Master Thesis: The (hidden) Benefits of Monitoring

Hauke C. Roggenkamp

2018-11-07

Contents

Chapter 1

Introduction

There seems to be a common wisdom that monitoring affects the workers' output negatively. A popular rationale is that it signals distrust and triggers psychological costs which are reciprocally passed back to the managers. As a consequence, the managers risk to suffer losses if they increase the level of attention they pay to their workers. These detrimental effects can be labeled as the hidden costs of monitoring and are subject to a rich body of empirical and theoretical work. Most of these studies identify monitoring as a management practice that is perceived as unkind in some sense and design or model it as such. Because we believe that monitoring is neither a bad nor a good management practice *per se*, we suggest a more nuanced contemplation. To this end we designed a laboratory experiment where monitoring can be perceived as kind or unkind. We hypothesize that workers reciprocate this perceived (un)kindness via their labor supply. Whether a worker perceives monitoring as kind or unkind is expected to depend on her productivity: Productive workers benefit from monitoring and perceive it as kind. Unproductive workers, in contrast, suffer from monitoring and perceive it as unkind. The underlying idea is easily understood in real-world applications such as in human resource managers' decisions that concern wage re-negotiations or promotions, for example. In such scenarios, monitoring is likely to help the manager to make informed decisions on the basis of relevant metrics (such as a worker's productivity). If the manager lacks assessments of the worker's productivity, she has to rely on inferior, if not arbitrary, characteristics that are easily observable, such as a worker's tenure. A young, ambitious and talented worker will likely benefit from monitoring as it implies that the manger's decision is based on work samples. Slow-going workers who already spent some years in the company would, in contrast, prefer the promotion decision to be based on the tenure. After all, that measure is unrelated to their work and improves their chances to be promoted. We expect the unproductive workers to express their discomfort of being monitored while the productive workers express their gratitude. We designed a laboratory experiment where the workers' labor supply is the only channel to express these emotions. Applying an intention-based reciprocity model in this setup, I theoretically predict one and the same action (monitoring) to have unsuspected costs *and* benefits. The results demonstrate that the workers' behavior cannot be explained by a standard model that assumes them to be purely self-interested. Using OLS regressions, OLS-based simulations, Fisher's exact test as well as a regression discontinuity design, I find mixed results whose sum I interpret as supporting evidence for the intention-based reciprocity predictions: The OLS's regression coefficients (that highly depend on three extreme values) are insignificant which indicates that there is no relationship between the workers' working morale, their productivity and monitoring; the simulations indicate that monitoring spoils the workers' working morale; Fisher's exact test finds that productive workers who were unobserved reduced their workload significantly more often than unproductive workers who were not observed while the opposite holds true for monitored workers; and the regression discontinuity design turns out to be impractical to analyze our data. Because the OLS' and OLS-based simulations' results are sensitive to only three observations and because the workload (analyzed with Fisher's exact test) is a good predictor of

> the workers' initial intention to work, I interpret the sum of the results as follows: Workers, who were unproductive and disliked to be monitored, punished the monitoring manager by working less. Simultaneously, productive workers who were monitored suffered less and therefore lowered their effort provision by less than their unproductive colleagues. Importantly, we do not find any hidden benefits of monitoring, that is, workers, who perceive the managers' intentions to monitor them to be kind, do not work harder than they would if they had no emotions. It simply appears as if this group of workers either perceived the managers' intentions as neutral or as if they had no channel to express kindness in practice: I suspect that we would have found hidden benefits if it was easier for workers to work more or better. I therefore suggest to adjust the experimental design accordingly. Regardless of whether monitoring triggers hidden benefits or whether the hidden costs disappear, this thesis suggests that the application of monitoring as a management practice needs to be assessed in a more nuanced way than the current literature suggests. The workers productivity appears to be a good candidate to unravel these nuances. The *empathetic* monitoring of workers might then explain parts of the persistent performance differences across seemingly similar firms.

How does monitoring affect the agents' working morale? There is a wide-spread belief that monitoring spoils the working morale and therefore, comes at hidden costs. It might, for instance, be perceived as a lack of trust and trigger psychological costs. These psychological costs are then reciprocally passed back to the agent through a decreased effort provision (Dickinson and Villeval 2008). Experiments (such as Falk's and Kosfeld's (Falk and Kosfeld 2006)) that support this claim are often designed in a way that the data could not possibly support other conclusions.[1] While Falk and Kosfeld designed an experiment in which the principals restrict the agents' autonomy, there also is a growing number of studies that model several other control devices, namely "managerial attention", "supervision", "verification" or "monitoring" (Schulze and Frank 2003; Guerra 2002; Dickinson and Villeval 2008). Their experimental setup is similar to the one of Falk and Kosfeld as they allow the application of these devices to only be perceived as a breach of trust.

While we acknowledge that all these management practices (to which I henceforth refer as "monitoring") may have detrimental effects on the agents' working morale, we also believe that they might be perceived as fair so that could also motivate the agent intrinsically. We argue that monitoring helps to receive high quality signals of the agent's performance and therefore minimizes the problem of incorrectly receiving bad performance signals. Monitoring can thus also be seen as a good management practice (Halac and Prat 2016) that helps to make legitimate decisions relating to the assignment of prestigious projects, salary negotiations, promotions *et cetera*. Furthermore, employees might be more motivated if they know that these decisions are made on the basis of a valid performance assessment instead of arbitrary characteristics such as the employee's tenure (Bloom and Van Reenen 2007, 1356). In our view, the belief that monitoring, if anything, spoils the working morale is too narrowly considered since it does not cover this positive dimension sufficiently. That is not to say that we believe monitoring to have positive effects *per se*. Instead, we argue that the circumstances determine whether monitoring is perceived as kind or unkind. To study a more nuanced view, we designed a laboratory experiment that allows for both (kind and unkind) perceptions. While this enables us to investigate hidden costs in form of a low level of effort provision (due to perceived unkindness that is reciprocally passed back to the principal) the design also allows us to search for hidden *benefits* of monitoring – at least to some extend.

The idea that the same action is perceived as legitimate in some scenarios while it is perceived as unjust in others is not new. A price increase, for instance, seems to be only perceived as unkind if the producer's costs did not increase (Okun 2011). Similarly, it is perceived as unjust to cut wages if the employer's wellbeing is not at stake. If it was at risk however, employees might accept it (Kahneman, Knetsch, and Thaler 1986). Finally, and more related to this thesis, the employer's control seems to crowd-out the employees intrinsic motivation to supply labor if it is not legitimate, that is, if it does not prevent antisocial behavior (Schnedler and Vadovic 2011). The restriction of internet or social media access in a small sized family business may be perceived as unjust as it signals distrust. The same policy might, however, be perceived as neutral in the setting of a large multi-national firm, where the inter-personal ties are weaker and the risk of selfish behavior

[1] See (Siemens 2013; Schnedler and Vadovic 2011; Masella, Meier, and Zahn 2014) for further references that analyzed their experimental design.

is higher. Similarly (Barkma 1995) and (Frey 1993) find suggesting evidence that the monitoring of hours worked can crowd-out the workers' motivation (and performance) if the monitoring principal was their own CEO while it has positive effects on their performance if the principal was a distant parent company. These studies, amongst others, suggest that one and the same action can be moderated by another variable that determines how this particular action is perceived. In our experiment, we identify the agent's productivity as such a variable.

To put it in a nutshell, our intention was to create situations in which overachievers, in contrast to layabouts, appreciate to be monitored and reciprocate this sense of appreciation. We therefore analyze the effect of the interaction of monitoring and the agents' productivity on the agent's working morale. To do so, we implemented an experimental principal-agent game in which the agent supplied labor in a real-effort task. This was costly for her but generated profit for the principal. Importantly, the principal, who paid the agent's salary was not able to directly observe the agent's output. Instead, the principal chose one out of two available mechanisms that we interpret as attention technology and thus, as monitoring. The chosen mechanism generated the agent's salary. Under both mechanisms, the agent's salary consisted of a flat wage and the chance to also receive a bonus payment. While the random mechanism flipped a virtual coin to determine whether the agent receives the bonus, the performance-based mechanism was more likely to pay out the bonus the higher the agent's performance was. Choosing the performance-based mechanism was, in a metaphorical sense, like observing the movements of the agent to make the bonus decision – the better her performance, the better the principal's impression of her work, the likelier it becomes that she receives the bonus if she worked well. The choice of the random mechanism is, in contrast, interpreted as a complete lack of monitoring: the principal had no impression of her work such that the agent's earnings must be determined randomly. The random mechanism therefore sent completely arbitrary performance signals that determined the agent's earnings and were likely to be incorrect. Monitoring consequently was valuable to the principal as it incentivized the agent to exert effort. In addition, it was beneficial for agents, who expected to perform well because it yielded better chances to earn the bonus than the coin flip. In contrast, it was disadvantageous for layabouts, who could hope for a lucky outcome of the the random mechanism's coin flip.

An important feature of our design is that both the principal and the agent received an objective assessment of the agent's *productivity* (or "talent") before the principal decided whether to monitor the agent. This assessment was based on the exact same task, which the two players executed in a previous stage of the experiment. In addition, both mechanisms avoided ex post hold-up problems because they were a function of a performance signal. A principal could thus, only decide on the signal's quality but not on the eventual payment she had to offer the agent. Consequently, we modeled a situation of complete contracts. Another important feature was that the agent learned the principal's choice of the mechanism before she worked for the principal, that is, whether the principal paid attention or not. We exploit these two features to analyze whether an agent provided more effort than her productivity would suggest if the principal chose the mechanism that was beneficial to the agent. In addition, we are interested in the agents' behavior if the principal chose the mechanism that was disadvantageous to them. In the latter case, we hypothesized them to provide less effort than one would expect (given their productivity). If this was the case, the "wrong" monitoring decision would have detrimental effects on the agent's working morale – *hidden costs*. In contrast, the "right" choice would yield *hidden benefits*. The rational principal might then have an incentive not to pay attention to the agent's work, that is, to choose the random mechanism, albeit provoking a moral hazard.

The results are mixed. They do not support the hypothesis that agents react kindly to monitoring.[^This might, however, be due to the experimental design: While it was easy to exert low levels of effort, it was hard to go beyond one's boundaries which were set by the individual productivity.] We therefore find no evidence for hidden benefits of monitoring. The hidden costs of monitoring, however, are present in our data. We find that the average unproductive agent decreases her effort provision by about three to twelve percentage points, if monitored. Although it should not be interpreted as a causal effect of monitoring, the data also show that principals who monitored the agents realized higher payoffs. The discrepancy between hidden costs on the one hand and higher payoffs on the other hand cannot be explained by higher performances. Instead, it might be a result of chance (despite the significant p-value). Furthermore, and as predicted, the data also suggest that these hidden costs disappear for productive agents who benefit from monitoring. The latter observation, however, depends on the subset of data and methods applied. After all, the fraction of productive agents,

who refused to supply effort while being monitored is relatively low. The unfirm robustness of the results calls for a continued data collection as well as an additional, refined treatment.

Investigating a more nuanced picture of the hidden effects of control helps to investigate management styles and their effect on the firms' productivities, profitabilities and survival rates (Bloom and Van Reenen 2007). (Gibbons and Roberts 2012, ch.~17) as well as (Bartelsman and Doms 2000) and (Syverson 2011) review a variety of studies and conclude that there are persistent performance differences across seemingly similar enterprises that may, in part, be explained by managerial skills and practices. Micromanagement might, for instance, have detrimental effects by eroding the workers' motivation (Foss 2003). We aim to identify monitoring as a management practice that affects the agent's working morale conditional on her characteristics. As such, we investigate whether monitoring is a practice that (1) may explain some of the performance differences across firms and that (2) requires skilled managers who are able to identify who benefits or suffers from their attention.

Chapter 2

Design

Our experiment consisted of two stages and a questionnaire which you can find in Appendix . The idea was to create an environment in which both the principal's decision to monitor and her omission of monitoring can be perceived as kind or unkind by the agent subject to her productivity.[1] We therefore implemented a real-effort task to measure an agent's productivity, disclosed this information to the agent as well as to a matched principal, let the principal choose between two options and observed the agent's reaction to the principal's choice. The principal's choice and the agent's reaction were interdependent, that is, their actions affected not only their own, but also the other player's earnings. Importantly, agent's who I'll later classify as unproductive preferred the principal's option which I interpret as the omission of monitoring, while the productive agents preferred the alternative option, which I interpret as monitoring. As this section explains, we made it easy for the agent's to form beliefs about the intentions of the principal in a, for us, comprehensible manner. Our design therefore allows me to identify who should feel treated kindly or unkindly to observe whether the perceived (un-)kindness is reciprocated.

2.1 Overview

The experiment consisted of two stages. One independent (or "individual") stage was played one-shot and followed by an interdepend stage in which we implemented an one-shot principal-agent setting.

The design of the first stage is illustrated in Figure XY, where i denotes any participant and l her effort provision (or "labor supply"). The player 0 is an artificial and thus uninterested[2] player to whom one can also refer as *chance*} as she only conducts an explicit randomization subject to i's effort provision.

```
include_graphics("images/20171127_GameTree")
```

A participant's effort provision in this task affected her, and only her, earnings in a simple way: The higher her effort provision, the higher her chances to earn a bonus payment of $b = 75$ Danish kroner (DKK) in addition to a flat wage of $w = 150$ DKK. The possible effort provision ranged between 0 and 100 percent. With each additional percentage point of provided effort, the participant's chances to earn the bonus payment increased by one percentage point as well. I'll refer to this mechanism as *"performance-based"* in what follows. As the effort a participant provided in this stage did not involve any strategic considerations, I'll refer to it as *"productivity"* and use it as a proxy to measure a participant's initial ability (which, in turn, can be described by an individual costs of effort function).

[1] There is one special case in which agents are neither classified as productive nor as unproductive. These agents are indifferent between the two options and thus perceive the principal's choice as neutral. All the other agents, however, prefer one of the two options such that they can feel treated kindly or unkindly.

[2] The player did not receive any payments and acted randomly.

This stage served two purposes: First, participants familiarized themselves with a performance-based payment mechanism which is an important element of the second stage as well. Second, they got a good understanding of the difficulty of the task as well as an objective assessment of their own productivity since we informed each participant about her effort provision in that stage.[3] This is important as the first stage's performance is, as we believe, likely to be used to evaluate another player's actions in the subsequent stage.

At the beginning of each session, participants were randomly assigned to be either a principal or an agent. In Stage 2, the agents had to work on the same real-effort task as before and faced similar incentives to supply effort. While their work environment was similar, it differed substantially from the first stage because the agents' decision to supply labor became strategically. The game that was played in the second stage is depicted in Figure XY and is described as follows:

Firstly, participants found themselves to be either in the role of a principal or an agent who I henceforth denote as j and i respectively. Each agent was matched with one principal. Only the agents were engaging in the real-effort task in this stage. To distinguish the effort provision in the first stage from the effort provision that followed in the second stage, I'll henceforth refer to the latter one as *"performance"*. The agents' performance affected both their own and the matched principal's payment function. Hence, instead of only playing a two-player game with the uninterested chance player, the real-effort task was now embedded into a three-player game (principal, agent and chance).

This game included, secondly, more actions than just the performance in the task. To begin with, the principal, who was not exerting any effort in this stage, had to choose the mechanism that determined the agent's earnings. More precisely, the principal was prompted to choose whether the agent's earnings are determined by a performance-based mechanism (φ) such as in the first stage or by a "*random*" mechanism(ρ). The performance-based mechanism can be found on the right branches (following history h^2) in Figure XY. As before, the agent's performance determined the probability with which the chance player 0 draws the bonus payment. The random mechanism on the left (following history h^1) looks similar and differs in only one important aspect: the probability $q \equiv 0.5$ with which an agent received the bonus payment was independent of her performance in Stage 2.

Before the principal chose the mechanism, she learned the matched agent's productivity. Without any ambiguity or uncertainty, both participants therefore knew how much effort the agent was willing to supply under the first stage's incentives. Importantly, the agent knew that the principal received this information.

After the principal made her decision, the agent was asked to choose her workload n in histories h^1 and h^2. In particular, she was asked to indicate on how many screens, that is, on how many repetitions, she intended to work in this stage's real-effort task. All participants knew that choosing, say, 80% of the screens would have spoiled their chance of achieving a performance of 100%; the best performance they could accomplish when choosing to only work on 80% of the screens was 80%. Subsequently, the agent exerted effort in the real-effort task subject to the workload she chose. Finally, chance executed its explicit randomizations – subject to the agent's performance l or by tossing a fair coin, ($q = \frac{1}{2}$).

```
include_graphics("images/20171127_GameTree")
```

Lastly, the second stage's game was different with respect to its payments, simply because the game evolved to a three-player game with two interested players. The artificial player chance was still uninterested. Furthermore, the agent was facing the same set of possible payments as in the first stage. The principal's payment function is designed such that she accounted for the agent's salary. In return, the principal earned one DKK for each percentage point of the agent's performance (if the agent's performance was, say, $0.65 = 65\%$, the principal earned 65 DKK) in addition to a flat wage of $\varepsilon \equiv 340$ DKK. Also the principal's choice was costly to her since the random and the performance-based mechanism came at the expense of $c_\rho \equiv 20$ or $c_\varphi \equiv 25$ DKK respectively. Given this parameterization, the principal's material payoff was increasing in the agent's effort provision for any of the two mechanisms. (This is not as obvious as it sounds since the

[3] Prior to running the incentivized box-clicking task, each participant engaged in a short unincentivized trial round. As a consequence, participants knew how the task looks like. However, they did not know how long the incentivized will take and did not learn how well they performed in the trial round.

principal had to pay the agent's expected bonus payment, which also increased in her performance in Stage 2.) Consequently, it was in the principal's best (material) interest to induce a positive level of effort provision.

To sum up, the second stage can be described as follows: Two participants were assigned to be either an agent, i, or a principal, j. Both faced an artificial, uninterested chance player denoted as 0. The agent's productivity was public knowledge to both human players. The principal's only action was to choose a payment mechanism that determined the agent's earnings in this stage. Since the principal accounts for the agent's earnings, this decision also affected her own earnings. The agent was informed about the principal's choice and chose a workload before she exerted effort. Finally, chance determined the earnings of the agent (and thus of the principal) in Stage 2. See Figure XY for (another) visual representation of the the second stage's timeline.

```
include_graphics("images/20180103_Timeline")
```

2.2 The Real-Effort Task

Participants worked on a tedious box-clicking task in which they faced a number of screens displaying dozens of randomly ordered black boxes. The participants indirectly earned money by clicking on these boxes, which caused the boxes to vanish. There was a timer running down from, say, eleven seconds. Each time the timer counted zero, the screen with all the black boxes that were left vanished and a new screen with a new set of randomly ordered boxes appeared. This usually happened before all boxes of the current screen were "clicked away". The participants' performance was then measured by the total number of boxes they clicked on, divided by the number of boxes they could have clicked away.

To ensure that we eventually observe a heterogenous group of agents who differ in their productivity, we manipulated the difficulty of the box-clicking task exogenously. More specifically, the time each screen with boxes was displayed differed between sessions but not between stages or within sessions. We implemented two different difficulties with either seven[4] or twelve seconds on average per screen. The idea was that less time per screen made it more difficult to click away a certain percentage of boxes. The maximum number of screens as well as the number of boxes per screen remained unchanged between sessions. In conclusion, one could expect (1) the average effort provision to be higher in the eleven-seconds sessions and (2) the eleven-seconds sessions to take a little longer.

We are confident that the task induced a positive cost of effort for participants since it was exhausting and boring. In addition, the task itself was pointless, such that we can also be confident that participants had no motive to spend any additional effort as a gesture of kindness towards the experimenters (for instance, to reciprocate the payments they offered).

Even though we implemented one and the same task twice (for the agents) we expect that neither fatigue nor learning and thus, a nonseparability of effort costs (or ability) across time confounded the design. First, because the task itself did not require any specific knowledge or skills that can be trained during the session. Even if there was a learning effect, it might be negligible because each subject participated in a trial round. In these rounds, the subject could have gained all the knowledge and skills that were to be learned. Second, participants read instructions and answered control questions in between the two box-clicking periods such that there was an extensive break to recover.

2.3 Implications

The design of both the first and the second stage was intended to be interpreted as a principal-agent setting in which the principal decides whether she wants to monitor the agent: By choosing the performance-based mechanism, the principal can either get a positive or a negative impression of the agent's work. The better the agent's performance, the higher the likelihood that the principal's impression of the agent's work is a

[4] 6.92 to be more precise.

positive one. To prevent truth telling problems the agent is paid according to the impression the principal has – a positive impression *automatically* leads to a bonus payment. The random mechanism resembles the omission of any monitoring such that the principal is forced to toss a virtual coin to make her bonus decision.

Given any level of performance except for $l = 0.5$, agents materially preferred one of the two options the principal could choose: Agents with a performance lower than 0.5 (to whom I'll refer as *"unproductive"*) were best off under the random mechanism while agents with a performance of 0.5 were indifferent and agents with a performance higher than 0.5 (the *"productive"* ones) materially preferred the performance-based mechanism as l was higher than q such that the expected earnings under the performance-based mechanism were higher as well. Hence, which choice of the principal would make the agent materially best off depended on the agents performance in Stage 2. This was known to the participants, as we asked several control questions that addressed these scenarios.

As explained above, the principals made their monitoring decisions before the agent exerted effort. As such, neither the principal nor the agent knew the agent's performance in the second stage when the principal chose to either monitor or disregard the agent. However, both knew that the real-effort task will be the same as in Stage 1 and as difficult as in Stage 1. Furthermore, both were informed about the agent's productivity in Stage 1. Assuming that the agent believes the principal to believe that the agent can replicate her effort provision from the first stage in the second stage, the presented design allows us to identify agents who should feel treated kindly or unkindly (this is a core-assumption I will elaborate in the following chapter). Having identified those who should feel treated kindly or unkindly, we can observe their performance in the second stage to investigate reciprocity: Those who faced the performance-based mechanism found themselves in the exact same real-effort task as before, except that their effort also generated profit for a second participant. If, say, the agent felt treated unkindly, she was given the opportunity to pass back the unkindness by reducing her performance relative to her productivity. This would generate a profit for the principal lower than what the agent could have given her and come at the cost that the agent's expected earnings would be lower as well. Likewise she could have passed back kindness with an increased effort provision.

2.4 Procedural Details

Each game was played one-shot and the whole experiment was framed in a neutral manner[5]. The experiment was computerized using a software developed by *Andreas Gotfredsen* and modified by me. Participants were randomly allocated a role upon arrival at the laboratory.[6] We made sure that no principal was seated next to her matched agent such that it was not possible to identify the person a participant was matched with by her clicking behavior. All sessions where conducted by at least one of two lab-assistants and supervised by me to ensure that there were no differences between sessions (session effects). Each participant read instructions and answered control questions before entering a stage. At several occasions during the whole experiment, a participant had to wait for the other participant to whom she was matched until she finished a certain part of the stage (such as the control questions). As a consequence, the experiment included extensive waiting periods in some cases.

The 186[7] subjects who participated in the experiment were students from Copenhagen based Universities studying various majors.[8] Using ORSEE (Greiner 2015), we recruited students with little experience, that

[5]We named agents and principals as "Person B" and "Person A" respectively. We avoided value loaded terms as well as terms related to natural employment relations. The only employment related term we used was "workload".

[6]Each participant drew a seat number. We placed login information composed of a unique username and a unique password on each seat. The usernames thereby referred to either of the roles.

[7]We ran additional sessions in the end of January, 2018. The results stemming from the complete data set are not further discussed in this thesis as the data collection was to close to the submission of this document. However, I ran the analysis that I conduct for the 186-participant data set also for the complete set and report the corresponding results in Appendix XY (without commenting them).

[8]Most of them (33 percent and 12 percent respectively) stated to study Economics and Business or Social Sciences (which includes Economics). Only 2 percent stated to study Psychology and one percent (two subjects) actively stated not to be a student.

is, who had participated in no or few experiments before. As a result, the median experience in economic[9] experiments was two sessions. Between 16 and 28 subjects participated in each of the 8 sessions, resulting in 96 and 90 participants in the fast and slow treatment respectively. Including the time needed to read instructions, to answer the control questions and to pay the participants eventually, the experiment took about 80 minutes on average. Since the payments were designed such that no participant would earn less than 90 DKK (which, at that time, corresponded to 13.5 USD), we did not pay any fixed show-up fee.[10] While payments ranged from 100 to 240 DKK (15 to 36 USD), participants earned 181 DKK (27 USD) on average for their participation in the experiment.

[9]The Psychology Department has a laboratory as well. However, there subject pool is organized by another software than ORSEE, such that we do not know whether subjects have participated in psychological experiments before.

[10]Those subjects who showed-up but were rejected to participate, for instance, due to overbooking, received 50 DKK (about 7.5 USD).

Chapter 3

Behvioral Predictions

> "To create a model, then, we make choices about what's important enough to include, simplifying the world into a toy version that can be easily understood and from which we can infer important facts and actions." — (O'Neil 2017)

We implemented a simple real-effort task where a participant's actions affected only her own payments in the first stage. We then adapted the game to a principal-agent setting in Stage 2. Because we are interested in social preferences, I focus on the second stage in what follows (and touch on the first stage whenever it eases the comprehension).

The behavioral predictions for our experiment depend on the subjects' preferences and the corresponding assumptions. I consider two cases: One in which agents are self-interested, that is, they are only interested in maximizing their own utilities, and one where they have social preferences that are described by models of intention-based sequential reciprocity. I predict that self-interested agents, who are exposed to the performance-based mechanism, supply similar levels of effort in Stage 2 as in Stage 1. Reciprocal agents, in contrast, are predicted to deviate from their first stage effort provision. Note that we are not interested in the standard case in and of itself – it simply serves as a reference point to contrast the intention-based reciprocity predictions.

The principals' preferences do not affect the empirical analyses of the agents' behavior much. For this reason, I assume them to be self-interested throughout the whole analysis and do not focus on their decisions in this thesis. I derive the predictions for the standard case, the self-interested preferences, first before I move to the reciprocity driven social preferences. To reiterate, the setup in Stage 2 depicted in Figure XY is the following: The principal (j) decides whether she monitors the agent by choosing a mechanism that determines her payment. I refer to this variable as $\mu \in$ (random, performance) where I abbreviate the random mechanism with ρ and the performance-based mechanism with φ to improve the readability of the formal expressions in what follows. The agent, by contrast, has two choice variables: n, which is her workload measured as the number of screens she intends to work on, and her performance $l \in [0, 1]$, with a ceiling determined by her choice of n with $\{n \in \mathbb{R}^+ | 1 \leq n \leq 25\}$. $c(l)$ describes her costs of providing effort. Agents are paid a fixed salary w and might receive an additional bonus payment b. In case the payment is not performance-based ($\mu \neq \varphi$), the agent receives a payment which is determined in a random procedure ($\mu = \rho$), where she receives the bonus payment with an exogenously set probability of $q \equiv \frac{1}{2}$. For each percentage point of the total number of boxes clicked away (which is the exact definition of l), the principal receives one DKK such that she will be paid a *relative* "piece rate" of l DKK. In addition, she receives a fixed payment of $\varepsilon \equiv 340$ DKK to avoid bankruptcies.

Since the last mover is *0* (*chance*), the game's final actions are explicit randomizations. I assume that the other two (human) players will not solely focus on the specific realizations of payoffs but calculate their *expected* monetary payoffs to develop their behavioral strategies. Stylizing this thought, one can imagine a reduced form *two*-player game as illustrated below in Figure XY. This assumption ultimately has an attribution theory style implication as participants who think in expected payoffs do not blame chance for

particularly low outcomes that might occur. Instead, agents hold their matched principal accountable for the relatively high or low *expected* outcome they are facing.

```
include_graphics("images/20171127_GameTree")
```

To describe the agent's behavior (I focus on the eventual effort provision l and neglect the workload decision n), I consider a standard model of effort provision with a utility function, that is separable in the subject's utility from her payment $\pi \in (w, w+b)$, her costs $c(l)$ stemming from her effort provision and her intrinsic motivation $\sigma \in (0,1)$ she derives from working on the task.

3.1 The self-interested Agent's Expected Utility

I start by deriving an agent's motives to exert effort by analyzing the strategic environment every participant (agents and principals) faces in the first stage. For simplicity, I focus on a representative (that is, homogenous) agent's (i's) motives as they can easily be transferred to a principal. Like in Stage 2, each agent is rewarded with a flat wage. Whether the agent receives the bonus is determined by a performance-based mechanism that is identical to the one the principal can choose in Stage 2. The first stage's effort provision $l^{1^{st}}$ is an independent measure of effort provision as it stems from a two-player game where only one human participant interacts with an artificial chance player. As mentioned, I refer to $l^{1^{st}}$ as *productivity*. In conclusion, one can stylize the game in Stage 1 as follows: the higher a participant's productivity, the higher the likelihood of receiving the bonus payment. Formally:

$$\mathbb{E}[\pi_i^{1^{st}}(l)] = l(w+b) + (1-l)w = w + l \cdot b$$

Considering a participant's costs of effort as well as her intrinsic motivation one can derive her utility function and solve the maximization problem:

$$U_i(l, c(\cdot), \sigma) = w + l \cdot b - c(l) + \sigma \cdot l \Rightarrow c_l(l^{1^{st}}) = b + \sigma \Leftrightarrow l^{1^{st}} = c_l^{-1}(b+\sigma)$$

$c_l(\cdot)$ thereby denotes the derivative of the cost function with respect to the effort level l (the marginal costs of effort) and $c_l^{-1}(\cdot)$ denotes the inverse of the marginal cost function. Because $c(l)$ is assumed to be convex, $c_l(l)$ is increasing and so is its inverse $c_l^{-1}(\cdot)$. From here it follows that

1. $l^{1^{st}}$ increases in the intrinsic as well as the variable extrinsic motivation and that
2. a self-interested subject chooses effort up to the point where the sum of both (intrinsic and variable extrinsic motivation) equals her marginal costs of effort.

In the second stage, the agent's expected monetary payoff is slightly more complex since it does not only depend on her own effort in Stage 2, l^*, (which I call her *performance*) but also on another variable: the principal's binary monitoring decision. We'll therefore have to consider two cases resulting from either a random (ρ) or a performance-based mechanism (φ).

$$\mathbb{E}[\pi_i(l)|\rho] = q(w+b) + (1-q)w = w + q \cdot b \mathbb{E}[\pi_i(l)|\varphi] = \mathbb{E}[\pi_i^{1^{st}}(l)] = w + l \cdot b$$

These two functions, along with the principal's expected monetary payoff, are visualized in an interactive *ShinyApp* I programmed and archived here.[1] Adding the costs of effort as well as the intrinsic motivation yields the following first-order conditions:

$$\frac{\partial U_i}{\partial l} = \begin{cases} \sigma - c_l(l) \ \Rightarrow \sigma = c_l(l_\rho^*) & \Leftrightarrow l_\rho^* = c_l^{-1}(\sigma) \\ b + \sigma - c_l(l) \ \Rightarrow b + \sigma = c_l(l_\varphi^*) & \Leftrightarrow l_\varphi^* = c_l^{-1}(b+\sigma) = l^{1^{st}} \end{cases}$$

[1] If you focus on the principal's (Person A's) earnings, you will see that the principal's earnings were strictly increasing in the agent's performance so that she had a monetary incentive to induce effort.

Hence, the agent will choose effort up to the point where the sum of her intrinsic and variable extrinsic motivation, if any, equals her marginal costs of effort. Because $c_l^{-1}(\cdot)$ is increasing and because $b > 0$, it follows that $l^*_\varphi \geqslant l^*_\rho$. In summary, I predict that:

```
include_graphics("images/00_Self_Prediction.pdf")
```

> A purely-self interested agent's performance, given a performance-based mechanism, will equal her productivity.
>
> A purely-self interested agent will perform better (that is, she will click on more boxes) if she faces the performance-based instead of the random mechanism.[2] "'

These predictions are conceptualized in Figure XY where the yellow line could have any non-negative slope (which is defined by σ). However, they hinge on the implicit assumptions that (1) an agent does neither learn (thus improve her ability to perform the box clicking task) nor fatigue (thus worsen her ability) and that (2) σ does not depend on the mechanism μ such that $\sigma(\rho) = \sigma(\varphi) = \sigma$. This translates into the assumption that the mechanism itself does not crowd out an agent's intrinsic motivation.[3]

Our design does not allow us to test any of these assumptions which classifies them as *postulates*. While I already argued that neither learning nor fatigue should be a concern here, the latter assumption deserves attention given the literature on the crowding-out effect of intrinsic motivation due to monetary incentives (see (Frey and Oberholzer-Gee 1997) for a general overview, (Bénabou and Tirole 2003) for a theoretical discourse or (Dickinson and Villeval 2008) as well as (Frey 1993) for papers that are closely related to this thesis) since the monetary incentive scheme is the key difference between the two mechanisms ρ and φ.

Explanations for the prevalence of a crowding-out effect due to monetary incentives require factors such as close relations between principals and agents, the prevalence of a less knowledgeable agent (compared to the principal) or agents who generate profit for the principals and have concerns about how the profit is distributed. None of these factors seem to confound the predictions in our setting as the relation between principals and agents is abstract and impersonal, as the agent has more and better information about herself as well as her performance in the two box-clicking tasks and because the distributional concerns (regarding the principal's income) are driven by reciprocity, which is the very subject of the following subsection and this thesis in general.

3.2 The Reciprocal Agent's Expected Utility

The basic intuition of the notion of reciprocity that I apply in this paper is that people respond kindly (unkindly) if they perceive actions of others as kind (unkind). As before, I will focus on the agent in our setting and apply this notion of reciprocity formally. To be more precise, I will base my considerations on the model of (Dufwenberg and Kirchsteiger 2004). Even though there is an uninterested chance player incorporated in our design, I do not need to involve her in the analysis as Sebald's (Sebald 2010) model would allow me to do. I focus on the expected outcomes as illustrated in Figure XY and omit the chance player.

Like (Dufwenberg and Kirchsteiger 2004) as well as (Sebald 2010), I denote b_{ij} as player i's belief about player j's strategy (first-order belief) and c_{iji} as player i's belief about player j's belief about player i's strategy (second-order belief). Players update their first- and second-order beliefs and strategies as soon as they learn the other player's actions which is why they depend on the history h. $a_i(h)$ describes the (updated) behavioral strategy that prescribes the same choices as a_i except for the choices player i has already made at h (since they are consequently made with probability 1). Incorporating the intrinsic motivation again, the agent's utility function, is assumed to look as follows:

$$U_i(a_i(h), (b_{ij}(h))_{j\neq i}, (c_{iji}(h))_{j\neq i}) = \pi_i(a_i(h), (b_{ij}(h))_{j\neq i}) + Y_{ij} \cdot \kappa_{ij}(a_i(h), (b_{ij}(h))_{j\neq i}) \cdot \lambda_{iji}(b_{ij}(h), (c_{iji}(h))_{j\neq i}) - c_i(a_i(h)) + \sigma \cdot a_i($$

[2] Note that her effort provision will, in the case of a random mechanism, only equal zero if her intrinsic motivation she derives from clicking boxes is zero as well.

[3] The subsequent section will show how *the choice of* the mechanism can crowd out intrinsic motivation.

According to this function, the agent's utility consists of four components: her expected material payoff, her psychological payoff, her costs of effort as well as her intrinsic motivation. The psychological payoff (the second term) includes a non-negative reciprocity parameter Y_{ij} describing her sensitivity towards the matched principal's (un)kindness, her (un)kindness towards the principal κ_{ij} as well as her perceived (un)kindness of the principal towards her λ_{iji}. Note that a reciprocity parameter of zero would describe a special case where an agent is not motivated by (intention-based) social preferences. In other words, a utility function with $Y_{ij} = 0$ would equal the purely self-interested case from above.

Before I derive explicit predictions concerning the reciprocal agent's behavior, I will focus on the elements that represent the psychological payoff. The original model's kindness function κ_{ij} implies that an agent evaluates her kindness towards the principal by comparing the payoff she grants the principal by her chosen action compared to what she could have given her – and she applies a similar mindset when evaluating the perceived kindness of the principal towards her (λ_{iji}). Formally,

$$\kappa_{ij}(a_i(h), (b_{ij}(h))_{j\neq i}) = \pi_j(a_i(h), (b_{ij}(h))_{j\neq i}) - \pi_j^{e_i}((b_{ij}(h))_{j\neq i})$$

where $\pi_j^{e_i}(\cdot)$ describes a j's equitable payoff that is affected by i. In the original paper, it is defined as

$$\pi_j^{e_i}((b_{ij}(h))_{j\neq i}) = \frac{1}{2}\Big[max\big\{\pi_j(a_i(h), (b_{ij}(h))_{j\neq i}) \mid a_i(h) \in (0,1)\big\} + min\big\{\pi_j(a_i(h), (b_{ij}(h))_{j\neq i}) \mid a_i(h) \in (0,1)\big\}\Big]$$

which basically means that the equitable payoff is a virtual average payoff that i can grant j.[4] If the eventual payoff j receives due to i's action is higher than this average, i considers herself as kind towards j.

I generally agree with the concept of an equitable payoff as a reference point and apply it later to evaluate the agents' *perceived* kindness. However, for the agent's evaluation of her own kindness towards the principal, I deviate from Dufwenberg and Kirchsteiger's (Dufwenberg and Kirchsteiger 2004) approach to determine it in the following way, since I believe that the original model does not fit into our setting:

$$\pi_j^{e_i}((b_{ij}(h))_{j\neq i}) \equiv \pi_j(l_i^{1^{st}}, (b_{ij}(h))_{j\neq i})$$

with $l_i^{1^{st}} \in [0,1]$ as the agent's productivity measured in the first stage. The rationale behind this is simple: I believe that agents are heterogeneous with respect to their productivity (their cost functions) and that an agent's inherent productivity is the best predictor of how well a particular agent can perform in the future. In other words, I expect an agent to be able to more or less replicate the effort provision from the first stage if she faces an identical strategic environment. Most importantly, I assume that subjects hold the belief that they could replicate their own effort and that they hold the same beliefs about others (j believes that i can easily replicate i's productivity from Stage 2 in Stage 1). Given this, Dufwenberg and Kirchsteiger's (Dufwenberg and Kirchsteiger 2004) definition of an equitable payoff does not make much sense since it would translate into an equitable payoff resulting from a performance of $\frac{1}{2} \cdot (0+1)$ irrespective of the idea that an agent could not possibly bring forth a performance of 100% due to a low productivity. Because it does seem even less intuitive and somehow arbitrary that an agent considers a payoff resulting from a performance of half her productivity $\frac{1}{2} \cdot (0 + l_i^{1^{st}})$ as equitable, I suspect $\pi_j(l_i^{1^{st}}, (b_{ij}(h))_{j\neq i})$ to be the best candidate for the fairness norm the equitable payoff was intended to represent.

This assumption is quite important as it sets the course for our analysis of kind or unkind behavior: kindness (unkindness) is identified as an increased (decreased) effort provision between the productivity in the first stage, $l_i^{1^{st}}$, and the performance in the second stage, $l(h)$:

$$\kappa_{ij}(l(h), (b_{ij}(h))_{j\neq i}) = \pi_j(l(h), (b_{ij}(h))_{j\neq i}) - \pi_j(l_i^{1^{st}}, (b_{ij}(h))_{j\neq i})$$

[4] In fact, the original equitable payoff is slightly different since it conditions the strategies to be part of a efficient space. There are, however, no inefficient strategies in our setting which is why I changed the corresponding formula slightly.

where I substituted \$a_i(h) = l(h) \$. This implies that an agent first chooses her effort at history h^1 or h^2 and then chooses her workload n subject to her effort decision.

Remember that the principal's earnings consisted of several components. In particular, her material payoff was designed as follows: $\pi_j \equiv \varepsilon + l - \pi_i(l,\mu) - c(\mu)$ where ε is a constant that is commonly known. Because the principal chooses μ before the agent makes her first move, the agent knows with certainty which mechanism was chosen by the principal when she evaluates her kindness at h^1 or h^2. She thus knows the principal's costs $c(\mu)$ and is able to infer the expected salary which she will receive from the principal ($\pi_i(l,\mu)$). Consequently, she knows each of the components that constitute the principals earnings. It thus suffices to only consider the agent's effort provision in either h^1 or h^2 to form $\pi_j^{e_i}$ and κ_{ij} as everything else cancels out. This means that the agent's effort provision is the only channel to exhibit kindness or unkindness. Given that the subgame of the second stage where the principal chooses the performance-based mechanism is very similar to the first stage, I understand an increased effort provision ($l_\varphi^* - l_i^{1^{st}} < 0$) as an expression of kindness and a decreased effort provision ($l_\varphi^* - l_i^{1^{st}} > 0$) as an expression of unkindness.

Similarly to $\kappa_{ij}(\cdot)$ in the original paper, the *perceived* kindness $\lambda_{iji}(\cdot)$ is expressed difference between an equitable payoff and the actual payoff – the two functions are, *prima facie*, mathematically equivalent.

$$\lambda_{iji}(b_{ij}(h),(c_{iji}(h))_{j\neq i}) = \pi_i(b_{ij}(h),(c_{iji}(h))_{j\neq i}) - \pi_i^{e_j}((c_{iji}(h))_{j\neq i})$$

In contrast to $\pi_j^{e_i}$, I find it practical to form the equitable payoff the agent can receive from the principal ($\pi_i^{e_j}$) as in the original paper because the principal has a binary set of actions $\mathcal{A}_j = \{\rho,\varphi\}$.

$$\pi_i^{e_j}((c_{iji}(h))_{j\neq i}) = \frac{1}{2}\Big[\big\{\pi_i(\rho,(c_{iji}(h))_{j\neq i}) \mid c_{iji}(h)_{j\neq i} \in (0,1)\big\} + \big\{\pi_i(\varphi,(c_{iji}(h))_{j\neq i}) \mid c_{iji}(h)_{j\neq i} \in (0,1)\big\}\Big]$$

Assuming an agent's performance not to equal one half, the two choices yield two different expected payoffs for the agent. Because the equitable payoff is the average of both of them, there will always be one action that leads to a payoff that is higher than the equitable payoff while the opposite choice will lead to a payoff that is lower. As a consequence, the agent will eventually perceive one action as kind while she will perceive the other one as unkind. Formally:

$$\pi_i^{e_j}((c_{iji}(h))_{j\neq i}) = w + \frac{1}{2}\cdot b\cdot(c_{iji}(h)_{j\neq i} + q) \Rightarrow \lambda_{iji}(\rho(h^1),(c_{iji}(h^1))_{j\neq i}) = w + q\cdot b - w - \frac{1}{2}\cdot b\cdot(c_{iji}(h^1)_{j\neq i} + q) = \frac{1}{2}\cdot b\cdot(q - c_{iji}(h^1)_{j\neq i}$$

Which action an agent perceives as kind (unkind) therefore depends on the agent's second-order belief, $c_{iji}(h)$ – what the agent believes the principal to believe about the agent's performance in the second stage.

The divisive question now is, how this second-order belief is formed. Given that the agent knows that the principal learned her productivity in the first stage, I find it most intuitive to set $c_{iji}(h) \equiv l_i^{1^{st}}$. This implies that the agent believes that the principal makes her decision expecting the agent to replicate her effort provision from the first stage. At this point, it is important to note that this belief is only reasonable for the performance-based mechanism (φ) as the choice of φ puts the agent into a similar strategic environment with identical material incentives as in the first stage. The important difference is that the principal is responsible for the subgame the agent finds herself in – but note that I assume the second-order beliefs to neglect this difference: by setting $c_{iji}(h) \equiv l_i^{1^{st}}$, I implicitly assume that the agent does not expect the principal to consider the impact of her decision on the agent's psychological payoff. (Incorporating reciprocity considerations into the second-order beliefs would, however, not affect the predictions much as I will show below.)

Alternatively, the material incentives between the first and the second stage differ starkly if the principal chooses ρ. It is therefore hard to make inferences about $c_{iji}(h^1)$. Sure, a smart principal would anticipate that the agent has no material incentive to exert effort, and assume that she exerts effort up to the point

where the absolute value of her costs of effort equal her intrinsic motivation and, perhaps, she would even take her psychological payoff into account. But would the agent believe the principal to have such elaborated beliefs about the agent's effort provision? After all, the agent knows that the principal neither has isolated information about her intrinsic motivation nor about her reciprocity parameter Y_{ij}. Due to the lack of information, I neglect the agents who are facing the random mechanism and concentrate on those who exert effort under the performance-based mechanism. Doing so, I consider ρ only as an alternative to φ which allows us to trigger emotions of kindness or unkindness because the principal's choice of φ could have been better (or worse) for an agent with a particular productivity.

Consider λ_{iji} in the branches that follow history h^2 for two agents with different productivities, $\underline{l}_n^{1^{st}} < q < \bar{l}_m^{1^{st}}$. The less productive agent n will perceive the choice of the performance-based mechanism as unkind while m will perceive it as kind because

$$\lambda_{njn}(\varphi(h^2), \underline{l}_n^{1^{st}}(h^2)) = \frac{1}{2} \cdot b \cdot (\underline{l}_n^{1^{st}} - q) < 0 \text{and} \lambda_{mjm}(\varphi(h^2), \bar{l}_m^{1^{st}}(h^2)) = \frac{1}{2} \cdot b \cdot (\bar{l}_m^{1^{st}} - q) > 0.$$

Conversely, they will perceive the choice of the random mechanism as kind and unkind, respectively. Importantly, one and the same mechanism can therefore be perceived as kind or unkind. This is the main feature of our design which we want to exploit to investigate hidden costs *and* benefits of monitoring.

As the psychological payoff is the product of Y_{ij}, $\kappa_{ij}(\cdot)$ and $\lambda_{iji}(\cdot)$, it is easy to see that a negative $\lambda_{iji}(\cdot)$ must be met by a negative $\kappa_{ij}(\cdot)$ to maximize this product if $Y_{ij} > 0$. Likewise, a positive $\lambda_{iji}(\cdot)$ must be met by a positive $\kappa_{ij}(\cdot)$. These two insights mirror the basic notion of reciprocity – tit for tat.

Putting all these pieces together, the utility function of agents who faced the performance-based mechanism looks as follows

$$U_i(l_i|\varphi) = w + b \cdot l_i + Y_{ij} \cdot [l_i - l_i^{1^{st}}] \cdot [\frac{1}{2} \cdot b \cdot (l_i^{1^{st}} - q)] - c(l_i) + \sigma \cdot l_i$$

and is solved by $l_i^* = c_l^{-1}(b + \sigma + Y_{ij} \cdot [\frac{1}{2} \cdot b \cdot (l_i^{1^{st}} - q)])$. Note that the equilibrium effort provision under the performance-based mechanism in the second stage looks similar to the first stage's equilibrium effort provision ($l_i^{1^{st}} = c_l^{-1}(b+\sigma)$). The only difference is that the perceived kindness now is a part of the first-order condition. Remember that $c_l^{-1}(\cdot)$ is assumed to be an increasing function (due to the convex cost function) such that $l_i^* > l_i^{1^{st}}$ if $l_i^{1^{st}} > q$ and if $Y_{ij} > 0$. Similarly, $l_i^* < l_i^{1^{st}}$ if $l_i^{1^{st}} < q$ and if $Y_{ij} > 0$. To put it more verbally, I predict that:

> Reciprocal agents with a productivity lower than $q = \frac{1}{2}$ perform worse in the second stage than they did before if their matched principal chooses the performance-based mechanism. That is, the principal suffers hidden costs of monitoring.
>
> Reciprocal agents with a productivity higher than q perform better in the secondstage than they did before, if their matched principal chooses the performance-based mechanism (such that the principal gains hidden benefits of monitoring).

```
include_graphics("images/00_Social_Prediction.pdf")
```

Assuming $c_l^{-1}(\cdot)$ to be a linear increasing function, these predictions are outlined in Figure XY which is based on Figure XY. The graph contains dashed and solid lines. The colored dashed lines mirror the predictions of the previous subsection which concern purely self-interested agents. The solid red line illustrates the predicted behavior of reciprocal agents who face the performance-based mechanism. Comparing the different predictions (that is, the solid and the dashed red lines) one recognizes the hidden costs to the left as well as the hidden benefits of monitoring on the right of $q = \frac{1}{2}$ (the vertical dashed line) as the solid line appears to be rotated counter-clockwise.

Consider now the case where the agent has more sophisticated second-order beliefs where she assumes the principal to be mindful of her psychological payoff and denote this second-order belief as $\tilde{l}_i$. We already know that an agent with $\underline{l}_i^{1^{st}} < q$ perceives the choice of φ as unkind and thus decreases her effort provision

(because l_i^* is increasing in $\lambda_{iji}(\cdot)$). An agent who believes that the principal anticipates this behavior would then perceive the principal's choice of φ as even less kind (or "more unkind") because she would believe that the principal believes that she would exert an effort of $\tilde{l}_i < \underline{l}_i^{1^{st}} < q$. In the end, low performances worsen the chance to receive the bonus payment (especially compared to the chances the same agent would have under the choice of ρ). This would, however, not make much sense as the agent knows that it would also be against the interest of the principal to decrease the agent's effort provision. In contrast, an agent with $\bar{l}_i^{1^{st}} > q$ considers the choice of φ as kind because it improves her chance to receive the bonus payment in the case where the psychological payoff was incorporated. If the agent believes the principal to believe that the agent would exert $q < \bar{l}_i^{1^{st}} < \tilde{l}_i$, it would result that $\lambda_{iji}(\varphi(h^2), \bar{l}_i^{1^{st}}(h^2)) < \lambda_{iji}(\varphi(h^2), \tilde{l}_i(h^2))$. Because the equilibrium effort provision increases in $\lambda(\cdot)$ a high $\tilde{l}_i$ goes hand in hand with a high l_i^*. Incorporating the psychological payoff into the second-order beliefs would therefore result either in an unreasonable belief (which might very well be replaced by $c_{iji}(h) = l_i^{1^{st}}$) or in a belief which reinforces itself.

Importantly, the original model does not incorporate the intrinsic motivation i draws from her work on the effort task. Instead, it only considers a material and a psychological payoff. The latter only depends on the material payoffs and a set of first- and second-order beliefs. Even if the intrinsic motivation is stable and not affected by the principal's (j's) choice μ, it would be difficult to incorporate σ into the fairness considerations of the psychological payoff. The problem is that the model would require the agent to form second-order beliefs about her intrinsic motivation and her equilibrium effort provision under different mechanisms to come up with $\pi_i^{e_j}$. This aggravation alone would blow up the model such that its predictive power would be reduced. Since we, as the researchers, as well as the participants do not have any isolated information about an agent's intrinsic motivation, I keep the model simple and retrain from considering the intrinsic motivation within the psychological payoff.

The most important caveat of this chapter is not that it is so rich in assumptions but, if anything, that it lacks assumptions one would need to make quantitative predictions. In particular, I made rather vague yet reasonable and therefore popular assumptions concerning the agents' costs of effort by stating that they are convex, bijective, increasing and equal to zero if the level of effort provided is zero as well. This allows me to analyze the inverse of the marginal cost function: As $c(\cdot)$ is convex and increasing, its derivative $c_l(\cdot)$ is non-negative and increasing. As a consequence, $c_l^{-1}(\cdot)$ is increasing and non-negative as well. However, I do not know (or do not assume to know) whether $c_l^{-1}(\cdot)$ is convex, linear or concave.

To understand the implication the curvature has on my predictions, imagine a concave inverse of the marginal cost function as illustrated in Figure XY.

```
include_graphics("images/09_Prediction_Problem.pdf")
```

Note that it illustrates an agent who finds herself in three different scenarios on the horizontal axis: a situation in which the agent feels treated unkindly, a situation in which she is purely self-interested (or neither treated kindly nor unkindly) as well as a situation in which she feels treated kindly (from left to right). You find the corresponding equilibrium levels of effort provision on the vertical axis where a corresponds to the unkind scenario, b to the neutral one and c to the one in which she feels treated kindly. It is easy to see that the increase of effort provision is smaller than the absolute value of the decrease, $c - b < |b - a|$, despite the fact that the perceived unkindness ($-\lambda \cdot Y_{ij}$) is exactly as strong as the perceived kindness ($\lambda \cdot Y_{ij}$).[5] The implication of this observation is that two opposing fairness perceptions of one and the same strength ($\pm\lambda \cdot Y_{ij}$) might result in two different effects that vary in their magnitude – or to put it more graphically: the red line in Figure XY could very well be concave (steeper to the left and flatter to the right) such that it looks as if it was harder to reciprocate kindness than unkindness (as I sketch it in Figure XY below).

[5] It is straightforward to imagine the cases where the inverse is linear or convex. I therefore skip further examples.

3.3 Interim Conclusion

The two previous sections have illustrated how different assumptions (pure self-interest versus reciprocity) lead to different predictions. In very broad terms, one could summarize the difference as follows: Agents who are purely self interested only care about their material payoffs while reciprocal agents, in contrast, also focus on the intentions of principals. As a consequence, self-interested agents exert the exact same effort in Stage 2 (given the performance-based mechanism) as in Stage 1 while reciprocal agents deviate.

Imagine a treatment in which an agent is matched with an artificial principal who makes random decisions. According to the model in the previous chapter, such a treatment would not allow for a non-zero psychological payoff because the agent would know that the principal would not have any intentions such that the perceived kindness would be zero. Alternatively one could argue that the agent would have a reciprocity parameter (towards the principal) of $Y_{ij} = 0$. In both cases, I would predict that the agent behaves the same way as a purely self-interested agent.

In conclusion, the actual and the hypothetical treatment are distinguished by the fact that reciprocity could potentially exist in the former treatment. To put it differently, subjects in the actual treatment are potentially *exposed* to reciprocity. Sketching a similar picture as before, Figure XY illustrates the effect of this exposure as a red-shaded area.

```
include_graphics("images/10_Treatment_Effect_Prediction.pdf")
```

If one uses the thought experiment and relies on my predictions, one would call this red area the treatment effect or the causal effect of reciprocity on performance. It seems, however, impossible to *observe* this difference, since our experimental design does not contain a treatment like the one I just described. It is the aim of the next chapter to describe how one can nevertheless *estimate* the causal effect of reciprocity to ultimately test, whether my predictions of chapter XY bear empiricism.

Chapter 4

Empirical Strategy

> It is my opinion that an emphasis on the effects of causes rather than on the causes of effects is, in itself, an important consequence of bringing statistical reasoning to bear on the analysis of causation and directly opposes more traditional analyses of causation. — (Holland 1986)

While the previous chapter derived predictions about the causes of effects, this chapter deals with the statistical measurement of effects of these causes. In particular, it describes how one can screen the data that were generated in our experiment to identify and to describe reciprocal behavior, if there is any. The corresponding Appendix XY introduces the empirical workhorse model and derives the *identifying assumptions* formally. It shows how one can make *causal* inferences using the experimental data. This chapter mainly argues why the assumptions are reasonable to make in our experimental setting and reports the strategy that results from the formal derivations.

The first strategy I present is called the *scientific solution*[1] (see Appendix XY) which is based on the idea that one can use the experiment's first stage and use it as a control condition which one then compare with the second stage as a treatment condition. In this sense, the treatment can be understood as the exposure to reciprocity. This *within*-subject design seems reasonable as the game that was played in the first stage is similar to the second stage's subgame where the agent faces the performance-based mechanism: in both (sub)games, the participants engage in the same real-effort task and are paid proportional to their effort provision. The two (sub)games only differ with respect to (1) the workload decision as well as (2) the social component: The agents' payoffs from the second stage depend on the principals' decision (and the principals' earnings depend on the agents' decisions). While the latter argument (2) is exactly what should differ between the two treatment conditions, the former (1) should not be much of a problem if the agents chose their optimal level of effort provision in the sequence *"find optimal level, then choose workload and perform accordingly"*. I believe that this is a reasonable assumption to make, especially because we designed the control questions such that participants must have understood this particular decision to answer them. Hence, they were aware of the decision's consequences and were forced to choose their effort provision at that point of time. Under this assumption, the workload decision becomes irrelevant for self-interested agents and was just a commitment device for reciprocal agents who intended to punish the principal. Consequently, the workload decision, if anything, is expected to strengthen the effect of reciprocity. In conclusion, I argue that the first stage is as good as *ceteris paribus* comparable to the respective subgame of the second stage.

We intend to interpret the observed difference between the productivity in Stage 1 and the performance in Stage 2 as the causal effect of reciprocity. To do so, we have to rule out all factors that can possibly cause a difference. Because the experiment was designed such that each within comparison is based on

[1] These postulates are the reason the section was coined as the "scientific solution" as the natural sciences proceeded far by making these assumptions. If you, for instance, throw a stone within an absolute vacuum to make inferences about the effect of the vacuum on some variable as the distance and then compare it to a comparable throw under "normal" conditions, you have to make these two assumptions. You assume that the stone would land at the same distance no matter at which point of time you throw it (moon phases do not affect anything here) and that throwing the stone in the vacuum does not change its flying characteristics for a later throw.

measurements that were conducted in the same sequence (*"first control then treatment"*), we have to rule out that time did not confound the observed difference. After all, it might be that subjects have been tired after running through the box clicking task for the first time. Alternatively, they could have improved their ability to click on boxes as fast as possible during the first stage. One could then argue that the observed difference is not driven by reciprocity, but by learning or fatigue effects. The analysis therefore depends on two postulates called *causal transience* and *temporal stability*.[2] In broad terms, they mean that the effect the control condition might have on the effort provision in any stage is reversible and that the immediate effect of the control condition is stable (that is, the same at every point in time). These two identifying assumptions are powerful because they allow me to interpret the observed differences as causal as long as they are reasonable. (Note that there is no omitted variable bias because one and the same observation is exposed to the control and the treatment condition.) Whether they are reasonable, is hard to say. The task itself was designed such that no knowledge is needed to complete it. As a consequence, there was no knowledge to gain during the first stage. Also, each participant was exposed to a short trial round which allowed them to acquire the skills even before the first stage began. In addition, there was an extensive break between the two stages due to the control questions. This gave the subjects time to recover. The problem we have is, that we cannot test whether there were learning or fatigue effects which is what qualifies them as postulates. I thus assume them to hold true and leave it to further considerations to design an experimental environment to test them.

The important question then is, whether the effect, if we observe one, matches my predictions. The following equation, while considering the specific subset of agents that were exposed to the performance-based mechanism in Stage 2 exclusively, describes the observed differences (which we intend to interpret as causal):

$$\Delta Y_u = \alpha + \beta Y_{u1} + \upsilon_u$$

I use the subscript u to denote agents of this subset and define the left-hand side as the observed differences: $\Delta Y_u \equiv Y_{u2} - Y_{u1}$. Importantly, Y_{uT} describes what I earlier referred to as the productivity (in $T = 1$) or the performance (in $T = 2$) and which I denoted as l in Chapter XY. Y_{uT} is thus, not to be confused with the reciprocity parameter Y_{ij} from the previous section.

To understand the regression expression, revisit Figure XY. ΔY_u describes the difference between the red curve and the red dashed (45°-) line and Y_{u1} constitutes the horizontal axis. The predictions describe a negative difference to the left of $Y_{u1} = 0.5$ and a positive difference on the right of this threshold. Consider now Figure XY, which illustrates the same elements as Figure XY but explains ΔY_u at the vertical axis.

```
include_graphics("images/11_OLS_Strategy.pdf")
```

Here, the red line, which I intend to estimate, is predicted to cross the horizontal axis at the threshold ($Y_{u1} = 0.5$) such that $\Delta Y_u = 0$ at this particular point. Considering the regression, this translates into a negative constant ($\alpha < 0$) as well as a positive slope of $\beta = |2 \cdot \alpha|$ (if one expects the causal effect to be linear).

While the theory of Chapter XY would be supported by data that are best described by parameters that correspond to these predictions ($\alpha < 0$ and $\beta \approx |2 \cdot \alpha|$), there are, of course, some other scenarios one can think of: If, for instance, ΔY_u (and thus α and β) equal zero at any point, one would conclude that reciprocity did not affect the working morale at all and reject my predictions. The second case, where ΔY_u is non-zero, is a little more complex to evaluate. If $\alpha \neq 0$ and $beta = 0$, one could reject the predictions as well. (In addition, one might be tempted to reject the assumptions of causal transience and temporal stability: $\alpha < 0$ together with $\beta = 0$, for instance, implies that, no matter the productivity, the agents are expected to perform worse in Stage 2 compared to Stage 1 – and this could be explained by fatigue. It could, however also mean that participant's dislike being monitored by a real person.) As a negative β stands in stark contrast to my predictions, one could conclude that my predictions turn out to be wrong if $\beta \leqslant 0$, no matter the constant.

If, however, $\alpha < 0$ and $0 < \beta < |2 \cdot \alpha|$ or $0 < |2 \cdot \alpha| < \beta$, the intersection between the horizontal and the regression line would be to the left or to the right of $Y_{u1} = 0.5$. Would that mean that my predictions were

[2] These assumptions reflect what I called "separability of effort costs" before.

wrong? Not necessarily, as this could be explained by the non-linearity that I described in the end of Section XY and in Figures XY and XY. In some cases, it might therefore become a little vague to judge whether the data actually supports my predictions or proves them wrong.

To sum up, I have argued that the first stage as well as the second stage (under the performance-based mechanism) only differ in a social dimension that I interpret as the exposure to reciprocity. Given the postulates, I suggest to run a simple OLS regression to estimate the average treatment effect of reciprocity on the agents' working morale given any observed productivity level. I furthermore indicated which realizations of α and, more importantly, of β would prove my predictions wrong and concluded that it is less clear-cut for some realizations of these parameters to judge whether they actually support the data.

If the mentioned identifying assumptions or postulates, however, are unreasonable, one has to apply the so called *statistical solution* and compare different subsets of the population of agents with each other. This section, broadly speaking, argues that one can compare agents that share similar, yet not identical, productivities with each other to make inferences about the average effect reciprocity has on their working morale. As before, I use an agent's productivity as the main explanatory variable. The treatment variable, however, is defined differently since it indicates whether agents are expected to feel treated kindly or unkindly. The agents' performance in Stage 2 will serve as the response variable in what follows.

Having that stated, an RDD intends to find a discontinuous jump around the defined threshold. Focusing on the subset of agents who faced the performance-based mechanism one expects the performance of agents with productivities marginally higher than one half to be discontinuously higher than of agents with productivities that is marginally lower. One can therefore say that the agents' measured productivity assigns them into two treatment conditions which one can call perceived kindness and perceived unkindness. The identifying assumption then is that agents, even while having some influence, are unable to precisely manipulate variable that assigns them into one of the treatment groups.

I claim that agents do not have perfect control about their productivity. I therefore argue that it is reasonable to make inferences using a RDD strategy. As this is the most important claim concerning this strategy, it deserves some support: First, note that the threshold is arbitrary. Besides its property to assign agents to their treatment condition, it has no further meaning than any of the other values in the neighborhood of q. Also, agents did not know about the importance of this particular value. They, consequently, had no incentive to deliberately manipulate their productivity correspondingly. Second, each participant worked on 25 screens with 35 boxes per screen so that q corresponds to 437.5 boxes that were clicked away in either 275 or 175 seconds. Due to the large number of boxes and the fact that they were ordered randomly[3], it seems highly doubtful that participants were aware of their score during the task. Hence, even if participants intended to manipulate their productivity to end up just above the threshold, it would have been extremely difficult for them. One might then, however, argue that participants who clicked away 438 boxes differed in some latent or omitted characteristic from those who clicked away 437 boxes. But as clean as a lab environment might be, I believe that there are still some environmental factors that affected this quantum leap-sized difference. Take the computer mice, the tables' textures, the sunlight or the air quality during the sessions as an example. On an individual level, the smallest lag of the computer, a sneezer or the sunlight that might interfere with the graphics on the screen at some corners of the laboratory can make out the difference between productive and unproductive agents. All these factors might make out the difference and cannot be controlled by the participants. So even if some are especially likely to have productivity values near one half, each of these agents would have approximately the same probability of being productive (slightly above q) or unproductive (slightly below q) – similar to a coin-flip experiment. As such, assignment into treatment is as good as random (around the threshold). Consequently, agents with a productivity of $q \pm \varepsilon$ with $\varepsilon \to 0$ are, on average, expected to be comparable – they should not differ systematically in any characteristic that could confound my analysis. Note that this assumption is not a postulate, that is, one can test whether it is reasonable. We can, for instance look at the covariates of those who are just below and just above the threshold. One also has to plot the distribution of Y_1 to spot whether the values are distributed unevenly around q. If all these variables are distributed smoothly, the identifying assumption is likely to be met.

Depending on what the data will look like eventually, a concern might be that a possible discontinuity

[3]The arrangement of boxes differed between the screens but was identical for all participants, given any specific screen.

around the threshold is unaccounted-for non-linearity. %: The jump in upper panel of Figure XY, is likely to disappear if one takes into account that the data is censored at zero. To contest such a concern one can run different specifications (including polynomials) or focus on the *"discontinuity sample"* – that is, focus on observations close to the threshold (Angrist and Lavy 1999) as explained above. As one does not need any polynomials or specifications that are complex in another sense, one can also describe the latter approach as non-parametric. Another robustness check I suggest is to run *"placebo RDDs"*. The idea here is to choose some random productivity values (maybe in advance to seeing the data) and to pretend these values to be the threshold. If the resulting RDDs detect discontinuities at these random values, one might doubt the original discontinuity to be caused by reciprocity.[4] To run these checks, I programmed a ShinyApp.

In summary, this strategy deviates from the theoretical predictions as it assumes the causal effect of reciprocity not to be a smooth function. If this was true, and if there was a causal effect in the first place, it should be identified using a non-parametric RDD as described above. Compared to the linear OLS specification, it has the advantage that it allows us to not only focus on the agents who were exposed to the performance-based mechanism, but also to narrow in on the agents who faced the random mechanism. After all, the theory from Chapter XY also applies to those subjects: they should perceive the choice of the random-mechanism as kind (unkind) if they were unproductive (productive). A second advantage is that it can, in principle, be applied even if causal transience and temporal stability are unreasonable to assume. This comes, however, at the costs that the analysis might be labeled as explorative since the discontinuity clashes with the predictions I derived earlier. In addition, there have to be enough data points in the neighborhood of q, which is not yet the case with our data.

[4] The placebo approach is often used in Diff-in-Diff Designs.

Chapter 5

Analysis

Chapter 6

Conclusion

In a simple real-effort laboratory experiment, we tested whether monitoring has hidden effects on the agents' working morale. Intention-based reciprocity models predict that unproductive agents dislike being monitored and suffer psychological costs that they pass back to the monitoring principal, even if this is costly for them. Productive agents, in contrast, are predicted to benefit from the principals' attention and to put more effort into a productive task to express their gratitude. The standard model that assumes agents to be purely self-interested yields the prediction that the agents' performance should not be affected if they were monitored in our experimental setting.

The data we gathered in the first eight sessions do not find any apparent net-costs or net-benefits of monitoring. It furthermore rejects predictions that are based on the standard model because agents who were monitored perform worse than expected. Interestingly, they do not perform any better than those agents who were not monitored and thus, did not face any material incentives to exert effort at all. While we do not find hidden benefits of monitoring, the data suggests that monitoring triggers hidden costs and that these costs are moderated by an agent's productivity.All in all, one can thus conclude that our data is neither perfectly in line with my predictions nor do they strongly dissent from them. That we do not find any hidden benefits that mirror the predicted pattern might also be a flaw of the experimental design: While it was relatively easy to restrain the labor supply in Stage 2, it was more difficult to excel (due to a kind principal or "good management practices").

But even without hidden benefits, our results contribute to understanding the adverse effects of monitoring. They support findings from the crowding-out literature only partly and demonstrate that monitoring does not trigger psychological costs *per se*: Yes, monitoring spoils some agents' working morale *but only* under the condition that they feel disadvantaged by the attention. This thesis thus objects the idea that monitoring is perceived as a lack of trust that triggers psychological costs and is reciprocated. Likewise, it casts doubt that intrinsic motivation (other than the psychological payoff discussed in this thesis) is crowded out. Instead, monitoring appears to be one of those actions that are perceived as legitimate under certain conditions while they seem unjustly otherwise. It occurs to be a management practice that requires skilled managers, who can assess who is likely to suffer from monitoring and who is unlikely to do so. Under careful considerations, a skilled manager could then minimize the hidden costs of monitoring that other (less talented) managers do not see. A nuanced employment of monitoring might then be one of the differences of successful firms that are seemingly comparable to the less successful ones.

Recall that these conclusions are based on data that is strongly influenced by only three data points. Whether additional data stemming from an identical experimental design can remove the ambiguities remains unclear. A post-hoc power analysis suggests that it might be reasonable to collect more data. However, it might make more sense to evaluate the experiment's setup to identify design flaws. Without changing the main features of the experiment, the design can easily be adjusted for future research to get a more comprehensive understanding of whether and how reciprocity affects the working morale. I conclude this thesis with several suggestions:

Comparative Statics. The theory I derived in Chapter XY is based on several exogenous factors that can be manipulated by the experimenters. This allows us to follow a comparative statics approach: We change one of these factors and observe whether the results move in the same directions as the outlined theory predicts. We could, for instance, manipulate q which I interpret as the important threshold that assigns agents to the productive or unproductive group. A variation in this parameter would thus change the definition of agents who felt treated kindly and unkindly. If the newly generated data was in line with the corresponding prediction, we would end up with further support. Likewise we could change the principals' payment function. This would affect the leverage of expressing reciprocity.[1]

Control Condition. Because we did not run a control condition with a separate set of participants, the analysis is based on postulates that allow a within-subject design. By definition, these postulates cannot be tested with our data at hand, which is why we can never be sure that they are reasonable. However, one could generate new data to either reject these postulates as unreasonable or to support them by letting participants play the first stage twice. This way, one would measure their productivity under identical conditions twice. This would allow us to test whether the ability, productivity or costs of effort (I use these terms interchangeably) are indeed separable across time. A second and more expensive approach would be to design a control treatment that is identical to the second stage except that each participant slips into the role of an agent and plays against an artificial principal. As the principal then has no intentions, reciprocity cannot emerge. The advantage of the latter suggestion is that we would end up with *ceteribus paribus* comparisons. The disadvantage is that it prunes observations (because we are not yet interested in the behavior following the choice of the random mechanism).

Comprehension. The experiment was framed in a neutral and thus abstract way. As it took more than 40 minutes for some participants to read (and hopefully understand) the instructions to answer the control question, one can reasonably suspect that not all of the participants understood the strategic environment they later found themselves in. Without changing the neutrality of the instructions' framing, one can adjust the instructions in at least two ways: First, one can conduct the sessions in a laboratory that manages a native subject pool and translate the instructions into the corresponding language. Second, the instructions could be framed a little more abstract, yet more visually. One could, for instance, describe the performance-based mechanism as a process in which a ball is drawn from a bin that contains red and green colored balls. If the drawn ball is red, the agent receives 225 DKK and 150 DKK otherwise. The performance of the agent then determines how many green balls are located within the container.

Regression Discontinuity Dedign. A fourth suggestion applies to the less conclusive empirical strategy applied in this thesis – the regression discontinuity design. One can argue that it did not yield any insights because the theory did not predict any discontinuities that could potentially be exploited. More importantly, there is not enough data around the threshold that could be exploited. The scatterplots indicate that we can influence the agents' productivity fairly well by manipulating the time each screen is displayed.[2] This means that we could manipulate the screen time such that there are many observations around the threshold. In addition, we could re-design the material incentives such that we would expect a discontinuous jump around the threshold. Because the discontinuity should only affect the psychological payoff and not the material payoff, this becomes a little more tricky however: If we changed the payoffs following the performance-based mechanism, for instance, we would design the material incentives so that they are discontinuous. One could then argue that a discontinuity in the observed behavior was predicted by the standard model as well and necessarily caused by reciprocity. If we only changed the material payoff of the random mechanism instead, we would indeed alter the perceived kindness without touching the material incentives of the performance-based mechanism. However, we would end up without the predicted discontinuity, given the fairness norm the standard model assumes. Hence, to predict a discontinuity, we might have to adjust the fairness norm correspondingly. Anotherdownside of this approach is that it complicates the interpretation of the principal's choice as monitoring even further.

Expression of Kindness The current design allows agents to reduce their workload. We have seen that this is a powerful tool as none of those agents who chose to work on the maximum amount of screens actually

[1] In the extreme case, a principal's earnings were not affected by the agent's performance. This would resemble the "artificial principal" suggestion I made above but would be even more expensive as we also had to pay the principal.

[2] If you focus on the productive agents in Figure XY you will see that they are scattered around $Y_1 \simeq 0.6$.

decreased their effort provision. It might therefore be a commitment device to follow through on impulsive and reciprocal strategies. While it was easy to reduce the performance by not supplying any effort, it might have been hard to increase it by the same amount (as illustrated in Figure XY. Even determined agents might thus benefit from the opportunity to increase their workload as a response to the principal's choice. It is not clear-cut what the standard model would predict the productive agents to do in this case.[3]. But as the standard model would predict all agents to behave similarly, it would stand in contrast to the intention-based reciprocity model, which would predict none of the unproductive agents to increase their workload. Because the design would make it easier for agents to express kindness ("on the right side of the threshold") while it does not affect the expression of unkindness ("on the left side of the threshold") the regression line depicted in Figure XY is expected to become steeper. In fact, I already implemented this suggestion into the code which can be found in the online Appendix. To avoid ambiguities with respect to the standard model's predictions, one could adjust the experiment a little further: One could design the extended workload such that it does not affect the agents' prospects to earn the bonus payment. In a real-world setting, one could translate such a design as unpaid overtime.[4]

Angrist, Joshua D, and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *The Quarterly Journal of Economics* 114 (2). MIT Press: 533–75.

Barkma, HarryG. 1995. "Do Top Managers Work Harder When They Are Monitored?" *Kyklos* 48 (1). Blackwell Publishing Ltd: 19–42. https://doi.org/10.1111/j.1467-6435.1995.tb02313.x.

Bartelsman, Eric J., and Mark Doms. 2000. "Understanding Productivity: Lessons from Longitudinal Microdata." *Journal of Economic Literature* 38 (3): 569–94. https://doi.org/10.1257/jel.38.3.569.

Bénabou, Roland, and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *The Review of Economic Studies* 70 (3): 489–520. https://doi.org/10.1111/1467-937X.00253.

Bloom, Nicholas, and John Van Reenen. 2007. "Measuring and Explaining Management Practices Across Firms and Countries*." *The Quarterly Journal of Economics* 122 (4): 1351–1408. https://doi.org/10.1162/qjec.2007.122.4.1351.

Dickinson, David, and Marie-Claire Villeval. 2008. "Does Monitoring Decrease Work Effort?: The Complementarity Between Agency and Crowding-Out Theories." *Games and Economic Behavior* 63 (1). Elsevier: 56–76.

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. "A Theory of Sequential Reciprocity." *Games and Economic Behavior* 47 (2). Elsevier: 268–98.

Falk, Armin, and Michael Kosfeld. 2006. "The Hidden Costs of Control." *American Economic Review* 96 (5): 1611–30. https://doi.org/10.1257/aer.96.5.1611.

Foss, Nicolai J. 2003. "Selective Intervention and Internal Hybrids: Interpreting and Learning from the Rise and Decline of the Oticon Spaghetti Organization." *Organization Science* 14 (3): 331–49. https://doi.org/10.1287/orsc.14.3.331.15166.

Frey, Bruno S. 1993. "Does Monitoring Increase Work Effort? The Rivalry with Trust and Loyalty." *Economic Inquiry* 31 (4). Wiley Online Library: 663–70.

Frey, Bruno S., and Felix Oberholzer-Gee. 1997. "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding- Out." *The American Economic Review* 87 (4). American Economic Association:

[3] One could argue that the additional workload makes it easier for the agents to exert effort. This would translate into lower costs of effort and a higher equilibrium effort provision. In contrast, one could also argue that the equilibrium (predicted by the standard model) should not be affected by an extended workload as the agents already supplied their optimal level of effort.

[4] Another approach might be to not only think about the quantity of the agent's labor supply but also about its *quality*. Suppose that the principal sells the agent's labor supply in the form of some good. The higher the quality of the good, the higher the principal's earnings. This means that we could give the agent the possibility to determine the amount of money the principal earns with each percentage point of boxes the agent clicked away. Agents who intend to reciprocate kindness but fail to provide more effort than in the first stage could then easily express their kindness by increasing the quality (worth) of their effort. This would also increase the ease of expressing unkindness. However, this adjustment might make the estimation of reciprocity more blurry if agents choose qualities and quantities that offset each other. In addition, one has to think about the agents' costs of the quality choice so that one can translate it into a convincing story.

746–55.

Gibbons, Robert, and John Roberts. 2012. *The Handbook of Organizational Economics*. Princeton University Press.

Greiner, Ben. 2015. "Subject Pool Recruitment Procedures: Organizing Experiments with Orsee." *Journal of the Economic Science Association* 1 (1). Springer: 114–25.

Guerra, Gerardo A. 2002. "Crowding Out Trust: The Adverse Effects of Verification. An Experiment." Economics Series Working Papers 98. University of Oxford, Department of Economics.

Halac, Marina, and Andrea Prat. 2016. "Managerial Attention and Worker Performance." *American Economic Review* 106 (10): 3104–32. https://doi.org/10.1257/aer.20140772.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396). [American Statistical Association, Taylor & Francis, Ltd.]: 945–60.

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler. 1986. "Fairness as a Constraint on Profit Seeking: Entitlements in the Market." *The American Economic Review* 76 (4). American Economic Association: 728–41.

Masella, Paolo, Stephan Meier, and Philipp Zahn. 2014. "Incentives and Group Identity." *Games and Economic Behavior* 86: 12–25. https://doi.org/https://doi.org/10.1016/j.geb.2014.02.013.

Okun, Arthur M. 2011. *Prices and Quantities: A Macroeconomic Analysis*. Brookings Institution Press.

O'Neil, Cathy. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

Schnedler, Wendelin, and Radovan Vadovic. 2011. "Legitimacy of Control." *Journal of Economics and Management Strategy* 20 (4): 985–1009.

Schulze, Günther G., and Björn Frank. 2003. "Deterrence Versus Intrinsic Motivation: Experimental Evidence on the Determinants of Corruptibility." *Economics of Governance* 4 (2): 143–60. https://doi.org/10.1007/s101010200059.

Sebald, Alexander. 2010. "Attribution and Reciprocity." *Games and Economic Behavior* 68 (1). Elsevier: 339–52.

Siemens, Ferdinand A. von. 2013. "Intention-Based Reciprocity and the Hidden Costs of Control." *Journal of Economic Behavior and Organization* 92 (Supplement C): 55–65. https://doi.org/https://doi.org/10.1016/j.jebo.2013.04.017.

Syverson, Chad. 2011. "What Determines Productivity?" *Journal of Economic Literature* 49 (2): 326–65. https://doi.org/10.1257/jel.49.2.326.