FACULTY OF SOCIAL SCIENCES
Department of Economics
University of Copenhagen

Master's Thesis

Hauke Carsten Roggenkamp

The (Hidden) Benefits of Monitoring

How Managerial Attention Affects the Employees' Working Morale

Supervisor: Alexander Christopher Sebald

ECTS points: 30

Date of submission: 11/02/2018

134,702 Keystrokes ≈ 57 Standard pages

Abstract

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. A worker's 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.

Keywords: Intention-based reciprocity, Monitoring, Managerial attention, Hidden costs of control

Acknowledgments

No thesis writes itself. Especially not this one. During the whole process of designing, programming and writing it, I was standing on the shoulders of a handful of people whose support was invaluable to me. First and foremost, I want to express my profound gratitude to my supervisor, Alexander Sebald, whose office —despite any delay— was always open whenever I ran into a trouble spot or had a question about my research or writing. Patiently, he gave me the freedom to delve deep into the various elements of this research project, but steered me in the right the direction whenever he thought I needed it. Most importantly, I want to thank him for his trust and for allowing me be a part in this research project. A special thanks goes to Andreas Gotfredsen, who not only shared his software with me but also spent a lot of time adjusting it to our needs and answering the many questions I had. It is fair to say that this thesis would not exist without him. I also owe a special thanks to Änne and Kett as well as the staff of the Centre for Experimental Economics in Copenhagen for helping me to run the experimental sessions as smooth as possible. I am grateful to Frauke for her thought-provoking impulses. Her sharp comments helped me to get back on track whenever I lost the bigger picture. I want to thank her as well as Ben and Christine for their encouragement, patience and attention. Thank you to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of writing this thesis. This accomplishment would not have been possible without you.

Statutory Declaration

Herewith, I solemnly declare that unless otherwise indicated in the text or references, or acknowledged, this thesis is entirely the product of my own scholarly work. Any inaccuracies of fact are my own and accordingly I take full responsibility. This thesis was not used in the same or in a similar version to achieve an academic grading or is being published elsewhere.

Hauke C. Roggenkamp

Style

Because the empirical project on which the thesis is built was conducted in a group consisting of Alexander Sebald, Georg Kirchsteiger and myself, I cannot do otherwise but refer to the group and to myself. I therefore use personal pronouns (even though it seems not to be common in the economics literature) and use “I” when I mean “I” and “we” when I mean “we”.

Note that parts of the following analysis were pre-specified. Likewise, parts of this thesis were written before I had data at hand — not only because of the time management, but, more importantly, due to the aim of an unbiased writing process. As some last-minute adjustments were necessary, I had to rewrite the thesis after seeing the data. This is why parts of the thesis sound ambiguous about the results and the data (such as parts of the empirical strategy) while chapters (such as the conclusion) refer to specific results. Thus, I ask the reader to excuse the variation in style.

As I do not expect the reader to go through this thesis in one session I tried increase the convenience by writing the chapters as “stand-alone” sections where one does not have to search the rest of the document for the necessary details. This is why you might encounter some repetitions.

Narrative

Imagine working on an assembly line as a summer job performing a tedious task repeatedly. At the end of the summer you will be paid a fixed salary. On top of that you might get a bonus payment. Your supervisor has to decide whether she will pay you the bonus. To make this decision, she can either monitor you to get an impression of your work or flip a coin. If she monitors you, she will take the impression you made as a basis: The better the impression, the more likely it is that you will receive the extra money.

As a motivated and skillful handyman who knows that you are delivering a good working performance, you might want her to pay as much attention as possible to ensure that she is aware of your performance. If you notice that she decided not to monitor you, you might be afraid that you are not being recognized and that you will end up without the bonus. On the contrary, you'll be happy to see her not monitoring you if you were a clumsy lazy man. After all, a 50% chance is not bad given that it is a higher likelihood of receiving the bonus than you could expect if you were monitored. To put it differently, if you were clumsy and she was paying attention to your work, you'll realize that your chances of a bonus are relatively low.

Since you know that her salary depends on the quantity of items you screwed together on the assembly line, what would you do as a response to her monitoring decision?

Contents

List of Figures

List of Tables

Introduction

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 (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, p. 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

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

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 et al., 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 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.[2] 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.

The thesis is organized as follows: I provide details of the experimental design in Chapter 2 before I discuss behavioral predictions in Chapter 3. Chapter 4 describes the empirical strategy and Chapter 5 reports the pre-specified analysis[3] before Chapter 6

[2] 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.

[3] You can retrieve the corresponding analysis plan as well as some additional information following this link: `https://howquez.github.io/The-hidden-Benefits-of-Monitoring/`.

concludes.

Design

Our experiment consisted of two stages and a questionnaire which you can find in Appendix B. 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.[4] 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 2.1, where i denotes any participant and l her effort provision (or "labor supply"). The player 0 is an artificial and thus uninterested[5] player to whom one can also refer as *"chance"* as she only conducts

[4] 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.

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

an explicit randomization subject to i's effort provision.

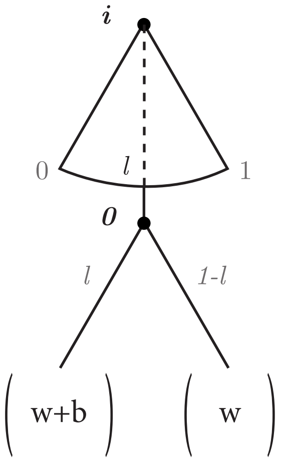

Figure 2.1: Experimental Design—Stage 1

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).

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.[6] This is important as the first

[6] 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.

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 2.2 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 2.2. 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, ($q = 1/2$).

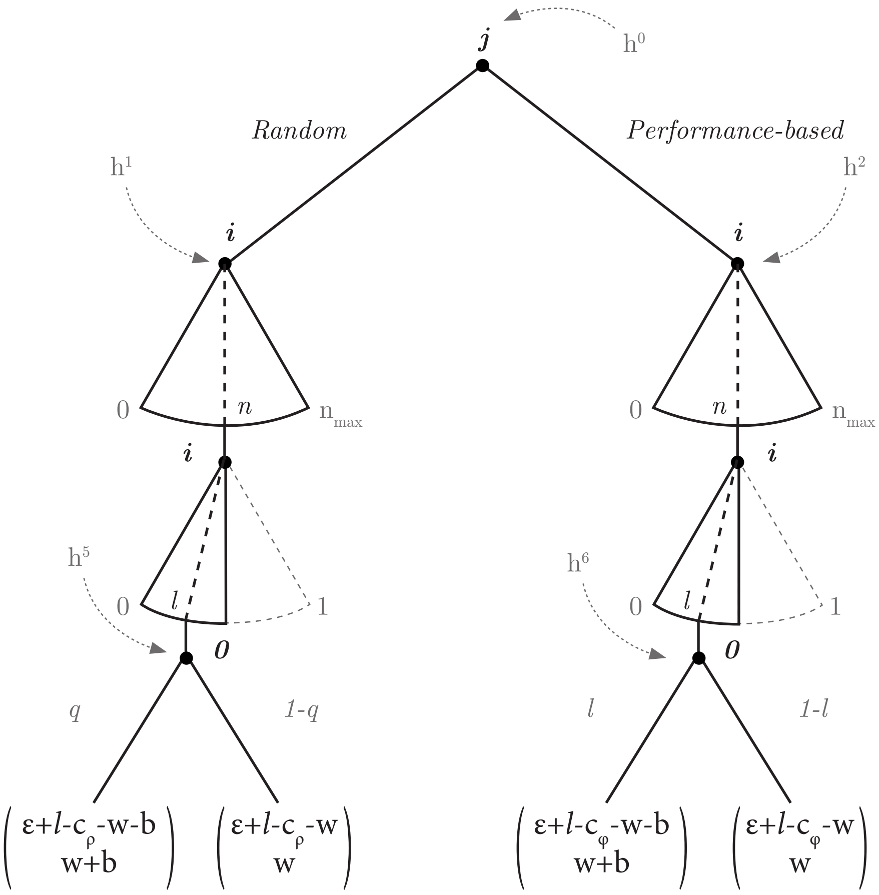

Figure 2.2: Experimental Design—Stage 2

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 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 2.3 for (another) visual representation of the the second stage's timeline.

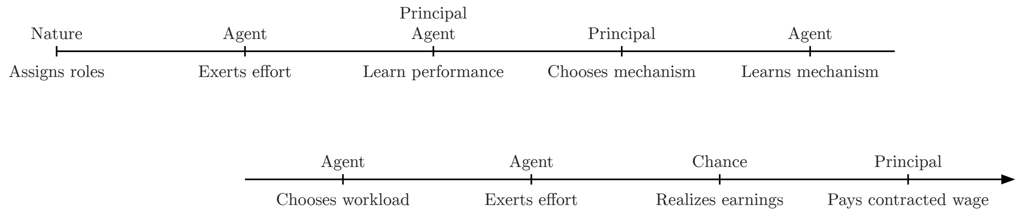

Figure 2.3: Sequence of Events—Stage 2

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[7] 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.

[7] 6.92 to be more precise.

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 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[8]. 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.[9] 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[10] subjects who participated in the experiment were students from Copenhagen based Universities studying various majors.[11] Using ORSEE (Greiner, 2015), we recruited students with little experience, that is, who had participated in no or few experiments before. As a result, the median experience in economic[12] experiments was

[8] 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".

[9] 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.

[10] 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 C (without commenting them).

[11] 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.

[12] 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

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.[13] 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.

psychological experiments before.

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

Behavioral 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 2.2 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 $q \equiv {}^1\!/_2$. 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 3.1. 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.

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. Her costs are assumed to be bijective, convex and increasing in l with $\frac{\partial c}{\partial l} > 0$, $\frac{\partial^2 c}{\partial l^2} > 0$ and $c(0) = 0$. On top of that, I conveniently assume risk-neutrality.

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.

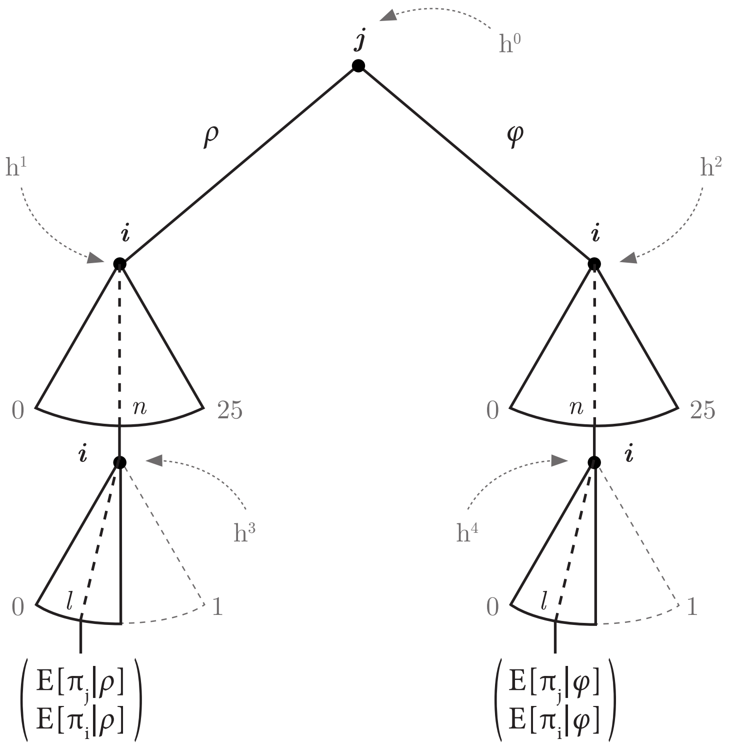

Figure 3.1: Reduced Form —Stage 2

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:

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

$c_l(\cdot)$ 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).

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

These two functions, along with the principal's expected monetary payoff, are visualized in an interactive *ShinyApp* I programmed and archived here[14].[15] Adding the costs of effort as well as the intrinsic motivation yields the following first-order conditions:

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

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 $c_l^{-1}(\cdot)$

[14] `https://roggenkamp.shinyapps.io/shiny_expectations/`

[15] 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.

is increasing and because $b > 0$, it follows that $l^*_\varphi \geqslant l^*_\rho$. In summary, I predict that:

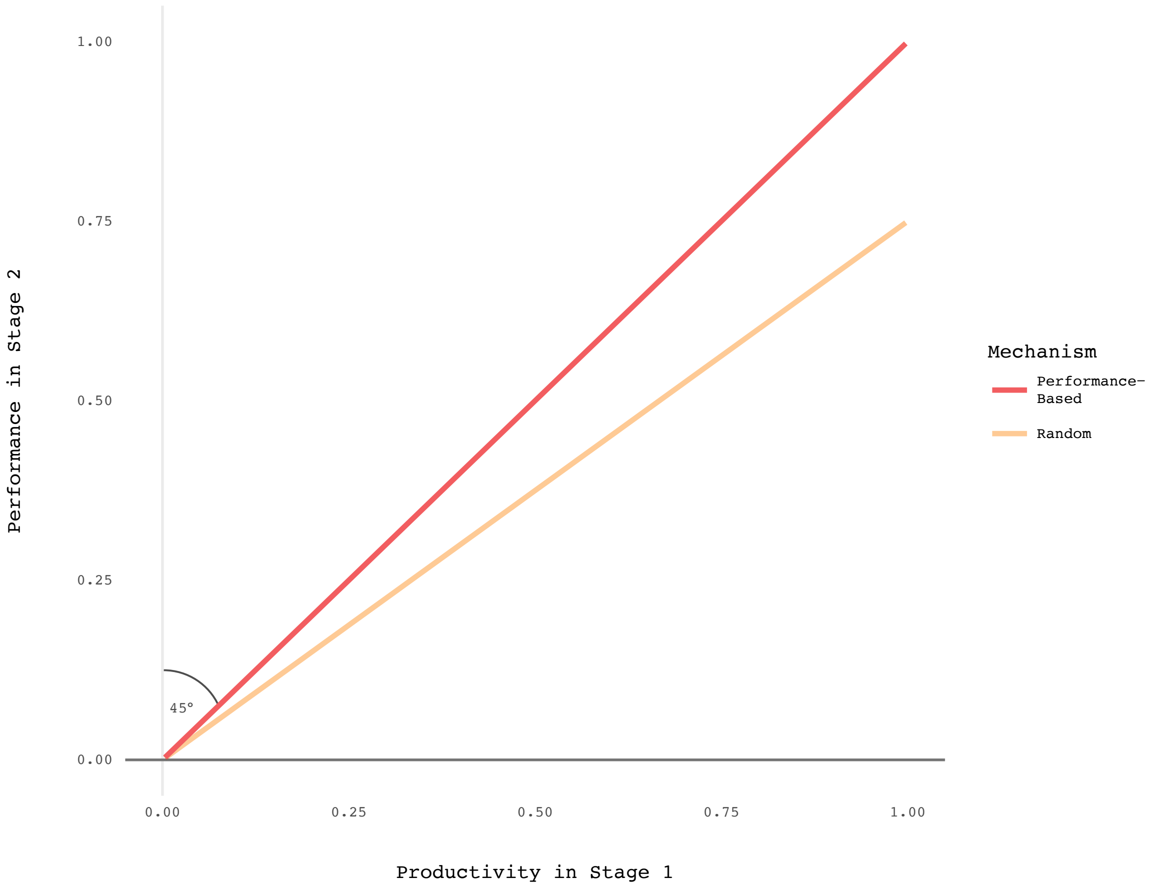

Figure 3.2: Predicted behavior for purely self-interested agents

Prediction 1 *A purely-self interested agent's performance, given a performance-based mechanism, will equal her productivity.*

Prediction 2 *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.*[16]

These predictions are conceptualized in Figure 3.2 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

[16] 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.

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.[17]

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 (2010) model would allow me to do. I focus on the expected outcomes as illustrated in Figure 3.1 and omit the chance player.

Like Dufwenberg and Kirchsteiger (2004) as well as Sebald (2010), I denote b_{ij}

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

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:

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

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

$$\pi_j^{e_i}((b_{ij}(h))_{j\neq i}) = {}^1\!/\!_2\Big[max\big\{\pi_j(a_i(h),(b_{ij}(h))_{j\neq i}) \mid a_i(h) \in (0,1)\big\} + min\big\{\pi_j(a_i(h),(b_{ij}(h))_{j\neq i}) \mid a_i(h) \in (0,1)\big\}\Big]$$

which basically means that the equitable payoff is a virtual average payoff that i can grant j.[18] 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 (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 (2004) definition of an equitable payoff does not make much sense since it would translate into an equitable payoff resulting from a performance of ${}^1\!/\!_2 \cdot (0 + 1)$ 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 ${}^1\!/\!_2 \cdot (0 + l_i^{1^{st}})$ 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

[18] 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.

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, $l(h)$:

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

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.[19] 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.

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

[19] Consider, for instance, the situation at h^1. At h^1, an agent know that the principal chose the random mechanism such that $b_ij(h^1) = \rho$. The hypothetical agent also knows that her effort provision does not affect her own earnings as they are determined by a fair coinflip (q). She can thus infer that $\mathbb{E}[\pi_j] = \varepsilon + l - w - q \cdot b - c(\rho)$. In this scenario, the agent can only affect the principal's material payoff via l which means that her effort provision is the only channel to express (un)kindness. The best the agent can do for the principal is an effort provision of $l = 1$ while the lowest material payoff she can grant the principal is $l = 0$. It follows that $\pi_j^e(\rho(h^1)) = {}^1/_2\big[\varepsilon + 0 - w - q \cdot b - c(\rho) + \varepsilon + 1 - w - q \cdot b - c(\rho)\big]$ such that $\pi_j^e(\rho(h^1) = \varepsilon - w - q \cdot b - c(\rho) + {}^1/_2$. To evaluate her own kindness towards the principal, the agent would subtract this value from the material payoff she eventually grants her$(\varepsilon + l - w - q \cdot b - c(\rho))$ which yields $\kappa_{ij}(l(h^1),\rho(h^1)) = \varepsilon + l - w - q \cdot b - c(\rho) - \varepsilon + w + q \cdot b + c(\rho) - {}^1/_2 = l - {}^1/_2$. As one can see, everything cancels out, except for the difference between her effort provision and the fairness norm $(l - {}^1/_2)$.

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 $\mathcal{A}_j = \{\rho, \varphi\}$.

$$\begin{aligned}\pi_i^{e_j}((c_{iji}(h))_{j\neq i}) = {}^1\!/\!_2\Big[&\big\{\pi_i(\rho, (c_{iji}(h))_{j\neq i}) \mid c_{iji}(h)_{j\neq i} \in (0,1)\big\} \\ &+ \big\{\pi_i(\varphi, (c_{iji}(h))_{j\neq i}) \mid c_{iji}(h)_{j\neq i} \in (0,1)\big\}\Big]\end{aligned}$$

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:

$$\pi_i^{e_j}((c_{iji}(h))_{j\neq i}) = w + {}^1\!/\!_2 \cdot b \cdot (c_{iji}(h)_{j\neq i} + q)$$

$$\begin{aligned}\Rightarrow \lambda_{iji}(\rho(h^1), (c_{iji}(h^1))_{j\neq i}) &= w + q \cdot b - w - {}^1\!/\!_2 \cdot b \cdot (c_{iji}(h^1)_{j\neq i} + q) \\ &= {}^1\!/\!_2 \cdot b \cdot (q - c_{iji}(h^1)_{j\neq i})\end{aligned}$$

$$\begin{aligned}\Rightarrow \lambda_{iji}(\varphi(h^2), (c_{iji}(h^2))_{j\neq i}) &= w + c_{iji}(h^2)_{j\neq i} \cdot b - w - {}^1\!/\!_2 \cdot b \cdot (c_{iji}(h^2)_{j\neq i} + q) \\ &= {}^1\!/\!_2 \cdot b \cdot (c_{iji}(h^2)_{j\neq i} - q)\end{aligned}$$

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

$$\lambda_{njn}(\varphi(h^2), \underline{l}_n^{1^{st}}(h^2)) = 1/2 \cdot b \cdot (\underline{l}_n^{1^{st}} - q) < 0$$

and

$$\lambda_{mjm}(\varphi(h^2), \bar{l}_m^{1^{st}}(h^2)) = 1/2 \cdot b \cdot (\bar{l}_m^{1^{st}} - q) > 0.$$

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

$$\begin{aligned}U_i(l_i|\varphi) = {} & w + b \cdot l_i \\ & + Y_{ij} \cdot [l_i - l_i^{1^{st}}] \cdot [{}^1\!/\!_2 \cdot b \cdot (l_i^{1^{st}} - q)] \\ & - c(l_i) \\ & + \sigma \cdot l_i\end{aligned}$$

and is solved by $l_i^* = c_l^{-1}(b + \sigma + Y_{ij} \cdot [{}^1\!/\!_2 \cdot b \cdot (l_i^{1^{st}} - q)])$. 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:

Prediction 3 *Reciprocal agents with a productivity lower than $q = {}^1\!/\!_2$ 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.*

Prediction 4 *Reciprocal agents with a productivity higher than q perform better in the second stage than they did before, if their matched principal chooses the performance-based mechanism (such that the principal gains hidden benefits of monitoring).*

Assuming $c_l^{-1}(\cdot)$ to be a linear increasing function, these predictions are outlined in Figure 3.3 which is based on Figure 3.2. 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 $q = {}^1\!/\!_2$ (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

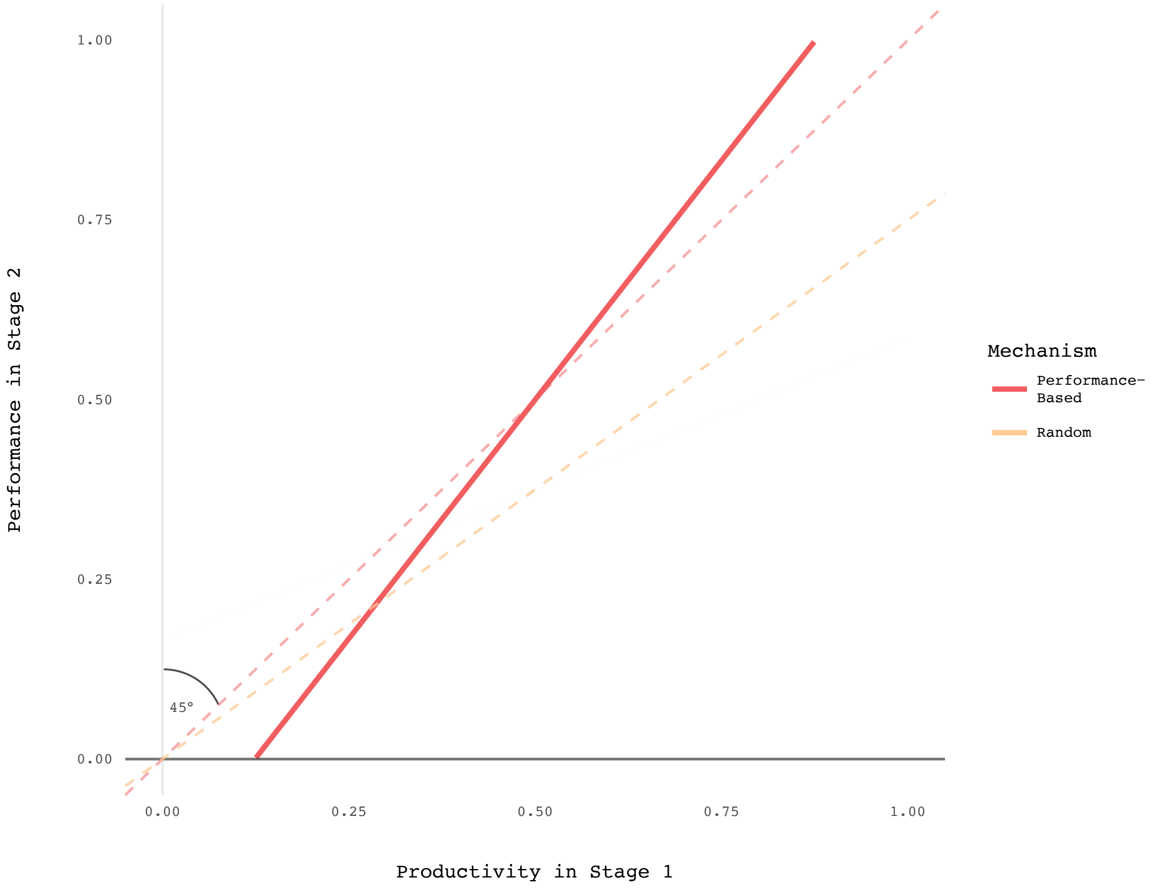

Figure 3.3: Predicted behavior for reciprocal agents

$\lambda_{iji}(\cdot))$. 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 3.4. 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,

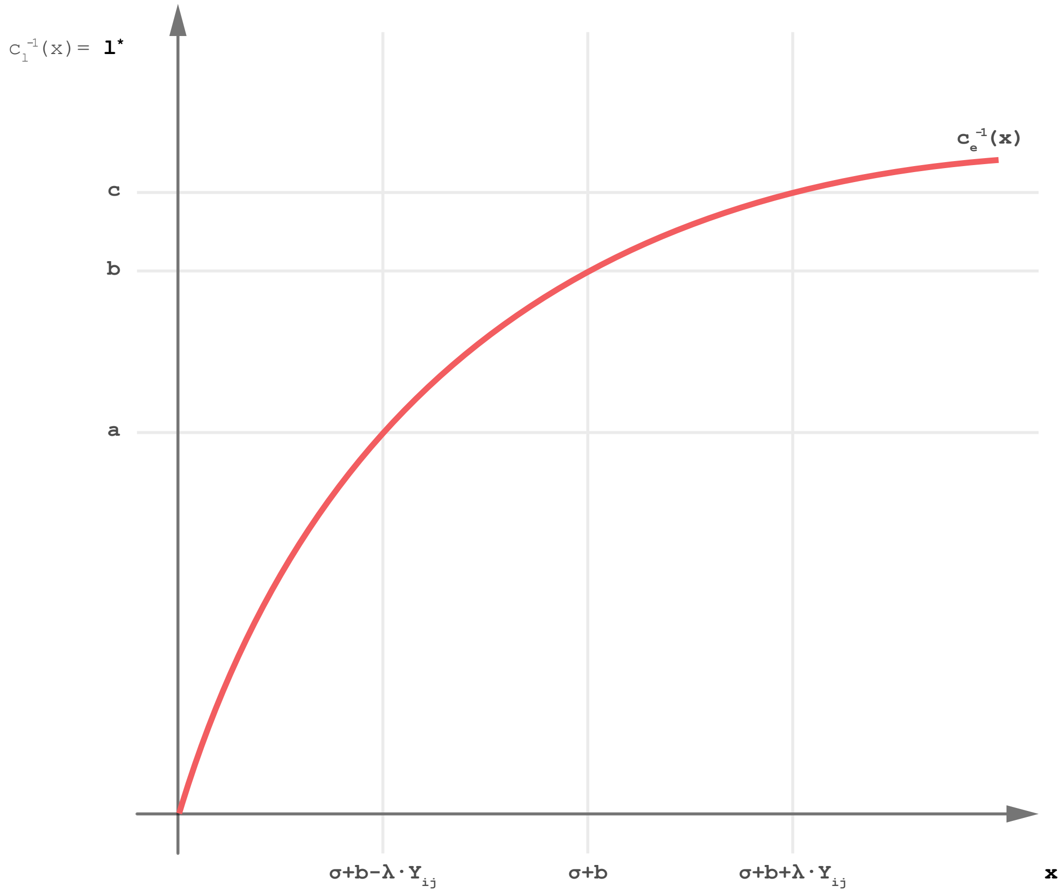

Figure 3.4: Hypothetical inverse of the marginal cost function

$c - b < |b - a|$, despite the fact that the perceived unkindness $(-\lambda \cdot Y_{ij})$ is exactly as strong as the perceived kindness $(\lambda \cdot Y_{ij})$.[20] 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 3.3 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 3.5 below).

[20] 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 3.5 illustrates the effect of this exposure as a red-shaded area.

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 3.2 bear empiricism.

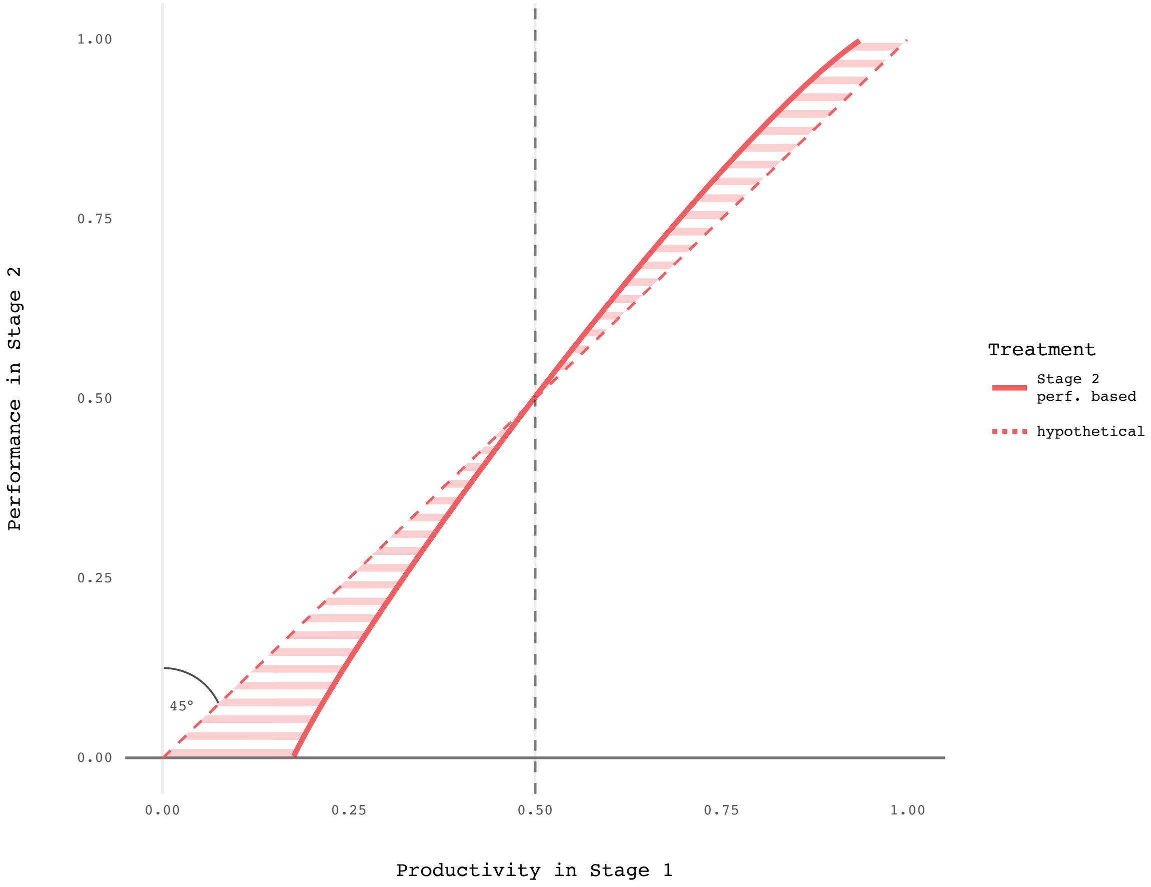

Figure 3.5: Illustration of the causal effect of reciprocity on the agents' performances

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, p. 947)

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 A 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*[21] (see Appendix A.2) 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:

[21] 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.

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 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.*[22] 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

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

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 3. 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 3.5. 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 4.1, which illustrates the same elements as Figure 3.5 but explains ΔY_u at the vertical axis. 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 3 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

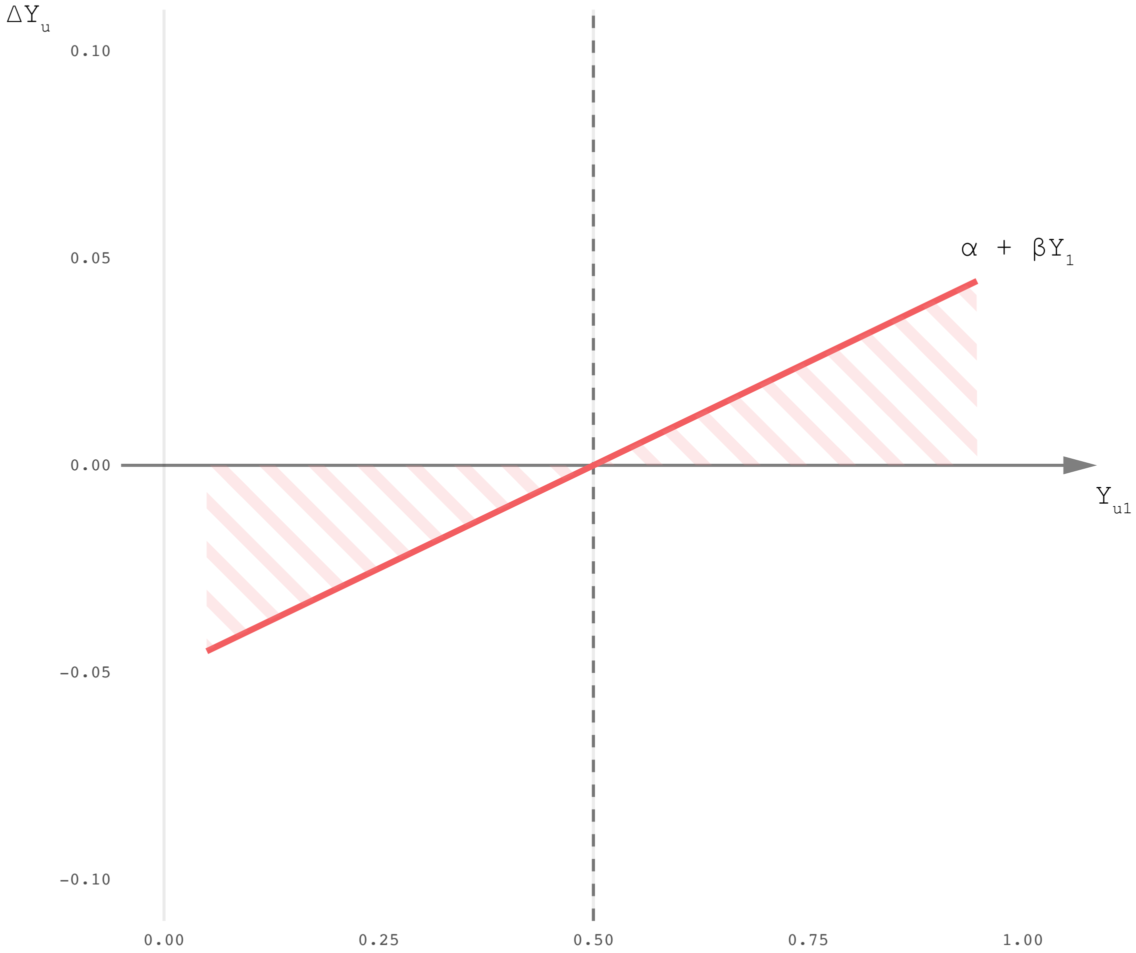

Figure 4.1: Predicted treatment effect for reciprocal agents

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 wrong? Not necessarily, as this could be explained by the non-linearity that I described in the end of Section 3.2 and in Figures 3.4 and 3.5. 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.

Because one uses different subsets of agents to observe differences, the main empirical concern is an omitted variable bias. That is, an over- or understated effect which is driven by the causal effect of another variable, that correlates with the control variables but is not measured and thus omitted. In fact, we did not ask the subjects for demographics and have no perfect measure for their ability. However, I argue that it is, in principle, possible to estimated the causal effect of reciprocity anyways. To do so, I run a non-parametric regression discontinuity design (RDD).

I thereby exploit the threshold of $q = {}^1\!/\!_2$ which is the most important parameter for our predictions. Recall, that an agent who faced the performance-based mechanism was predicted to feel treated kindly if her productivity measured in Stage 1 was higher than this threshold. This agent is predicted to reciprocate the perceived kindness with a high performance in Stage 2. The exact opposite is predicted for agents with productivities lower than q. As such, a productivity of exactly one half is an ideal threshold for the application of RDDs. The only caveat is that I do not predict discontinuities. An agent with a productivity marginally lower or higher than one half should reciprocate only mildly. This is why one can label the RDD strategy as exploratory. But as the strategy was pre-specified, I consider it as useful and unproblematic.

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[23], 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

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

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 around the threshold is unaccounted-for non-linearity. 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.[24] To run these checks, I programmed a ShinyApp[25] that allows the reader to zoom in on the threshold and to set arbitrary thresholds.

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.

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

[25] See `https://roggenkamp.shinyapps.io/ShinyRDD/`

After all, the theory from Chapter 3 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.

Analysis

"If we torture the data for long enough, they will confess."
— Statisticians' saying

This chapter mainly presents the results and concludes that, given the current number of observations, there is no reason to expect hidden benefits of monitoring. Hidden costs (that is, a spoiled working moral), however, are likely to prevail. An important question that remains unanswered is, whether the agents' productivity moderates these costs. One (rather dubious) subset strongly suggests that the hidden costs only emerge if the agents are unproductive.

As I intended to pre-specify the whole analysis, I will comment on this topic in a section I named *research integrity.* Most of the analysis was designed *after* the data became available so that a discussion of pre-specification lacks practical relevance for this thesis. I nonetheless comment on this topic as it is an interesting and important step towards more reliable empirical research. The subsequent sections will then present the results of the analysis that were derived above.

5.1 Research Integrity

> *I am aware of the fact that this thesis is no platform to communicate this project's history of thought. Nevertheless, I want to emphasize that the share of pre-specified analyses (relative to the whole analysis body) should have been larger than it eventually is. In other words: it was planed to be an important asset of this thesis and took a lot of consideration. Due to a last-minute change in the experimental design, however, parts of my archived R scripts became useless (even though they are functioning). While the current version of this thesis is focusing on the second stage performance of the agents who were exposed to a performance-based mechanism, the previous design led to predictions and an analysis that also considered agents facing the random mechanism. The design's new features no doubt improved the experimental design immensely and pointed towards some weaknesses in the previous empirical strategy. These improvements, however, come at some costs: Unfortunately, I was not able to adjust the code to the new design. As a consequence, I had to discard many of the analyses I pre-specified, which makes the results less credible than intended.*

The many analyses of Lalonde's (1986) data[26] illustrate that different modeling assumptions and specifications can yield very different alleged *"causal"* inferences.[27] As it is easier to publish significant results, researchers have an incentive to try their preferred specification first, change the control variables in a second run, try a different functional form in the third, vary the measure of the dependent variable in the fourth run, try a different subset and so on until they find a significant effect (Ho et al., 2007). Likewise, there are no incentives to report the different analyses one conducted until one found an effect. This is a major problem because it leads to a biased pool of *"findings"* in many scientific disciplines. As a consequence, it becomes harder and sometimes impossible to judge whether the results one finds in journal articles are robust and replicable.[28] Especially because the robustness checks, that are conducted to ensure

[26] See for instance Ashenfelter (1978), Dehejia and Wahba (2002, 1999), Smith and Todd (2005).

[27] See this fivethirtyeight article for a nice illustration, where you can *'Hack your way to glory'*: `https://fivethirtyeight.com/features/science-isnt-broken/`

[28] In addition, there is also a problem inherent in the methods we use: false positives. These are simply a result of chance and as such not affected by the researchers themselves. Eventually, however,

that the analysis is free of p-hacking attempts might be hand picked as well.

One expensive way to quantify the magnitude of the problem within the experimental literature is to replicate published findings. Doing exactly this, Camerer et al. (2016) found a significant effect in the same direction as in the original studies in only about 60% of their replications. This indicates that more than one in three of the considered studies were not replicable. Given this problem concerning the reproducibility of results in behavioral sciences and the resulting credibility problem, open access and pre-specifications appear to be one good approach to reduce the degree of researcher digression (or *"researcher degrees of freedom"*) described above. I do not claim that these approaches eliminate the possibility of reporting false-positives. It only restrains the researcher from searching (un-)consciously for significant results that are not there — it thus keeps us away from (deliberately) fabricating false-positives.

Prior to looking at the data from the actual experiment, I analyzed a simulated test data set and publicly archived the resulting analysis online.[29] The archive also enables the interested reader to browse through the files' editing process and shows, which changes were made at which point in time. Most importantly, the archive also contains a link to the eventual analysis, that is, the analyses I ran after we changed the design. Once the reader has the data at hand, the whole analysis, including tables and figures, is easy to reproduce. As the previous chapter already indicated, I am applying different strategies and analyses, which yield different results. The critical reader might object that interpreting these results still gives me some degree of freedom. But since all of the results are reported or easy to reproduce, the reader can make her own informed decisions regarding the quality and credibility of these results.

I follow most of Simmons et al.'s (2011) requirements of transparent research. In particular, I published the rule for terminating data collection before the data collection began and reported this rule; I list all variables that were collected; I report a previous design of this experiment, that was tested 18 months earlier and did not lead to the expected results; I use different subsets of the data but also report the coefficients of the complete data set.

To put it in a nutshell, I was hoping to provide a good analysis in the sense of Murphy

they are over represented in scientific journals because they might appear more surprising and thus, more interesting.

[29] You can retrieve the corresponding files following this link: `https://howquez.github.io/The-hidden-Benefits-of-Monitoring/`

(1993)[30] when I registered the pre-specified code. That is, I aimed to achieve a high level of consistency, which is better understood as honesty: I wanted to credibly signal that I ran the best analysis I could think of at that time without modifying the analysis in some non-transparent way before being presented to the public. This procedure would not have affected the accuracy of the analysis since it allows for adjustments which would have been indicated in the text. I thereby tried to follow the customs widely applied in medical randomized controlled trials which are rarely seen in (experimental) economic research.[31] Since we modified the experimental setup while I did not adjust the analysis, the share of pre-specified analyses is negligible small. The subsequent sections therefore present results that are obtained with specifications I came up with after seeing the data.

5.2 Results

Chapter 4 introduced two different strategies, an OLS regression as well as a regression discontinuity design. This section will be organized correspondingly: I will report the OLS results before I proceed with the RDD. Before I do so, however, I will comment on the data.

5.2.1 Data

The data I analyze in this chapter stems from the first 9 sessions we ran in 2017. We ran three additional sessions deploying the same experimental design in the end of January, 2018. I will not consider the new sessions in this chapter (as the data collection was too close to the submission of this document) but print the same figures and tables I report here in Appendix C.

Using the exact specifications I introduced, I analyze different subsets of the data — especially in the OLS section. As the experiment does not offer many covariates to control for (or to omit), I use this subsetting method as a mean to perform robustness checks. The different samples and subsets are defined as follows:

[30] As interpreted by Nate Silver (2012, p. 149).

[31] An example from the subfield of policy evaluation which actually inspired me to use this approach was the evaluation of the Oregon Health Insurance Experiment conducted by Amy Finkelstein and colleagues (see for instance Finkelstein et al. (2012, 2016), Taubman et al. (2014), Baicker et al. (2014)).

FULL: This sample includes all *observations*. The units of observation generally are agents as we are interested in their behavior. However, as their behavior depends on the matched principals' actions, an observation also contains information about their behavior.

full: This subset includes only observations where the agents faced the performance-based mechanism. Since I mainly analyze how agents respond to this specific mechanism, this subset is the most important one.

recip.: This subset is based on the *full* subset but excludes all those observations where the corresponding agents have a below-median propensity to reciprocate according to their answers in the final questionnaire.

clean: This subset is based on the *full* subset but excludes three observations which one might label as extreme values. In particular, I excluded observations who were productive in the first stage (they clicked away more than 50 percent of the boxes) but refused to supply labor in the second stage.

smart: During the sessions, the experimenters took notes about participants who took a long time to answer the control question and who were likely to not understand the strategic environment they found themselves in. These judgments were, of course, rather subjective and are based not only on the time the participants needed but also on the questions they asked and the experimenters' impressions of how well they might have understood the questions. This subset is based on the *full* subset but excludes these specific observations.

double: This subset is a combination of *full*, to which I added some noise, and *smart*. It represents a thought experiment of what would happen if we collected the same data once again, but this time without the extreme values.

Some of the subsets (*smart*, *clean* and *double* need to be justified for the estimation as they were generated by pruning observations in a way that is either subjective or untraceable to the reader. However, the comparison between the *full* and other subsets provides insights into the robustness of my estimation. I will first report the results before I comment on the subsets that turn out to be important. Table 5.1 provides summary statistics for all these subsets and shows that the performance in Stage 2 differs consistently from the productivity in Stage 1 across all subsets. This applies to both the standard deviation as well as the mean: the average productivity is higher than the average performance and is distributed more closely to this value (as indicated by a lower standard deviation). Also, the range (as indicated by the difference between max. and min. values) in which the productivity is distributed is narrower than the performance's range. Yet, the medians of the two variables are quite similar, which indicates that the tails are shifting. The table also shows, that while most people chose the maximum workload, some decided to minimize their workload (which has a big impact on the variable's average). The *Reciprocity Parameter* in column 4 ranges between 1.7 and 5 which already indicates a tendency to the mean (with a range of 1 = *"not reciprocal at all"* to 6 = *"very reciprocal"*) and becomes apparent if one plots the distribution. Finally, the table shows that the number of available observations is small (the most important subset only counts 60 observations). In conclusion, all subsets exhibit a relatively large variance in the performance (compared to the productivity) that is to be explained. The self-proclaimed propensity to reciprocate appears to be a bad candidate to do so.

5.2.2 OLS

Before reporting the hidden effects of monitoring, it should be stated that principals who monitored the matched agents realized higher payoffs[32]. More precisely, the monitoring principals' earned 17.2 DKK more (standard error = 8.5 and p = 0.046)[33] than those who did not monitor. Even though these results are significant, I expect the higher payoffs to be causally unrelated to the mechanism or its effect on the agents' performances. After all, the higher earnings could stem from three different routes. First, the chance player could have drawn realizations of the agents' payoffs that came in favor of the principals. If the principals were lucky, the chance player realized no bonus payment for the agents which would have been paid by the principals. Second, the agents could have worked harder under the performance-based mechanism. This would translate into (hidden) benefits of monitoring as the hard work pays out for

[32] The experiment was designed such that each Stage yielded a *payoff*. At the end of the experiment, one of these payoffs was randomly drawn to be paid out. Here, I refer to the principals' payoffs from the second stage (where it was generated by the agents).

[33] This applies to the following two OLS specifications where the standard error as well as the p-value are almost identical: $Pay = \alpha + \beta\mu + \upsilon$ and $Pay = \alpha + \beta\mu + \gamma Y_1 + \upsilon$.

Table 5.1: Summary statistics

	Productivity	Performance	Workload	Recip. Parameter
FULL sample of agents ($N = 93$)				
Mean	0.467	0.371	19.097	3.674
St. Dev.	0.122	0.208	8.964	0.752
Min	0.275	0.000	1	1.667
Median	0.423	0.391	25	3.667
Max	0.711	0.720	25	5.000
full subset of agents ($N = 63$)				
Mean	0.466	0.393	19.810	3.735
St. Dev.	0.123	0.213	8.701	0.767
Min	0.286	0.000	1	1.667
Median	0.423	0.407	25	3.667
Max	0.711	0.720	25	5.000
recip. subset of agents ($N = 31$)				
Mean	0.455	0.402	20.774	4.376
St. Dev.	0.116	0.202	8.317	0.363
Min	0.287	0.000	1	4.000
Median	0.434	0.423	25	4.333
Max	0.677	0.710	25	5.000
clean subset of agents ($N = 60$)				
Mean	0.458	0.411	20.750	3.744
St. Dev.	0.120	0.201	7.789	0.780
Min	0.286	0.000	1	1.667
Median	0.411	0.416	25	3.833
Max	0.711	0.720	25	5.000
smart subset of agents ($N = 58$)				
Mean	0.470	0.391	19.534	3.718
St. Dev.	0.122	0.218	8.943	0.779
Min	0.287	0.000	1	1.667
Median	0.429	0.405	25	3.667
Max	0.711	0.720	25	5.000
double subset of agents ($N = 123$)				
Mean	0.488	0.428	20.268	3.740
St. Dev.	0.124	0.207	8.248	0.770
Min	0.286	0.000	1	1.667
Median	0.450	0.432	25	3.667
Max	0.761	0.770	25	5.000

the principal.[34] OLS regressions do not support the claim that agents supplied more effort under the performance-based mechanism than under the random mechanism (p = 0.137 for the naive regression and p = 0.090 when controlled for productivity).[35] Finally, it might be that principals anticipated the adverse effect of monitoring on the agents' working morale which I predicted. In that case, principals would only monitor agents who are productive. Because productive agents are likely to perform better than unproductive ones in Stage 2, the average performance of monitored agents should be higher than of those who were not monitored (who faced the random mechanism). The monitoring principals higher earnings could thus be endogenous. If this was the case, principals should have been more likely to choose the random mechanism while being matched with an unproductive agent. This is pattern, however, is not present in our data.[36] I therefore suspect the higher earnings to be a result of chance. One can therefore conclude that:

Result 1 *Agents did not perform better if they faced the performance-based instead of the random mechanism. (Rejection of the pure self-interest prediction 2.)*

Result 2 *Upon closer examination, the monitoring device tested in our experiment (that is, the performance-based mechanism) did not cause any apparent net-costs or net-benefits.*

That we do not find any apparent benefits of monitoring suggests that the benefits are offset by hidden costs and/or that the agents were mostly motivated by their intrinsic motivation. Agents who faced the random mechanism had no material incentive to exert effort and behaved similar as the agents who faced the performance-based mechanism. For this reason, I expect the intrinsic motivation[37] to be the main driver of the agents' labor supply.

Having that stated, the remainder of the thesis investigates whether there are hidden costs or benefits beneath the zero net-effect: Table 5.2 demonstrates that it is

[34] Recall that the principal's earnings were strictly increasing in the agent's effort provision, even though a higher performance increased the likelihood with which the principal had to pay the agent's bonus payment.

[35] I ran the similar specifications as before, when I analyzed the principals' payoffs. In addition, I simulated the data using an algorithm I explain in Footnote 38 and come to the same conclusion.

[36] I ran a logit regression explaining the Mechanism with the Productivity Dummy.

[37] Note that the motivation might just have been the desire to pass the time. It was not necessarily fun to work on the task but at least more fun than to just wait.

Table 5.2: OLS estimates of the effect of reciprocity on the agent's working morale

	Response variable: Performance-Productivity (ΔY_u)				
	(1)	(2)	(3)	(4)	(5)
Subsets:	*full*	*recip.*	*clean*	*smart*	*double*
β : Productivity	-0.01	-0.03	0.32^{**}	0.01	-0.01
	(0.18)	(0.27)	(0.14)	(0.20)	(0.18)
α : Constant	-0.07	-0.04	-0.19^{***}	-0.08	-0.07
	(0.09)	(0.13)	(0.06)	(0.10)	(0.10)
Is there an effect[1]?	No	No	Yes	No	No
Observations	63	41	60	58	123
R^2	0.00	0.01	0.08	0.00	0.01
F Statistic	0.001	0.59	5.35^{**}	0.001	1.02

*p<0.1, **p<0.05, ***p<0.01. [1]Estimated as the observed difference (ΔY_u) $\neq 0$. Standard errors in parentheses. The unit of observation is a participant who was assigned the role of an agent ('Person B'). The *full* sample includes only agents who where exposed to the performance-based from the subset of agents, who were exposed to the performance-based mechanism in Stage 2. The *recip.* subset includes all the observations from the *full* sample without those, who are below-median reciprocal according to the final questionnaire. The *clean* subset includes all the observations from the *full* sample without the three extreme-values. The *smart* subset includes all the observations from the *full* sample without the observations who attracted attention during the sessions. The *double* subset is a combination of *clean* and *full.*

questionable whether there are hidden costs of monitoring while the OLS rejects the hidden benefits hypothesis. The first column is based on the most important, the *full* subset, and presents coefficients that are not distinguishable from zero. The constant of -0.07, if significant, would have suggested that agents dislike monitoring (or are exhausted) and therefore decrease their effort provision between the first and the second stage by 7 percentage points on average. The negative coefficient of beta would then have suggested that these hidden costs are larger for more productive agents — and would thus translate into the opposite behavior that I predicted. However, the standard errors exceed the coefficients' magnitude. As a consequence, the regression does not find general observed differences and thus, neither hidden benefits nor hidden costs one can explain with reciprocity. One might argue that there are two types of agents, namely, reciprocal and selfish agents. To account for this heterogeneity, I used the questionnaire to distinguish between the two types. Subsetting the data to run the specification only over those observations that indicated that they are reciprocal in everyday life does, however, not yield any significant coefficients either (in the second column). This can by explained by two things: either, the questionnaire we used to assess the participants' reciprocity parameter is impractical to elicit the participants' inherent propensity to reciprocate (recall that questionnaires are not strategyproof, that is, the participants had no incentives to reveal their private information truthfully) or reciprocity has no influence on the agents' working morale. The *smart* subset's coefficients in column 4 look no different. I therefore conclude that the subjects who seemingly struggled with the instructions did not drive the zero-result. The third column exhibits significant results that look as predicted: monitoring spoils the unproductive agent's working morale by 19 percentage points. The positive coefficient of productivity then implies that these hidden costs of monitoring become weaker the more productive an agent is until they turn zero or even turn into hidden benefits (where the agent exerts more effort in Stage 2 than in Stage 1). These positive results are, however, no surprise as the subset that is considered is cleaned out of "unpleasant" extreme values. The final column suggests that we, even if we collected data that was as good as the *clean* subset, would not find any significant results in the resulting data set. All these null results do not change using robust standard errors and are underlined by low R^2's: the fraction of variance that is explained by the productivity ranges from virtually zero to 0.8 in our most favorable subset.

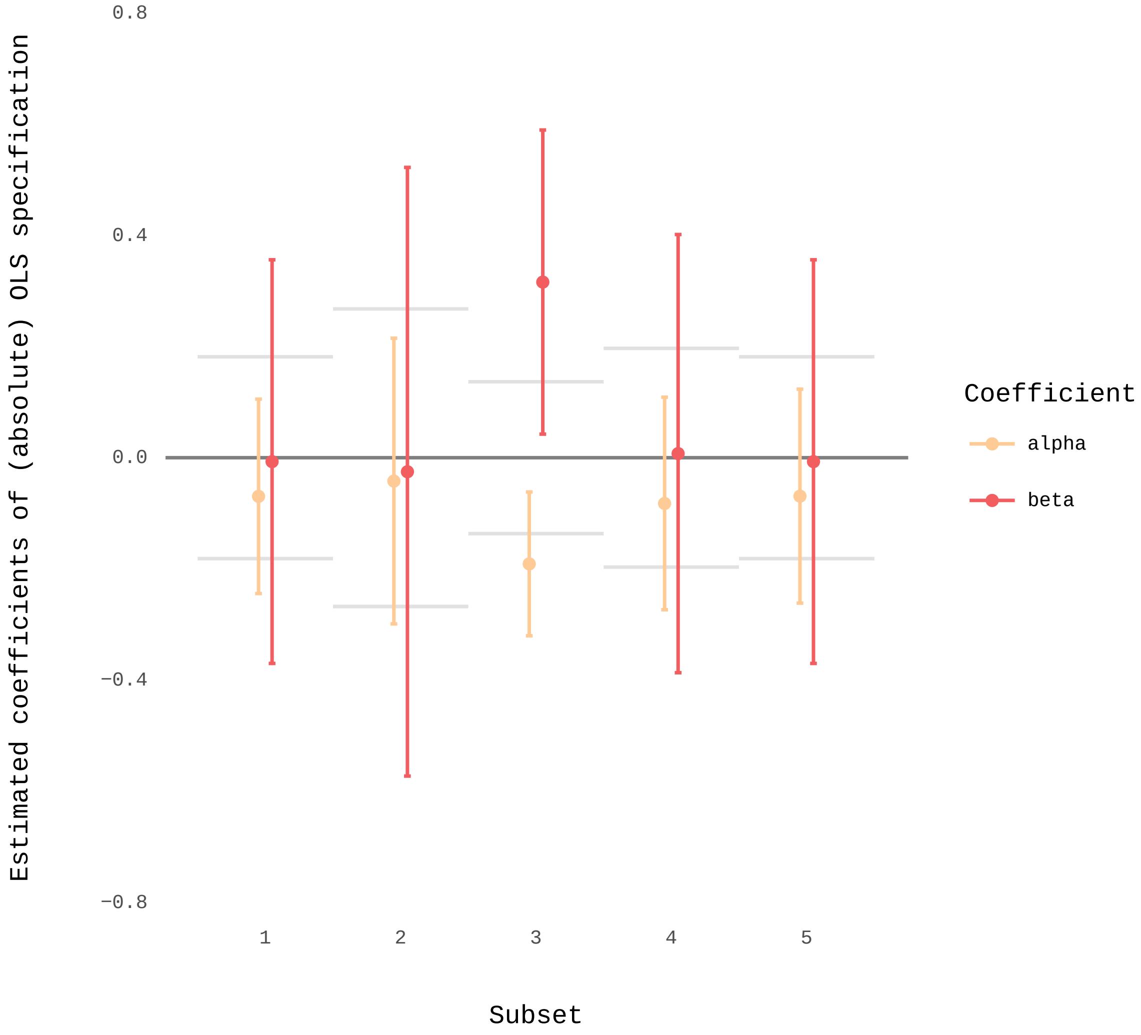

Figure 5.1: 95-percent confidence intervals of the coefficients estimated in Table 5.2

Standard errors are visualized as gray horizontal lines. The unit of observation are groups of a principal and the matched agent. This Figure focuses on the agents' behavior. *alpha* refers to the constant estimated in the OLS specification used throughout this chapter and *beta* refers to the slope coefficient of Y_1. (1) refers to the *full* subset, (2) to the *recip.* subset containing observations with above average (self-proclaimed) reciprocity parameters, (3) to the *clean* subset that excludes the extreme values, (4) to the *smart* subset who is expected to have understood the instructions and (5) to the *double* subset that combines (1) and (3). The gray vertical lines represent the slope coefficients' (β) standard errors.

In addition, the confidence intervals of each of the estimated coefficients are larger than 0.24 (remember that the agents' action space reaches only from 0 to 1). Referring to the results printed in Table 5.2, Figure 5.1 visualizes this point in the spirit of *the new statistics* (Cumming, 2014). The subsets are numerated as in Table 5.2 so that the most left bars illustrate the *full* subset's coefficients. As a substitute to the conventional tests for significance, this plot also shows how far away most of the coefficients are from being confidently distinguishable from zero. It also shows that, relative to the wide range of the interval, the *clean* subsets are relatively close to zero. Taken together, Table 5.2 and Figure 5.1 suggest that there are neither hidden costs, nor hidden benefits of monitoring.

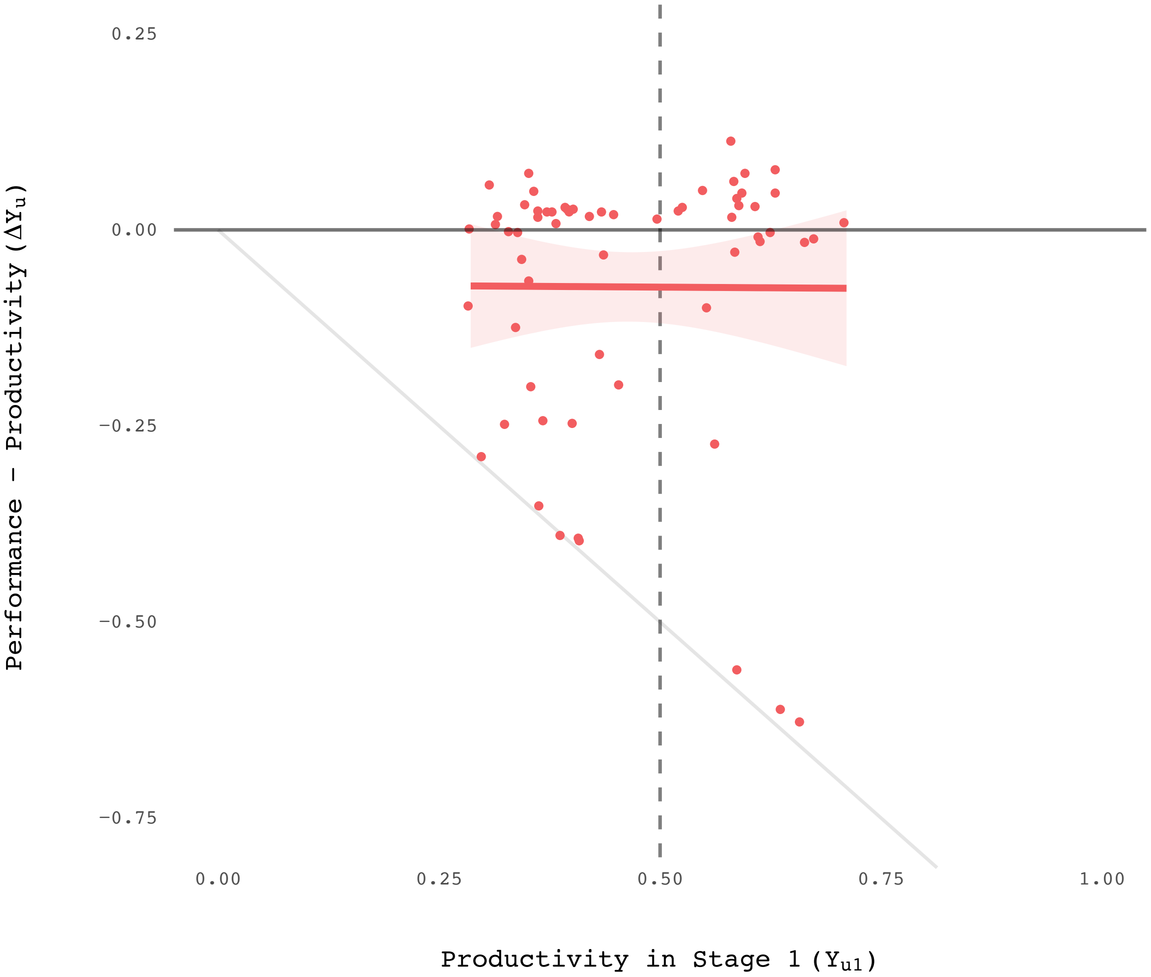

Figure 5.2: OLS Specification for the full *performance-based subset with 95% confidence interval*

The unit of observation are groups of a principal and the matched agent. This Figure focuses on the agents' behavior. The red regression line corresponds to the first column of Table 5.2. The vertical dashed line prints the threshold that splits the data into productive and unproductive. The downward sloped line censors the data as a data point cannot possible be lower than this value.

The scatterplot in Figure 5.2 plots each data point (of the *full* subset) and adds further nuances to the analysis. It also illustrates the regression of the table's first column as a red regression line surrounded by the corresponding 95-percent confidence interval. The line looks as expected: it is located below the zero-difference axis and has a negligible slope. Focusing on the data points instead of the regression line, one finds that there is a large fraction of agents who held their effort provision constant or even increased it between Stage 1 and Stage 2. As Table 5.2 suggests, this observation holds true for productive ($Y_{u1} > y_0$) and unproductive ($Y_{u1} < y_0$) agents.

Using the results printed in column 1 of Table 5.2, I also ran a computational simulation algorithm[38] proposed by King et al. (2000) that allows me to interpret the results in the following way: With 95-percent confidence, I estimate that agents with a productivity of 0.37 (which is the average productivity of unproductive agents) decreased their effort provision by 2.8 to 11.7 percentage points. Simultaneously, I estimate with the same confidence that agents with a productivity of 0.60 (which is the average productivity of productive agents) decreased their effort provision by 1.8 to 12.6 percentage points. My best guess for the unproductive and productive agents is a downward-deviation of 7.2 and 7.3 percentage points respectively. I present these claims in the form of density estimates (which are smooth versions of histograms) in Figure 5.3.

The graph shows plainly that both the distributions of ΔY as well as the simulated expected values (indicated by the vertical lines) are almost identical but smaller than zero. In fact the distributions' tails hardly exceed the vertical zero-difference line. Referring to meaningful productivity levels (the means), the simulations therefore support the notion of hidden costs of monitoring which are, however, not conditional on an agents productivity.

Even though the regressions reported in Table 5.2 tell a consistent story, the simulation and visualization thereof show that one should not conclude too hasty. The simulations have demonstrated that there are hidden costs if one refers to particular

[38] In the first step, I ran the OLS specification I described above to learn the coefficients as well as the corresponding variance-covariance matrix. Having these information (coefficients as well as their variance-covariance matrix), I simulated a set of these coefficients from a Multivariate Normal Distribution for 1000 times. The result was a 1000 × 2 matrix as I only estimate two coefficients. I then set the explanatory variable (that is, the productivity) to its mean to calculate $\tilde{\alpha} + \tilde{\beta} \cdot \overline{Y_1}$ for each of the simulated values of β (indicated by the tilde). The numbers in the 25th and the 976th positions of the resulting vector represented the upper and lower bounds of a 95-percent confidence interval. Averaging over all 1000 cells yields the simulated expected value.

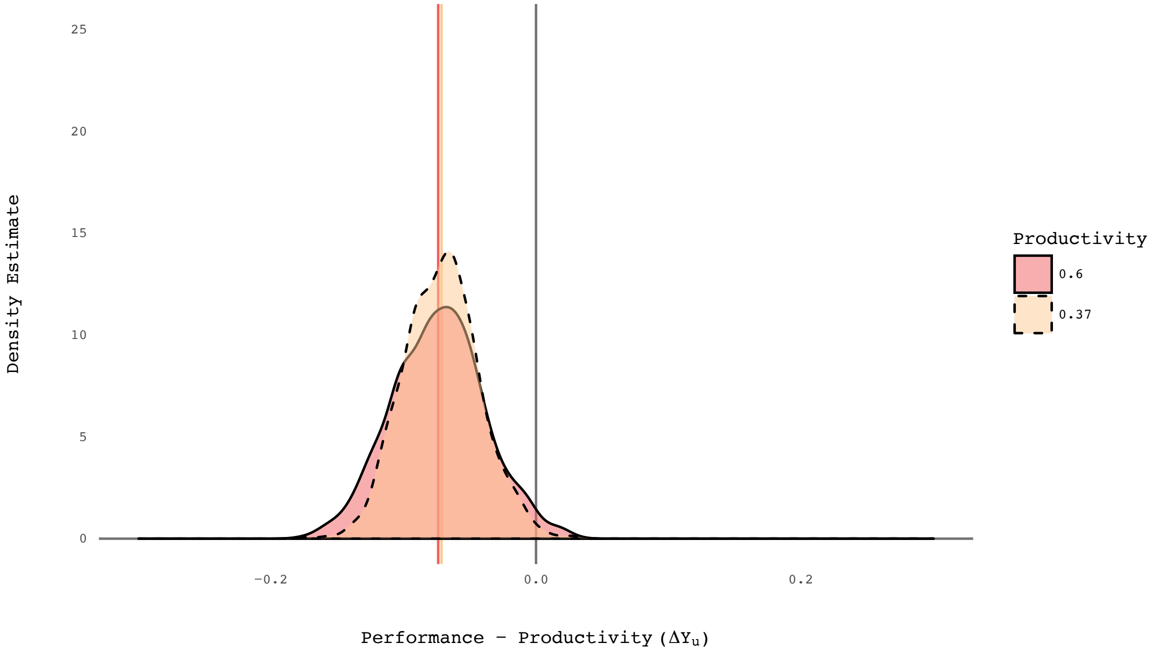

Figure 5.3: Simulated levels of productivity based on regression coefficients from the full *performance-based subset*

The unit of observation are simulations of agents with a productivity of either 0.6 or 0.37. These levels correspond to the means of productive and unproductive agents in the *full* subset. The simulations are based on coefficients estimated in the first column of Table 5.2. The density estimate is based on 1000 simulations for each productivity level. The Simulation process is described in Footnote 38.

productivity levels. This is an observation one would not have made by solely focusing on regression coefficients. It thus seems as if the regression and the simulation lead to conflicting evidence — even though the latter is based on the former. This is, however not the case. The difference between the two analytical means is that simulations translate the regression results back into meaningful statistics — they simply add information. In our case, there is only one covariate, the productivity, of which I chose its mean for two different subsets. This way, I can draw conclusions about the average productive agent and the average unproductive agent. The simulated distributions (which represent the uncertainty) are then based on only two data points and not as universal as a regression line that fits every productivity level. Having stated that the only difference between the two approaches is additional information, I conclude that:

Result 3 *The average productive as well as the average unproductive agent, reduced their labor supply as a response to the performance-based mechanism. (Rejection of pure self-interest prediction 1.)*

Table 5.3: Contingency table for occurrences of non-maximal workload decisions

Mechanism	Unproductive	Productive	Total
Performance-based	15	5	20
Random	6	10	16
Total	21	15	36

The unit of observation are groups of a principal and the matched agent. This Figure focuses on the agents' behavior. A non-maximal workload decision is measured as choices smaller than 25. A threshold of 0.5 is used to distinguish between high and low productivity. The null hypothesis, that non-maximal workload choices are distributed evenly across the agents' productivity and the mechanism they faced, can be rejected ($p = 0.041$).

Note that this does not necessarily mean that these agents decrease their effort (=labor supply) as a response to monitoring. This interpretation is only valid if the two identifying postulates (causal transience and temporal stability or "separability of effort costs" for short) hold.

Figure 5.2 exemplifies that there are many unproductive agents who decreased their effort substantially but only few productive agents who did so. This can also be demonstrated by Fisher's exact test (which I pre-specified) that counts the occurrences of non-maximal workload decisions to the left and separately to the right of the vertical threshold. The null hypothesis, that non-maximal workload choices are distributed evenly across the agents' productivity and the mechanism they faced, can be rejected ($p = 0.041$). Table 5.3 shows that unproductive agents were more likely to decrease their effort if they faced the performance-based mechanism. The first row reports that, of all the 20 agents who faced the performance-based mechanism and who reduced their workload, only 5 can be classified as productive. To put it differently, three quarters of the agents considered in the *full* subset (that is, who faced the performance-based mechanism) who deliberately reduced their workload were unproductive. Likewise, fewer of the unproductive agents chose to reduce their workload under the random mechanism. The opposite pattern can be found for productive agents: few declines in the workload for the performance-based mechanism and a relatively great number of declines for agents who faced the random mechanism.

The pattern described in Table 5.3 is very much in line with my predictions as I expected productive agents to find the choice of the performance-based mechanism as kind while unproductive agents should perceive the same choice as unkind. If one extends the predictions to the agents who faced the random mechanism, one would expect the random mechanism to be perceived as the kind (unkind) choice for unproductive (productive) agents. This is important as one can show that, considering the performance-based subset, all the major downward deviations in the effort provision strongly correlate with a non-maximal workload choice (Kendall's rank correlation tau $= -0.63$ with $p < 0.00$ and where the correlation coefficient (tau) is comprised between -1 and 1). Figure 5.4 visualizes this property nicely. Only few of the yellow observations (who chose the maximal workload) decreased their performance and none of them did so drastically. The observations of agents who chose to reduce their workload in contrast, deviated by higher magnitudes. Even though this is not surprising, it shows that the workload variable is a reasonable proxy for a spoiled working morale.

Result 4 *The fraction of unproductive agents who reduced their labor supply (or their workload) is considerably larger than the fraction of productive agents who did so, albeit at a small number of observations.*

One reason why the scatter plot in Figure 5.2, in combination with Table 5.3 and Figure 5.4, shows some support in favor of my predictions while the regression table rejects them is the extreme values' leverage: Note that the downwards deviation of effort was censored as an agent cannot deviate by more than her productivity in Stage 1. This is indicated by the downward sloped line in Figures 5.2 and 5.4. Points which are on this line exhibit a Stage 2 performance of zero. The deviation to a zero-performance, is larger for productive agents than for unproductive agents. The three productive agents who refused to exert effort in the second stage therefore have a strong influence on the OLS estimation. In addition, these specific agents differ from all the other productive agents because they haven't exerted any effort at all. If one compares the scatterplot with the resulting regression line, it becomes clear that the estimated slope is affected by these three points to the right. This is why I consider them as extreme values. It is these extreme values which I exclude in the *clean* subset.

Had I not observed the three extreme values, the predictions stemming from the simulations[39] would be quite different as well: I would then estimate that, with 95-percent confidence, agents with a productivity of 0.37 decreased their effort provision by 4.2 to 10.6 percentage points. With the same confidence, I would estimate that agents

[39] These simulations are based on the coefficients presented in the third column of Table 5.2. They are calculated for the same productivity means I used before before.

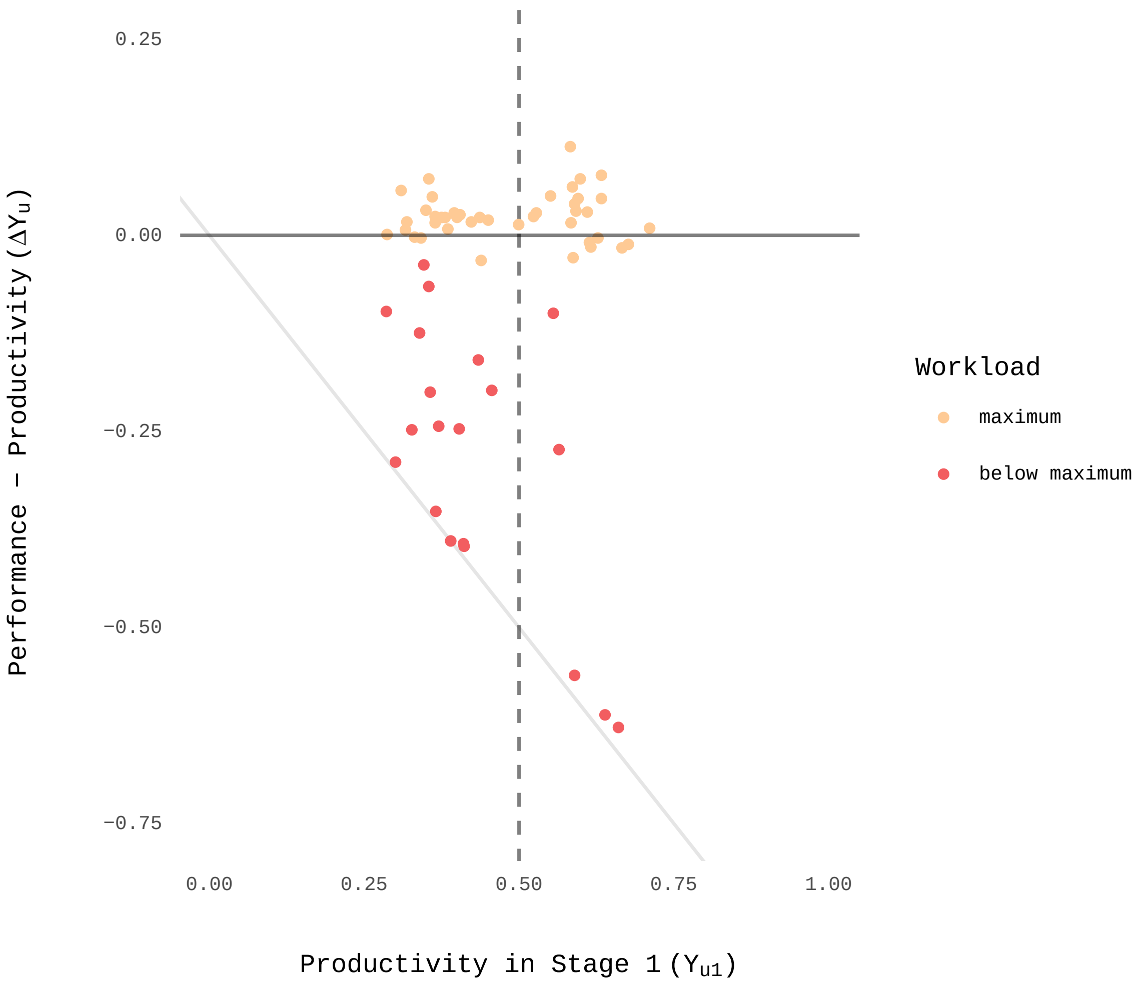

Figure 5.4: Scatterplot of the clean *performance-based by workload decisions*

The unit of observation are groups of a principal and the matched agent. This Figure focuses on the agents' behavior. The workload is either at its maximum of 25 screens (red) or below this value (yellow). The vertical dashed line prints the threshold that splits the data into productive and unproductive. The downward sloped line censors the data as a data point cannot possible be lower than this value.

with a productivity of 0.60 increased their effort provision by 4.4 percentage points or decreased it by the same amount. My best guess for the unproductive and productive agents is a downward-deviation of 7.4 and 0.0 percentage points respectively.[40] These

[40] It is counterintuitive that the estimates for the unproductive agents differ even though the considered subsets only differ with respect to the productive agents. This stems from the way the data was simulated. The simulation is based on the OLS coefficients I present in Table 5.2. The different subsets not only yield different coefficients, but also a different variance-covariance matrix — and both of these information are used to run the simulation.

estimates are visualized in Figure 5.5.

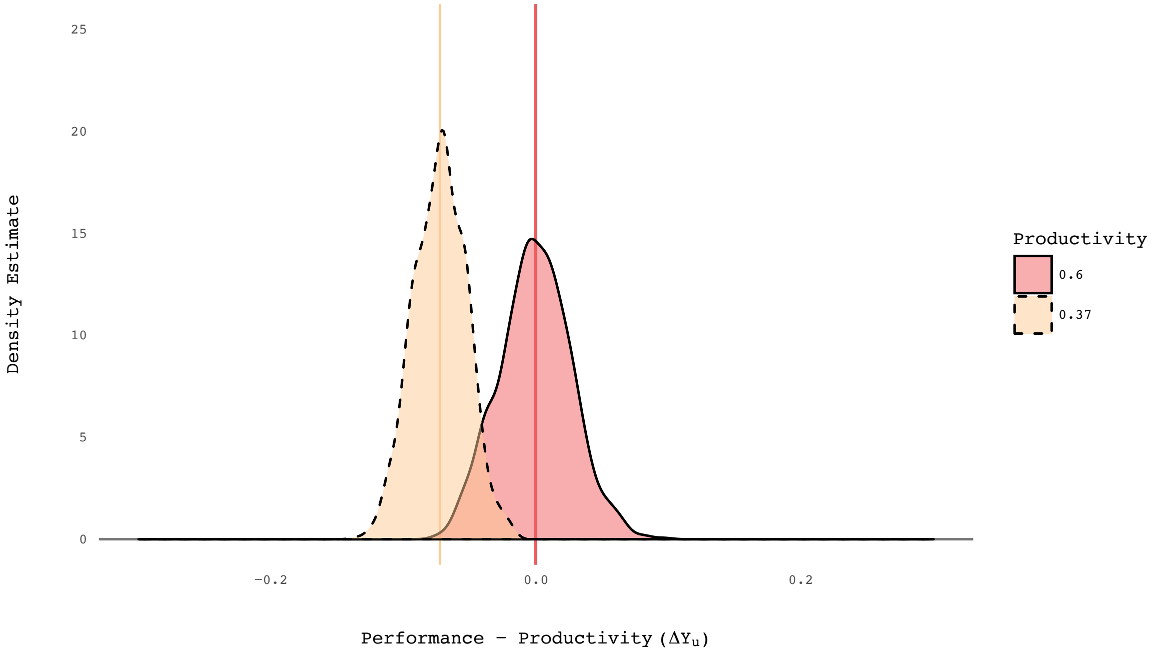

Figure 5.5: Simulated levels of productivity based on regression coefficients from the clean *performance-based subset*

The unit of observation are simulations of agents with a productivity of either 0.6 or 0.37. These levels correspond to the means of productive and unproductive agents in the *clean* subset. The simulations are based on coefficients estimated in the third column of Table 5.2. The density estimate is based on 1000 simulations for each productivity level. The Simulation process is described in Footnote 38.

As the reader can see, this graph draws a distinguishable effect of productivity: while the more productive agent tries to replicate her effort provision from Stage 1 in Stage 2, the less productive agent works less hard. Figure 5.5 would thus show that there are hidden costs of monitoring, yet only for unproductive agents. The performance of productive agents, in contrast, would not appear to be affected. One would thus conclude that monitoring only spoils the working morale of unproductive agents and reject the hidden benefits hypothesis. Yet, the crucial question is whether it is legit to evaluate the *clean* subset in the first place. After all it can be interpreted as a desperate try to *p-hack my way to glory.*

Result 5 *If it is reasonable to prune three extreme values, one can find a reduced labor supply only under the condition that the agent is unproductive.*

I already mentioned that the productive agents have more leverage than the unproductive agents. I took this into account and adjusted the response variable by dividing it by the initial productivity such that the estimations concern $\frac{\Delta Y}{Y_1}$. One would then interpret the deviation between Stage 1 and 2 relatively ("The agents worked only 90 percent as hard as before"). I adjusted the OLS specification accordingly and found the same qualitative results: none of the coefficients are significantly different from zero in the *full* subset, while both of them are as predicted in the *clean* subset. The simulations for the two mean productivity levels yield the same pictures as illustrated in Figures 5.3 and 5.5. Also the estimation within the *recip.* and *smart* subsets are unaffected by the adjustment of the response variable. The only difference between the analyses of two response variables are the results for the *double* subset. In contrast to the results in the fifth column of Table 5.2, both coefficients are significantly different from zero ($\alpha = -0.38$ and $\beta = 0.5$). With regard to the simulation, this translates into indistinguishable hidden costs for the absolute response variable (ΔY) versus distinguishable hidden costs for the relative response variable ($\frac{\Delta Y}{Y_1}$).

The underlying idea of the *double* subset was the question of what would happen if we collected similar data again, given that the extreme values were indeed outliers (that is, given that we would not observe values as far off as they are a second time). The answer to that question is ambiguous. Given the absolute response variable, I conclude that we could hope for a small negative effect (hidden costs) that does not depend on the agents' productivities. Given the relative response variable, we would still find no hidden benefits, yet hidden costs that diminish in the agents' productivities.

To conclude, the data leaves a mixed impression as different subsets, methods and transformations of the response variable convey slightly different messages: While some analyses show null results, other indicate that there are hidden costs and still others suggest that these hidden costs are decreasing in the agents' productivity (such that one could hope for hidden benefits if the data offered full support over the full range of theoretically possible productivity levels). What all of them have in common is that we cannot find hidden benefits of monitoring in the data. An inspection of the workload variable shows that agents' active choice of their workload reflects hidden costs of monitoring which are in line with my reciprocity predictions. The simulations, which provide meaningful statistics that are easy to understand, lead to the same conclusion. The question, of whether additional data can add clarity depends on whether the extreme values are outliers.[41] If and how one should pursue this research question is discussed in the final chapter of this thesis.

5.2.3 RDD

Recall that the RDD relies on rich data that contains many observations which are distributed around the threshold. In our setting this threshold was a productivity level of 0.5 which is denoted by a vertical dashed line in most of the figures of this paper. As the scatterplots in the previous section have shown, our current data does not satisfy this requirement. This is the reason why Figure 5.6 is not suited to investigate the reciprocity predictions. Even though the regression lines in the upper panel illustrate a discontinuity around the threshold, the 95-percent confidence intervals overlap. On top of that, the discontinuity is likely to be driven by the three extreme values as one can image the regression line to be much steeper at the right-hand side if one omitted them.

The lower panel of Figure 5.6 analyzes the subset of agents who faced the random mechanism and thus, tests an extension of my predictions. It shows a similar picture as the performance-based sample though: wide confidence intervals and almost no data at the threshold.

Result 6 *Due to the lack of the data around the threshold, the non-parametric RDD cannot yield any additional insights and does not allow me to make any inferences.*

This makes the ShinyApp[42] I programmed obsolete. It should, however, be stressed that an RDD would have been the most powerful approach to analyze the data as it relies on weaker assumptions. It already is a powerful strategy for quasi-experiments and should be considered as an even more powerful for laboratory experiments where we, as experimenters, can define and manipulate the threshold.

[41] This is, however, a question which should not be quantified into a single metric as this only adds a pseudo-scientific appearance. I propose to collect more data and to find ways (such as surveys) to understand these observations if there are more extreme values to come.

[42] See `https://roggenkamp.shinyapps.io/ShinyRDD/`.

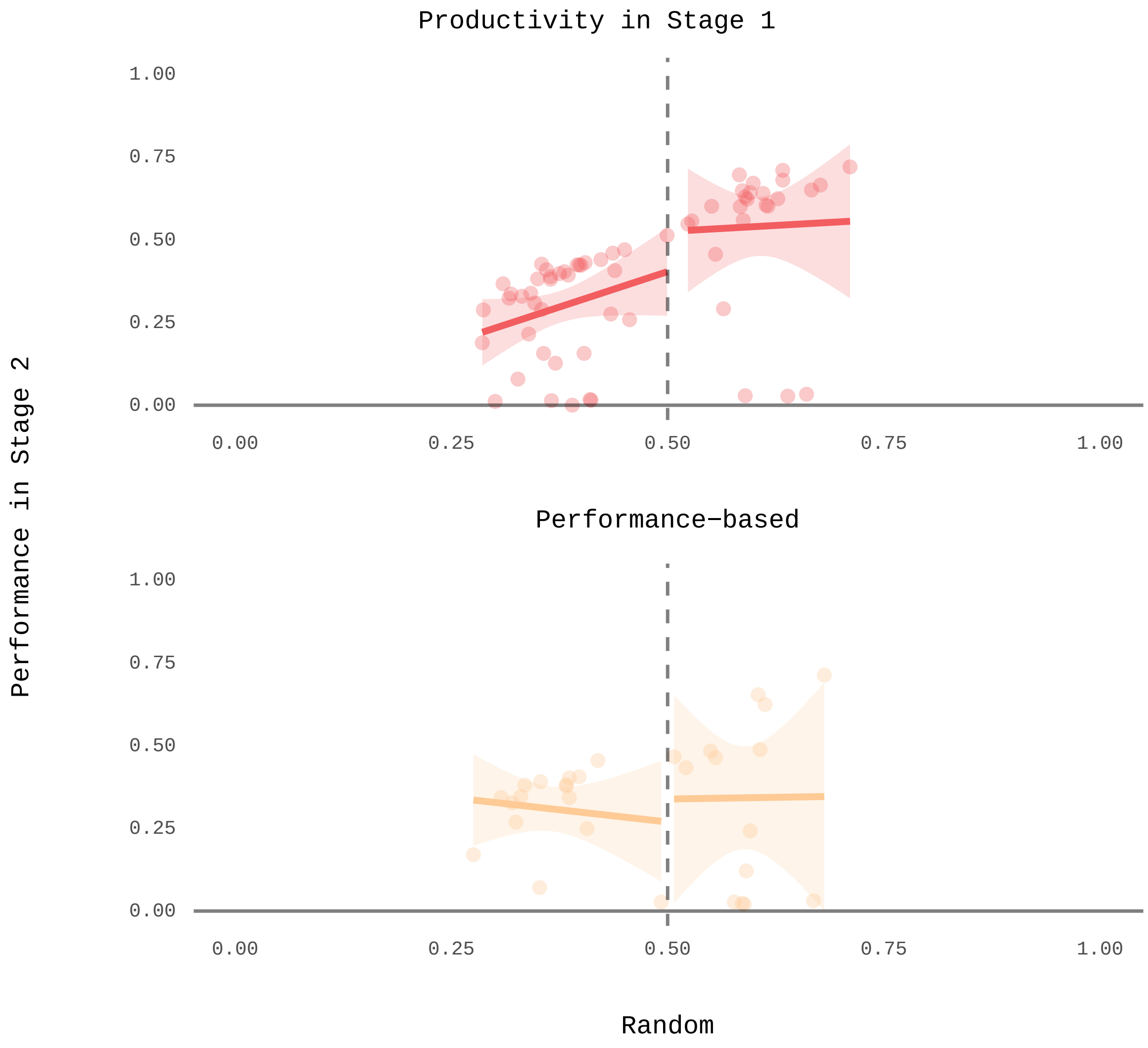

Figure 5.6: RDD for the FULL *data set with 95% confidence intervals*

The unit of observation are groups of a principal and the matched agent. This Figure focuses on the agents' behavior. The lower panel plots all the observations where the principal chose the random mechanism while the upper one plots the *full* subset. The vertical dashed line prints the threshold that splits the data into productive and unproductive. It is thus, the arbitrary value that assigns the observations' treatment statuses (in the RDD logic).

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 3 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.[43]

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

[43] 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.

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.

RDD. 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.[44] 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,

[44] If you focus on the productive agents in Figure 5.2 you will see that they are scattered around $Y_1 \simeq 0.6$.

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. Another downside 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 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 3.4). 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.[45]. 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 5.2 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

[45] 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.

as unpaid overtime.[46]

[46] 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.

Bibliography

Angrist, J. D. and Lavy, V. (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement, *The Quarterly Journal of Economics* **114**(2): 533–575.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings, *The Review of Economics and Statistics* **60**(1): 47–57.

Baicker, K., Finkelstein, A., Song, J. and Taubman, S. (2014). The impact of medicaid on labor market activity and program participation: Evidence from the oregon health insurance experiment, *The American economic review* **104**(5): 322–328.

Barkma, H. (1995). Do top managers work harder when they are monitored?, *Kyklos* **48**(1): 19–42.

Bartelsman, E. J. and Doms, M. (2000). Understanding productivity: Lessons from longitudinal microdata, *Journal of Economic Literature* **38**(3): 569–594.

Bénabou, R. and Tirole, J. (2003). Intrinsic and extrinsic motivation, *The Review of Economic Studies* **70**(3): 489–520.

Bloom, N. and Van Reenen, J. (2007). Measuring and explaining management practices across firms and countries*, *The Quarterly Journal of Economics* **122**(4): 1351–1408.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M. and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics, *Science* **351**(6280): 1433–1436.

Cumming, G. (2014). The new statistics: Why and how, *Psychological Science* **25**(1): 7–29. PMID: 24220629.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs, *Journal of the American statistical Association* **94**(448): 1053–1062.

Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies, *Review of Economics and statistics* **84**(1): 151–161.

Dickinson, D. and Villeval, M.-C. (2008). Does monitoring decrease work effort?: The complementarity between agency and crowding-out theories, *Games and Economic behavior* **63**(1): 56–76.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity, *Games and economic behavior* **47**(2): 268–298.

Falk, A., Becker, A., Dohmen, T. J., Huffman, D. and Sunde, U. (2016). The preference survey module: A validated instrument for measuring risk, time, and social preferences, *IZA Discussion Paper* (9674).

Falk, A. and Kosfeld, M. (2006). The hidden costs of control, *American Economic Review* **96**(5): 1611–1630.

Finkelstein, A. N., Taubman, S. L., Allen, H. L., Wright, B. J. and Baicker, K. (2016). Effect of medicaid coverage on ed use—further evidence from oregon's experiment, *New England Journal of Medicine* **375**(16): 1505–1507.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K. and Group, O. H. S. (2012). The oregon health insurance experiment: evidence from the first year, *The Quarterly journal of economics* **127**(3): 1057–1106.

Foss, N. J. (2003). Selective intervention and internal hybrids: Interpreting and learning from the rise and decline of the oticon spaghetti organization, *Organization Science* **14**(3): 331–349.

Frey, B. S. (1993). Does monitoring increase work effort? the rivalry with trust and loyalty, *Economic Inquiry* **31**(4): 663–670.

Frey, B. S. and Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding- out, *The American Economic Review* **87**(4): 746–755.

Gibbons, R. and Roberts, J. (2012). *The handbook of organizational economics*, Princeton University Press.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee, *Journal of the Economic Science Association* **1**(1): 114–125.

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

Halac, M. and Prat, A. (2016). Managerial attention and worker performance, *American Economic Review* **106**(10): 3104–32.

Ho, D. E., Imai, K., King, G. and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference, *Political Analysis* **15**(3): 199–236.

Holland, P. W. (1986). Statistics and causal inference, *Journal of the American Statistical Association* **81**(396): 945–960.

Kahneman, D., Knetsch, J. L. and Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market, *The American Economic Review* **76**(4): 728–741.

King, G., Tomz, M. and Wittenberg, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation, *American Journal of Political Science* **44**(2): 347–361.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data, *The American economic review* **76**(4): 604–620.

Lee, D. S. (2008). Randomized experiments from non-random selection in u.s. house elections, *Journal of Econometrics* **142**(2): 675 – 697. The regression discontinuity design: Theory and applications.

Lee, D. S. and Lemieux, T. (2010). Regression Discontinuity Designs in Economics, *Journal of Economic Literature* **48**(2): 281–355.

Masella, P., Meier, S. and Zahn, P. (2014). Incentives and group identity, *Games and Economic Behavior* **86**: 12 – 25.

Murphy, A. H. (1993). What is a good forecast? an essay on the nature of goodness in weather forecasting, *Weather and forecasting* **8**(2): 281–293.

Okun, A. M. (2011). *Prices and quantities: A macroeconomic analysis*, Brookings Institution Press.

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*, Broadway Books.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies., *Journal of educational Psychology* **66**(5): 688.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate, *Journal of educational Statistics* **2**(1): 1–26.

Schnedler, W. and Vadovic, R. (2011). Legitimacy of control., *Journal of Economics and Management Strategy* **20**(4): 985 – 1009.

Schulze, G. G. and Frank, B. (2003). Deterrence versus intrinsic motivation: Experimental evidence on the determinants of corruptibility, *Economics of Governance* **4**(2): 143–160.

Sebald, A. (2010). Attribution and reciprocity, *Games and Economic Behavior* **68**(1): 339–352.

Silver, N. (2012). *The signal and the noise: why so many predictions fail–but some don't*, Penguin.

Simmons, J. P., Nelson, L. D. and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant, *Psychological science* **22**(11): 1359–1366.

Smith, J. A. and Todd, P. E. (2005). Does matching overcome lalonde's critique of nonexperimental estimators?, *Journal of econometrics* **125**(1): 305–353.

Syverson, C. (2011). What determines productivity?, *Journal of Economic Literature* **49**(2): 326–65.

Taubman, S. L., Allen, H. L., Wright, B. J., Baicker, K. and Finkelstein, A. N. (2014). Medicaid increases emergency-department use: evidence from oregon's health insurance experiment, *Science* **343**(6168): 263–268.

von Siemens, F. A. (2013). Intention-based reciprocity and the hidden costs of control, *Journal of Economic Behavior and Organization* **92**(Supplement C): 55 – 65.

Appendices

Identification 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, p. 947)

While the previous chapter derived predictions about the causes of effects, this chapter deals with the statistical measurement of effects of causes. In particular, it describes how one can screen the data that were generated in our experiment to describe and to identify reciprocal behavior, if there is any. In other words, I derive the conditions under which the estimated effect can be interpreted as a *causal* effect. In addition, I describe what this effect should look like to support the predictions derived in the previous chapter.

A.1 The Empirical Workhorse Model

In what follows, I will rely on the model as well as the corresponding notation introduced in Holland (1986). Using his notation has the advantage that the reader can easily find answers to questions this chapter might evoke. The downside, however, is that the notation applied in this chapter will differ from the notation I used to derive the predictions.

As I only predict (and test) the behavior of a specific subset of agents, the model starts with a population U of agents who were exposed to the performance-based mechanism in Stage 2. These agents will be denoted as u. There are several real-valued functions, called variables, that are defined on each agent in U. A measurement process assigns a number to a variable of a specific agent — the measurement of an agent's effort provision in Stage 1, for instance, assigns a value to a variable, which is called productivity and which is defined on this specific agent.

The reason we ran this experiment is that for each agent u in U, there is a variable we intend to explain — a so called *response variable* Y with values $Y(u)$. As this thesis is devoted to understand the agent's behavior in the second stage, $Y(u)$ describes the performance of a specific agent as a numerical value.[47] In what follows, all probabilities and expectations for variables such as Y are computed over U. As such, the expected value of Y is the average of that particular variable over all agents in U. Likewise, regressions represent a Conditional Expectation Function (CEF) that calculate the average value of Y for all agents that share some value of another variable that is used in that regression. The statistical analyses will thus make inferences about parameters that link the response variable Y to some other variables on the basis of data we gathered in our experiment about Y and these " other variables" from agents in U.

Following Holland (1986) (and Rubin (1974, 1977)) further, I'll refer to the possible cause of an effect as *treatment*. This might, again, be confusing as we solely manipulated the difficulties of the box clicking task between subjects. We therefore do not present a treatment in the classical experimental sense. Think, instead, of the treatment as the exposure to reciprocity. The following sections explain how one can interpret the two conditions more precisely. To sum up there are only two different treatment statuses which I denote as c (the control condition) and t (the treatment condition). Let S be the variable that assigns the treatment status to each agent in U.

The role of the response variable Y is to measure the effect of the treatment. It thus needs to be a measured after the exposure to a treatment. To measure the *effect* of a treatment one needs not a single response variable, but two variables Y_t and Y_c. One would then know the value of the response variable of u if the unit were exposed to the treatment $Y_t(u)$ and the value of the response variable of the same agent u if she was not exposed to the treatment $Y_c(u)$.

> *The difference $Y_t(u) - Y_c(u)$ is understood as the causal effect of the treatment on the specific agent u.*

The problem is that it is impossible to *observe* this difference as any agent, within a period of time, is either treated (that is, exposed to reciprocity) or not. However, as Holland (1986, p. 947) stresses, the fact that one cannot observe $Y_c(u)$ and $Y_t(u)$ simultaneously, does not mean that it is impossible to make inferences with a causal

[47] Previously, I denoted this variable as the effort provision, performance or labor supply l.

flavor.

The best solution to the so called fundamental problem of causal inference would be to run a few sessions with the hypothetical treatment I introduced in Chapter 3.3 (the introduction of an artificial principal). One would then argue that the assignment of participants into one of the two treatments was random and calculate the *observed differences* $\mathbb{E}[Y(u)|S(u) = t] - \mathbb{E}[Y(u)|S(u) = c]$. This, however, might induce some bias if subjects could affect their treatment status — if, for instance, subjects with a higher ability joined the control sessions they would exert a higher performance than those in the treatment sessions if they had not been treated. Or formally:

$$\begin{aligned}\underbrace{\mathbb{E}[Y(u)|S(u) = t] - \mathbb{E}[Y(u)|S(u) = c]}_{\text{Observed Differences}} &= \underbrace{\mathbb{E}[Y_t(u)|S(u) = t] - \mathbb{E}[Y_c(u)|S(u) = t]}_{\text{Average Treatment Effect on Treated}} \\ &\quad + \underbrace{\mathbb{E}[Y_c(u)|S(u) = t] - \mathbb{E}[Y_c(u)|S(u) = c]}_{\text{Selection Bias}}\end{aligned}$$

$\mathbb{E}[Y_t(u)|S(u) = t]$ is the expected value of the response variable (the performance) of all those agents in U who have been treated ($S(u) = t$) if they actually have been treated (which is indicated by Y's subscript t). $\mathbb{E}[Y_t(u)|S(u) = c]$ in contrast, is the expected value of the response variable of the same subset of agents who have been treated($S(u) = t$) had they not been treated (which is indicated by Y's subscript c). We cannot observe the counterfactual expectations where we think of control (treated) agents if they have (not) been treated. However, we can estimate it as the randomized treatment assignment implies that $(Y_c, Y_t) \perp\!\!\!\perp S(u)$. It follows that $\mathbb{E}[Y_c(u)|S(u) = t] = \mathbb{E}[Y_c(u)|S(u) = c]$ such that the selection bias is zero (meaning that all those who have been treated eventually, perform, on average, as well as those who have not been treated, if they had not been treated). It further follows that

$$\begin{aligned}\mathbb{E}[Y_t(u)|S(u) = t] - \mathbb{E}[Y_c(u)|S(u) = t] &= \mathbb{E}[Y_t(u) - Y_c(u)|S(u) = t] \\ & \underbrace{\mathbb{E}[Y_t(u) - Y_c(u)]}_{\text{Average Treatment Effect}} \quad .\end{aligned}$$

As a consequence, actually running a few of the hypothetical control sessions would allow us to apply the so called *statistical solution* and calculate the average causal effect of the treatment — even though we cannot observe it on an individual level. The problem however is, that we did not run these sessions such that we, without any further assumptions, do not have data on $\mathbb{E}[Y(u)|S(u) = c]$.

The remainder of this Appendix deals with two ways to approach this problem and is thus split into two sections. Section A.2 describes the *scientific solution* and results in an OLS regression. Subsequently, Section A.3 is based on less restrictive assumptions that allow me to apply a more complex version of the *statistical solution*: a Regression Discontinuity Design (RDD).[48]

A.2 The Scientific Solution: Ordinary Least Squares

The first approach is to focus on the first stage of the experiment and use it as a control condition. In Chapter 4 I argued that the first stage is *as good as ceteris paribus* comparable to the respective subgame of the second stage. Measuring the response variable before a population of agents is exposed to reciprocity (that is, in Stage 1) and post exposure (in Stage 2), to estimate the effect of S on Y is called *within*-subject design. The remainder of this section derives how it can be applied to our setting.

As every agent in the set of U was exposed to both treatment statuses in the sequence of *'first c then t'*, one could also interpret S as a variable that describes the time of measurement. But as time becomes important, it will be indexed by another variable called $T \in \{1, 2\}$. One can now describe what the data allows us to observe more formally:

$$Y_{t2}(u) - Y_{c1}(u)$$

The identification problem here is, that the minuend and the subtrahend differ in two characteristics, S *and* T. It thus, might be the case that my estimation is confounded by some term I denote as *sequential bias.* The following formula describes how measuring the observed differences might result in two terms, the average treatment effect in $T = 2$

[48] Admittedly, the RDD is a rather explorative strategy in our setting as none of my predictions foresees any discontinuities. But as this particular strategy as well as the corresponding analysis itself is pre-specified, I believe an RDD can either enrich our understanding of reciprocity or serve as a robustness check.

as well as the unwanted sequential bias.[49]

$$\begin{aligned}\mathbb{E}[Y(u)_t|T(u)=2] - \mathbb{E}[Y(u)_c|T(u)=1]& \\ = \underbrace{\mathbb{E}[Y(u)_t|T(u)=2] - \mathbb{E}[Y(u)_c|T(u)=2]}_{\text{Average Treatment Effect in } T=2}& \\ + \underbrace{\mathbb{E}[Y(u)_c|T(u)=2] - \mathbb{E}[Y(u)_c|T(u)=1]}_{\text{Sequential Bias}}&\end{aligned}$$

To understand the sequential bias, imagine that an agent u improves her abilities to perform the real-effort task while she undertakes it — imagine she is *learning.* This would mean that her performance, no matter the treatment status, would be better when she executes the task a second time: $Y_{c2}(u) - Y_{c1}(u) > 0$ and $Y_{t2}(u) - Y_{t1}(u) > 0$. I would thus overstate the treatment effect if this was the case because the sequential bias would be positive. In contrast, If the task was exhausting such that agents perform worse when the execute the task a second time, I would understate the treatment effect. Estimating a treatment effect by calculating the observed difference therefore only make sense if

$$Y_{c1}(u) = Y_{c2}(u) = Y_c(u),$$

that is, if $Y_c(u) \perp\!\!\!\perp T(u)$. Note that this does not mean that the treatment status is independent of the timing. This would make no sense as the treatment status can be described as a deterministic function of the timing. Instead, the condition means that the value of the response variable (that is, the individual effort provision) is independent of the timing. As our design does not allow me to test whether this actually holds true, I have to assume it. In fact, this postulate breaks down into two identifying assumptions:

Identifying Assumption 1 *Causal Transience: $Y(u)$ is not permanently affected by the exposure of u to the control condition (such that c's effect is reversible).*

Identifying Assumption 2 *The effect of the control condition is the same at every point in time (now and in the future).*

If the same postulate is reasonable to assume for the exposure to the treatment, that is, if $Y_t(u) \perp\!\!\!\perp T(u)$, I am able to estimate the average treatment effect using the

[49] Note that there is no selection bias, as one compares the two measurements of the response variable of the same set of agents.

observed difference because:

$$\begin{aligned}\mathbb{E}[Y(u)|S(u) = t, T(u) = 2] - \mathbb{E}[Y(u)|S(u) = c, T(u) = 1]&\\ = \mathbb{E}[Y_{t2}(u) - Y_{c2}(u)|T(u) = 2]&\\ = \mathbb{E}[Y_t(u) - Y_c(u)].&\end{aligned}$$

To sum up, I can estimate the causal effect reciprocity has on the response variable, that is, the performance in the real-effort task, if the reader agrees with the assumptions I stated above.[50] Chapter 4 describes what the OLS specification looks like and which results one would expect given the theoretical predictions of Chapter 3.

A.3 The Statistical Solution: Regression Discontinuity Design

The previous section dealt with two postulates that allowed for a simple OLS regression. If these identifying assumptions, however, are unreasonable, one has to apply the 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 will use an agent's productivity Y_1 as the main explanatory variable. The treatment variable S, however, will be defined differently as it will indicate whether agents are expected to feel treated kindly or unkindly. The agents' performance in Stage 2 (Y_2) will serve as the response variable in this section.

RDDs[51] are based on a special kind of selection-on-observables story where the value of an agent's explanatory variable $Y_1(u)$ has not only a direct effect on the response variable $Y_2(u)$[52] but also on an agent's treatment assignment $S(u)$. Remember that the predictions (concerning reciprocal agents) depend on an agent's productivity and

[50] Note that it is relatively simple to test the assumption empirically by designing some additional sessions where the first stage is played (at least) twice.

[51] See (Lee, 2008) for a famous example of this method applied to study the incumbency effect in political economics.

[52] An agent's productivity is affected by her inherent ability (or costs of effort). This ability is likely to also affect the agent's performance in Stage 2. It is therefore reasonable to assume that an agent who is not able to click away 90 percent of the boxes will neither do so in Stage 1 nor in Stage 2. As such the ability, proxied by the productivity, affects an agent's response variable.

the exogenously set probability of $q = {}^1\!/\!_2$ (the probability of receiving the bonus payment under the random mechanism). If an agent is exposed to the performance-based mechanism while she was productive (that is, she clicked away more than 50 percent of the boxes in the first stage) she perceived the principal's choice as kind and exerts more effort than before. If she was, however, unproductive (she clicked away less than 50 percent of the boxes), she is predicted to perceive the principal's choice of the performance-based mechanism as unkind and exerts less effort in Stage 2 than before. Based on Y_1 the predictions therefore state who should feel treated kindly and unkindly in the following way:

$$S_u(u) = \begin{cases} h & \text{if} \quad Y_1(u) < {}^1\!/\!_2 \\ n & \text{if} \quad Y_1(u) = {}^1\!/\!_2 \\ k & \text{if} \quad Y_1(u) > {}^1\!/\!_2 \end{cases}$$

with h for *harshly* (or " unkind"), n for *neutral* and k for *kind.* A RDD is a useful strategy if there is no smooth transition (and thus a discontinuous jump) in the response variable between agents who perceived their matched principal's choice as kind and those who perceived it as unkind. Imagine the extreme case where agents respond to perceived kindness or unkindness by increasing or decreasing (respectively) their effort provision by a fixed amount of $|{}^\delta\!/\!_2| > 0$ as illustrated in Figure A.1. If one compares the Figure to Figure 3.3 it becomes clear that this discontinuity does not reflect my predictions entirely: Sure, the agents left to the threshold perceive the principals' choices of the performance-based mechanism as unkind while the agents to the right perceive them as kind but they are not expected to respond at $Y_1 = 0.51$ as strong as at $Y_1 = 0.75$. That is, $|Y_2 - Y_1|$ is predicted to grow continuously (and linearly) in $|Y_1 - {}^1\!/\!_2|$. This section neglects the continuity, as well as the neutral treatment status and explains how discontinuities can be exploited to identify a causal effect (I re-define $S(u)$ such that $S(u) = k$ if $Y_1(u) \geq {}^1\!/\!_2$). To make causal inferences, I rely on another, yet weaker, identifying assumption (Lee and Lemieux, 2010) than those I described above:

Identifying Assumption 3 *Agents, even while having some influence, were unable to precisely manipulate variable that assigns them into one of the treatment groups* Y_1.

In Chapter 4, I claimed that agents do not have perfect control about their productivity and argued that it is reasonable to make inferences using an RDD approach.

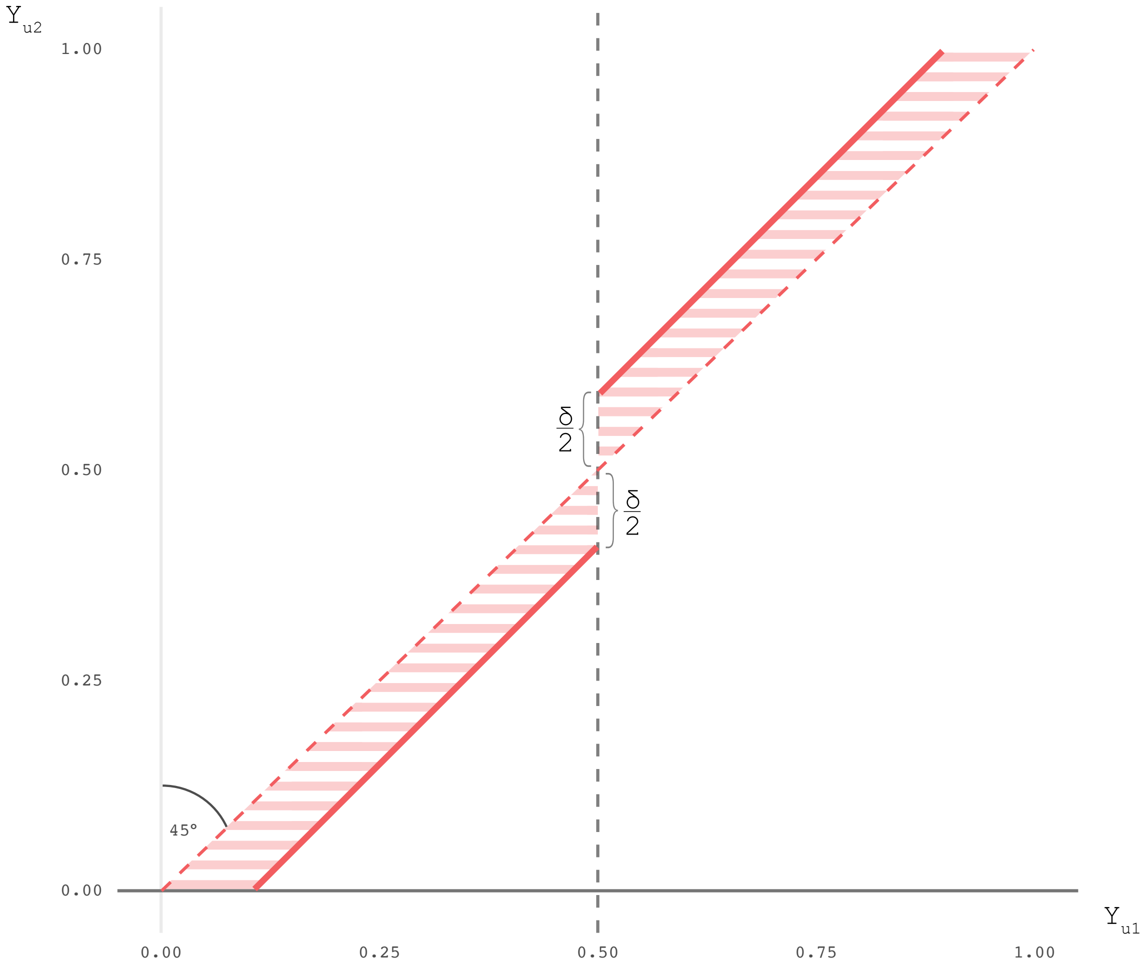

Figure A.1: Illustration of a discontinuous response to perceived (un-)kindness

Let y_0 instead of q denote the threshold and $Y_{k2}(u)$ the performance in Stage 2 as the response variable of agents if they were treated kindly (irrespective of their actual treatment status). $Y_2(u)$ denotes u's Stage 2 performance no matter to which of the two conditions she was exposed to. Further, let $Y_1(u)$ be an agent's productivity in the first stage. One can then demonstrate the unconfoundedness formally:

$$\mathbb{E}[Y_2(u)|y_0 - \varepsilon < Y_1(u) \leq y_0] \simeq \mathbb{E}[Y_{h2}(u)|Y_1(u) = y_0]$$
$$\mathbb{E}[Y_2(u)|y_0 + \varepsilon > Y_1(u) \geq y_0] \simeq \mathbb{E}[Y_{k2}(u)|Y_1(u) = y_0]$$

so that

$$\begin{aligned} &\lim_{\varepsilon \to 0} \mathbb{E}[Y_2(u)|y_0 - \varepsilon < Y_1(u) \leq y_0] - \mathbb{E}[Y_2(u)|y_0 + \varepsilon > Y_1(u) \geq y_0] \\ &\quad = \mathbb{E}[Y_{h2}(u) - Y_{k2}(u) \mid Y_1(u) = y_0]. \end{aligned}$$

A possible problem is that, by narrowing the neighborhood of y_0, I also prune a lot of data and might end up with too few observations to consider for my estimation.

If, however, there is enough data, the RDD story translates into the regression language using u as a subscript again as follows:

$$Y_{u2} = \alpha + \beta Y_{u1} + \delta S_u + \upsilon_u$$

with $S_u = \mathbb{1}_{Y_{u1} \geq y_0}$. Note that this regression's response variable is different from the previous strategy: the previous regression described properties of the observed differences which I aimed to interpret as causal. This regression, however, aims to explain the agents' behavior in Stage 2 (that is, their performance). Using the treatment assignment as an explanatory variable, I estimate the causal effect which I identify as δ.

Re-translating the regression into conditional expectations, one obtains:

$$\begin{aligned} \alpha + \beta Y_{u1} &= \mathbb{E}[Y_{uh2}|Y_{u1}] \\ Y_{uk2} &= Y_{uh2} + \delta \\ \Rightarrow \delta &= Y_{uk2} - Y_{uh2} \\ \Leftrightarrow \delta &= \mathbb{E}[Y(u)|S(u) = k, T(u) = 2] - \mathbb{E}[Y(u)|S(u) = h, T(u) = 2] \end{aligned}$$

I argue that δ estimates the (constant) causal effect of kindness relative to unkindness on the performance in Stage 2. In other words, I argue that the estimate is not biased because, as discussed earlier, the assignment into treatment is as good as random around the threshold.

Applying an RDD in our setting is best done using a graph such as the one illustrated in Figure A.1 where we see a discontinuity around y_0. Focusing on the solid lines, I sketched a jump of about 10 percentage points at the threshold. The resulting interpretation of such a result would be that agents who are predicted to feel treated unkindly performed, on average, 10 percentage points worse in Stage 2 than those agents who are predicted to feel treated kindly. Because agents to the left and to the right of Y_0 do not differ systematically, the best candidate to explain such a pattern would be kindness to the right of the threshold or unkindness to the left — or reciprocity for short. Note that Figure A.1 illustrates an extreme case with $\alpha = 0$ and $\beta = 1$ and without an interaction (which implies that the slopes on both sides are identical). I just used this parameterization for illustrative purposes. It might very well be that α and β take on different values (for instance, as predicted in the previous section). Eventually, this strategy is just about spotting the discontinuity. In fact, the regression line could also be non-linear apart from the discontinuity.

Questionnaire

The following questionnaire stems from Falk et al. (2016). Variable names in parentheses where the second letter denotes the participant's role (A denotes "Person A" who is the principal, B therefore denotes "Person B", the agent). The answers' enumerations reflect the variables' coding.

Q1 (RA1; RB1) Imagine the following situation: you are shopping in an unfamiliar city and realize you lost your way. You ask a stranger for directions. The stranger offers to take you with his car to your des na on. The ride takes about 20 minutes and costs the stranger about 200 DKK in total. The stranger does not want money for it. You carry six bottles of wine with you. The cheapest bottle costs 50 DKK, the most expensive one 300 DKK. You decide to give one of the bottles to the stranger as a thank-you gift. Which bottle do you give?

1. 50 DKK worth
2. 100 DKK worth
3. 150 DKK worth
4. 200 DKK worth
5. 250 DKK worth
6. 300 DKK worth

Q2 (RA2; RB2) How do you see yourself: If I am treated very unjustly, I will take revenge at the first occasion, even if there is a cost to do so.

1. Completely untrue of me
2. Mostly untrue of me
3. Slightly more untrue than true

4. Slightly more true than untrue

5. Mostly true of me

6. Describes me perfectly

Q3 (RA3; RB3) How do you see yourself: When someone does me a favor I am willing to return it.

1. Completely untrue of me

2. Mostly untrue of me

3. Slightly more untrue than true

4. Slightly more true than untrue

5. Mostly true of me

6. Describes me perfectly

Q4 (RA4; RB4) How do you see yourself: I assume that people have only the best intentions.

1. Completely untrue of me

2. Mostly untrue of me

3. Slightly more untrue than true

4. Slightly more true than untrue

5. Mostly true of me

6. Describes me perfectly

New Data

This appendix reports the data as well as the corresponding analysis for data that was collected in eleven sessions. Three of them were conducted in the last week of January in 2018. In total, the data stems from 117 observations, that is, from 234 participants.

I will not comment on these results. As they are similar to the ones reported in Chapter 5, however, you can read that chapter with this appendix as an accompanying booklet.

Table C.1: Summary statistics

	Productivity	Performance	Workload	Recip. Parameter
FULL sample of agents ($N = 117$)				
Mean	0.475	0.385	19.530	3.621
St. Dev.	0.123	0.210	8.785	0.773
Min	0.275	0.000	1	1.667
Median	0.445	0.398	25	3.667
Max	0.711	0.720	25	5.000
full subset of agents ($N = 79$)				
Mean	0.469	0.411	20.570	3.650
St. Dev.	0.121	0.206	8.093	0.768
Min	0.286	0.000	1	1.667
Median	0.439	0.424	25	3.667
Max	0.711	0.720	25	5.000
recip. subset of agents ($N = 35$)				
Mean	0.461	0.419	21.257	4.362
St. Dev.	0.116	0.200	7.931	0.356
Min	0.287	0.000	1	4.000
Median	0.437	0.424	25	4.333
Max	0.677	0.710	25	5.000
clean subset of agents ($N = 76$)				
Mean	0.462	0.426	21.342	3.654
St. Dev.	0.119	0.195	7.225	0.778
Min	0.286	0.000	1	1.667
Median	0.435	0.425	25	3.667
Max	0.711	0.720	25	5.000
smart subset of agents ($N = 74$)				
Mean	0.472	0.411	20.405	3.631
St. Dev.	0.120	0.210	8.274	0.776
Min	0.287	0.000	1	1.667
Median	0.439	0.423	25	3.667
Max	0.711	0.720	25	5.000

Table C.2: OLS estimates of the effect of reciprocity on the agent's working morale

	Response variable: Performance-Productivity (ΔY_u)			
	(1)	(2)	(3)	(4)
Subsets:	*full*	*recip.*	*clean*	*smart*
β : Productivity	0.06	0.03	0.32***	0.08
	(0.15)	(0.24)	(0.11)	(0.16)
α : Constant	−0.09	−0.05	−0.19***	−0.10
	(0.07)	(0.11)	(0.05)	(0.08)
Observations	79	35	76	74
R^2	0.002	0.0004	0.10	0.003
F Statistic	0.16	0.01	8.39***	0.22

*p<0.1, **p<0.05, ***p<0.01. [1]Estimated as the observed difference (ΔY_u) $\neq 0$. Standard errors in parentheses. The unit of observation is a participant who was assigned the role of an agent ('Person B'). The *full* sample includes only agents who where exposed to the performance-based from the subset of agents, who were exposed to the performance-based mechanism in Stage 2. The *recip.* subset includes all the observations from the *full* sample without those, who are below-median reciprocal according to the final questionnaire. The *clean* subset includes all the observations from the *full* sample without the three extreme-values. The *smart* subset includes all the observations from the *full* sample without the observations who attracted attention during the sessions.

Table C.3: Contingency table for occurrences of non-maximal workload decisions

Mechanism	Unproductive	Productive	Total
Performance-based	17	5	22
Random	6	13	19
Total	23	18	41

The unit of observation are groups of a principal and the matched agent. This Figure focuses on the agents' behavior. A non-maximal workload decision is measured as choices smaller than 25. A threshold of 0.5 is used to distinguish between high and low productivity. The null hypothesis, that non-maximal workload choices are distributed evenly across the agents' productivity and the mechanism they faced, can be rejected ($p < 0.005$).

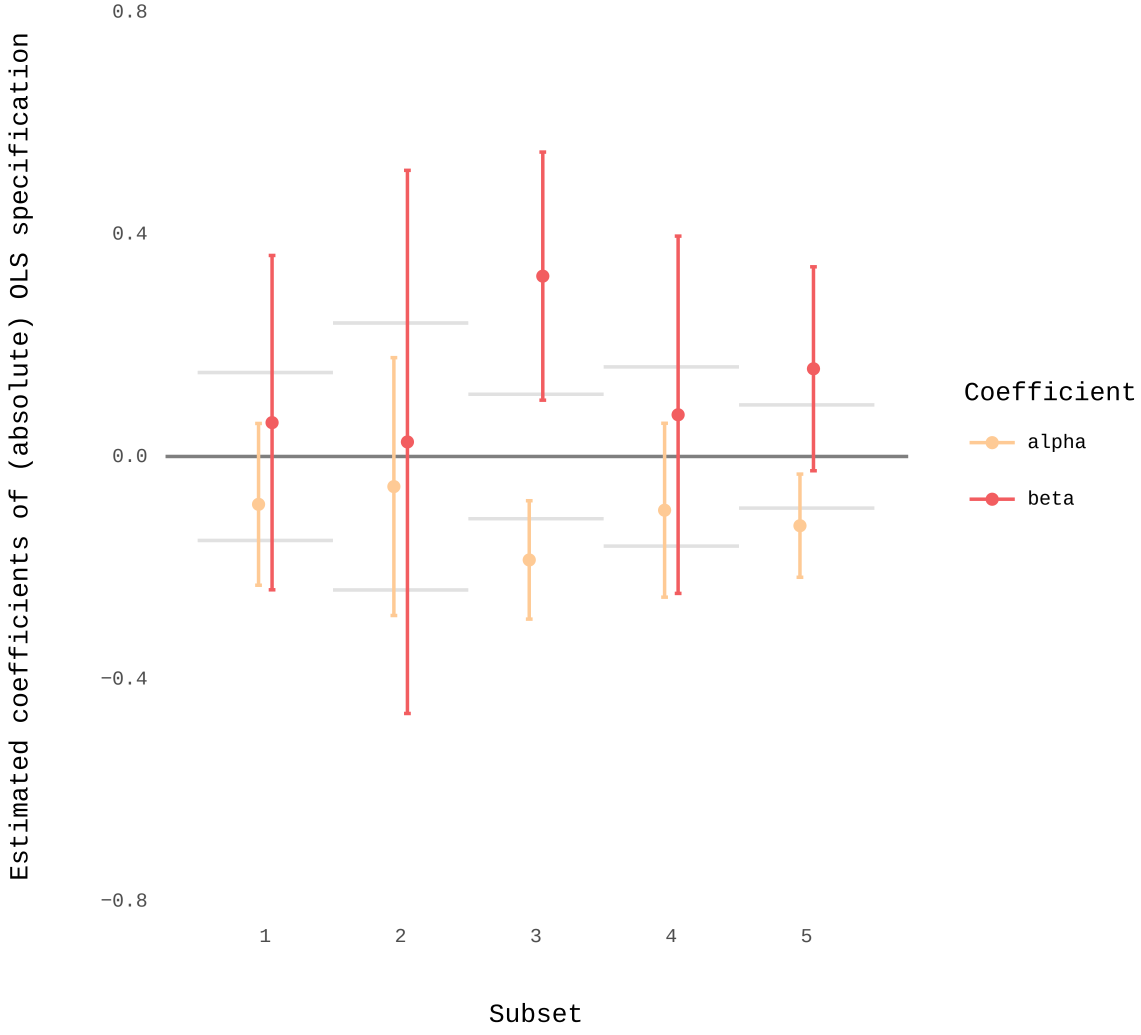

Figure C.1: 95-percent confidence intervals of the coefficients estimated in Table 5.2

The unit of observation are groups of a principal and the matched agent. This Figure focuses on the agents' behavior. *alpha* refers to the constant estimated in the OLS specification used throughout this chapter and *beta* refers to the slope coefficient of Y_1. (1) refers to the *full* subset, (2) to the *recip.* subset containing observations with above average (self-proclaimed) reciprocity parameters, (3) to the *clean* subset that excludes the extreme values, (4) to the *smart* subset who is expected to have understood the instructions and (5) to the *double* subset that combines (1) and (3). The gray vertical lines represent the slope coefficients' (β) standard errors.

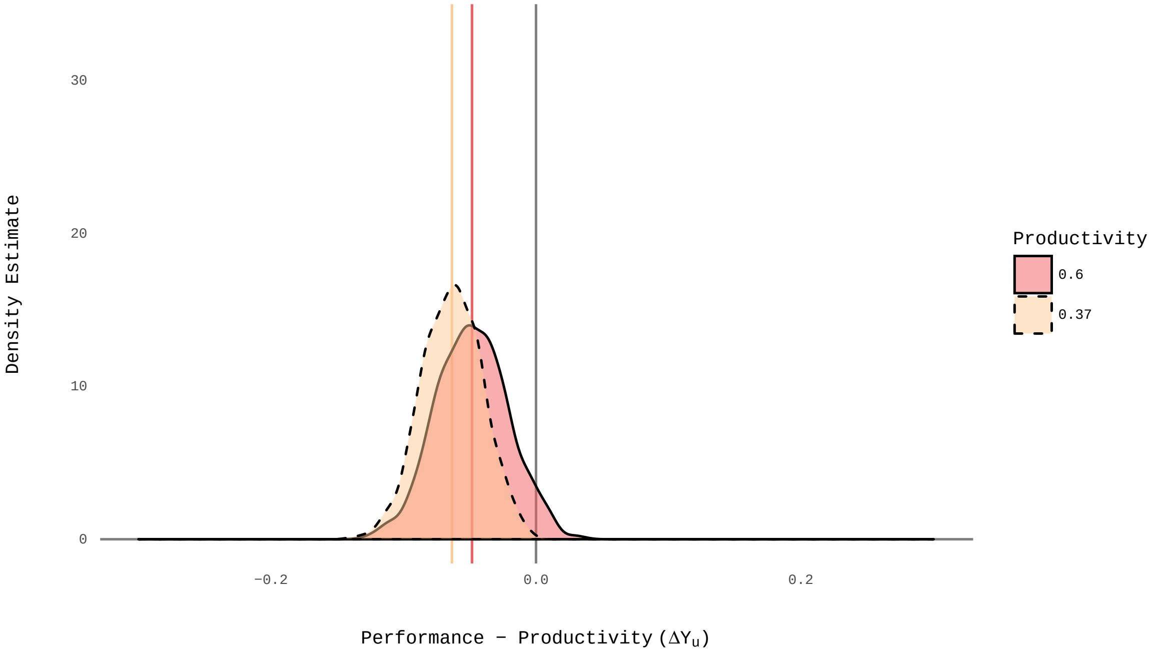

Figure C.2: Simulated levels of productivity based on regression coefficients from the full *performance-based subset*

The unit of observation are simulations of agents with a productivity of either 0.6 or 0.37. These levels correspond to the means of productive and unproductive agents in the *full* subset. The simulations are based on coefficients estimated in the first column of Table 5.2. The density estimate is based on 1000 simulations for each productivity level. The Simulation process is described in Footnote 38.

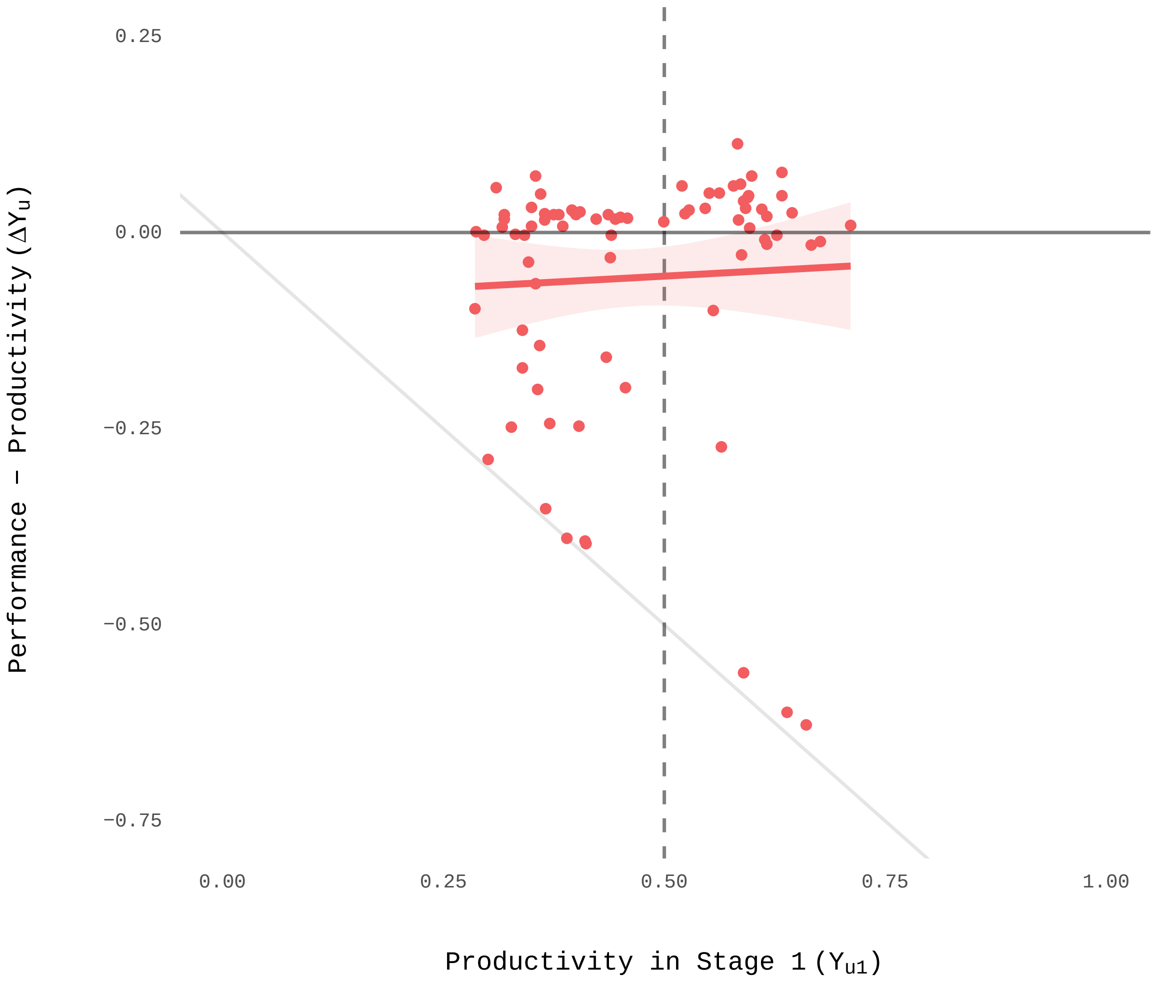

Figure C.3: OLS Specification for the full *performance-based subset with 95% confidence interval*

The unit of observation are groups of a principal and the matched agent. This Figure focuses on the agents' behavior. The red regression line corresponds to the first column of Table 5.2. The vertical dashed line prints the threshold that splits the data into productive and unproductive. The downward sloped line censors the data as a data point cannot possible be lower than this value.

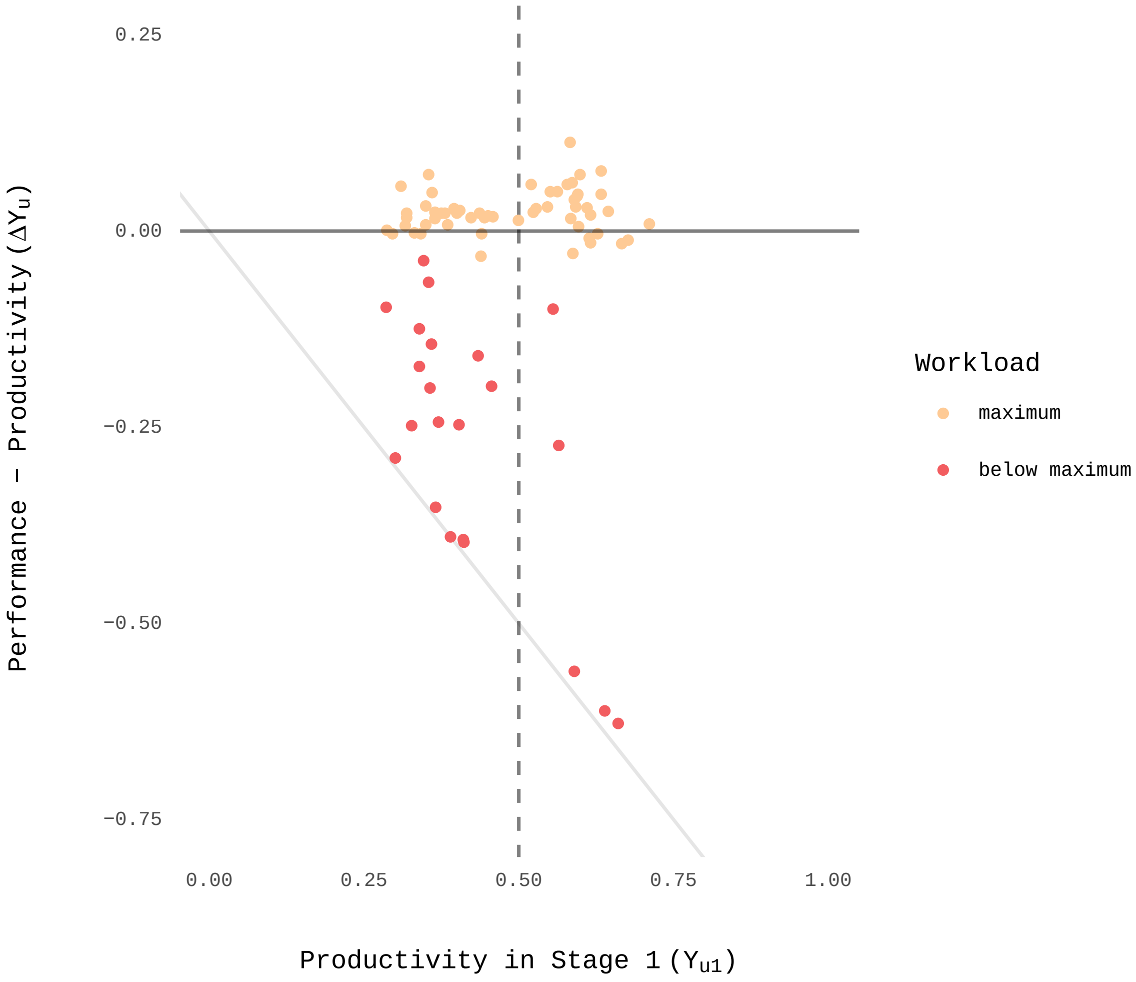

Figure C.4: Scatterplot of the clean *performance-based by workload decisions*

The unit of observation are groups of a principal and the matched agent. This Figure focuses on the agents' behavior. The workload is either at its maximum of 25 screens (red) or below this value (yellow). The vertical dashed line prints the threshold that splits the data into productive and unproductive. The downward sloped line censors the data as a data point cannot possible be lower than this value.

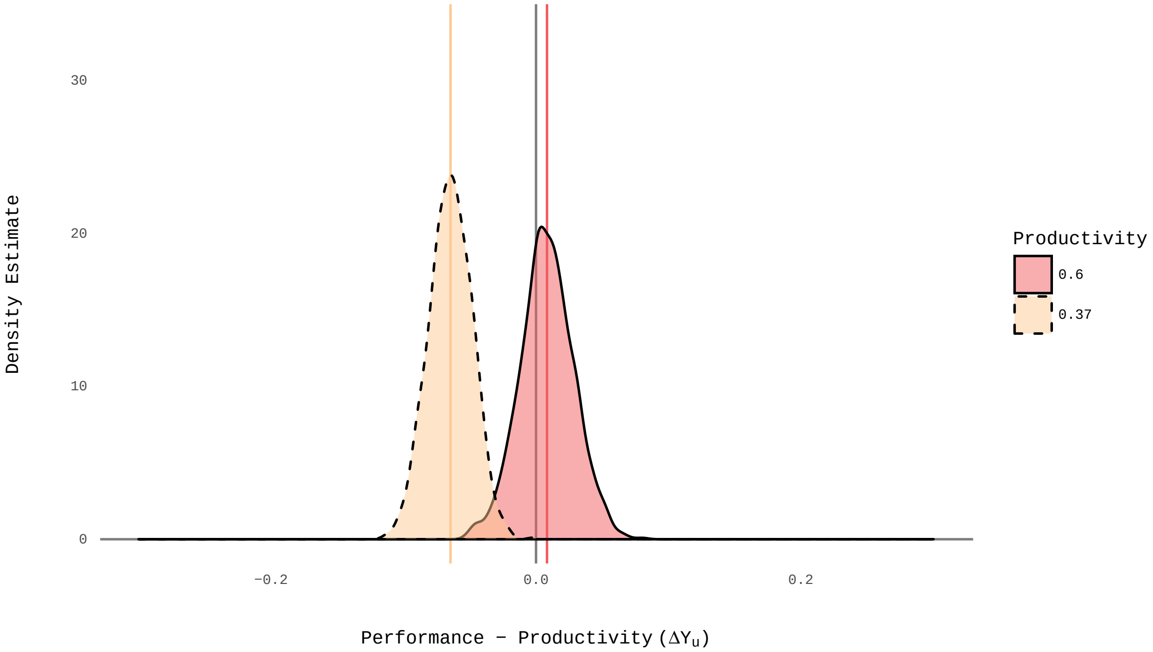

Figure C.5: Simulated levels of productivity based on regression coefficients from the clean *performance-based subset*

The unit of observation are simulations of agents with a productivity of either 0.6 or 0.37. These levels correspond to the means of productive and unproductive agents in the *clean* subset. The simulations are based on coefficients estimated in the third column of Table 5.2. The density estimate is based on 1000 simulations for each productivity level. The Simulation process is described in Footnote 38.

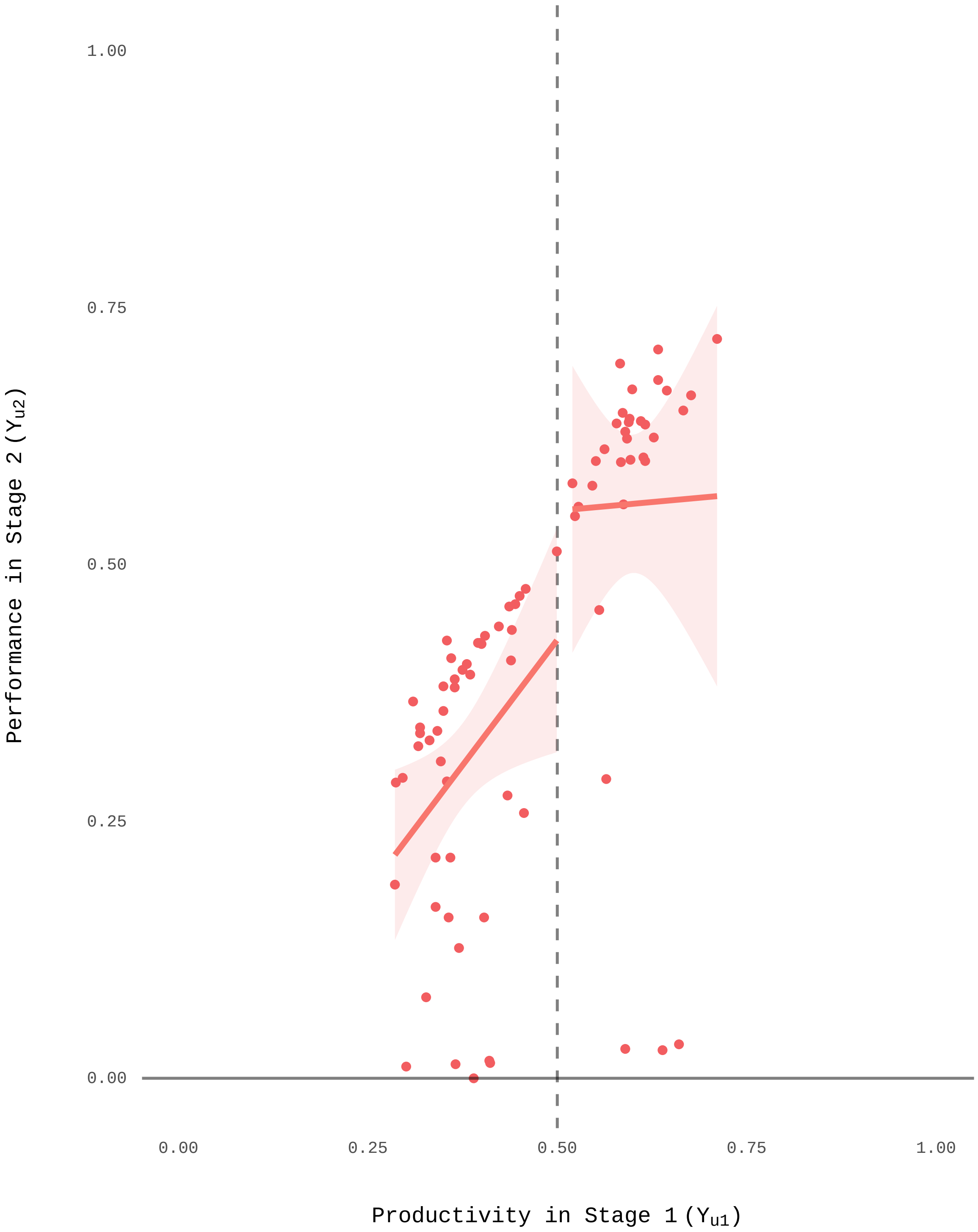

Figure C.6: RDD for the FULL *data set with 95% confidence intervals*

The unit of observation are groups of a principal and the matched agent. This Figure focuses on the agents' behavior. The lower panel plots all the observations where the principal chose the random mechanism while the upper one plots the *full* subset. The vertical dashed line prints the threshold that splits the data into productive and unproductive. It is thus, the arbitrary value that assigns the observations' treatment statuses (in the RDD logic).