lagged appointments

separately for each month $m = 2, ..., 12$ controlling for (log) branch size, call center appointments as well as lagged profits ($x_{bm-1}$) in the respective product category. Our coefficient of interest is $\delta_m$, which estimates whether there is a difference in the predictive power of lagged appointments between treatment and control branches in the considered month $m$. Figure 3 plots the coefficients $\delta_m$ as well as their 90% confidence bands for each regression.While there is quite some noise, elasticities with regard to past appointments indeed tend to be larger in July in the treated branches in four out of the five categories. The interaction term $Appoint_{bm-1} \times TreatmentBranch_b$ is positive and at least weakly significant in the July regressions for loans ($p = 0.093$), investment products ($p = 0.094$), credit cards ($p = 0.088$) and total profits ($p = 0.046$). In these categories the interaction term is not significantly different from zero for all other months. For savings plans,

however, July does not differ from the other months in this regard.[43]

We repeated this excercise to see whether there are also structural differences in the predictive power of appointments in the same month (i.e. replacing the interaction term with $Appoint_{bm} \times TreatmentBranch_b$). If employees shift June sales to July we may observe that appointments in June should be less predictive for sales in June for the treatment group.

Here we see a significant negative difference only for credit cards ($p = 0.024$). The point estimate is also negative for investment products but this is not significantly different from zero ($p = 0.182$). We do not see a negative difference for loans and total profits. A potential explanation for the weaker reduction in the "conversion rate" of current appointments in June could be that employees strategically made additional appointments with relatively profitable customers at the end of June, thus raising sales in July without lowering the "conversion rate" of appointments in June for sales in June substantially.[44]

In section 4.3 we have seen that the size of the branch has a substantial impact on the financial returns from objective performance measurement. We gave an explanation suggesting the existence of structural differences between smaller and larger branches in the way the work is processed. The conclusion was that the lack of division of labor leads to more multitasking problems in smaller branches. Here it is also instructive to investigate whether the role of appointments differs according to branch size. The respective regression results are shown in Table A2 in the Appendix. We find that employees in larger branches and branches of intermediate size do not significantly increase the number of self-initiated appointments under the treatment intervention. However, those in the smallest 30% of the branches increase appointments by almost 19%. In light of our estimates of

[43] For savings plans for building societies the only significant difference is in February, for which we have no specific explanation. A possible conjecture is that savings plans need longer to process (recall that we saw in the above that here appointments in the same month are not predictive for sales). We checked whether here, $Appoint_{bm-2}$ may predict differently in July, but this is not the case.

[44] We thank an anonymous referee for suggesting the analysis of sales in June as well as for pointing out this explanation.

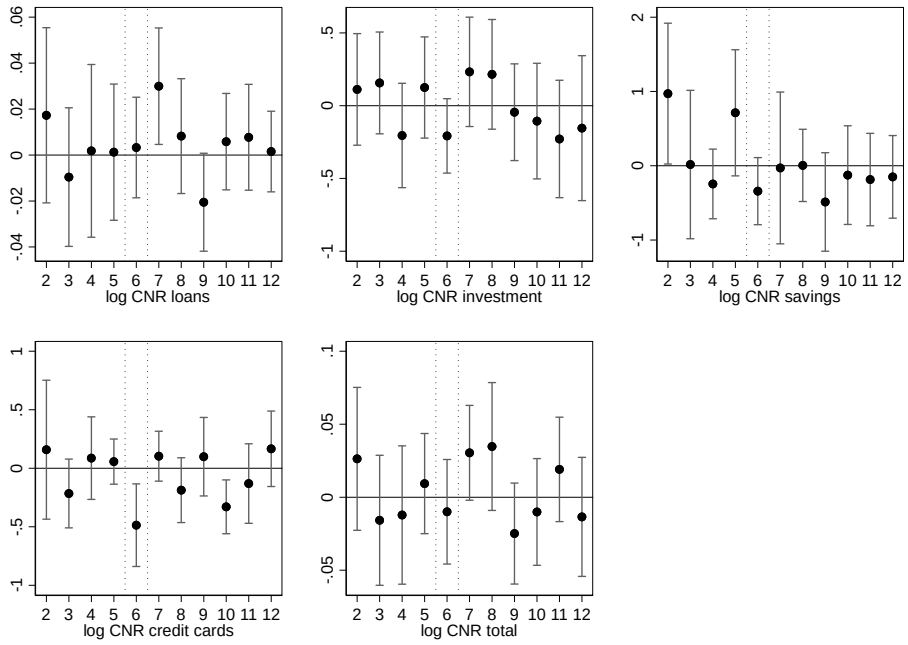



Figure 4: Treatment differences by month in predictive power of appointments in the same month

the treatment effects on profits of small branches (reported in Table 5) and the appointment elasticity of sales, this supports the conclusion that employees of smaller branches shift attention towards selling investment products. The sale of these products can actively be triggered by (time-consuming) appointments, which reduces the time available to care for customers visiting the branches on their own initiative, for instance, to obtain a loan. The question remains why there is no significant effect on appointments in larger branches.[45] In our view, the most likely interpretation is again the stronger division of labor, in particular with respect to investment products. To understand this, first note that sales of investment products increase to a weaker extent in larger branches (by 15% instead of 28% in the smaller ones

[45] Note that while the point estimate for the treatment effect in the "middle" 60% of the branches is not significantly different from zero, it is still 6% in Table A2.

- see Table 6). When some sales agents specialize in investment products, the model suggests that they should indeed sell more of these products as the marginal returns to their efforts increase. This effect, however, should be weaker than for less specialized agents in the smaller branches, who can also shift efforts from the core product loans towards investment products. Hence, the additional time spent on appointments to sell investment products can be larger in smaller branches with less specialized agents.

### 4.6 Persistence

Finally, it is important to study the persistence of the effects. It is conceivable that the availablity of objective performance measures causes a "Hawthorne" effect and leads to a short-term increase in efforts, and thus profits, which then fall back towards the initial level. In order to check for this, we again consider the number of self-initiated appointments and the total CNR. The results of our analysis are reported in Table 8.

Columns (1) and (2) of Table 8 show fixed-effects regression results with dummies "Information" and "Treatment" and the interaction of "Treatment" and the months since the start of the new system.[46] We do not find an effect for the time passed since the introduction of the regime, neither for the number of appointments nor for the total CNR. Another specification reported in columns (3) and (4) interacts quarter-dummies with the treatment and also shows that in the last quarter there is a (weakly) significant positive performance effect of the treatment intervention that is very close to that of quarter 3. Thus, at least for the six months after the introduction, we do not find evidence for the occurrence of a Hawthorne effect.[47]

[46] Recall that, as in the models reported above, the dummy "Information" takes value 1 for branches in the treatment group in month 5 and 6 when employees in the treatment group had already been informed about the new system but it was not yet in place.

[47] We also split the sample into small and large branches as in section 4.3, now adding quarter interactions to the regressions to see whether the size-dependent patterns also remain stable over time. The results are reported in Table A3 in the appendix. Interestingly, the effort shift from loans to investment products in the small branches seems to even get stronger over time. For larger branches the results are rather stable. The positive impact of the intervention only diminishes for credit cards in the fourth quarter, which may indicate a saturation of demand.

Table 8: Persistence of treatment effect over time

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | *Appoint.* | *log CNR* total | *Appoint.* | *log CNR* total |
| Information | 29.19*** | 0.0118 | 29.19*** | 0.0122 |
| | (7.660) | (0.00785) | (7.660) | (0.00785) |
| Treatment | 24.82** | 0.0266*** | | |
| | (10.17) | (0.0101) | | |
| Treatment x | -0.00935 | -0.00139 | | |
| months since start | (2.001) | (0.00217) | | |
| Treatment x Q3 | | | 22.46** | 0.0223** |
| | | | (9.367) | (0.0102) |
| Treatment x Q4 | | | 27.13*** | 0.0238* |
| | | | (9.760) | (0.0122) |
| Constant | 258.6*** | 5.613*** | 258.6*** | 5.613*** |
| | (2.926) | (0.00418) | (2.926) | (0.00418) |
| $R^2$ | 0.326 | 0.394 | 0.326 | 0.394 |

"Month since start" takes value 1 in August in the treatment branches and then increases by 1 in each month in these branches; it is zero for all control branches and in the months prior to the intervention. Q3 and Q4 are dummy variables indicating whether an observation is from the third and fourth quarter of the year. Branch fixed-effects, month dummies and call center initiated appointments included. Robust standard errors clustered on branches, *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Of course, we cannot rule out completely that the intervention did not have potentially detrimental longer term effects.[48] But we note that the bank tracked cancellation rates and the system was rolled out to all branches and is still in place more than 14 years after the experiment.

# 5 Conclusion

We study the interplay between subjective performance evaluations and multitasking incentives, analzying a formal economic model and data from a natural field experiment in a bank. First, we found that giving supervisors access to a comprehensive set of objective performance indicators in the bank increased not only employee efforts, as measured by self-initiated customer appointments, but also the financial success of the treated branches. It is notable that these profit increases came at virtually no direct costs for the bank because the individual performance measures were tracked by existing software systems. As a result, the bank rolled out the system to all branches one year after the experiment.

More importantly, the natural experiment allows us to study incentive distortions caused by subjective evaluations in multitasking environments. Our formal model shows that a direct extension of a standard framework, whereby the supervisor's information acquisition is endogenized in a multitasking setting, naturally leads to specific distortions that carry direct consequences for incentives. If a supervisor is interested in assessing profit contributions accurately but has a limited capacity to monitor different agents and tasks: (i) subjective evalutions lead to lower powered incentives as compared to a situation where precise objective performance indicators are available; (ii) this incentive loss increases with the span of control; and, (iii) incentives are biased towards the more profitable tasks as these receive more attention from supervisors.

An empirical analysis of the data from the natural experiment provides

[48] It is at least in principle conceivable that the intervention, for instance, led sales agents to sell investment products with risk-return profiles which may have reduced customer well-being and thus long term profits.

evidence in line with these patterns. A key observation is that the overall performance effect is driven by larger branches, which benefited strongly from the availability of objective performance measures, increasing profits by about 5%. The performance gains are entirely due to higher powered incentives for products that were previously not within the main focus of supervisors. While the bank's main product, loans (where profits per transaction are substantially larger than for the other products) did hardly benefit, there were large performance gains for the other more fringe product categories, such as an 18% increase in the sales of investment products. In smaller branches, where there is no division of labor, the use of objective performance measures even shifted sales efforts away from loans to investment products and significantly reduced loan sales.

While the paper shows that granting access to objective performance information raises profits, we caution that this does not necessarily imply that objective performance measurement always dominates subjective assessments. First of all, supervisors in the experiment had access to objective performance measures but they still had leeway to include other sources of information. Hence, the experiment does not show that using objectives measures instead of subjective assessment raised profits but that *granting access* to this information was beneficial. Second, in our experiment supervisors in the treated group had access to a rather comprehensive set of objective performance measures. It is well conceivable that the use of a subset of this information may have led to other distortions.

We believe that the experiment provides insights for the design subjective performance evaluations, which are very pervasive in real-world organizations. To the best of our knowledge, the natural field experiment is among the very few experiments on incentives that cover white collar workers in regular jobs. Apart from that, a treatment intervention is tracked for half a year and thus much longer than in most previous (lab and field) experiments on incentive design.

Subjective performance evaluations are among the most controversially discussed HR practices, and firms are continuously struggling to improve

their appraisal processes.[49] Among the key issues discussed in the debate are the costs and benefits of different appraisal processes. The results of our study therefore yield insights that may help to improve the design of performance evaluations in practice. First of all, obtaining a comprehensive set of objective performance measures can indeed raise performance. Second, the returns to objective performance measures should be stronger when there are larger spans of control. Third, the benefits from collecting balanced, objective measures should be larger when there are stronger asymmetries in the importance of tasks.

Finally, several large firms such as Microsoft, Yahoo, Accenture, or Bosch have recently very prominently announced dramatic changes in their appraisal processes.[50] The fact that so many companies are redesigning their appraisal systems every couple of years indicates that there is likely still substantial uncertainty about the optimal design of these systems. As the example of the retail bank we study shows, firms could, instead of dramatically redesigning their system from time to time, start by varying the process in a subset of the organizational units thus learning more about the true causal effects of specific design elements for specific job types. If more and more firms follow such an example, the combined pieces of evidence should lead to much faster and more precise learning about how performance should be measured and rewarded in an optimal manner.

[49] Just compare articles in the popular press with headlines such as "The Push Against Performance Reviews" (New Yorker, July 24, 2015), "Study finds that basically every single person hates performance reviews" (Washington Post, January 27, 2014), or "Performance Reviews: Many Need Improvement" (New York Times, September 10, 2006).

[50] See, for instance, "Yahoo is ranking employees. When Microsoft did that, it was a disaster" (Washington Post, November 12, 2013), or "Accenture CEO explains why he's overhauling performance reviews" (https://www.accenture.com/ma-en/company-accenture-ceo-performance-review.aspx).

## References


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

Angrist, J. D. and J.-S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Baker, G. (2002). Distortion and risk in optimal incentive contracts. *Journal of human resources*, 728–751.

Baker, G., R. Gibbons, and K. J. Murphy (1994). Subjective performance measures in optimal incentive contracts. *Quarterly Journal of Economics 109*, 1125–56.

Bandiera, O., I. Barankay, and I. Rasul (2011). Field experiments with firms. *The Journal of Economic Perspectives 25*(3), 63–82.

Barankay, I. (2012). Rank incentives: Evidence from a randomized workplace experiment. *Working Paper Wharton School*.

Bénabou, R. and J. Tirole (2006). Incentives and prosocial behavior. *American Economic Review 96*(5), 1652–1678.

Bentley MacLeod, W. (2003). Optimal contracting with subjective evaluation. *The American Economic Review 93*(1), 216–240.

Berger, J., C. Harbring, and D. Sliwka (2013). Performance appraisals and the impact of forced distribution: An experimental investigation. *Management Science 59 (1)*, 54–68.

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

Bordalo, P., N. Gennaioli, and A. Shleifer (2012). Salience theory of choice under risk. *The Quarterly Journal of Economics 127*(3), 1243–1285.

Bretz, R. D. J., G. T. Milkovich, and W. Read (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management 18*, 321–352.

DeGroot, M. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

Delfgaauw, J. and M. Souverijn (2016). Biased supervision. *Journal of Economic Behavior & Organization 130*, 107–125.

Engellandt, A. and R. T. Riphahn (2011). Evidence on incentive effects of subjective performance evaluations. *Industrial and Labor Relations Review 64*, 241–257.

Englmaier, F., A. Roider, and U. Sunde (2016). The role of communication of performance schemes: Evidence from a field experiment. *Management Science*.

Fehr, E. and K. M. Schmidt (2004). Fairness and incentives in a multi-task principal–agent model. *The Scandinavian Journal of Economics 106*(3), 453–474.

Feltham, G. and J. Xie (1994). Performance measure congruity and diversity in multi-task Principal/Agent relations. *The Accounting Review 69*, 429–453.

Friebel, G., M. Heinz, M. Krüger, and N. Zubanov (2017). Team incentives and performance: Evidence from a retail chain. *Forthcoming: American Economic Review*.

Gibbs, M., K. A. Merchant, W. A. van der Stede, and M. E. Vargus (2003). Determinants and effects of subjectivity in incentives. *The Accounting Review 79*, 409–436.

Giebe, T. and O. Gürtler (2012). Optimal contracts for lenient supervisors. *Journal of Economic Behavior & Organization 81*(2), 403–420.

Golman, R. and S. Bhatia (2012). Performance evaluation inflation and compression. *Accounting, Organizations and Society 37*(8), 534–543.

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *Review of Economic Studies 66*, 169–182.

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

Hong, F., T. Hossain, J. A. List, and M. Tanaka (2013). Testing the theory of multitasking: Evidence from a natural field experiment in chinese factories. *NBER Working Paper No. 19660*.

Kampkötter, P. and D. Sliwka (2018). More dispersion, higher bonuses? on differentiation in subjective performance evaluations. *Journal of Labor Economics 36*(2), 511–549.

Keller, G. (2011). Brownian motion and normally distributed beliefs. Technical report, Working Paper, University of Oxford.

Kőszegi, B. and A. Szeidl (2013). A model of focusing in economic choice. *The Quarterly Journal of Economics 128*(1), 53–104.

Larkin, I. (2014). The cost of high-powered incentives: Employee gaming in enterprise software sales. *Journal of Labor Economics 32*(2), 199–227.

Levin, J. (2003). Relational incentive contracts. *The American Economic Review 93*(3), 835–857.

Levitt, S. D. and S. Neckermann (2014). What field experiments have and have not taught us about managing workers. *Oxford Review of Economic Policy 30*(4), 639–657.

List, J. A. and I. Rasul (2011). Field experiments in labor economics. *Handbook of labor economics 4*, 103–228.

List, J. A., S. Sadoff, and M. Wagner (2011). So you want to run an experiment, now what? some simple rules of thumb for optimal experimental design. *Experimental Economics 14*(4), 439.

Murphy, K. R. and J. N. Cleveland (1995). *Understanding Performance Appraisal*. Thousand Oaks: Sage.

Murphy, K. R., R. A. Jako, and R. L. Anhalt (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology 78*(2), 218.

Ockenfels, A., D. Sliwka, and P. Werner (2014). Bonus payments and reference point violations. *Management Science*.

Oyer, P. (1998). Fiscal year ends and nonlinear incentive contracts: The effect on business seasonality. *Quarterly Journal of Economics*, 149–185.

Prendergast, C. and R. Topel (1996). Favoritism in organizations. *Journal of Political Economy 104*, 958–978.

Prendergast, C. J. (1999). The provision of incentives in firms. *Journal of Economic Literature 37*, 7–63.

Prendergast, C. J. (2002). Uncertainty and incentives. *Journal of Labor Economics 20*, 115–37.

Schnedler, W. (2008). When is it foolish to reward for a while benefiting from b? *Journal of Labor Economics 26*(4), 595–619.

Takahashi, S., H. Owan, T. Tsuru, and K. Uehara (2014). Multitasking incentives and biases in subjective performance evaluation. Technical report, Institute of Economic Research, Hitotsubashi University.

Wolfstetter, E. (2002). *Topics in Microeconomics*. Cambridge University Press.

Zabojnik, J. (2014). Subjective evaluations with performance feedback. *The RAND Journal of Economics 45*(2), 341–369.

# 6 Appendix:

**Continuous time interpretation:**

Let $t_{ij} \in \mathbb{R}_0^+$ be the time spent on the performance of agent $i$ for task $j$. The supervisor again has an overall time budget $T$ that she can allocate on the different tasks and agents such that $\sum_{i=1}^{n}\sum_{j=1}^{m} t_{ij} = T$. The "observable" performance $S_{ij}(t)$ follows a Brownian Motion with drift $\pi_{it}$ and volatility $\sigma_\varepsilon$ such that $S_{ij}(t) = t\pi_{ij} + \sigma_\varepsilon \cdot W_t$ where $W_t$ is a standard Wiener process such that $W_t \sim N(0,t)$. Assume now that a manager investing time $t_{ij}$ observes the "average slope" of this process in a time frame $\tau_{ij}$

$$\begin{aligned} s_{ij}(t_{ij}) &= \frac{1}{t_{ij}} S_{ij}(t_{ij}) \\ &= \pi_{ij} + \frac{\sigma_\varepsilon \cdot W_{\tau_{ij}}}{t_{ij}}. \end{aligned}$$

Then $s_{ij}(\tau_{ij})$ is normally distributed with

$$E[s_{ij}(t_{ij})] = E[\pi_{ij}] = \mu_{ij} + e_{ij}^*$$

and

$$V[s_{ij}(t_{ij})] = \sigma_a^2 + \frac{1}{t_{ij}}\sigma_\varepsilon^2$$

which again yields (1). Note that the model is also consistent with the assumption that the manager observes the evolution $dS_{it}(t)$ over time and continuously updates beliefs about $\pi_{ij}$ in a Bayesion fashion, as analyzed in Keller (2011). The conditional expectation (2) then directly follows from expression (1) and (2) in Keller (2011).

**Conditional expectation on the profit contribution:**

$$\begin{aligned} \tilde{\pi}_{ij} &= E[\pi_{ij}|\, s_{ij}] = E\left[\pi_{ij}|\, \pi_{ij} + \frac{1}{t_{ij}}\sum_{\tau=1}^{t_i}\varepsilon_{ij\tau}\right] \\ &= \left(\mu_{ij} + e_{ij}^*\right) + \frac{Cov\left[\pi_{ij}, \pi_{ij} + \frac{1}{t_{ij}}\sum_{\tau=1}^{t_i}\varepsilon_{ij\tau}\right]}{V\left[\pi_{ij} + \frac{1}{t_{ij}}\sum_{\tau=1}^{t_i}\varepsilon_{ij\tau}\right]}\left(s_{ij} - \left(\mu_{ij} + e_{ij}^*\right)\right) \\ &= \frac{\sigma_\varepsilon^2\left(\mu_{ij} + e_{ij}^*\right) + t_{ij}\sigma_a^2 s_{ij}}{t_{ij}\sigma_a^2 + \sigma_\varepsilon^2} \end{aligned}$$

**Ex-ante expected disutility of misreporting:**

$$\sum_{i=1}^{n} E\left[\left(\sum_{j=1}^{m} b_j \cdot \left(\frac{\sigma_\varepsilon^2\left(\mu_{ij}+e_{ij}^*\right)+\sigma_a^2 t_{ij}\left(\pi_{ij}+\frac{1}{t_{ij}}\sum_{\tau=1}^{t_i}\varepsilon_{ij\tau}\right)}{\sigma_\varepsilon^2+t_{ij}\sigma_a^2}-\pi_{ij}\right)\right)^2\right]$$

Using $E\left[X^2\right]=V\left[X\right]+\left(E\left[X\right]\right)^2$ this is equivalent to

$$\begin{aligned}\sum_{i=1}^{n} \quad & \left[V\left[\sum_{j=1}^{m} b_j \cdot \left(\frac{\sigma_\varepsilon^2\left(\mu_{ij}+e_{ij}^*\right)+\sigma_a^2 t_{ij}\left(\pi_{ij}+\frac{1}{t_{ij}}\sum_{\tau=1}^{t_i}\varepsilon_{ij\tau}\right)}{\sigma_\varepsilon^2+t_{ij}\sigma_a^2}-\pi_{ij}\right)\right]+\right. \\ & \left.+E\left[\sum_{j=1}^{m} b_j \cdot \left(\frac{\sigma_\varepsilon^2\left(\mu_{ij}+e_{ij}^*\right)+\sigma_a^2 t_{ij}\left(\pi_{ij}+\frac{1}{t_{ij}}\sum_{\tau=1}^{t_i}\varepsilon_{ij\tau}\right)}{\sigma_\varepsilon^2+t_{ij}\sigma_a^2}-\pi_{ij}\right)\right]^2\right].\end{aligned}$$

But the expected value of the squared deviations is equal to zero as

$$\begin{aligned} & E\left[\sum_{j=1}^{m} b_j \cdot \left(\frac{\sigma_\varepsilon^2\left(\mu_{ij}+e_{ij}^*\right)+\sigma_a^2 t_{ij}\left(\pi_{ij}+\frac{1}{t_{ij}}\sum_{\tau=1}^{t_i}\varepsilon_{ij\tau}\right)}{\sigma_\varepsilon^2+t_{ij}\sigma_a^2}-\pi_{ij}\right)\right] \\ = & \sum_{j=1}^{m} b_j \cdot \left(\frac{\sigma_\varepsilon^2+\sigma_a^2 t_{ij}}{\sigma_\varepsilon^2+t_{ij}\sigma_a^2}\left(\mu_{ij}+e_{ij}^*\right)-\left(\mu_{ij}+e_{ij}^*\right)\right) \\ = & 0\end{aligned}$$

such that the disutility of misreporting is equal to :

$$\begin{aligned} & \sum_{i=1}^{n} V\left[\sum_{j=1}^{m} b_j \cdot \left(\frac{\sigma_\varepsilon^2\left(\mu_{ij}+e_{ij}^*\right)+\sigma_a^2 t_{ij}\left(\pi_{ij}+\frac{1}{t_{ij}}\sum_{\tau=1}^{t_i}\varepsilon_{ij\tau}\right)}{\sigma_\varepsilon^2+t_{ij}\sigma_a^2}-\pi_{ij}\right)\right] \\ = & \sum_{i=1}^{n}\sum_{j=1}^{m} b_j^2 V\left[\left(\frac{\sigma_a^2 t_{ij}-\sigma_\varepsilon^2-t_{ij}\sigma_a^2}{\sigma_\varepsilon^2+t_{ij}\sigma_a^2}\right)\pi_{ij}+\frac{\sigma_a^2}{\sigma_\varepsilon^2+t_{ij}\sigma_a^2}\left(\sum_{\tau=1}^{t_i}\varepsilon_{ij\tau}\right)\right] \\ = & \sum_{i=1}^{n}\sum_{j=1}^{m} b_j^2\left(\frac{\sigma_\varepsilon^4\sigma_a^2}{\left(\sigma_\varepsilon^2+t_{ij}\sigma_a^2\right)^2}+\frac{t_{ij}\sigma_\varepsilon^2\sigma_a^4}{\left(\sigma_\varepsilon^2+t_{ij}\sigma_a^2\right)^2}\right) \\ = & \sum_{i=1}^{n}\sum_{j=1}^{m} b_j^2\frac{\sigma_\varepsilon^2\sigma_a^2}{\sigma_\varepsilon^2+t_{ij}\sigma_a^2}\end{aligned}$$

**Proof of Proposition 1:**
First note that the marginal returns to attention must be equal for all tasks that receive positive attention in equilibrium, as otherwise the supervisor is better off shifting attention towards the task with higher marginal returns. Moreover, the marginal returns to attention are identical across all agents for a specific task $j$ such that $t_{ij} = t_j$ for all agents $i$. Hence, for any tasks $j$ and $j'$ that receive positive attention, their marginal returns must be identical which leads to the condition that

$$\frac{b_j}{b_{j'}} = \frac{\sigma_\varepsilon^2 + t_j \sigma_a^2}{\sigma_\varepsilon^2 + t_{j'} \sigma_a^2}. \tag{8}$$

which is equivalent to

$$t_{j'} = \frac{b_{j'}}{b_j}\left(\frac{\sigma_\varepsilon^2}{\sigma_a^2} + t_j\right) - \frac{\sigma_\varepsilon^2}{\sigma_a^2}.$$

Summing up this expression across the first $m$ tasks we obtain that the total time spend must be equal to

$$T = n \sum_{j' \leq m} \left(\frac{b_{j'}}{b_j}\left(\frac{\sigma_\varepsilon^2}{\sigma_a^2} + t_j\right) - \frac{\sigma_\varepsilon^2}{\sigma_a^2}\right)$$

$$\Leftrightarrow \left(\frac{\sigma_\varepsilon^2}{\sigma_a^2} + t_j\right)\frac{\sum_{j' \leq m} b'_j}{b_j} - m \cdot \frac{\sigma_\varepsilon^2}{\sigma_a^2} = \frac{T}{n}.$$

Solving for $t_j$ yields that for each task $j$ for which $t_j > 0$ we must have

$$t_j = \frac{b_j}{\sum_{j' \leq m} b'_j}\left(\frac{T}{n} + m \cdot \frac{\sigma_\varepsilon^2}{\sigma_a^2}\right) - \frac{\sigma_\varepsilon^2}{\sigma_a^2}. \tag{9}$$

Now, note that the marginal returns to attention for a task $j$ at $t_j = 0$ is equal to $b_j^2 \frac{\sigma_a^4}{\sigma_\varepsilon^2}$. As the objective function is strictly concave in each $t_j$ and because $b_{j+1} < b_j$ it must be the case that if $t_j = 0$ then $t_{j+1} = 0$. Suppose that $j$ is the last task that is actively monitored (i.e. $t_j > 0$ but $t_{j+1} = 0$). The marginal return of the last unit of monitoring task $j$ is then equal to

$$\frac{b_j^2 \sigma_a^4 \sigma_\varepsilon^2}{\left(\sigma_\varepsilon^2 + t_j \sigma_a^2\right)^2} = \frac{b_j^2 \sigma_a^4 \sigma_\varepsilon^2}{\left(\sigma_\varepsilon^2 + \left(\frac{b_j}{\sum_{j'=1}^{j} b'_j}\left(\frac{T}{n} + j \cdot \frac{\sigma_\varepsilon^2}{\sigma_a^2}\right) - \frac{\sigma_\varepsilon^2}{\sigma_a^2}\right)\sigma_a^2\right)^2} = \frac{\left(\sum_{j'=1}^{j} b'_j\right)^2 \sigma_a^4 \sigma_\varepsilon^2}{\left(\frac{T}{n}\sigma_a^2 + j \cdot \sigma_\varepsilon^2\right)^2}$$

Hence, it is not worthwhile spending time on monitoring task $j+1$ if

$$\frac{\left(\sum_{j'=1}^{j} b'_j\right)^2 \sigma_a^4 \sigma_\varepsilon^2}{\left(\frac{T}{n}\sigma_a^2 + j \cdot \sigma_\varepsilon^2\right)^2} > b_{j+1}^2 \frac{\sigma_a^4}{\sigma_\varepsilon^2}$$

which is equivalent to condition (4). ■

**Proof of Corollary 2:**
By substituting the first derivatives of the cost function $C(e_1, e_2) = \frac{c_1}{2}e_1^2 + \frac{c_2}{2}e_2^2 + c_{12}e_1e_2$ in (6) and solving for $(e_1, e_2)$ we obtain the optimal efforts under objective performance measurement

$$\begin{aligned} e_1^O &= \beta \frac{c_2 b_1 - c_{12} b_2}{c_1 c_2 - c_{12}^2}, \\ e_2^O &= \beta \frac{c_1 b_2 - c_{12} b_1}{c_1 c_2 - c_{12}^2}. \end{aligned}$$

By substituting in (6) we analogously obtain the efforts under subjective evaluation

$$\begin{aligned} e_1^{c_{12}} &= \frac{\beta}{c_1 c_2 - c_{12}^2}\left(c_2\left(\frac{b_1 \frac{T}{n}\sigma_a^2 + (b_1 - b_2)\sigma_\varepsilon^2}{\frac{T}{n}\sigma_a^2 + 2\sigma_\varepsilon^2}\right) - c_{12}\left(\frac{b_2 \frac{T}{n}\sigma_a^2 + (b_2 - b_1)\sigma_\varepsilon^2}{\frac{T}{n}\sigma_a^2 + 2\sigma_\varepsilon^2}\right)\right), \\ e_2^{c_{12}} &= \frac{\beta}{c_1 c_2 - c_{12}^2}\left(c_1\left(\frac{b_2 \frac{T}{n}\sigma_a^2 + (b_2 - b_1)\sigma_\varepsilon^2}{\frac{T}{n}\sigma_a^2 + 2\sigma_\varepsilon^2}\right) - c_{12}\left(\frac{b_1 \frac{T}{n}\sigma_a^2 + (b_1 - b_2)\sigma_\varepsilon^2}{\frac{T}{n}\sigma_a^2 + 2\sigma_\varepsilon^2}\right)\right). \end{aligned}$$

By comparing $e_1^{c_{12}}$ with $e_1^O$ and rearranging terms we obtain that $e_1^{c_{12}} > e_1^O$ iff $c_{12} > c_2$ which completes the proof. ■

Table A1: Diff-in-Diff Analysis

| | (1) *log CNR loans* | (2) *log CNR investment* | (3) *log CNR savings* | (4) *log CNR credit cards* | (5) *log CNR total* |
|---|---|---|---|---|---|
| *Diff-in-Diff estimation* | | | | | |
| Treatment | 0.00578 | 0.182*** | 0.391** | 0.0887 | 0.0191** |
| | (0.00908) | (0.0421) | (0.174) | (0.0548) | (0.00898) |
| Test phase | 0.0427*** | 0.208*** | -0.290*** | 0.393*** | 0.0577*** |
| | (0.00342) | (0.0175) | (0.0317) | (0.0194) | (0.00363) |
| Test branch | 0.0220 | -0.106 | -0.299 | 0.0900 | 0.0137 |
| | (0.0914) | (0.134) | (0.186) | (0.143) | (0.0911) |
| $R^2$ | 0.003 | 0.024 | 0.026 | 0.084 | 0.006 |
| *Month fixed effects* | | | | | |
| Treatment | 0.00578 | 0.181*** | 0.394** | 0.0887 | 0.0191** |
| | (0.00910) | (0.0422) | (0.176) | (0.0549) | (0.00900) |
| Test branch | 0.0220 | -0.106 | -0.302 | 0.0900 | 0.0137 |
| | (0.0916) | (0.134) | (0.188) | (0.143) | (0.0912) |
| $R^2$ | 0.004 | 0.133 | 0.046 | 0.126 | 0.008 |
| *Month fixed effects and branch fixed effects* | | | | | |
| Treatment | 0.00578 | 0.181*** | 0.414** | 0.0887 | 0.0191** |
| | (0.00909) | (0.0422) | (0.177) | (0.0549) | (0.00900) |
| $R^2$ | 0.305 | 0.382 | 0.085 | 0.377 | 0.394 |

Test phase is dummy variable and takes value "1" in months 7-12. Treatment branch is a dummy variable and take value "1" when a branch is in the treatment group. The panel at the top includes no further controls, the panel in the middle includes month dummies and the panel at the bottom includes both month dummies and branch fixed effects. Robust standard errors clustered on branches, *** p<0.01, ** p<0.05, * p<0.1.

Table A2: Heterogeneous treatment effects: Appointments

| | *log Appointments* | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Treatment | 0.0828** | 0.0593 | 0.0386 |
| | (0.0400) | (0.0483) | (0.0682) |
| Information | 0.0944** | 0.0944** | 0.0944** |
| | (0.0401) | (0.0401) | (0.0401) |
| Treatment x branch size (centered) | -0.00960 | | |
| | (0.0107) | | |
| Treatment x 20% smallest branches | | 0.127** | |
| | | (0.0577) | |
| Treatment x 20% largest branches | | 0.0299 | |
| | | (0.0692) | |
| Treatment x 30% smallest branches | | | 0.146** |
| | | | (0.0655) |
| Treatment x 30% largest branches | | | 0.0271 |
| | | | (0.0715) |
| $R^2$ | 0.346 | 0.346 | 0.347 |

Branch size is the number of (full time equivalent) employees in a branch centered at the mean. Branch fixed-effects and month dummies included. Robust standard errors clustered on branches, *** p<0.01, ** p<0.05, * p<0.1.

Table A3: Persistence and branch size

| | (1) *log CNR loans* | (2) *log CNR investment* | (3) *log CNR savings* | (4) *log CNR credit cards* | (5) *log CNR total* |
|---|---|---|---|---|---|
| *Small branches (*$\leq$ 6 full time equivalent employees*)* | | | | | |
| Treatment x Q3 | -0.0267** | 0.182* | -0.110 | 0.0102 | -0.0143 |
| | (0.0129) | (0.0914) | (0.273) | (0.0689) | (0.0124) |
| Treatment x Q4 | -0.0307*** | 0.369*** | 0.155 | 0.0317 | 0.00189 |
| | (0.0111) | (0.0610) | (0.311) | (0.175) | (0.0151) |
| $R^2$ | 0.168 | 0.351 | 0.090 | 0.303 | 0.245 |
| *Large branches (*$>$ 6 full time equivalent employees*)* | | | | | |
| Treatment x Q3 | 0.0126 | 0.175*** | 0.448** | 0.143** | 0.0268*** |
| | (0.00879) | (0.0630) | (0.189) | (0.0647) | (0.00926) |
| Treatment x Q4 | 0.0176 | 0.129** | 0.578** | 0.0640 | 0.0240** |
| | (0.0113) | (0.0568) | (0.233) | (0.0650) | (0.0116) |
| $R^2$ | 0.448 | 0.403 | 0.098 | 0.435 | 0.533 |

Branch fixed-effects, month dummies and call center initiated appointments included. Robust standard errors clustered on branches, *** p<0.01, ** p<0.05, * p<0.1.

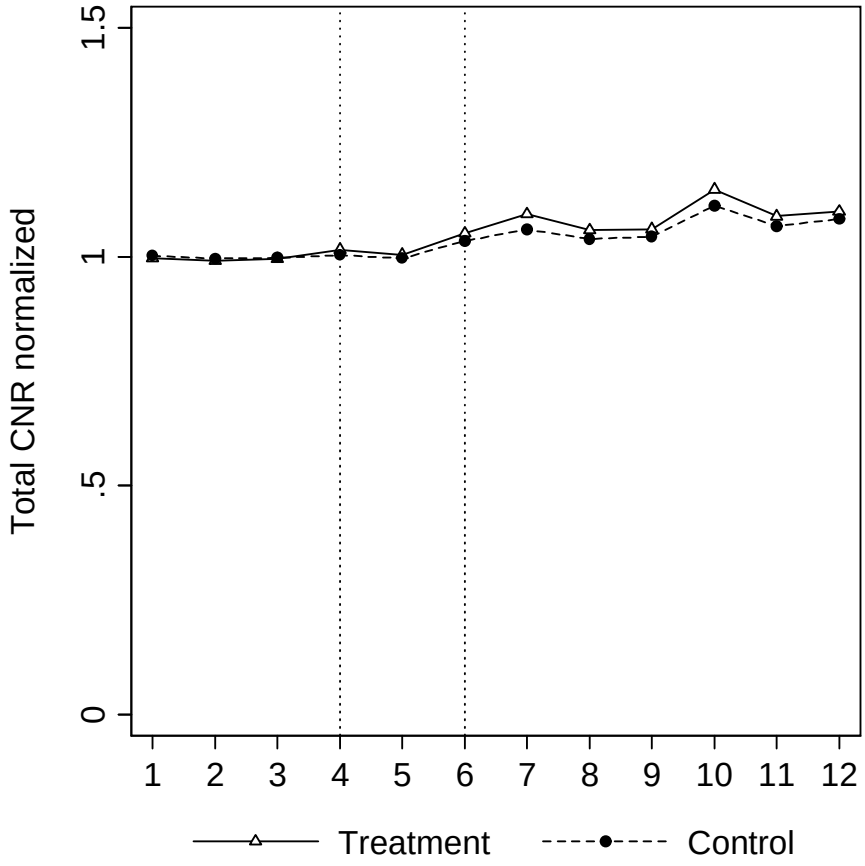



Figure A1: Normalized profits ("Customer Net Revenue") over time

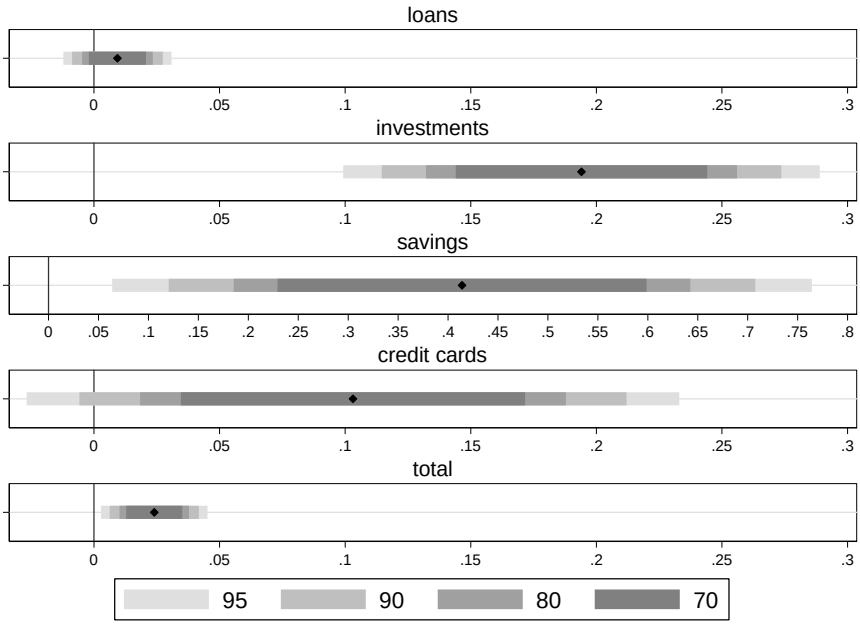



Figure A2: Treatment effects and confidence bands

Figure A3: Branch size in month 3 (Full Time Equivalents)