# The Hidden Costs of Whistleblower Protection*

Niklas Wallmeier†

January 24, 2023

**Abstract**

I conduct a laboratory experiment to analyze the influence of whistleblower protection on the cooperative behavior between a manager and an employee. Before taking part in a trust game with her employee, a manager has the opportunity to embezzle money at the expense of a third party. The employee observes her decision and may trigger an investigation by blowing the whistle. The treatments vary with respect to immunity and anonymity for the whistleblower. I compare misbehavior, reporting, and the cooperative behavior across the treatments. The results suggest that whistleblower protection could deter wrongdoing, but could also have a detrimental effect on cooperation if it makes it harder for the employee to signal trustworthiness.

**JEL-Codes**: C91, D73, K42, M51

**Keywords**: corporate fraud, whistleblowing, business ethics, laboratory experiment.

*I would like to thank Ralph Bayer, ohn DeNew, Shahar Dillbary, Leonie Gerhards, Michael Kurschilgen, Christos Litsios, Gerd Muehlheusser, Petra Nieken, Andreas Roider, Karl Schlag, Fanny Schories, Simeon Schudy, Kathrin Thiemann and Tom Wilkening as well as seminar and conference participants in Cologne (LEOH), Essen (RWI), Hamburg (UHH) and Melbourne (UoM) for their valuable comments and suggestions. Pamela Mertens and Olaf Bock provided excellent research assistance. Financial support by the Fritz Thyssen Foundation (Grant 10.13.2.097) is gratefully acknowledged.

†Department of Economics, University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany; niklas.wallmeier@uni-hamburg.de.

# 1 Introduction

In an era of corporate fraud causing severe damages, whistleblowing is a major source of fraud detection (see e.g., Dyck et al., 2010). Consequently, whenever insiders unveil a large corporate scandal, there emerges public demand to support whistleblowers by providing legal protection.[1] This paper investigates experimentally the behavioral effects of protection for whistleblowers in the form of immunity and anonymity. It is the first paper that analyzes how changes in the legal framework affect the cooperative climate between an employer and an employee. The results suggest that laws, which increase the frequency of whistleblowing, do not only drive down managerial wrongdoing, but also lead to a decline in productive cooperation, as it becomes more difficult for the employee to signal trustworthiness.

Becoming a whistleblower comprises a non-negligible trade-off for an employee. The fear of retaliation, e.g., a dismissal or a denied promotion, is a major obstacle that often thwarts whistleblowing (see e.g., Near and Miceli, 1986; Alford, 2001; Cassematis and Wortley, 2013). Therefore, whistleblowers might be encouraged to come forward by legally protecting them from retaliation. To this end, international organizations (see e.g., OECD, 2016) requested protection for whistleblowers. Some legislators already made an effort to increase the legal certainty. The most-frequent features of whistleblower protection are *immunity*, which means to guarantee income (see e.g., Kohn et al., 2004, pp. 97), and *anonymous reporting* (see e.g., Thüsing and Forst, 2016).

However, these legal approaches are discussed controversially, since the support of whistleblowers might come at a cost. Whistleblowing laws often condition the protection only on a "reasonable belief" (Kohn et al., 2004, pp. 92). While obviously unfounded complaints are deterred with this standard,[2] an adverse effect may be nevertheless an increase in false claims. That means an employee blows the whistle to be protected from a dismissal, although there was no misbehavior by her employer (see e.g., Callahan and Dworkin, 1992; Howse and Daniels, 1995; Givati, 2016). Such false claims would cause financial, or reputational, damage from investigations for the organization and increase the costs of the authorities to screen claims for their adequacy (see e.g., Mechtenberg et al., 2020).

Moreover, the efficient work within an organization relies on productive cooperation, which requires a sufficient level of trust.[3] Employees might jeopardize this trust if they use their insider status to report the employer. While a report could be beneficial for altruistic

[1] I focus on whistleblowing as organization members disclosing illegitimate practices under the control of their employers to organizations that may be able to effect action defined by Near and Miceli (1985).

[2] Buccirossi et al. (2021) shows theoretically how to deter unfounded reports by sufficiently high fines.

[3] For example, Bloom et al. (2012) find that higher levels of trust allow for a lager degree of delegation and therefore a larger firm size.

reasons (see e.g., Bartuli et al., 2016; Heyes and Kapur, 2009), or financial reasons if reporting is incentivized, refraining from a report could demonstrate loyalty and let the employee appear more trustworthy. In consequence, the employer may not dismiss the whistleblower to punish her, but perceive whistleblowing as a signal of the employee being less trustworthy such that future collaboration appears non-profitable.[4] Therefore, if a whistleblower protection law encourages more reporting, it becomes more difficult for the manager to receive a signal indicating the trustworthiness of her employee. This would cause the manager to distrust her employees more often, which has detrimental effects on cooperation (see e.g., Dworkin and Near, 1997; Vinten, 1994; Walters, 1975).

To study the relationship between whistleblower protection and the cooperative climate, I create a lab experiment in which an employee can report wrongdoing of her manager who subsequently decides whether to cooperate with the employee. At the beginning of a period, the manager has the opportunity to embezzle money and increase her payoff at the expense of a third party. The employee observes her choice. While the embezzlement does not affect the employee's payoff, she can become a whistleblower and trigger an investigation by reporting misbehavior to an authority. In contrast to other studies (see e.g., Mechtenberg et al., 2020), I model the authority to respond perfectly to a report, which reflects the standard of a reasonable belief. In consequence, the manager can tell from an investigation that the whistle was blown. If the employee files a report, the manager suffers a cost from the investigation and has to pay a fine if she had embezzled. At the end of a period, the manager and the employee interact in a modified version of the *trust game* (Berg et al., 1995). As the sender, the manager decides first which amount to send to her employee or to take from her employee's endowment. In this respect, the game is similar to the *moonlighting game* (Abbink et al., 2000). If she sends a positive amount, productive cooperation takes place. That means the amount is multiplied, and the employee can decide which fraction she wants to return. If the manager takes some of the endowment of her employee (i.e., beneficial cooperation does not take place), the amount is simply transferred.

Compared to a baseline treatment without protected whistleblowing ($B$), the employees are protected by either immunity ($I$), or anonymity ($A$), or both ($AI$). Immunity means that the manager cannot take any of the employee's endowment if the employee filed a report. Recent experimental economics literature provides broad evidence that monetary incentives increase the willingness to report of potential whistleblowers (see e.g., Mechtenberg et al., 2020; Schmolke and Utikal, 2018; Butler et al., 2020). Anonymity allows the employee to report without revealing her action prior to the trust game to the manager.

[4] In this regard, findings from experimental economics show that pre-play experience significantly influence the level of trust in subsequent interactions (see e.g., Bracht and Feltovich, 2009; Fehrler and Przepiorka, 2016; Gambetta and Székely, 2014; Heyes and List, 2016).

Therefore, the manager cannot condition her cooperation on the employee's reporting decision. There is evidence from laboratory experiments that managers retaliate against known whistleblowers (see e.g., Mir Djawadi and Nieken, 2019; Reuben and Stephenson, 2013).

I derive my hypotheses from a model that assumes that a manager faces an employee of an unknown "loyalty type". A loyal type suffers moral costs rather from being disloyal to the manager, for example not reciprocating an investment, and less from the damage caused by the embezzlement by the manager. For a disloyal type, it would be the other way around. Consequently, a manager wants to cooperate with a loyal employee, but not with a disloyal employee. The intuition of the model is as follows: If the reporting decision of the employee would perfectly reveal the loyalty type, the manager would not cooperate if she witnesses a report. If this holds, a loyal type would not report the manager, since the moral costs from being disloyal to her outweigh those from undetected embezzlement. The disloyal type, however, would weigh higher the moral costs from the undetected embezzlement and report her manager. Therefore, embezzlement allows the manager to screen the type of her employee. Immunity for a whistleblower would make reported embezzlement more costly for the manager. In consequence, there would be less embezzlement and therefore less screening of the employees, which may lead to less cooperation. With anonymous reporting, the manager cannot identify the type of the employee and has to make her decision on cooperation based on her belief about the probability to face a loyal type.

In the context of whistleblowing, a laboratory approach has two major advantages compared to the field. First, only detected misbehavior is observable in actual organizations, such that the true amount of misbehavior remains unknown. Second, we only observe reporting behavior conditional on misbehavior. That means we can account for truthful reporting when there is misbehavior and for false reporting in there is none, but not for the hypothetical behavior in the state that has not been realized. In addition, a number of studies show a high out of lab correlation in unethical behavior (see e.g., Abeler et al., 2019).

The results show that whistleblower protection increases honest reporting and, in turn, reduces embezzlement. At the same time, it provokes adverse incentives for the employees and lead to an increase in false whistleblowing. For the managers' willingness to cooperate, I find a positive influence of unreported embezzlement. As embezzlement is mostly deterred in treatment *AI*, where protection features both immunity and anonymity, unreported embezzlement does not occur, which drives down cooperation significantly.

The remainder of the paper is organized as follows: Section 2 reviews the related literature, Sections 3 and 4 present the experimental design and the behavioral predictions, while Section 5 and 6 present and discuss the results.

# 2 Related Literature

As this study investigates the relation of whistleblowing and cooperation, it contributes to the literature on the effectiveness of whistleblowing and whistleblower protection. Recent studies have used laboratory experiments to test how potential whistleblowers respond to protection in the form of incentives. Schmolke and Utikal (2018) find that fines for non-reporting, rewards, and commands increase the probability of whistleblowing. Moreover, reporting is more likely, if the misconduct affects the whistleblowers themselves, or the enforcement authority. Butler et al. (2020) investigate the effect of monetary rewards on whistleblowing in the presence of potential crowding out of intrinsic motivation. They find an enhancing effect of monetary rewards on the willingness to report and no evidence for substantial crowding out of non-monetary motivations. In a field experiment, however, Fiorin (2023) finds employees in the education system are less willing to blow the whistle on their peers' absence once it is incentivized and if there are consequences for the wrongdoer.

Furthermore, studies have found mixed evidence on the effect of whistleblowing on the efficiency within organizations. In a theoretical model, Friebel and Guriev (2012) show that the possibility for whistleblowing might harm a firm's productive efficiency if wrongdoers "bribe" other members of the organization as this could undermine effort incentives. Felli and Hortala-Vallve (2016) provide a model in which incentivized whistleblowing can prevent opportunistic behavior that takes the form of collusion or blackmail between supervisors and employees.

Mechtenberg et al. (2020) consider both the effectiveness and the efficiency of whistleblower protection. They investigate the effects of whistleblower protection on reporting and on the efficiency of law enforcement in a theory-guided lab experiment. Their findings show that when the legal protection provokes false reporting, whistleblowing becomes a less informative signal such that more reports do not necessarily materialize in more investigations. Since the employees are externally heterogeneous with respect to their productivity, a dismissal could be driven either by efficiency concerns or by preferences for retaliation.

I complement these studies by considering the connection of the effectiveness of whistleblower protection and the efficiency of whistleblowing within the organization. This study features the effect of protected whistleblowing on the reporting behavior and deterrence as well. Moreover, I add a dimension that captures the effect of whistleblowing on efficient cooperation. More precisely, the reporting decision is a possibility to signal trustworthiness to the organization, which is crucial for cooperation to take place. Therefore, by deterring misbehavior, whistleblower protection affects the frequency of reportable misbehavior, which may make it harder for employees to signal trustworthiness.

# 3 Experimental Design

## 3.1 The Game

To investigate the influence of whistleblower protection on misbehavior, reporting behavior, and cooperative behavior, I combine a *whistleblowing game* with a modified trust game. The subjects take either the role of a manager, an employee, or a third party. While the third party is completely passive, both the other roles have to make up to two decisions. In the whistleblowing game, the manager decides in the first stage whether to comply with the law ($e = 0$) or to embezzle money ($e = 1$), which generates a revenue for her and a cost for the third party. In stage two, the employee decides whether to stay silent ($r = 0$) or to file a complaint ($r = 1$). She makes this decision conditional on the embezzlement decision of the manager. This means that the employee decides about reporting truthfully ($r^t$) in case the manager embezzles, and about reporting falsely ($r^f$) in case the manager complies.[5]

The trust game starts in stage three. The manager decides whether cooperation takes place by choosing the level $c \in [-30, 60]$. She can choose a negative amount, which means that she would take some of the employee's endowment. If she trusts her employee (i.e., $c$ is positive), this amount is multiplied by three and transferred to the employee. In stage four, the employee can return an amount $t$ to her manager, if $c$ was positive. If the employee has reported in stage two, an investigation takes place at the end of a period, which is costly for the manager. If this investigation reveals embezzlement, the manager has to pay a fine. Moreover, the damage for the victim is partly recovered.

**Cost and reward parameters** There are four possible combinations of the decisions on embezzlement and reporting $(e, r)$. These can be ranked in terms of social welfare $\pi_S(e, r)$ if three assumptions hold: (i) compliance is better than embezzlement, (ii) detected embezzlement is better than undetected embezzlement, and (iii) in case of compliance, the employee should not report. The order is given by

$$\pi_S(e = 0, r = 0) > \pi_S(e = 0, r = 1) > \pi_S(e = 1, r = 1) > \pi_S(e = 1, r = 0).$$

I chose the cost and reward parameters (in parentheses) such that these assumptions hold. The intuition is as follows: The most preferred outcome would be to have no embezzlement and no report ($e = 0, r = 0$). In this case, there is neither damage from embezzlement nor from an investigation, which leaves all players with just their endowment ($\Delta\pi_S = 0$). The

[5] Using the strategy method (Selten, 1967) allows to keep track of the reporting behavior independent of the compliance behavior. Brandts and Charness (2011) suggest that using the strategy method should not yield different results if the decision maker is not directly affected, which applies in this context.

least favorable outcome is undetected embezzlement ($e = 1, r = 0$). Here, the manager earns a benefit (50), which is outweighed by the cost for the third party (90). This would result in a social net loss ($\Delta\pi_S = -40$). A preferable outcome would be detected embezzlement ($e = 1, r = 1$). The manager would have to pay a fine (60), which exceeds her benefit from embezzlement, and the costs of the investigation (10). On the other side, the third party partially recovers her loss $R$ (80), such that social welfare loss is lower ($\Delta\pi_S = -30$). The fourth possibility is a false claim ($e = 0, r = 1$). This means that there is neither a damage for the third party nor a benefit for the manager, but it creates an investigation cost (10) for the manager ($\Delta\pi_S = -10$).

For the trust game, I impose a range from $-30$ to 60 (with discrete steps of length ten) on $c$. A manager, who does not want to cooperate, because she does not expect this to be beneficial, could just choose $c = 0$. However, choosing a negative amount for $c$ would indicate that the manager punishes an employee she does not trust. This decision could be interpreted as a dismissal or a denied promotion. Furthermore, the gradations of $c$ give the manager the opportunity to differentiate whether she wants to recover the damage the employee caused–that is the loss from a false report ($c = -10$), or from a true report ($c = -20$)–or whether she wants to maximize her payoff ($c = -30$). These different values reflect that a manager could choose a very strict or a rather mild punishment in a real world setting, e.g., she could offer a more or less generous severance pay when she dismisses the employee. For positive values of $c$ the upper bound is set to 60. This guarantees that the employee cannot punish the manager stronger by keeping the entire investment than by reporting. The endowment is set sufficiently high (100) that neither party could make a loss nor is restricted in her choice set. Below, the payoffs for the three roles in a period, given the decisions of the subjects, are summarized.

$$\pi_{Manager} = 100 + e \times (50 - (60 \times r)) - r \times 10 - c + t \tag{1}$$

$$\pi_{Employee} = 100 + \begin{cases} c \times 3 - t & if\ c > 0 \\ c & if\ c \leq 0 \end{cases} \tag{2}$$

$$\pi_{3rdParty} = 100 - e \times (90 - (80 \times r)) \tag{3}$$

## 3.2 Treatments

I vary the legal environment in the treatments in two dimensions: i) immunity, which means an insurance against a monetary loss and ii) anonymity, which means that the employee has not to reveal her reporting decision to the manager. This results in four treatments. These differ with respect to the choice set for the manager in the trust game conditional on the reporting decision of the employee (immunity) and the date when the

manager is informed about the reporting decision (anonymity).

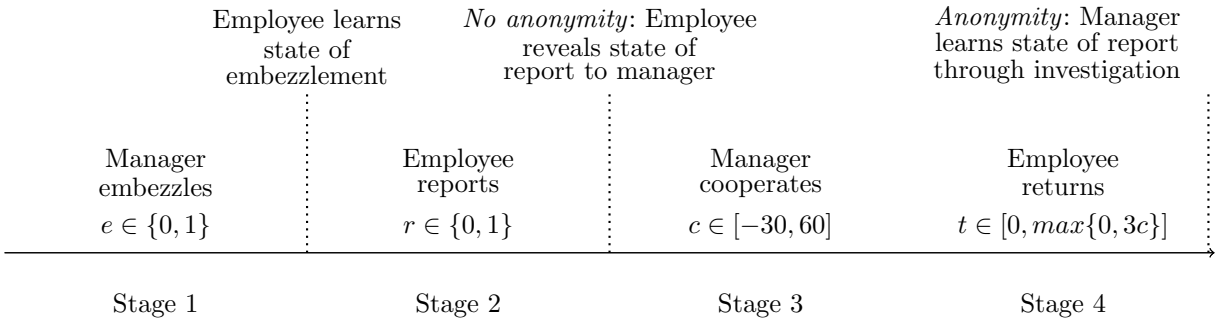


Figure 1: Timing in a period with and without anonymity

**Baseline Treatment (*B*)** In the baseline treatment, the manager knows after stage two about the employee's reporting decision, i.e., *before* she chooses $c$. Further, she is free to choose a negative $c$ independent of the reporting decision (compare to Figure 1).

**Immunity Treatment (*I*)** In treatment $I$, in which only immunity is introduced, the manager knows the employee's reporting decision after stage two as well. In this treatment, immunity is modeled such that by filing a report the employee can guarantee her status quo payoff. That means, if there has been a report, truthful or false, $c$ has to be at least zero.

**Anonymity Treatment (*A*)** In treatment $A$, in which only anonymity is granted, the information about whistleblowing is disclosed only after stage four through the investigation, i.e., *after* the manager chose $c$. The choice of $c$ is again unrestricted for any reporting decision. This change in the timing guarantees that the manager cannot condition cooperation on the actual behavior of the employee.

**Anonymity and Immunity Treatment (*AI*)** In treatment $AI$, with both immunity and anonymity, the manager knows only after her choice of $c$ whether the employee reported. In case the manager chose a negative $c$, it is set to ex-post.

## 3.3 Implementation

**Session design** The decision whether to implement these treatments with a between-subject or a within-subject design contains several trade-offs. Between designs are more conservative, but may have limitations in relation to testing several variations. On the other hand, within designs are more powerful, but can suffer from confounds (for discussion, see e.g., Charness et al., 2012; Moffatt, 2015). A deciding factor for the design choice

is the research question at hand and its practical implications. This study is motivated by the debate on supporting whistleblowers by introducing whistleblower protection, for example in the form of immunity and anonymity. Therefore, a natural design appears to be a within variation to observe a change in behavior after the whistleblower protection is introduced. Of course, confronting the subjects with four different treatments would pronounce the disadvantages of a within design, for example, the issue of order effects. Consequently, I chose to have just one of the dimensions varied for the same subject. Introducing anonymity means a larger variation, since it changes the information structure within a period, while immunity only changes the choice set for the trust game. Therefore, I model the introduction of immunity as a within-subject variation, while I vary anonymity in a between-subject design.

Still, the within design has to be implemented carefully as the treatment before the intervention may influence the results in the treatment after the intervention. To mitigate the influence of the treatments before the intervention on those after, I made two design choices: First, there was short break between the treatments, where the subjects received the new instructions. In this way, I tried to separate the treatments as effectively as possible. Moreover, I chose a relatively large number of periods per treatment (eight per treatment, 16 per session, see Figure 2). Consequently, the subjects have time to gain experience about the behavior of the other subjects in a given treatment. For example, a manager learns how often she is reported and how a report relates to the return behavior in the trust game. After immunity is introduced, the manager needs to learn how this protection translates into reporting behavior and to update how this possibly different reporting behavior is related to the return behavior. However, these measures do not rule out a confound of the treatments. Therefore, it is crucial to control for experience as well for period fixed effects in the analysis. In addition, I provide figures for the decisions over time, which allow identifying patterns of potentially confounding factors (see Figures D.1-7).

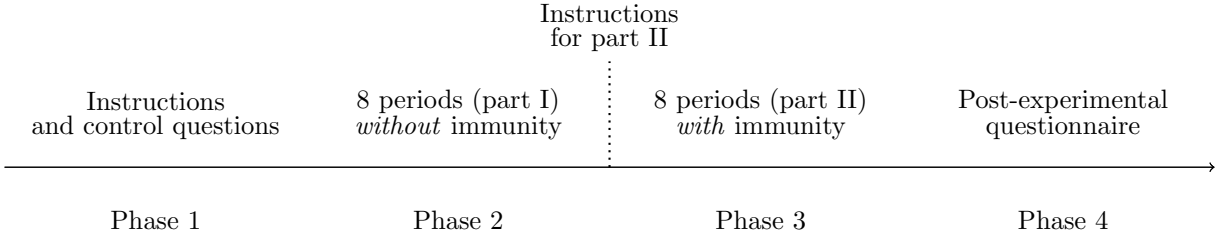


Figure 2: Timing in a session

**Framing** In experimental economics, there is a discussion on the conditions under which a neutral or a loaded framing is more appropriate.[6] In this regard, I chose a compromise similar to Mechtenberg et al. (2020). I framed the experiment in a workplace context and spoke of employers and employees to support the subjects in understanding the hierarchical relation between the players. Furthermore, I gave a reminder that the alternative corresponding to embezzlement means a violation of the law. Thereby I model an important feature of unethical decision-making in the real world, since wrongdoers are clearly aware of that such decisions are illegal. The employee's decision about a report was phrased as 'filing a complaint' to make them aware of the social undesirability of embezzlement. Drawing attention to unethical behavior may influence the subjects' decisions, which would be appropriate for this specific research question, though. Although a mixed framing may not do as well as a purely loaded framing in terms of clarifying the instructions, I phrased the choice about embezzlement in a neutral way (alternatives: CIRCLE or TRIANGLE). This was motivated by the possibility that subjects may bring in their individual perception of the severeness of a specific misbehavior. In this regard, embezzlement might be perceived as rather mild or rather serious misbehavior. Therefore, a neutral framing should prevent that the individual perception of embezzlement influences the behavior. Moreover, I used payoff tables (see Appendix B.1) and control questions (see Appendix B.2) to ensure that the precise consequences for all players are understood.

**Procedural details** At the beginning of the experiment, the subjects received instructions, which explained the game described above. They were informed that that this game will be played for eight periods before they receive instructions for the second part of the experiment.[7] Furthermore, they were told that managers keep their role throughout the experiment, while the other two roles are reshuffled after each period. The motivation for reshuffling the roles of the employee and the third party was twofold: First, it should make the harm of embezzlement more salient. Second, this procedure allows for a larger number of independent observations and reduces the likelihood that managers and employees face each other multiple times. Before a period started, groups of three were randomly formed with one subject of each role. The subjects face a stranger matching and cannot infer any information about their group members from previous periods.

While I asked the control questions at the start of a session, subjects completed a (non-incentivized) questionnaire in which I elicited socio-demographic information (e.g., age,

[6] Alekseev et al. (2017) survey a wide range of experimental literature with respect to the instructions and find that meaningful language could be useful for understanding the environment. For the context of unethical behavior see Abbink and Hennig-Schmidt (2006); Barr and Serra (2009).

[7] The second instructions only added that for the upcoming eight periods the managers could not choose a negative amount for $c$ if the employee filed a report.

| Treatment | Managers | | Employees | |
|---|---|---|---|---|
| | subjects | decisions | subjects | decisions |
| $B$: | 30 | 240 | 60 | 480 |
| $I$: | | 240 | | 480 |
| $A$: | 19 | 152 | 38 | 304 |
| $AI$: | | 152 | | 304 |

Table 1: Observations per Decision

gender, and field of study), risk preferences (via the "100,000 euro" question of Dohmen et al., 2011), and their attitudes towards revealing misbehavior (measured on a five-level Likert scale) at the end. With these questions, I wanted to elicit whether the attitudes differ between the subjects across the treatments, since this may influence the results. However, comparing subjects with different roles within a treatment, and subjects with same role across the treatments, I do not find statistically significant differences for the reported attitudes. (see Table C.1 for the average characteristics).

The experiment was programmed with the software z-Tree (Fischbacher, 2007). It was conducted in the laboratory of the University of Hamburg, June 2016, and I used hroot for recruitment (Bock, Baetge, and Nicklisch, 2014). I ran five sessions with a total number of 147 student subjects (65% female, average age: 25 years, the majority of the subject were enrolled in economics or business programs). Four sessions had 30 participants (ten groups per period), one had 27 participants (nine groups per period). The number of subjects per role and treatment as well as the total number of obervations is summarized in Table 1.

To keep the incentives identical for every period over the entire experiment, after the questionnaire has been completed, one period was randomly drawn for payout. The subjects received payments between 5.50 and 18.50 euro (including a show-up fee of 5 euro) with an average of 10.07 euro.

# 4 Behavioral Predictions

In this section, I establish the behavioral predictions on the employees' willingness to report–truthfully and falsely–and their return behavior as well as on the decision of the managers to embezzle and to cooperate for the different treatments. I derive the behavioral predictions from the equilibrium analysis of a simplified version of the game described in Section 3, which is formally spelled out in Appendix A (see Propositions 1 - 4).

**Summary of the theoretical model** In contrast to the game played in the experiment, I assume that the trust game, played in the stages three and four, consists of two binary decisions: The manager decides to cooperate or not, the employee returns either a high amount or a low amount. The high amount is larger than the investment of the manager, while the low amount is smaller. Further, I assume that there are two types of employees: the "loyal" type has a relatively high moral cost from being disloyal to the manager–that is, to return less than the manager invested in the trust game, or to report falsely–and relatively low moral cost from undetected embezzlement. For the "disloyal" type, it is the other way around. The manager does not know which type she is facing. She would like to cooperate if she faces a loyal employee since she could expect that cooperation would be profitable. Analogously, she would refrain from cooperation if she faces a disloyal type. The crucial scenario is when the manager embezzles: Assuming that the reporting decision would perfectly reveal the type of the employee, the manager would not cooperate if the employee reports, but would cooperate if the employee does not report. In this case, the loyal employee would not report as the profit from cooperation outweighs her costs from undetected embezzlement. For the disloyal employee the moral costs from undetected embezzlement are higher than the profit from cooperation such that she would report. If the manager does not embezzle, the reporting decision would not reveal the type and the decision to cooperate depends on the manager's expectation about the share of loyal employees.

**Predictions on truthful whistleblowing ($r^t$)** From the equilibrium described above, it follows that disloyal (loyal) employee will (not) report in treatment $B$. This equilibrium also holds in treatment $I$ as well. However, it becomes more difficult to sustain, as there is a reward for reporting and therefore the opportunity costs from not reporting for a disloyal employee become larger. Consequently, I expect to be the empirical willingness to report to be at least as high in treatment $I$. In the treatments $A$ and $AI$, reporting does not convey information about the type of the employee. Therefore, the reporting decision does not influence the manager's cooperation decision and both types report embezzlement (see Prediction $r^t$).

**Prediction** $(r^t)$**.** $\quad r^t_B \leq r^t_I, \quad r^t_B < r^t_A, \quad r^t_I < r^t_{AI}, \quad r^t_A = r^t_{AI}.$

**Predictions on false whistleblowing ($r^f$)** The model considered the case where the reward for reporting does not outweigh the moral costs from false reporting. Consequently, both types would not report falsely in any treatment. This assumption does not necessarily hold in reality. Therefore, it is worthwhile to discuss possible deviations. If this assumption is relaxed and, for example, the moral costs are smaller than the reward for both types, there would still be no false reports in the treatments $B$ and $A$, as

there is no reward for reporting, and both types would still be better off by avoiding the moral cost from a false report. However, in the treatments where the manager have to compensate whistleblowers, $I$ and $AI$, the scenario would be different. Both types would have an incentive to report falsely if they expect the manager not to cooperate. While in treatment $I$ a report would make cooperation to happen less likely, in treatment $AI$ it cannot affect the cooperation decision of the manager. Therefore, false reports should occur more frequently in treatment $AI$ (see Prediction $r^f$).

**Prediction** $(r^f)$**.** $\quad r_B^f \leq r_I^f, \quad r_B^f = r_A^f, \quad r_I^f \leq r_{AI}^f, \quad r_A^f \leq r_{AI}^f.$

**Predictions on embezzlement (*e*)** The decision to embezzle in treatment $B$ depends on the manager's belief about the share of loyal employees. It must hold that separating the types by embezzlement is more profitable than basing the cooperation decision on the belief about the share of loyal employees. The higher the share of loyal employees the more profitable becomes embezzlement: the probability to earn a profit from both cooperation and embezzlement increases, while the probability to be fined for embezzlement decreases. In treatment $I$, the threshold for embezzlement to be profitable is higher since the manager has to compensate the "disloyal" type. Therefore, screening becomes more expensive and the frequency of embezzlement should be lower. In treatments $A$ and $AI$, any employee will report embezzlement, but the manager cannot learn about the type. Therefore, embezzlement should not occur (see Prediction $e$).

**Prediction** (**e**)**.** $\quad e_B > e_I, \quad e_B > e_A, \quad e_I > e_{AI}, \quad e_A = e_{AI}.$

**Predictions on the frequency of cooperation (*c*)** Concerning the willingness to cooperate, the behavioral predictions are ambiguous. First, the comparison between treatments with and without anonymity depends on the share of loyal employees. As the manager cannot screen the employees in the treatments with anonymity, she would always cooperate if she expects the share to be high enough to make a profit on expectation. Vice versa, she would not cooperate in the treatments $A$ and $AI$, if she expects the share as too low. In the treatments without anonymity, the manager can identify the type of employee and would not cooperate with disloyal employees. Therefore, if a manager believes that the share of loyal employees is sufficiently high to cooperate without a signal about the trustworthiness, the frequency of cooperation would be lower in the non-anonymous treatments. In contrast, if a manager believes the share of loyal employees is sufficiently low (i.e., she would not cooperate under anonymity), the frequency of cooperation would increase in treatments without anonymity, as she could identify employees she wants to cooperate with. Between the anonymity treatments $A$ and $AI$, the model predicts no difference in the willingness to cooperate. For the comparison between the

treatments $B$ and $I$, there is less embezzlement in treatment $I$, such that there are less cases where the manager can identify the type of her employee. That means cooperation would be more frequent in treatment $I$ if the manager expects a sufficiently large share of loyal employees, and vice versa. Furthermore, there are opposite effects in treatment $I$. On the one hand, it is more difficult for the loyal type to remain silent (compare to the prediction on truthful reporting). On the other hand, it is more expensive for the manager not to cooperate with employees who reported since she would have to compensate them. The model does not allow making a claim which effect might outweigh the other. Taken together, it is a priori difficult to compare the frequency of cooperation across treatments since it largely depends on the belief of then manager about the share of loyal employees (see Prediction $c$). While this belief is unknown to the experimenter, it is likely shaped by the experience from previous periods, which has to be considered in the analysis.

**Prediction (c).** $c_B \lesseqgtr c_I, \quad c_B \lesseqgtr c_A, \quad c_I \lesseqgtr c_{AI}, \quad c_A = c_{AI}.$

**Predictions on the frequency of high returns ($t$)** The model allows making predictions about the frequency of employees returning an amount larger than the investment. These predictions depend on the cooperation decision of the manager. Recall that only the loyalty type of the employee influences the return decision. As the manager cannot differentiate between the employee types in the treatments with anonymity, the model predicts no difference between the treatments $A$ and $AI$. In the treatments without anonymity, the manager can identify the types in case she embezzles such that she does not cooperate with a disloyal type. The frequency of high returns should therefore be higher in the treatments without anonymity. As there is more embezzlement in treatment $B$, and therefore a higher degree of separation of the types, the frequency of high returns should be higher than in treatment $I$ (see Prediction $t$).

**Prediction (t).** $t_B > t_I, \quad t_B > t_A, \quad t_I > t_{AI}, \quad t_A = t_{AI}.$

# 5 Results

In this section, I analyze the treatment differences to identify the effects of whistleblower protection on reporting, embezzlement, as well as on the sending and the return behavior in the trust game. In a first step, I present how the subjects decided on average across the four treatments.

To test for statistical significance, I follow Moffatt (2015) and use non-parametric tests with subject-role-level averages as observational units. For between-subject differences, I apply a Mann-Whitney $U$ test ($B$ vs. $A$, $I$ vs. $AI$), while I account for within-subject differences with a Wilcoxon signed-rank test ($B$ vs. $I$, $A$ vs. $AI$). To include the influence of

past decisions, I use panel regressions in addition, which account for the number of independent observations and allow controlling for period fixed effects. As for the behavioral predictions, I will first consider the reporting behavior of the employees.

| | | |
|---|---|---|
| $r_B^t$: 0.72 | $<^{***}$ | $r_I^t$: 0.87 |
| $\wedge$ | | $\wedge$ |
| $r_A^t$: 0.84 | $<^{***}$ | $r_{AI}^t$: 0.89 |

(a) Truthful reporting

| | | |
|---|---|---|
| $r_B^f$: 0.13 | $<^{***}$ | $r_I^f$: 0.31 |
| $\wedge$ | | $\wedge^{***}$ |
| $r_A^f$: 0.22 | $<^{***}$ | $r_{AI}^f$: 0.54 |

(b) False reporting

| | | |
|---|---|---|
| $e_B$: 0.41 | $<^{***}$ | $e_I$: 0.24 |
| $\vee$ | | $\vee^{**}$ |
| $e_A$: 0.32 | $>^{***}$ | $e_{AI}$: 0.08 |

(c) Embezzlement

| | | |
|---|---|---|
| $r_B$: 0.40 | $<$ | $r_I$: 0.46 |
| $\vee$ | | $\wedge^{**}$ |
| $r_A$: 0.39 | $<^{***}$ | $r_{AI}$: 0.58 |

(d) Total reports

Notes: The values in tables (a)–(c) report the average decisions of the subjects across the treatments. Table (d) reports the frequency of reports, i.e. the combinations of embezzlement and true reporting, and no embezzlement and false reporting. Significance levels : $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 2: Reporting and Embezzlement across Treatments

**Truthful whistleblowing ($r^t$)** Since I use the strategy method for the employee's decision to report, I track separately how the willingness to report truthfully and falsely differs across the treatments. Table 2a displays the fractions of employees that choose to report truthfully for each treatment. To evaluate the effect of the instruments on the willingness to report, I compare the outcome of the treatments $I$ and $A$ to treatment $B$. For treatment $I$, I find a significant increase from 72% to 87% ($p < 0.01$). In treatment $A$, the fraction rises to 84%, although this increase is not statistically significant ($p < 0.20$). I find the highest fraction of truthful reports in treatment $AI$ with 89%. This is a significant increase compared to treatment $A$ ($p < 0.01$, only two out of 38 subjects decrease the reporting frequency), but not compared to treatment $I$ ($p < 0.49$). These results provide evidence that both instruments, but especially immunity, affect the employee's willingness to report truthfully as intended and therefore support mostly the prediction $r^t$.

**False whistleblowing ($r^f$)** Analogously, Table 2b displays the willingness to conduct a false report. Surprisingly, I find 13% of the employees would blow the whistle although there was no misbehavior in treatment $B$.[8] In line with prediction $r^f$, I find that the

[8]Although the employees cannot target the managers whom behavior they disliked, some employees might still want to punish managers in general. Regression results indeed suggest that employees whose manager cooperated in the previous round are less likely to report falsely (Table D.1, columns 5 and 6), while it does not affect the willingness to report truthfully (columns 2 and 3).

share of false reporting does not increase significantly in treatment $A$ (22%, $p < 0.54$). However, introducing immunity in treatment $I$ leads to a significant jump in false reports to 31% ($p < 0.01$). These results indicate that a significant share of subjects expects the manager to take from their endowment instead of cooperating, and that the expected loss outweighs the moral cost from false reporting–other than assumed in model. In line with this, the share of employees willing to file a false claims peaks in treatment $AI$ with 54% ($p < 0.01$ compared to $A$ and to $I$). These findings support the prediction $r^f$ and suggest that subjects react also to the adverse incentives of whistleblower protection.

The results for truthful and false whistleblowing already provide evidence for costs as well as for benefits of whistleblower laws. Protection increases truthful reports, but provokes adverse effects in the form of false reports at the same time.

**Embezzlement ($e$)** The previous results indicate that under whistleblower protection embezzlement would be reported more often. Further, it is of interest whether managers anticipate these changes in reporting such that embezzlement is deterred (see Table 2c). Compared to treatment $B$ (41%), I find a significant drop in embezzlement when reporting is incentivized in treatment $I$ to 24% ($p < 0.01$). In treatment $A$, where the employee can report anonymously, also a lower share of 32% decides to embezzle money. However, this decline is not statistically significant ($p < 0.58$). In treatment $AI$, only 8% of the managers choose to embezzle money. This is a significant decline both from treatment $A$ ($p < 0.01$) and from treatment $I$ ($p < 0.03$). These results mostly support prediction $e$. Regression results indicate that the subjects anticipate the reporting behavior based on their experiences. The dummy for reported embezzlement in the previous period is negative and highly significant, while the controls for the treatments do not explain the embezzlement frequency (see Table D.2).

Interestingly, the number of overall reports is the highest in treatment $AI$ (58%), i.e. when the frequency of embezzlement is the lowest (see Figure 2d). While the number of reports are very similar in the treatments $B$, $A$, and $I$ ($B$ vs. $I$: $p < 0.46$, $B$ vs. $A$: $p < 0.80$), in treatment $AI$, it increases significantly compared to both treatments $A$ (39%, $p < 0.01$) and $I$ (46%, $p < 0.03$). While the managers anticipate the high tendency to report, and therefore embezzlement is mostly deterred, the remaining cases (8%) are entirely reported in treatment $AI$. That means there are no unreported cases of embezzlement and there is the highest frequency of false reports.

**Cooperation ($c$)** Having analyzed the reporting and the embezzlement behavior, I will evaluate the willingness to cooperate and the level of cooperation over the different treatments. The prediction $c$ pointed out that treatment differences are difficult to anticipate, since the cooperation decision depends on the expectation of the manager with respect to

the "loyalty type" she is facing. This expectation should be influenced by her latest experience with respect to the employees' reporting and return behavior. Therefore, it will be crucial to use regression analysis to investigate the mechanisms between reporting and cooperative behavior. First, I analyze the cooperative behavior on the aggregated level. Table 3a shows the share of managers who chose a positive $c$ across the treatments.

| | | |
|---|---|---|
| $c_B$: 0.30 | $>$ | $c_I$: 0.26 |
| $\wedge$ | | $\vee$ |
| $c_A$: 0.34 | $>^{**}$ | $c_{AI}$: 0.17 |

(a) Cooperation frequency

| | | |
|---|---|---|
| $c_B^l$: 0.43 | $>$ | $c_I^l$: 0.42 |
| $\vee$ | | $\|\|$ |
| $c_A^l$: 0.40 | $<$ | $c_{AI}^l$: 0.42 |

(b) Cooperation level

Notes: The values in the tables (a) and (b) report the average rate of (a) the frequency cooperation and (b) the level of cooperation across the treatments, conditional on $c$ being positive. Observations of cooperation level: $B$: 73, $I$: 62, $A$: 51, $AI$: 26. Significance levels : $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 3: Frequency and Level of Cooperation across Treatments

Considering treatment $B$, I find a fraction of 30% of the managers choosing to cooperate. Both in treatment $A$ (34%, $p < 0.86$), and in treatment $I$ (26%, $p < 0.26$), the willingness to cooperate is not significantly different. However, in treatment $AI$, only 17% of the managers decide to cooperate. While this is not significantly different from treatment $I$ ($p < 0.35$), it is a significant drop compared to treatment $A$ ($p < 0.02$). The comparison of the average cooperation over the treatments does not support the prediction $e$ as the model does not predict differences between the anonymity treatments.

To investigate in detail what could explain the treatment differences, it is important to recall what could drive the cooperation decision of a manager. In treatments without anonymity, the manager may receive a direct signal from the employee on her trustworthiness through the reporting decision. Precisely, if a manager embezzles and the employee does not report, it may lead the manager to cooperate more likely. In treatments with anonymity, the manager cannot infer any information about the trustworthiness from the reporting decision of the employee as it is not observed. However, since the decisions are made repeatedly, a manager may infer from reporting decisions from previous periods how likely it is to face a trustworthy employee. Moreover, in all treatments, the experience with respect to the return behavior of the employees may shape the expectation about the average trustworthiness.

To analyze the influence of the managers' experience, I use regressions that control for the embezzlement decision in the respective period, the most recent reporting decision and whether there was profitable cooperation in the last period. As the most recent reporting decision is different for subjects in the anonymous and in the non-anonymous treatments, I split the sample and conduct the regressions separately.

The regression results for the frequency of cooperation in non-anonymous treatments

| | Cooperation Frequency | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Treatment I | -0.0458 | 0.104 | 0.145 |
| | (0.0538) | (0.0872) | (0.0902) |
| LowReturn(lag) | | -0.245*** | -0.251*** |
| | | (0.0797) | (0.0901) |
| Embezzlement | | -0.228** | -0.238* |
| | | (0.112) | (0.125) |
| FalseReport | | 0.102 | 0.0377 |
| | | (0.156) | (0.168) |
| UnreportedEmbezzlement | | 0.356* | 0.348* |
| | | (0.212) | (0.201) |
| Constant | 0.304*** | 0.572*** | 0.366 |
| | (0.0452) | (0.112) | (0.233) |
| Period FE | No | No | Yes |
| $N$ | 480 | 131 | 131 |
| $N_{groups}$ | 30 | 28 | 28 |
| $R^2$ | 0.00260 | 0.128 | 0.156 |

Notes: The table reports results from a random-effects GLS regression where $N_{groups}$ is the number of individuals. Dependent variable: (1)-(3): cooperation decision of managers (0 or 1). LowReturn, Embezzlement, FalseReport, UnreportedEmbezzlement are all binary variables. (lag) indicates a lagged variable. ReportedEmbezzlement is omitted because of collinearity. Standard errors in parentheses are clustered on the individual level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 4: Regression Analysis: Cooperation without Anonymity

are reported in Table 4 (columns 1-3). First, I turn the attention to the experienced return behavior (column 2). The coefficient for a loss from cooperation in the previous period–that is, the employee returned less than the manager sent–is negative and highly significant. Moreover, the results indicate that also the reporting behavior plays an important role. The coefficient for having experienced unreported embezzlement is positive and significant, though only weakly. This suggests that the subjects adjust their expectations about the profitability of cooperation based on their cooperation experience and on whether embezzlement was reported. Controlling for period fixed effects, the coefficients in Table 4, column (3) suggest that the results do not change qualitatively.

Analogously, the regression results for the frequency of cooperation in the anonymous treatments are reported in Table 5. It is controlled for the same variables as before, with the exception that the most recent observed reporting is from the previous period. Therefore, the variables on reporting are lagged by one period. The results for the anonymous treatments provide a similar picture as before (see Table 5, column 2). The coefficient for having experienced a loss from cooperation in the previous period is negative and highly significant as well. Further, the results for anonymous reporting indicate that the most recent reporting behavior plays a role, even if it cannot be linked to the present employee.

| | Cooperation Frequency | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Treatment AI | -0.164*** | -0.113 | -0.0540 |
| | (0.0575) | (0.103) | (0.129) |
| LowReturn(lag) | | -0.538*** | -0.506*** |
| | | (0.126) | (0.171) |
| Embezzlement | | -0.311*** | -0.339*** |
| | | (0.115) | (0.0926) |
| FalseReport(lag) | | 0.158 | 0.154 |
| | | (0.140) | (0.145) |
| UnreportedEmbezzlement(lag) | | 0.344*** | 0.351* |
| | | (0.128) | (0.180) |
| ReportedEmbezzlement(lag) | | -0.0521 | -0.102 |
| | | (0.106) | (0.149) |
| Constant | 0.336*** | 0.810*** | 0.980*** |
| | (0.0671) | (0.0968) | (0.194) |
| Period FE | No | No | Yes |
| $N$ | 304 | 75 | 75 |
| $N_{groups}$ | 19 | 16 | 16 |
| $R^2$ | 0.0358 | 0.355 | 0.379 |

Notes: The table reports results from a random-effects GLS regression where $N_{groups}$ is the number of individuals. Dependent variable: (1)-(3): cooperation decision of managers (0 or 1). LowReturn, Embezzlement, FalseReport, UnreportedEmbezzlement, ReportedEmbezzlement are all binary variables. (lag) indicates a lagged variable. Standard errors in parentheses are clustered on the individual level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 5: Regression Analysis: Cooperation with Anonymity

The coefficient for having experienced unreported embezzlement is positive and highly significant. If I include the period fixed effect, the coefficient remains significant, but only weakly.

The regression results indicate that the decision to cooperate depends on the most recent experience with respect to the returning and the reporting behavior of the employees. This may explain the differences in cooperation between the anonymous treatments on the aggregate level. Note that in treatment *AI*, embezzlement is completely deterred. Therefore, cases, where an employee could refrain from a truthful report to signal trustworthiness, do not occur. While the theory predicts no difference for the frequency of embezzlement between the anonymous treatments, embezzlement is more prevalent in treatment *A* nevertheless (see Figure 2c). Moreover, in 19% of the cases, the employee does not report the embezzlement. That means, in treatment *A* the managers experience unreported embezzlement, which they may perceive as signal of trustworthiness, but not in treatment *AI*. In consequence, the drop in cooperation between the anonymous treatments may result from the different levels of deterrence.

In addition to the treatment differences, it interesting whether the cooperative behavior

of managers can be distinguished based on their prior behavior. Before the cooperation decision takes place, each manager already made decision whether to embezzle. The results from Tables 4 and 5 show that managers, who chose to embezzle, cooperate less often. Moreover, false reports seem to affect the cooperation decision. This suggests that mainly managers, who chose to embezzle, drive the lower cooperation rates.

Apart from the general decision to cooperate or not, the level of cooperation is of interest. Since the design allows varying the level of trust, managers may rather adjust the amount that is trusted to the employee instead of refraining from cooperation in general. To account for this, I consider only those managers who chose to cooperate and report the trusted share of their endowment (Table 3b). The results suggest that there are no treatment differences for the size of cooperation. Independent of the protection scheme, the trusted share lies within a range of 40 to 44 % of the endowment, which roughly corresponds to the average investment level across experimental studies (see e.g., Fehr and Schmidt, 2006).

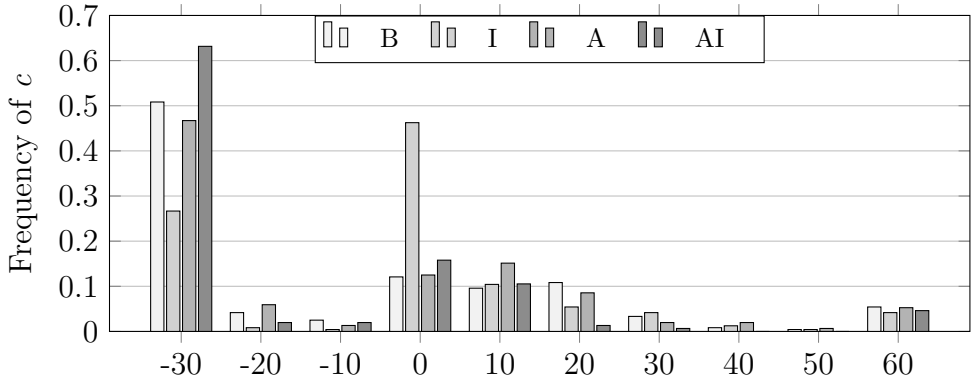


Figure 3: Distribution of amounts taken/sent

To obtain a more detailed picture of the cooperation level, Figure 3 illustrates the distribution of the realizations of $c$ across the treatments. Given the decision to send a positive or negative $c$, the results suggests that the choice of the level of $c$ is very similar across the treatments. Those managers, who cooperate, send predominantly a small $c$ (10, or 20) and sometimes the highest possible amount (60). Those managers, who do not cooperate, just send nothing in one out of six cases, while in almost all of the remaining cases, they choose the lowest possible $c$ (-30). However, there is a striking difference for the treatment *I*. As there is immunity if the employee reports, the manager cannot choose $c = -30$ as the minimum in this case, but only zero. Correspondingly, a notably higher rate (roughly two thirds) of those managers, who do not cooperate, (has to) choose zero. As the managers choose a negative $c$ when they do not expect the employee to be trustworthy, and false reports seem not to decrease the cooperation likelihood, immunity increases the earnings for employees.

**Return behavior** To conclude, I will turn the focus to the return behavior of the employees. In a similar vein as for the cooperation decision, I consider a binary decision–whether to return more or less than the manager sent–and the level of the amount the employees return. Table 6a shows the share of employees, who returned more that the manager sent (high return), conditional on the manager had chosen a positive $c$.

| | | |
|---|---|---|
| $c_B$: 0.49 | $>^{**}$ | $c_I$: 0.44 |
| $\wedge$ | | $\vee$ |
| $c_A$: 0.51 | $>$ | $c_{AI}$: 0.38 |

(a) High return

| | | |
|---|---|---|
| $t^l_B$: 0.76 | $>$ | $t^l_I$: 0.65 |
| $\vee$ | | $\vee$ |
| $t^l_A$: 0.65 | $>$ | $t^l_{AI}$: 0.57 |

(b) Level of return

Notes: The values in the tables (a) and (b) report the average rate of (a) employees returning more than what was sent and (b) the level of return across the treatments. Observations of return decisions: $B$: 73, $I$: 62, $A$: 51, $AI$: 26. Significance levels : $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 6: Return Behavior

Since the managers can only infer from the reporting behavior of the employee how likely it is to receive a high return, there should be no differences between the treatments $A$ and $AI$. Moreover, the non-anonymous treatments should have a higher share of employees, who send a high return. The share should be higher for treatment $B$ than for treatment $I$ (compare to prediction $t$). The comparison of the treatments $B$ and $I$ supports this prediction as the decrease from 49% to 44% is statistically significant ($p < 0.01$, just one of 32 subjects increased the frequency of returning more than what was sent). However, the results cannot support a higher frequency of high returns in the non-anonymous treatments ($B$ vs. $A$: $p < 0.69$, $I$ vs. $AI$: $p < 0.33$). Comparing the two anonymous treatments, the frequency appears to be higher in treatment $A$ (51%) than in treatment $AI$ (38%), but I cannot provide statistical significance for this difference as it results from a low number of subjects (N=13).

Furthermore, controlling for past behavior allows to identify whether the reporting decision of an employee corresponds to a certain return behavior. The results from a regression analysis illustrates that employees, who made a false claim, are less likely to return more than what the manager sent (see Table D.3). As false reports do not lower the frequency of cooperation, managers seem not to anticipate this return behavior. Moreover, unreported embezzlement is not associated with a higher frequency of high returns, although managers reward it more often with cooperation. Interestingly, these results suggest that managers may not react optimally to the observed reporting behavior.

Similarly to the cooperation decision of the managers, the employees may adjust rather the level of the amount they return. Figure 6b show that the employee return, on average, between 57 and 76% of what was sent to them. The tests do not report any significant difference between the treatments.

# 6 Discussion

With this paper, I shed light on the potential hidden costs of whistleblower protection. In a workplace setup, a manager could embezzle money at the expense of a third party, while her employee observes this and could report her manager before they play a trust game. I varied the framework in two dimensions to capture two prominent features of whistleblower protection laws: First, not revealing the reporting decision to the manager before the trust game allows the employee to report anonymously. Second, prohibiting the manager to take from the employee in the trust game conditionally on a report, enables the employee to insure herself against retaliation from the manager by blowing the whistle.

In line with the literature (see, e.g. Bartuli et al., 2016; Schmolke and Utikal, 2018; Butler et al., 2020), my results confirm that both instruments have the intended effects: I observe an increased willingness to report truthfully by the employees, which is anticipated by the managers who reduce embezzlement. This suggests that whistleblower laws offer a rich potential for fighting the damage of corporate fraud through both increased deterrence and detection. On the other hand, the findings demonstrate that whistleblower protection also provokes adverse effects. Since the incentives for reporting are not provided conditionally on a successful investigation, these do not only increase truthful reporting, but also trigger false whistleblowing by the employees.

A novel finding of this paper relates to the costs associated with the deterrence of misbehavior. Beyond the negative direct impact from reports in the form of costs from the investigation for authorities or the organization, I point out the importance of whistleblowing for the cooperative climate in an organization. Increasing the willingness to report leads to a high level of deterrence, therefore limiting the possibility to signal trustworthiness to the manager. This may create an "atmosphere of distrust", which hampers productive cooperation. In consequence, social welfare could be negatively affected by whistleblower protection although it deters misbehavior.

I chose a simple design for the whistleblowing game, where the employee has precise knowledge about the state of illegal behavior of her superior. Further, the employee does not face the risk of leaks under anonymity. In addition, an investigation and immunity are guaranteed consequences of a report. This captures the intended increase in legal certainty for the whistleblower. In reality, when not all of these assumptions are met, uncertainty may also influence the behavior under the different protection regimes and cause a lower responsiveness of employees (see e.g., Chassang and Miquel, 2019; Mechtenberg et al., 2020). Therefore, my results serve as a benchmark for future studies that relax these assumptions.

# References


Abbink, K. and H. Hennig-Schmidt (2006). Neutral versus loaded instructions in a bribery experiment. *Experimental Economics 9*(2), 103–121.

Abbink, K., B. Irlenbusch, and E. Renner (2000). The moonlighting game: An experimental study on reciprocity and retribution. *Journal of Economic Behavior & Organization 42*(2), 265–277.

Abeler, J., D. Nosenzo, and C. Raymond (2019). Preferences for truth-telling. *Econometrica 87*(4), 1115–1153.

Alekseev, A., G. Charness, and U. Gneezy (2017). Experimental methods: When and why contextual instructions are important. *Journal of Economic Behavior & Organization 134*, 48–59.

Alford, C. (2001). *Whistleblowers: Broken Lives and Organizational Power*. Cornell University Press.

Barr, A. and D. Serra (2009). The effects of externalities and framing on bribery in a petty corruption experiment. *Experimental Economics 12*(4), 488–503.

Bartuli, J., B. Mir Djawadi, and R. Fahr (2016). Business ethics in organizations: An experimental examination of whistleblowing and personality. *IZA Discussion Paper 10190*.

Berg, J., J. Dickhaut, and K. McCabe (1995). Trust, reciprocity, and social history. *Games and Economic Behavior 10*(1), 122–142.

Bloom, N., R. Sadun, and J. Van Reenen (2012). The organization of firms across countries. *The Quarterly Journal of Economics 127*(4), 1663–1705.

Bock, O., I. Baetge, and A. Nicklisch (2014). hroot: Hamburg registration and organization online tool. *European Economic Review 71*, 117–120.

Bracht, J. and N. Feltovich (2009). Whatever you say, your reputation precedes you: Observation and cheap talk in the trust game. *Journal of Public Economics 93*(9-10), 1036–1044.

Brandts, J. and G. Charness (2011). The strategy versus the direct-response method: A first survey of experimental comparisons. *Experimental Economics 14*(3), 375–398.

Buccirossi, P., G. Immordino, and G. Spagnolo (2021). Whistleblower rewards, false reports, and corporate fraud. *European Journal of Law and Economics 51*(3), 411–431.

Butler, J. V., D. Serra, and G. Spagnolo (2020). Motivating whistleblowers. *Management Science 66*(2), 605–621.

Callahan, E. S. and T. M. Dworkin (1992). Do good and get rich: Financial incentives for whistleblowing and the False Claims Act. *Villanova Law Review 37*, 273.

Cassematis, P. G. and R. Wortley (2013). Prediction of whistleblowing or non-reporting observation: The role of personal and situational factors. *Journal of Business Ethics 117*(3), 615–634.

Charness, G., U. Gneezy, and M. A. Kuhn (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization 81*(1), 1–8.

Chassang, S. and G. P. I. Miquel (2019). Crime, intimidation, and whistleblowing: A theory of inference from unverifiable reports. *The Review of Economic Studies 86*(6), 2530–2553.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association 9*(3), 522–550.

Dworkin, T. and J. Near (1997). A better statutory approach to whistle-blowing. *Business Ethics Quarterly 7*(1), 1–16.

Dyck, A., A. Morse, and L. Zingales (2010). Who blows the whistle on corporate fraud? *The Journal of Finance 65*(6), 2213–2253.

Fehr, E. and K. M. Schmidt (2006). The economics of fairness, reciprocity and altruism–experimental evidence and new theories. *Handbook of the Economics of Giving, Altruism and Reciprocity 1*, 615–691.

Fehrler, S. and W. Przepiorka (2016). Choosing a partner for social exchange: Charitable giving as a signal of trustworthiness. *Journal of Economic Behavior & Organization 129*, 157–171.

Felli, L. and R. Hortala-Vallve (2016). Collusion, blackmail and whistle-blowing. *Quarterly Journal of Political Science 11*(3), 279–312.

Fiorin, S. (2023). Reporting peers' wrongdoing: Experimental evidence on the effect of financial incentives on morally controversial behavior. *Journal of the European Economic Association*, forthcoming.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics 10*(2), 171–178.

Friebel, G. and S. Guriev (2012). Whistle-blowing and incentives in firms. *Journal of Economics & Management Strategy 21*(4), 1007–1027.

Gambetta, D. and Á. Székely (2014). Signs and (counter) signals of trustworthiness. *Journal of Economic Behavior & Organization 106*, 281–297.

Givati, Y. (2016). A theory of whistleblower rewards. *The Journal of Legal Studies 45*(1), 43–72.

Heyes, A. and S. Kapur (2009). An economic model of whistle-blower policy. *Journal of Law, Economics, and Organization 25*(1), 157–182.

Heyes, A. and J. A. List (2016). Supply and demand for discrimination: Strategic revelation of own characteristics in a trust game. *American Economic Review 106*(5), 319–23.

Howse, R. and R. Daniels (1995). Rewarding whistleblowers: The costs and benefits of an incentive-based compliance strategy. In R. Daniels and R. Morck (Eds.), *Corporate Decisionmaking in Canada*. Calgery: University of Calgary Press.

Kohn, S. M., M. D. Kohn, and D. K. Colapinto (2004). *Whistleblower law: A Guide to Legal Protections for Corporate Employees*. Greenwood Publishing Group.

Mechtenberg, L., G. Muehlheusser, and A. Roider (2020). Whistleblower protection: Theory and experimental evidence. *European Economic Review 126*, 103447.

Mir Djawadi, B. and P. Nieken (2019). Labor market chances of whistleblowers-potential drivers of discrimination. *Available at SSRN 3481126*.

Moffatt, P. G. (2015). *Experimetrics: Econometrics for Experimental Economics*. Macmillan International Higher Education.

Near, J. P. and M. P. Miceli (1985). Organizational dissidence: The case of whistle-blowing. *Journal of Business Ethics 4*(1), 1–16.

Near, J. P. and M. P. Miceli (1986). Retaliation against whistle blowers: Predictors and effects. *Journal of Applied Psychology 71*(1), 137.

OECD (2016). *Committing to Effective Whistleblower Protection*. Paris: OECD Publishing.

Reuben, E. and M. Stephenson (2013). Nobody likes a rat: On the willingness to report lies and the consequences thereof. *Journal of Economic Behavior & Organization 93*, 384–391.

Schmolke, K. U. and V. Utikal (2018). Whistleblowing: Incentives and situational determinants. *Available at SSRN 3198104*.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In *Beiträge zur experimentellen Wirtschaftsforschung*, pp. 136–168. Tübingen: JCB Mohr (Paul Siebeck).

Thüsing, G. and G. Forst (2016). *Whistleblowing–A Comparative Study*, Volume 16. Springer.

Vinten, G. (1994). Whistleblowing-fact and fiction. an introductory discussion. *Whistleblowing: Subversion or corporate citizenship*, 1–20.

Walters, K. D. (1975). Your employees right to blow whistle. *Harvard Business Review 53*(4), 26.

# A Theory

This Appendix is structured as follows: In Section A.1, the model is presented, and in Section A.2, I derive the equilibrium outcome for each treatment. The behavioral predictions of Section 4 are derived from the Propositions 1 - 4.

## A.1 Model

**The Game Played** I consider a model played by two players, a manager and an employee. The manager is matched with an employee of loyalty type $l \in \{\underline{l}, \bar{l}\}$, where $\underline{l} < \bar{l}$ ($l = \underline{l}$: $\underline{l}$-employee, $l = \bar{l}$: $\bar{l}$-employee), which is private information of the employee. The manager's belief that she faces a loyal employee is $\Pr(l = \bar{l}) = q(r)$ and she has a common prior $\Pr(l = \underline{l}) = 1 - \alpha$ and $\Pr(l = \bar{l}) = \alpha$. In stage 1, the manager decides whether or not to embezzle $e \in \{0, 1\}$, which is observed by the employee. Then, the employee decides whether or not to report $r \in \{0, 1\}$, which is observed by the manager in treatments without anonymity (treatments $B$ and $I$), but not in treatments with anonymity (treatments $A$ and $AI$). In stage 3, the manager decides whether or not to cooperate with the employee $c \in [0, 1]$. If the manager cooperates, the employee decides to return $t \in [t_0, t_1]$, where $t_0 < t_1 < 1$, back to the manager. Otherwise, the game ends after stage 3.

**Treatments**

- In treatment $B$, the manager observes the reporting decision before deciding on cooperation and there is no minimum payment to the employee if the manager does not cooperate.
- In treatment $I$, the manager observes the reporting decision before deciding on cooperation and the employee must get at least $x$ where $\underline{l} > x > 0$ when she reports and the manager does not cooperate.
- In treatment $A$, the manager does not observe the reporting decision before deciding on cooperation and there no minimum payment to the employee if the manager does not cooperate.
- In treatment $AI$, the manager does not observe the reporting decision before deciding on cooperation the employee must get at least $x$ when he reports and the manager does not cooperate.

**Payoffs** All payoffs (monetary and non-monetary) are summarized in Table A.1. First, the payoff of the manager depends on whether or not she embezzles, whether or not the employee reports, whether or not she cooperates and which $t$ is returned in case of cooperation. The manager's potential gain from embezzlement is the monetary payoff $E$. If the employee reports and there was embezzlement, the manager pays a net fine $F$. If

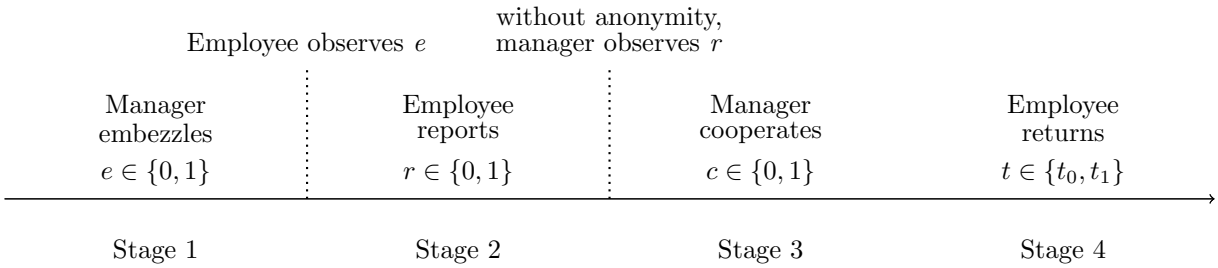


Figure A.1: Model

she chooses to cooperate, she pays an investment $I$ to create a pie of size 1, which the employee then distributes between her and the manager by sending either $t_0$ or $t_1$. In treatments with(out) immunity, the manager does (not) have to pay $x$ to the employee if she has reported.

Second, the payoff of the employee depends on whether or not the manager embezzles, whether or not she reports, whether or not the manager cooperates and whether she returns $t_1$ or $t_0$ in case the manager cooperates. If the manager embezzles and the employee does not report, she faces a moral cost from undetected embezzlement $\delta = 1 - l$. Note that $\underline{l} < \bar{l} \implies \underline{\delta} > \bar{\delta}$. If the manager does not embezzle, but she does report, she faces a moral cost $l$. If the manager cooperates and the employee returns $t_1$ that leaves her with a payoff of $1 - t_1$. If she returns $t_0$ instead, she faces a moral cost $l$, since she did not reciprocate the cooperation decision properly.

## A.2 Equilibrium Analysis

### A.2.1 Preliminaries

When deriving my predictions, I focus of Perfect Bayesian Equilibria (PBE) in pure strategies (i.e., all players choose best responses given their beliefs and given the strategies of the other players, where beliefs are formed in accordance with Bayes' Rule whenever possible). More precisely, I focus on separating equilibria where a $\underline{l}(\bar{l})$-employee does (not) report when the manager chooses to embezzle. This captures the trade-off between the detection of embezzlement and signalling trustworthiness to support productive cooperation.

**Assumption 1.** *$\underline{l} < t_1 - t_0 < \bar{l}$, i.e. the disutility of an $\underline{l}(\bar{l})$-employee from returning the low amount $t_0$ is smaller (larger) than the monetary gain from returning low amount $t_0$ instead of high amount $t_1$.*

**Assumption 2.** *$t_0 < I < t_1$, i.e. cooperation pays off for the manager only if the employee returns the high amount $t_1$.*

**Treatments without immunity**

| embezzlement | report | cooperation | return | employee | manager |
|---|---|---|---|---|---|
| 0 | 0 | 0 | n.a. | 0 | 0 |
| 0 | 0 | 1 | $t_0$ | $1-t_0-l$ | $t_0-I$ |
| 0 | 0 | 1 | $t_1$ | $1-t_1$ | $t_1-I$ |
| 0 | 1 | 0 | n.a. | $-l$ | 0 |
| 0 | 1 | 1 | $t_0$ | $1-t_0-2l$ | $t_0-I$ |
| 0 | 1 | 1 | $t_1$ | $1-t_1-l$ | $t_1-I$ |
| 1 | 0 | 0 | n.a. | $-\delta$ | $E$ |
| 1 | 0 | 1 | $t_0$ | $-\delta+1-t_0-l$ | $E+t_0-I$ |
| 1 | 0 | 1 | $t_1$ | $-\delta+1-t_1$ | $E+t_1-I$ |
| 1 | 1 | 0 | n.a. | 0 | $-F$ |
| 1 | 1 | 1 | $t_0$ | $1-t_0-l$ | $-F+t_0-I$ |
| 1 | 1 | 1 | $t_1$ | $1-t_1$ | $-F+t_1-I$ |

**Differences in treatments with immunity**

| embezzlement | report | cooperation | return | employee | manager |
|---|---|---|---|---|---|
| 0 | 1 | 0 | n.a. | $x-l$ | $-x$ |
| 1 | 1 | 0 | n.a. | $x$ | $-F\ -x$ |

Table A.1: Theory Payoffs

**Assumption 3.** $t_1 < \bar{l}$, *i.e. the disutility from embezzlement must be smaller than the profit from cooperation for an* $\bar{l}$*-employee.*

**Assumption 4.** $t_0 < I - x$, *i.e. receiving the low amount* $t_0$ *does not pay off for the manager compared to paying the reward for reporting.*

**Assumption 5.** $t_1 + x < \bar{l}$, *i.e. the sum of the disutility from embezzlement and the reward for reporting must be smaller than the profit from cooperation for* $\bar{l}$*-employee.*

### A.2.2 Return behavior: Equilibrium outcome

**Lemma 1. (Return behavior)** *In every equilibrium in all treatments, if the manager chose* $c = 1$*, a* $\underline{l}$*-employee always chooses* $t = t_0$ *and a* $\bar{l}$*-employee always chooses* $t = t_1$*. That is,*

$$t^*(c,l) = \begin{cases} t_1 & \text{if } l = \bar{l} \text{ and } c = 1, \\ t_0 & \text{if } l = \underline{l} \text{ and } c = 1, \\ n.a. & \text{if } c = 0. \end{cases}$$

*Proof.* First, the employee can only choose a transfer $t$ for $c = 1$, i.e. a pie of 1 is created. So assume that the utility of the employee in stage 4 is $U_4(t_0) = 1 - t_0 - l$ and $U_4(t_1) = 1 - t_1$. From $U_4(t_1) > U_4(t_0)$ follows $l > t_1 - t_0$. By assumption 1, a type $\bar{l}$ ($\underline{l}$) employee will choose $t_1$ ($t_0$). □

### A.2.3 Treatment *Baseline*: Equilibrium Outcome

**Lemma 2. (*Baseline*: Cooperation)** *The manager always (never) cooperates if she chose to embezzle and the employee does not (does) send a report. If the manager didn't embezzle, she only cooperates if the share of loyal employees is high enough. That is,*

$$c^*(r,e,\alpha) = \begin{cases} 1 & \text{if } e = 1 \text{ and } r = 0, \\ 1 & \text{if } e = 0 \text{ and } r = 0 \text{ and } \alpha > \alpha', \\ 0 & else, \end{cases}$$

*with* $\alpha' := \frac{I - t_0}{t_1 - t_0}$.

*Proof.* While the profit for the manager from not cooperating in stage 3 is zero, the expected profit from cooperation for the manager in stage 3 is $\pi(c = 1) = \Pr(\bar{l}) \cdot t_1 + 1 - \Pr(\bar{l}) \cdot t_0 - I$. For cooperation to be profitable it must hold that $\pi(c = 1) > \pi(c = 0) \iff \Pr(\bar{l}) \cdot t_1 + (1 - \Pr(\bar{l})) \cdot t_0 > I$. First, we consider the case where the manager chose $e = 1$. In the candidate separating equilibrium, the reporting decision perfectly reveals the employee's type. That means $r = 1 \implies \Pr(\bar{l}) = q^*(1) = 0$ and $r = 0 \implies \Pr(\bar{l}) = q^*(0) = 1$. Therefore, by assumption 2 the manager will choose $c^*(r = 0) = 1$ and $c^*(r = 1) = 0$, since $r = 1 \implies \pi(c = 1) = t_0 - I < 0$ and $r = 0 \implies \pi(c = 1) = t_1 - I > 0$. If the manager chose $e = 0$, reporting cannot be profitable for any type. Therefore, the employee not reporting does not reveal the employee's type, such that $\Pr(\bar{l}) = q^*(0) = \alpha$. Therefore, the manager cooperates only if $\alpha > \alpha' := \frac{I - t_0}{t_1 - t_0}$. □

**Lemma 3. (*Baseline*: Reporting)** *A $\underline{l}$($\bar{l}$)-employee always (never) reports if the manager chooses to embezzle. Both types do not report if manager does not embezzle. That*

*is,*

$$r^*(l,e) = \begin{cases} 1 & \text{if } l = \underline{l} \text{ and } e = 1, \\ 0 & \text{if } l = \bar{l} \text{ and } e = 1, \\ 0 & \text{if } e = 0. \end{cases}$$

*Proof.* The employee anticipates subsequent cooperation and payment decisions, as well as beliefs by the manager. First, we consider the scenario where the manager chose $e = 1$ and a $\bar{l}$-employee. Since the $\bar{l}$-employee would choose $t_1$ in stage 4 if the manager cooperates, her utility in stage 2 is $\bar{U_2}(r = 1) = 0$ and $\bar{U_2}(r = 0) = -\bar{\delta} + 1 - t_1$. From $\bar{U_2}(r = 0) > \bar{U_2}(r = 1)$ follows $-\bar{\delta}+1-t_1 > 0 \iff \bar{\delta} = 1-\bar{l} < 1-t_1$ and therefore $\bar{l} > t_1$, which holds by assumption 3. Second, we consider the scenario where the manager chose $e = 1$ and a $\underline{l}$-employee. Since the $\underline{l}$-employee would choose $t_0$ in stage 4 if the manager cooperates, her utility in stage 2 is $\underline{U_2}(r = 1) = 0$ and $\underline{U_2}(r = 0) = -\underline{\delta} + 1 - t_0 - \underline{l}$ From $\underline{U_2}(r = 1) > \underline{U_2}(r = 0)$ follows $-\underline{\delta} + (1 - t_0) - \underline{l} < 0$ and therefore $0 > -t_0$. As we consider a scenario where only loyalty costs occur in the case of a false report for both types, neither type reports when there is no embezzlement. In consequence, a type $\underline{l}(\bar{l})$-employee will optimally choose (not) to report if the manager embezzles and neither type reports if the manager does not embezzle. □

**Lemma 4. (*Baseline*: Embezzlement)** *A manager only chooses to embezzle if the share of loyal employees $\alpha$ is sufficiently high. That is,*

$$e^*(\alpha) = \begin{cases} 1 & \text{if } l = \alpha > \alpha' \text{ and } \alpha > {\alpha_B}'', \\ 1 & \text{if } l = \alpha < \alpha'' \text{ and } \alpha > {\alpha_B}''', \\ 0 & \text{else}, \end{cases}$$

*with $\alpha' := \frac{I-t_0}{t_1-t_0}$, ${\alpha_B}'' := \frac{t_0-I+F}{t_0-I+F+E}$ and ${\alpha_B}''' := \frac{F}{t_1-I+F+E}$.*

*Proof.* If the manager chooses $e = 0$, both types would not report and the decision about $c$ depends on the size of $\alpha$. If $\alpha > \alpha'$ the manager would cooperate and receive an expected payoff of $\pi(c = 1) = \alpha \cdot t_1 + (1-\alpha) \cdot t_0 - I > 0$. If $\alpha < \alpha'$ the manager would get $\pi(c = 0) = 0$. Therefore, whether $e = 1$ is profitable depends as well on the size of $\alpha$. Given $\alpha > \alpha'$, for embezzlement to be profitable it must hold that $\pi(e = 1) > \pi(e = 0) \iff \pi(c = 1) = \alpha \cdot (t_1 + E - I) + (1-\alpha) \cdot (-F) > \alpha \cdot t_1 + (1-\alpha)t_0 - I \iff \alpha > {\alpha_B}'' := \frac{t_0-I+F}{t_0-I+F+E}$. Given $\alpha < \alpha'$, for embezzlement to be profitable it must hold that $\pi(e = 1) > \pi(e = 0) \iff \pi(c = 1) = \alpha \cdot (t_1 + E - I) + (1 - \alpha) \cdot (-F) > 0 \iff \alpha > {\alpha_B}''' := \frac{F}{t_1-I+F+E}$. □

**Proposition 1. (*Baseline*: Equilibrium Outcome)** *The Baseline treatment has the following equilibrium outcome: (i) An $\underline{l}(\bar{l})$-employee always (never) reports if the manager chooses to embezzle. (ii) Any employee does not report if the manager does not embezzle. (iii) A manager never cooperates if the employee sent a report. (iv) A manager cooperates if she embezzled and the employee did not send a report or if she didn't embezzle and the share of $\bar{l}$-employees $\alpha$ is larger than $\alpha' := \frac{I-t_0}{t_1-t_0}$. (v) A manager does only embezzle, if $\alpha$ is larger than ${\alpha_B}'' := \frac{t_0-I+F}{t_0-I+F+E}$ if $\alpha > \alpha'$, or if $\alpha$ is larger than ${\alpha_B}''' := \frac{F}{t_1-I+F+E}$ if $\alpha < \alpha'$. (vi) An employee of type $\bar{l}(\underline{l})$ always chooses $t = t_1$ $(t = t_0)$.*

### A.2.4 Treatment *Immunity*: Equilibrium Outcome

**Lemma 5. (*Immunity*: Cooperation)** *The manager always (never) cooperates if she chose to embezzle and the employee does not (does) send a report. If the manager didn't embezzle she only cooperates if the share of loyal employees is high enough. That is,*

$$c^*(r, e, \alpha) = \begin{cases} 1 & \text{if } e = 1 \text{ and } r = 0, \\ 1 & \text{if } e = 0 \text{ and } r = 0 \text{ and } \alpha > \alpha', \\ 0 & \text{else}, \end{cases}$$

*with $\alpha' := \frac{I-t_0}{t_1-t_0}$.*

*Proof.* As in the baseline treatment, if the manager chose $e = 0$, reporting cannot be profitable for any type since $x < \underline{l} < \bar{l}$. Therefore, not reporting does not reveal the employee's type such that $\Pr(\bar{l}) = q^*(0) = \alpha$. Therefore, the manager cooperates only if $\alpha > \alpha' := \frac{I-t_0}{t_1-t_0}$. For $e = 1$, the cooperation decision has to be evaluated differently given the reporting decision of the employee, since reporting is incentivized. First, if the employee does not report the scenario is identical to the baseline treatment. Second, if the employee does report, the payoff for the manager from not cooperating is now $\pi(c = 0) = -x$. The expected profit from cooperation for the manager in stage 3 is still $\pi(c = 1) = \Pr(\bar{l}) \cdot t_1 + 1 - \Pr(\bar{l}) \cdot t_0 - I$. For cooperation to be profitable it must hold that $\pi(c = 1) > \pi(c = 0) \iff \Pr(\bar{l}) \cdot t_1 + (1 - \Pr(\bar{l})) \cdot t_0 > I - x$. In the candidate separating equilibrium, the reporting decision perfectly reveals the employee's type. That means $r = 1 \implies \Pr(\bar{l}) = q^*(1) = 0$ and $r = 0 \implies \Pr(\bar{l}) = q^*(0) = 1$. Therefore, by assumption 4 the manager will choose $c^*(r = 0) = 1$ and $c^*(r = 1) = 0$, since $r = 0 \implies \pi(c = 1) = t_1 - I > 0$ and $r = 1 \implies \pi(c = 1) = t_0 - I + x < 0$ (which is harder to sustain compared to the baseline treatment). □

**Lemma 6. (*Immunity*: Reporting)** *An employee of type $\underline{l}$ ($\bar{l}$) always (never) reports if the manager chooses to embezzle. Both types do not report if manager does not embezzle. That is,*

$$r^*(l) = \begin{cases} 1 & \text{if } l = \underline{l} \text{ and } e = 1, \\ 0 & \text{if } l = \bar{l} \text{ and } e = 1, \\ 0 & \text{if } e = 0. \end{cases}$$

*Proof.* Suppose assumption 4 holds: Again, we first consider the scenario where the manager chose $e = 1$ and an employee of type $l = \bar{l}$. Since she would choose $t_1$ in stage 4 if the manager cooperates, her utility in stage 2 is $\bar{U}_2(r = 1) = x$ and $\bar{U}_2(r = 0) = -\bar{\delta} + 1 - t_1$. From $\bar{U}_2(r = 0) > \bar{U}_2(r = 1)$ follows $-\bar{\delta} + 1 - t_1 > x \iff \bar{\delta} = 1 - \bar{l} < 1 - t_1 - x$ and therefore $\bar{l} > t_1 + x$, which holds by assumption 5. Second, we consider the scenario where the manager chose $e = 1$ and the employee is of type $l = \underline{l}$. Since she would choose $t_0$ in stage 4 if the manager cooperates, her utility in stage 2 is $\underline{U}_2(r = 1) = x$ and $\underline{U}_2(r = 0) = -\underline{\delta} + 1 - t_0 - \underline{l}$ From $\underline{U}_2(r = 1) > \underline{U}_2(r = 0)$ follows $-\underline{\delta} + (1 - t_0) - \underline{l} < x$ and therefore $x > -t_0$. As we consider a scenario where the reward $x$ is smaller than the than the loyalty costs–which occur in case of a false report–for both types, neither type reports when there is no embezzlement. In consequence, a type $\underline{l}(\bar{l})$ employee will optimally choose (not) to report if the manager embezzles and neither type reports if the manager does not embezzle. □

**Lemma 7. (*Immunity*: Embezzlement)** *A manager only chooses to embezzle if the share of loyal employees $\alpha$ is sufficiently high. That is,*

$$e^*(\alpha) = \begin{cases} 1 & \text{if } l = \alpha > \alpha' \text{ and } \alpha > {\alpha_I}'', \\ 1 & \text{if } l = \alpha < \alpha'' \text{ and } \alpha > {\alpha_I}''', \\ 0 & \text{else}, \end{cases}$$

*with $\alpha' := \frac{I - t_0}{t_1 - t_0}$, ${\alpha_I}'' := \frac{t_0 - I + F + x}{t_0 - I + F + E} + x$ and ${\alpha_I}''' := \frac{F + x}{t_1 - I + F + E + x}$.*

*Proof.* If the manager chooses $e = 0$, both types would not report and the decision about $c$ depends on the size of $\alpha$. If $\alpha > \alpha'$ the manager would cooperate and receive an expected payoff of $\pi(c = 1) = \alpha \cdot t_1 + (1 - \alpha) \cdot t_0 - I > 0$. If $\alpha < \alpha'$ the manager would get $\pi(c = 0)$. Therefore, whether $e = 1$ is profitable depends as well on the size of $\alpha$. Given $\alpha > \alpha'$, for embezzlement to be profitable it must hold that $\pi(e = 1) > \pi(e = 0) \iff \pi(c = 1) = \alpha \cdot (t_1 + E - I) + (1 - \alpha) \cdot (-F - x) > \alpha \cdot t_1 + (1 - \alpha) t_0 - I \iff \alpha > {\alpha_I}'' := \frac{t_0 - I + F + x}{t_0 - I + F + E + x}$. Given $\alpha < \alpha'$, for embezzlement to be profitable it must hold that $\pi(e = 1) > \pi(e = 0) \iff$

$\pi(c=1) = \alpha \cdot (t_1 + E - I) + (1-\alpha) \cdot (-F - x) > 0 \iff \alpha > \alpha_I''' := \frac{F+x}{t_1 - I + F + E + x}$. □

**Proposition 2. (*Immunity*: Equilibrium Outcome)** *The immunity treatment has the following equilibrium outcome: (i) An $\underline{l}(\bar{l})$-employee always (never) reports if the manager chooses to embezzle. (ii) Any employee does not report if the manager does not embezzle. (iii) A manager never cooperates if the employee sent a report. (iv) A manager cooperates if she embezzled and the employee did not send a report or if she didn't embezzle and the share of $\bar{l}$-employees $\alpha$ is larger than $\alpha' := \frac{I - t_0}{t_1 - t_0}$. (v) A manager does only embezzle, if $\alpha$ is larger than $\alpha_I'' := \frac{t_0 - I + F + x}{t_0 - I + F + E + x}$ if $\alpha > \alpha'$, or if $\alpha$ is larger than $\alpha_I''' := \frac{F+x}{t_1 - I + F + E + x}$ if $\alpha < \alpha'$. (vi) An employee of type $\bar{l}(\underline{l})$ always chooses $t = t_1$ ($t = t_0$).*

### A.2.5 Treatment *Anonymity*: Equilibrium Outcome

**Lemma 8. (*Anonymity*: Cooperation)** *The manager only cooperates if the share of loyal employees is high enough. That is,*

$$c^*(\alpha) = \begin{cases} 1 & \text{if } \alpha > \alpha', \\ 0 & else, \end{cases}$$

*with $\alpha' := \frac{I - t_0}{t_1 - t_0}$.*

*Proof.* In the anonymity treatment, the reporting decision does not convey any information about the type of the employee. The crucial condition for the cooperation decision of the manager is therefore: $\alpha \cdot t_1 + (1-\alpha) \cdot t_0 \lesseqgtr I$. First, we consider $\alpha < \alpha' := \frac{I - t_0}{t_1 - t_0}$. If the manager chose $e = 1$, both types will report the embezzlement, because they only avoid the disutility from undetected embezzlement. That means, the belief of the manager is $\Pr(\bar{l}) = q(1) \implies \alpha \implies c^* = 0$. As before, if the manager chose $e = 0$, reporting cannot be profitable for any type since $x < \underline{l} < \bar{l}$. Therefore, the employee not reporting does not reveal her type, such that $\Pr(\bar{l}) = q(0) \implies \alpha \implies c^* = 0$. Second, we consider $\alpha > \alpha' := \frac{I - t_0}{t_1 - t_0}$. If the manager chose $e = 1$, both types will report the embezzlement, because they avoid the disutility from undetected embezzlement and do not affect the cooperation decision of the manager. That means, the belief of the manager is $\Pr(\bar{l}) = q(1) \implies \alpha \implies c^* = 1$. For $e = 0$, still, reporting cannot be profitable for any type. Therefore, the employee not reporting does not reveal her type, such that $\Pr(\bar{l}) = q(0) \implies \alpha \implies c^* = 1$. □

**Lemma 9. (*Anonymity*: Reporting)** *Any employee reports if the manager chooses to*

*embezzle. Both types do not report if manager does not embezzle. That is,*

$$r^*(l) = \begin{cases} 1 & \text{if } e = 1, \\ 0 & \text{if } e = 0. \end{cases}$$

*Proof.* First, we consider $\alpha < \alpha' := \frac{I-t_0}{t_1-t_0}$. If the manager chose $e = 1$, the utility of the employee in stage 2 is $U_2(r = 1) = 0$ and $U_2(r = 0) = -\delta \implies r^*(l) = 1$. Second, we consider $\alpha > \alpha' := \frac{I-t_0}{t_1-t_0}$. If the manager chose $e = 1$, the utility of a $\underline{l}$-employee in stage 2 is $\underline{U}_2(r = 1) = 1-\underline{l}-t_0$ and $\underline{U}_2(r = 0) = -\underline{\delta}+1-t_0-\underline{l} \implies r^*(l) = 1$. For a $\bar{l}$-employee, the utility in stage 2 is $\bar{U}_2(r = 1) = 1 - t_1$ and $\bar{U}_2(r = 0) = -\bar{\delta} + 1 - t_1 \implies r^*(l) = 1$. As before, if the manager chose $e = 0$, reporting cannot be profitable for any type. In consequence, any employee will optimally choose (not) to report if the manager does (not) embezzle. □

**Lemma 10. (*Anonymity*: Embezzlement)** *A manager never embezzles. That is,* $e^* = 0$.

*Proof.* If the manager chooses $e = 1$, both types would report and she would make a loss for sure since her cooperation decision is independent of the reporting decision such that $\pi(e = 0) = \alpha \cdot t_1 + (1 - \alpha) \cdot t_0 - I < \alpha \cdot t_1 + (1 - \alpha) \cdot t_0 - I - F = \pi(e = 1) \iff -F < 0 \implies e^* = 0$. □

**Proposition 3. (*Anonymity*: Equilibrium Outcome)** *The anonymity treatment has the following equilibrium outcome: (i) Any employee does not report if the manager does not embezzle. (ii) A manager cooperates if the share of $\bar{l}$-employees $\alpha$ is larger than $\alpha' := \frac{I-t_0}{t_1-t_0}$. (iii) A manager never embezzles. (iv) An employee of type $\bar{l}(\underline{l})$ always chooses $t = t_1$ $(t = t_0)$.*

### A.2.6 Treatment *Anonymity and Immunity*: Equilibrium Outcome

Note that the only difference between treatments "Anonymity" and "Anonymity and Immunity" is that in the latter reporting is rewarded in the case where cooperation does not take place. Since both types of employees report if and only if the manager embezzled, embezzlement is already deterred by the provision of anonymity. In consequence, both types do not report and the reward does not come into effect. It follows that the respective equilibrium outcomes are the same in both treatments:

**Proposition 4. (*Anonymity and Immunity*: Equilibrium Outcome)** *In the treatments "Anonymity" and "Anonymity and Immunity", the equilibrium outcomes coincide.*

# B Supplementary Material

## B.1 Translated Instructions

Welcome to today's experiment! If you read the following instructions carefully, you can earn a significant payment - depending on your decisions.

Please note, that from now on and during the whole experiment no communication is allowed. If you have any questions, please direct these at one of the experimenters. Neglecting these rules result in exclusion from this experiment and all payments.

All your decisions during this experiment will remain anonymous and cannot be related to you by either the experimenters nor the fellow subjects. Your earnings will be accounted in points. The points you acquire during this experiment will be exchanged for euro at the end. The exchange rate is: **10 points = 50 eurocent**.

***General procedure:***

There are **three roles** in this experiment: *Manager*, *employee* and *a third party*. These roles are assigned randomly. If you are drawn into the role *manager*, you'll maintain this role throughout the entire experiment. If you start with one of the other two roles, your role will be drawn randomly before each period. In each period you are part of a group consisting of exactly one manager, one employee and one third party. Also the group composition will result from a random draw in every period.

The experiment is divided into two parts consisting of multiple periods. Beneath you find the procedure of a period in part 1. For the second part, you'll receive instructions on your screen immediately before it starts.

***Procedure of a period in part 1:***

Every subject is endowed with 100 points. After the roles are assigned, the manager chooses between two alternatives (CIRCLE or TRIANGLE). CIRCLE has no payoff consequences for any member of the group. TRIANGLE represents violating the law, resulting in a gain (50 points) for the *manager*, and a loss (90 points) for the *third party*. Again, there are no consequences for the *employee*.

After the manager has made her choice about CIRLCE and TRIANGLE, the employee has to decide whether she wants to file a complaint. This decision is taken separately for both alternatives (complaint if CIRCLE was chosen; complaint if TRIANGLE was chosen). Filing a complaint causes costs for the manager in any case (10 points). If CIRCLE has been chosen and complaint has been filed, the manager has to pay an additional fine (60 points). The third party receives partial compensation for her damage (80 points).

The table below displays all possible combinations of the decisions made by the manager and the employee as well as its respective payoffs for all group members.

| *Manager* chooses alternative | *Employee* files a complaint | Payoffs | | |
|---|---|---|---|---|
| | | *Manager* | *Employee* | *Third Party* |
| **Circle** | No | 0 | 0 | 0 |
| **Circle** | Yes | -10 | 0 | 0 |
| **Triangle** | No | 50 | 0 | -90 |
| **Triangle** | Yes | -20 | 0 | -10 |

Subsequently, all group members are informed about the chosen alternative[ and whether there has been a complaint].

To conclude a period the manager and the employee play an investment game. First, the manager chooses an amount x between -30 and 60 points. Negative figures mean that points are taken from the employee. Positive mean that points are sent to the employee. If the manager deducts points from the employee these points are transferred and the investment game ends. If the manager sends a positive amount to the employee, it will be multiplied by three. In this case, the employee chooses an amount y between 0 and $3 \cdot x$ which she would like to return to the manager. There are no consequences for the third party in the investment game.

Payoffs in the investment game:

Manager = - x + y points,

Employee = max(x, 3· x) - y points,

Third party = 0.

At the end of a period[ all of the group members are informed whether there was a complaint and] your surplus adds up from your **endowment** (100 points), **your revenue from the decisions made** (see table) and **your revenue from the investment game**.

Summary of a period in part 1

1. Manager chooses alternative CIRLCE or TRIANGLE (violation of law)
2. Employee decides upon reporting
3. Every member of a group learns about the chosen alternative [and the reporting decision]

4. Manager and employee engage in an investment game
(5. Every member of a group learns about the reporting decision)

5./6. The surplus is computed

After you have completed the second part and a questionnaire, **one period** is drawn for payout. You'll receive the points you earned in that period converted according to the exchange rate plus 5 euro as show up fee.

Thank you for participating and good luck!

## B.2 Control Questions

1. Do you keep your role through the entire experiment?
   ☐ Yes, always.
   ☐ No, my role is randomly drawn in each period.
   ☐ Yes, in case I am an manager. If I am an employee or the third party, it may change from period to period.
2. Do you have the same members in your group over several periods?
   ☐ No.
   ☐ Yes, in the second part of the experiment.
   ☐ Yes, always.
3. If the manager chooses TRIANGLE, ...
   ☐ she receives a profit and harms the employee as well as the third party.
   ☐ she does not receive a profit, but harms the employee as well as the third party.
   ☐ she receives a profit and harms the third party, but not the employee.
4. If the manager chooses CIRCLE and the employee files a report, ...
   ☐ all payoffs are unaffected.
   ☐ it causes a cost for the manager. Both the employee and the third party are not affected.
   ☐ it causes a cost for the manager. Both the employee and the third party receive a profit.
5. If the manager sends 30 points in the investment game, how many points does the employee receive?
   _____

## B.3 Questionnaire

**Demographics**

1. How old are you? _____
2. What is your sex? ☐ Male ☐ Female
3. What are you studying? _____

4. How much work expericence do you have?
   (a) Internships (in month): _____
   (b) Full-time (in month): _____
   (c) Student jobs (in month): _____

**Risk preferences**

1. Imagine you had won 100,000 euros in a lottery. Almost immediately after you collect, you receive the following financial offer from a reputable bank, the conditions of which are as follows: There is the chance to double the money within two years. It is equally possible that you could lose half of the amount invested. What fraction would you choose to invest?
   ☐ 0 ☐ 20,000 ☐ 40,000 ☐ 60,000 ☐ 80,000 ☐ 100,000

**Attitudes towards whistleblowing**

1. What is your opinion with respect to the following claims?
   (a) A person should be supported in disclosing serious misbehavior, even if this requires disclosure of insider information.
   ☐ Strongly agree ☐ Agree ☐ No opinion ☐ Disagree ☐ Strongly disagree
   (b) A person should be supported in disclosing already mild misbehavior, even if this requires disclosure of insider information.
   ☐ Strongly agree ☐ Agree ☐ No opinion ☐ Disagree ☐ Strongly disagree
   (c) I would disclose serious misbehavior, even it would cause disadvantages for me.
   ☐ Strongly agree ☐ Agree ☐ No opinion ☐ Disagree ☐ Strongly disagree
   (d) I would disclose already mild misbehavior, even it would cause disadvantages for me.
   ☐ Strongly agree ☐ Agree ☐ No opinion ☐ Disagree ☐ Strongly disagree
   (e) If the chance is larger that misbehavior is detected it could be deterred.
   ☐ Strongly agree ☐ Agree ☐ No opinion ☐ Disagree ☐ Strongly disagree
2. In your opinion, how acceptable are the following actions?
   (a) Disclosing insider information about serious misbehavior by person in authority of an organization.
   ☐ Very acceptable ☐ Acceptable ☐ Neither, nor ☐ Unacceptable ☐ Very unacceptable
   (b) Disclosing insider information about serious misbehavior by regular employees of an organization.
   ☐ Very acceptable ☐ Acceptable ☐ Neither, nor ☐ Unacceptable ☐ Very unacceptable
   (c) Disclosing insider information about serious misbehavior by a friend or family

member of an organization's member.

☐ Very acceptable ☐ Acceptable ☐ Neither, nor ☐ Unacceptable ☐ Very unacceptable

3. Imagine you had insider information about serious misbehavior in an organization you are a member of. How important was each of the following items for the decision to tell someone about it?
   (a) Persons in authority would support me.
   ☐ Very important ☐ Important ☐ Neither, nor ☐ Unimportant ☐ Very unimportant
   (b) I would be legally obliged to report.
   ☐ Very important ☐ Important ☐ Neither, nor ☐ Unimportant ☐ Very unimportant
   (c) Somebody would act to end the misbehavior.
   ☐ Very important ☐ Important ☐ Neither, nor ☐ Unimportant ☐ Very unimportant
   (d) Only people I choose would know my identity.
   ☐ Very important ☐ Important ☐ Neither, nor ☐ Unimportant ☐ Very unimportant
   (e) Apart from the people I contact, the information would remain confidential.
   ☐ Very important ☐ Important ☐ Neither, nor ☐ Unimportant ☐ Very unimportant
   (f) I would remain completely anonymous.
   ☐ Very important ☐ Important ☐ Neither, nor ☐ Unimportant ☐ Very unimportant

# C Descriptive Statistics and Survey Responses

Table C.1 displays the average characteristics of the subjects cut by treatment and role.

# D Regression Analysis

In this section, I present the regression results for the decisions on embezzlement, truthful and false reporting, and the return behavior, which are discussed in Section 5.

| | Anonymity | | No Anonymity | |
|---|---|---|---|---|
| | Manager | Employee | Manager | Employee |
| *characteristic* | | | | |
| age | 24.0 | 24.2 | 24.8 | 25.9 |
| female | 0.79 | 0.63 | 0.77 | 0.57 |
| risk | 1.16 | 1.34 | 1.20 | 0.95 |
| work experience | 0.17 | -0.15 | 0.01 | 0.06 |
| attitude reporting | -0.15 | 0.09 | -0.26 | 0.03 |
| attitude disclosure | -0.04 | 0.10 | -0.13 | -0.02 |
| attitude environment | 0.00 | -0.03 | -0.06 | 0.03 |
| No. of subjects | 19 | 38 | 30 | 60 |

Notes: The table reports the average characteristics of the subjects per treatment and role. *risk* is measured on a scale from 0-5, where 5 is extremely risk-loving. *work experience* is a standardized measure of the answers to question 4 in the "demographics" section of the questionnaire, where a higher score represents more month of work experience. *attitude reporting* is a standardized measure of the answers to question 1 in the "attidtudes towards whistleblowing" section of the questionnaire, where a higher score represents a stronger support for whistleblowing. *attitude disclosure* is a standardized measure of the answers to question 2 in the "attidtudes towards whistleblowing" section of the questionnaire, where a higher score represents a greater appropriateness of disclosing insider information. *attitude reporting* is a standardized measure of the answers to question 3 in the "attidtudes towards whistleblowing" section of the questionnaire, where a higher score represents a greater importance of the legal environment for the decision to become a whistleblower.

Table C.1: Average characteristics per role over treatments

| | Truthful reporting | | | False reporting | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Treatment A | 0.131** | 0.143** | 0.144** | 0.0893 | 0.111* | 0.111* |
| | (0.0591) | (0.0610) | (0.0610) | (0.0569) | (0.0586) | (0.0593) |
| Treatment I | 0.167*** | 0.189*** | 0.194*** | 0.194*** | 0.187*** | 0.206*** |
| | (0.0437) | (0.0466) | (0.0473) | (0.0463) | (0.0475) | (0.0500) |
| Treatment AI | 0.189*** | 0.213*** | 0.218*** | 0.421*** | 0.412*** | 0.433*** |
| | (0.0610) | (0.0629) | (0.0634) | (0.0661) | (0.0658) | (0.0674) |
| CooperatingManager(lag) | | 0.0373 | 0.0238 | | -0.0982** | -0.0873** |
| | | (0.0394) | (0.0389) | | (0.0431) | (0.0415) |
| Constant | 0.704*** | 0.675*** | 0.630*** | 0.118*** | 0.148*** | -0.00979 |
| | (0.0464) | (0.0530) | (0.0678) | (0.0299) | (0.0374) | (0.0728) |
| Period FE | No | No | Yes | No | No | Yes |
| $N$ | 784 | 735 | 735 | 784 | 735 | 735 |
| $N_{groups}$ | 98 | 98 | 98 | 98 | 98 | 98 |
| $R^2$ | 0.0343 | 0.0426 | 0.0502 | 0.104 | 0.113 | 0.129 |

Notes: The table reports results from a random-effects GLS regression where $N_{groups}$ is the number of individuals. Dependent variable: (1)-(3): willingness to report truthfully (0 or 1), (4)-(6): willingness to report falsely (0 or 1). Standard errors in parentheses are clustered on the individual level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table D.1: Regression Analysis: Reporting

| | Embezzlement | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Treatment A | -0.0967 | -0.0127 | -0.0103 |
| | (0.0855) | (0.102) | (0.106) |
| Treatment I | -0.171*** | -0.0134 | -0.0204 |
| | (0.0501) | (0.0600) | (0.0688) |
| Treatment AI | -0.334*** | -0.0657 | -0.0735 |
| | (0.0749) | (0.0890) | (0.0885) |
| Embezzlement(lag) | | 0.708*** | 0.738*** |
| | | (0.0815) | (0.0846) |
| Report(lag) | | 0.0721 | 0.102 |
| | | (0.0912) | (0.0962) |
| ReportedEmbezzlement(lag) | | -0.613*** | -0.655*** |
| | | (0.135) | (0.137) |
| LowReturn(lag) | | -0.0204 | -0.0447 |
| | | (0.0696) | (0.0701) |
| Constant | 0.412*** | 0.243*** | 0.166 |
| | (0.0673) | (0.0782) | (0.140) |
| Period FE | No | No | Yes |
| $N$ | 784 | 206 | 206 |
| $N_{groups}$ | 49 | 44 | 44 |
| $R^2$ | 0.0694 | 0.163 | 0.198 |

Notes: The table reports results from a random-effects GLS regression where $N_{groups}$ is the number of individuals. Dependent variable: embezzlement decision of managers (0 or 1). Standard errors in parentheses are clustered on the individual level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table D.2: Regression Analysis: Embezzlement

| | Returning more than sent | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Treatment A | -0.0407 | -0.0313 | -0.0417 |
| | (0.103) | (0.102) | (0.103) |
| Treatment I | -0.124** | -0.101* | -0.0982 |
| | (0.0603) | (0.0599) | (0.0651) |
| Treatment AI | -0.214* | -0.173 | -0.154 |
| | (0.122) | (0.132) | (0.135) |
| Embezzlement | | -0.0968 | -0.0927 |
| | | (0.0857) | (0.0782) |
| FalseReport | | -0.183** | -0.208** |
| | | (0.0902) | (0.0940) |
| UnreportedEmbezzlement | | -0.0121 | -0.0612 |
| | | (0.128) | (0.127) |
| Constant | 0.540*** | 0.586*** | 0.690*** |
| | (0.0627) | (0.0689) | (0.0913) |
| Period FE | No | No | Yes |
| $N$ | 212 | 212 | 212 |
| $N_{groups}$ | 87 | 87 | 87 |
| $R^2$ | 0.00633 | 0.0255 | 0.0504 |

Notes: The table reports results from a random-effects GLS regression where $N_{groups}$ is the number of individuals. Dependent variable: (1)-(3): employees return more than what the manager sent (0 or 1). ReportedEmbezzlement, FalseReport, UnreportedEmbezzlement are all binary variables. Standard errors in parentheses are clustered on the individual level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table D.3: Regression Analysis: Return Behavior