Whistleblower Protection: Theory and Experimental Evidence[*]

Lydia Mechtenberg[†] Gerd Muehlheusser[‡] Andreas Roider[§]

January 2020

Abstract

Whistleblowing by employees plays a major role in uncovering corporate fraud. Recent laws and global policy recommendations aim at facilitating whistleblower protection to enhance the willingness to report and to increase the detection and deterrence of misbehavior. We study these issues in a theory-guided laboratory experiment. As expected, protecting whistleblowers leads to more reporting of misbehavior. However, the predicted improvements in detection and deterrence do not materialize in the experiment. This is not solely driven by non-meritorious claims and by a lower informativeness of reports when protection is in place.

JEL-Code: C91, D83, D73, K42, M59.

Keywords: Corporate Fraud, Corruption, Whistleblowing, Business Ethics, Cheap-Talk Games, Laboratory Experiment

[*] We are grateful to the editor Steffen Huck, the associate editor, and two referees for their comments and suggestions. This paper is a substantial revision of a previous version, circulated as CEPR DP No. 11898, CESifo WP No. 6394, and IZA DP No. 10607. We gratefully acknowledge financial support from the Fritz Thyssen Foundation (Grant 10.13.2.097). We also thank Ralph Bayer, Shahar Dillbary, Matthew Ellman, Eberhard Feess, Leonie Gerhards, Talia Gillis, Luise Görges, Johann Graf Lambsdorff, Igor Legkiy, Debrah Meloso, Claudia Schwirplies, Eric Talley, Niklas Wallmeier, and seminar participants at Aarhus, Bonn, EUI Florence, Fribourg, Graz, Hamburg, Innsbruck, Konstanz, Marburg, National Taiwan University, Paris X Nanterre, Passau, Portsmouth, Regensburg, St. Gallen, Ulm, the 2018 Western Finance Association Meeting (San Diego), the 2018 American Law and Economics Association Meeting (Boston), the 2018 International Industrial Organization Conference (Indianapolis), the 2018 Society for Institutional and Organizational Economics Meeting (Montreal), the 2017 European Law and Economics Association Meeting (London), the 2017 Verein für Socialpolitik Annual Conference (Vienna), and the 2018 Workshop on "Empirical Methods in Legal Scholarship" (Columbia) for their comments and suggestions. Christos Litsios and Pamela Mertens provided excellent research assistance.

[†] University of Hamburg, Department of Economics, lydia.mechtenberg@uni-hamburg.de

[‡] University of Hamburg, Department of Economics, IZA, and CESifo, gerd.muehlheusser@uni-hamburg.de

[§] University of Regensburg, Department of Economics, CEPR, IZA and CESifo, andreas.roider@ur.de

1 Introduction

1.1 Motivation

Corporate fraud is a major challenge in both developing and advanced economies, and employee whistleblowers play an important role in uncovering it. Indeed, the issue of protecting employee whistleblowers looms high on the international anti-corruption agenda of the G20 group, the Council of Europe, and the OECD, and some form of whistleblower protection legislation is already in existence in countries such as the U.S. or the UK. To the best of our knowledge, this is the first paper to study whistleblower protection in a theory-guided laboratory experiment, where the decisions to commit fraud, to blow the whistle, to investigate reports, and to retaliate against whistleblowers are endogenous. Our results suggest that whistleblower protection indeed fosters the reporting of misbehavior, but that its effects on the detection and deterrence of misbehavior are more intricate.

The topicality of corporate fraud is exemplified by high-profile scandals at Volkswagen, Enron, or Worldcom. More systematic evidence is, for example, presented by Dyck, Morse, and Zingales (2014). Using a natural experiment, they estimate the average cost of both detected and undetected fraud in large U.S. corporations in the period 1996-2004 to be $360 billion per year.[1] This evidence suggests that uncovering existing fraud and deterring potential fraud should indeed be a high priority for legislators and policy makers.

In fact, in recent years, the importance of employee whistleblowers (who are not participating in the misbehavior) for uncovering fraud has become evident, primarily because of their access to crucial information (in particular, with respect to fraud involving company insiders).[2] For example, Dyck, Morse, and Zingales (2010) consider all reported cases of fraud in large U.S. corporations between 1996 and 2004. They find that in 17% of the cases they study, the fraud was uncovered by employee whistleblowers, thereby outnumbering other players such as the SEC, auditors, non-financial market regulators, or the media.[3] The importance of employee

[1] According to the Association of Certified Fraud Examiners (2014), the average loss of organizations due to fraud (which includes financial statement fraud, asset misappropriation, and corruption) is estimated to be 5% of annual revenues. Taken at face value, this number would extrapolate into a worldwide loss from fraud of up to $3.7 trillion. Furthermore, in the latest "Global Fraud Report" (Kroll, 2016), 75% of surveyed senior executives stated that their company had become a fraud victim in the previous year.

[2] Kroll (2016) finds that in 81% of all fraud cases where perpetrators were known at least one company insider was involved, and a substantial share of 36% of these perpetrators came from senior or middle management.

[3] In the notorious corporate fraud scandals of Enron and Worldcom, the misbehavior was uncovered by employee whistleblowers (see, e.g., Healy and Palepu, 2003). Miceli, Near, and Dworkin (2009) survey fraud cases unveiled by whistleblowers in more than 20 countries.

whistleblowers has lead to a broad consensus among scholars and practitioners alike that, in order to uncover and deter corporate fraud, more employee whistleblowing would be very desirable. While employees who do report fraud often feel a moral obligation to do so, the fear of retaliation from co-workers or management is often a strong countervailing factor.[4] As a result, the overall willingness of employees to report misbehavior is often perceived as low. For example, Dyck, Morse, and Zingales (2010, p.2245) argue that "the surprising part is not that most employees do not talk, but that some talk at all."

As a consequence, the best-practice recommendations of international bodies, such as the G20 group, the Council of Europe, and the OECD (Council of Europe, 2014; OECD 2011, 2016), urge for comprehensive legal protection from retaliation for whistleblowers. Such legislation is already in place in the U.S., the UK, and a number of other countries, where examples include the Sarbanes-Oxley Act (SOX), the Dodd-Frank Act, and the Public Interest Disclosure Act (see, e.g., Thüsing and Forst, 2016). For instance, U.S. law aims to shield whistleblowers from tangible employment actions (such as dismissal or demotion) and other forms of retaliation.

There is no doubt that employee whistleblowers deserve strong protection. A number of scholars, practitioners, and enforcement agencies have at the same time voiced the concern that current whistleblower-protection policies might also lead to dysfunctional responses in the form of non-meritorious claims. In particular, it has been argued that low-performing employees might have an incentive to lodge claims to seek shelter from unfavorable actions such as dismissal. For example, in their study of the New York City adminstration, Anechiarico and Jacobs (1996, p.69) state that "some disgruntled, incompetent, or otherwise poorly performing employees will file whistleblower claims in order to keep their jobs as long as possible or simply to harass their supervisors."[5] If such non-meritorious claims were indeed successful, the employee would either be retained, or would settle to the effect of receiving a severance payment in exchange for termination.[6]

[4] See e.g., Near and Miceli (1986) and Alford (2001). Gobert and Punch (2000, pp.33ff) provide a number of striking examples of whistleblowers suffering from various forms retaliation after coming forward.

[5] See also Gobert and Punch (2000, pp.32ff), Schmidt (2005, p.158), Bowen, Call, and Rajgopal (2010, p.1240), or Blount and Markel (2012, p.1042), who reiterate the arguments and cases of Anechiarico and Jacobs (1996, pp.67ff). Furthermore, a USA Today (2004) article on the practical consequences of SOX quotes practitioner statements such as "some of the more difficult problems I've had is whistleblowers who will raise issues in which we find some merit, but where they will raise them to gain personal protection for marginal performance".

[6] In a similar vein, it has been argued that monetary rewards for whistleblowers, as for example provided for in the Dodd-Frank Act, might also generate incentives to file non-meritorious claims, see e.g., Hartmann (2011, p.1303), Ebersole (2011, p.135), Blount and Markel (2012, p.1041), Hansberry (2012, p.196), or Rose (2014, p.1283).

Since such non-meritorious claims would be an unintended by-product in the quest for improving the protection of whistleblowers, this raises the question of their empirical relevance. In this respect, many of these policies (such as SOX and Dodd-Frank) and the abovementioned best practice recommendations do not seem to provide strong safeguards against non-meritorious claims. In particular, whistleblowers are often deliberately not required to provide conclusive proof of their allegations. Instead, they need to demonstrate a *reasonable belief* with respect to the presence of fraud. The rationale behind the use of this judicial standard is to set the bar for protection not too high in order to encourage whistleblowers to come forward (for a discussion, see e.g., Kohn, Kohn and Colapinto, 2004, pp.92ff). However, since claims often involve complex inside information, a claim's real merit is often hard and costly to assess for a judge from an ex ante perspective, and even allegations without merit might appear reasonable (see e.g., Schmidt, 2005, p.158, Ebersole, 2011, p.135). Therefore, while the reasonable belief standard will certainly prevent obviously unsubstantiated claims, it potentially leaves scope for (ex post) non-meritorious claims to result in protection.[7] Moreover, existing policies do usually not sanction whistleblowers whose claims pass the reasonable belief threshold ex ante, but then turn out to be non-meritorious ex post (again motivated by the desire not to deter whistleblowers from coming forward).[8]

From the viewpoint of the responsible authorities, if non-meritorious claims are indeed an empirically relevant issue, reports might in general be perceived to be less informative about underlying misbehavior. As it is costly to "separate the wheat from the chaff" (Fleischer and Schmolke, 2012, p.254), this might reduce the authorities' responsiveness to reports (see e.g., Ebersole, 2011, p.135 and Rose, 2014, p.1283).[9] In turn, this could hamper the detection and deterrence of misbehavior (see e.g., Casey and Niblett, 2014, pp.1208ff).

[7] See e.g., Rose (2014, p.1283), who also argues that unwarranted protection might in addition be aided by the vagueness of many legal provisions (see also Ebersole, 2001, p.135, Hansberry, 2012, pp.211ff, or Blount and Markel, 2012, p.1041). Claims that are easily recognizable as fraudulent could for example be punished with sufficiently high fines, and hence be deterred in the first place.

[8] For example, Thüsing and Forst (2016) compare whistleblower legislation in 23 countries and find that, once obtained, protection often remains intact even if, in the end, it turns out that there was no misbehavior. Two such cases are discussed in Anechiarico and Jacobs (1996, pp.67ff). Furthermore, sanctions for claims that turn out to be unsubstantiated are in general either mild or ruled out altogether, as is for example the case for all claims administered by the U.S. Department of Labor (see e.g., Kohn, Kohn and Colapinto, 2004, p.34).

[9] For example, out of the 27,921 cases determined by the U.S. Department of Labor's Occupational Safety and Health Administration (OSHA) over the time period 2006-2016, 56% were dismissed (see `https://www.whistleblowers.gov/factsheets_page/statistics`).

1.2 Research Question, Framework, and Results

Given the global scale of corporate fraud and the importance of employees in uncovering it, whistleblower protection plays an instrumental role. To achieve the aims of improved reporting, detection, and deterrence of misbehavior, the effects of legal whistleblower-protection policies need to be well understood. To this end, we conduct a theory-guided experiment where predictions are derived from a cheap-talk model in the spirit of Crawford and Sobel (1982). Our framework considers the interaction between an employer (who may misbehave), an employee (who may blow the whistle), and a prosecutor (who may act upon the employee's report). Moreover, the employer might retaliate against a non-protected whistleblower in the form of dismissal. We focus on whistleblower protection in the form of employment protection (i.e., a protected employee cannot be dismissed), which is consistent with common legal practice as discussed above.[10] We allow employees to be heterogenous with respect to their productivity. In the main part of the paper, the monetary incentive structure is such that the employer prefers to dismiss low-productivity employees, but to retain high-productivity employees. When protection is available, this might give low-productivity employees an incentive to file non-meritorious claims in order to be shielded from dismissal.

In our baseline setting, we compare two treatments: In *NoP*, whistleblower protection (i.e., employment protection) is not available. In *P-R*, protection is obtained by filing a report. This treatment is meant to capture in a stylized way real-world legal regimes (such as U.S. and UK law, and the G20 group's policy recommendation), where protection is granted when the employee can demonstrate a reasonable belief with respect to the presence of fraud. Throughout we focus on the case where all whistleblower claims satisfy this reasonable-belief criterion, i.e., from the prosecutor's perspective they are not obviously unsubstantiated ex ante.

Theoretical predictions for the baseline treatments are derived in the Appendix. The underlying model incorporates three behavioral motives that are not incentivized in the experiment: First, employees suffer a disutility from undetected misbehavior, which provides an incentive to report. Second, employers face a disutility when retaining a whistleblower, which provides an incentive to retaliate in the form of dismissal. Third, employers differ with respect to their net benefit from misbehavior. Apart from these three behavioral motives, we assume that

[10] Employment protection is the most common remedy in whistleblower cases (see, e.g., Kohn, Kohn and Colapinto, 2004, pp.97ff). Thereby, the stated aim is to *make whole* the whistleblower, i.e., to re-establish the attained employment status before becoming a whistleblower (see, e.g., Kohn, Kohn and Colapinto, 2004, pp.102ff).

players have standard preferences.[11] We consider equilibria in which the prosecutor investigates if and only if the employee sends a report. The model makes the intuitive predictions that whistleblower protection leads to more reporting, improves the detection and deterrence of misbehavior, but also leads to non-meritorious claims by low-productivity employees.

Many of the theoretical predictions are supported in the experiment, for example with respect to retaliation in the form of dismissals and with respect to the stronger incentive to report misbehavior when there is protection. But there are also interesting deviations. In particular, prosecutors exhibit a lower responsiveness to reports when protection is available. This directly hampers the deterrence of misbehavior, and it potentially contributes to explaining why the predicted positive effect of whistleblower protection on deterrence only rarely materializes in the experiment.

The lower responsiveness of prosecutors to reports under whistleblower protection is then investigated in more detail in two additional treatments *P-RI* and *P-RIM*. We find that non-meritorious claims cannot fully explain this phenomenon.

In our baseline comparison, employers have a strong incentive to dismiss low-productivity employees because they are less productive than outside replacements, which employers could hire instead. To investigate the robustness of our results, in two further treatments *NoP-Low* and *P-R-Low* we remove such productivity-based incentives to dismiss low-productivity employees. This leads to a richer incentive structure with respect to the reporting and dismissal decisions. In this alternative labor-market setting, the main effects of whistleblower protection are qualitatively similar to the baseline comparison.

Overall, our findings suggest that whistleblower protection indeed fosters the reporting of misbehavior. However, the aims of increasing the detection and deterrence of misbehavior seem more intricate to achieve. Hence, one implication of our analysis is that these two issues should be carefully taken into account when designing whistleblower-protection policies.

The remainder of the paper is structured as follows. Section 2 discusses the related literature. Section 3 introduces the experimental setup. Section 4 summarizes the theoretical predictions, provides the underlying intuition, and presents the experimental results for our baseline treatments *NoP* and *P-R*. The results for the additional treatments are presented in

[11] Of course, in the general context of whistleblowing also other behavioral motives might be relevant. Examples include crowding-out of intrinsic motivation through financial incentives (Butler, Serra, and Spagnolo, 2019; Benabou and Tirole, 2003; Gneezy, Meier, and Rey-Biel, 2011), social judgment and image concerns (Butler, Serra, and Spagnolo, 2019; Bénabou and Tirole, 2006), or lying aversion (see e.g., Abeler, Nosenzo, and Raymond, 2019). For the sake of tractability these motives are not incorporated in the model.

Sections 5 and 6. Section 7 concludes. Appendix A contains the theoretical analysis. Appendix B provides translations of the experimental instructions. Appendices C–E contain supplementary material for the empirical analysis.

2 Related Literature

Our paper contributes to four strands of literature: First, it complements empirical research on employee whistleblowing using field data.[12] For example, Dyck, Morse, and Zingales (2010) document that employee whistleblowers play an important role in uncovering corporate fraud. However, with field data, some important variables, such as the level of *undetected* misbehavior, are typically not observed.[13] This makes it difficult to evaluate a whistleblower-protection policy's effect on the detection and deterrence of (total) misbehavior. Such limitations are alleviated in an experimental setup, where these variables are observed by the experimenter. Moreover, in the laboratory one can also vary the institutional setting, thereby studying various features of whistleblower-protection policies, which is typically difficult in the field. Our experimental setup builds on these advantages to analyze how whistleblower protection affects the reporting, detection, and deterrence of misbehavior.[14] Thereby, we complement other recent experimental studies on misbehavior and whistleblowing which differ in focus and are discussed in more detail below.

Second, there is a theoretical literature on whistleblowing that analyzes the optimal responsiveness of prosecutors to reports. In particular, in Chassang and Padró i Miquel (2019) whistleblowing is fostered through investigation policies that generate "garbled" information. They show that, to shield a whistleblower from retaliation by his employer, the optimal investigation policy (to which the prosecutor can commit ex ante) must not be too well aligned with reporting behavior. The reason is that such a policy would reveal that whistleblowing has in fact occurred, which would then trigger retaliation. In turn, this would undermine the incentive to report in the first place.[15] Like the present paper, Chassang and Padró i Miquel

[12] There also exists a large body of research in disciplines such as psychology, sociology, organizational behavior, and business ethics that studies the motives for whistleblowers to come forward (see e.g., the overviews by Miceli and Near, 1992; Miceli, Dworkin, and Near, 2008; Mesmer-Magnus and Viswesvaran, 2005; Vadera, Aguilera, and Caza, 2009). For a recent incentivized laboratory experiment, see Bartuli, Djawadi, and Fahr (2016).

[13] An exception is the natural experiment exploited by Dyck, Morse, and Zingales (2014).

[14] In a related setup, Wallmeier (2019) experimentally analyzes the effect of whistleblower protection on the level of trust among the members of an organization.

[15] Benoît and Dubra (2004) and Muehlheusser and Roider (2008) show that, even in the absence of a threat of direct retaliation, reporting might not occur due to the fear of enforcement errors or future non-cooperation.

(2019) analyze a cheap-talk game in which the decisions to misbehave, to report, and to investigate are endogenous. Hence, from a theoretical perspective, their setup is the one most closely related to ours, but there are a number of important differences: We compare equilibrium behavior under different schemes with and without protection, and we allow for heterogeneity of workers with respect to productivity. Moreover, we focus on pure-strategy equilibria where the prosecutor has no commitment power (and hence decides on whether or not to investigate only after a report has arrived). Finally, we also empirically test our model predictions in a laboratory experiment. Using a different modeling approach, Heyes and Kapur (2009) analyze how the optimal responsiveness of investigations depends on different behavioral motives for whistleblowing such as conscience cleansing, social welfare considerations, or disgruntlement. Our model captures the first of these motives by assuming that potential whistleblowers suffer a disutility from undetected misbehavior.

Third, there is a literature that analyzes the role of monetary rewards in fostering whistleblowing, as for example implemented in the False Claims Act and the Dodd-Frank Act. Dyck, Morse, and Zingales (2010) and Zingales (2004) discuss the beneficial role of such rewards in uncovering fraud, while others point to potentially adverse effects such as fostering non-meritorious claims.[16] In recent laboratory experiments, Butler, Serra, and Spagnolo (2019) and Schmolke and Utikal (2016) find that financial rewards indeed lead to more whistleblowing. In our paper, we do not consider direct financial rewards. Rather, an employee's reporting decision might increase or decrease the likelihood of retention (through employment protection and retaliation, respectively). In this sense, the reporting decision may also affect a subject's monetary payoff (i.e., the wage payment received), and we find that a positive (negative) monetary incentive leads to more (less) reporting. Butler, Serra, and Spagnolo (2019) also consider the interaction of financial incentives with social judgement, crowding-out of intrinsic motivation, and the type of misbehavior.[17] However, they do not allow for non-meritorious claims, retaliation by the employer, and endogenous prosecutor behavior, which are key elements of our analysis.

Finally, there is a literature that considers settings where also the whistleblowers themselves

[16] See e.g., Givati (2016), Howse and Daniels (1995), Callahan and Dworkin (1992) and the references given in Footnote 6. Of the 27 countries surveyed in OECD (2016), 30% have incentives for whistleblowers in place (e.g., financial rewards, expediency of the process, or follow-up mechanisms). In a world-wide survey, the Association of Certified Fraud Examiners (2014) finds that 11% of organizations had a reward scheme in place.

[17] Potential crowding-out effects of financial rewards on intrinsic motivation to report are also studied by Fiorin (2019) and Farrar, Hausserman, and Rennie (2019) in a field experiment and in a vignette study, respectively.

might have participated in the misbehavior, which is not the case in our setup. For example, recent experimental work on leniency programs in anti-trust has analyzed the self-reporting of cartel members and how to foster the reporting of illegal activities (see, e.g., Apesteguia, Dufwenberg, and Selten, 2007, Hinloopen and Soetevent, 2008, and Feltovich and Hamaguchi, 2018). Furthermore, Reuben and Stephenson (2013) conduct an experiment comparing groups that are either exogenously or endogenously formed. They study the level of misbehavior and its reporting and find that misbehavior is less frequently reported in endogenously formed groups. Also in an experimental setting, Cotten and Santore (2016) analyze the impact of transparency and amnesty rules in the context of corporate fraud by criminal teams.

3 Experimental Setup

In this section, we explain the design of the experiment, the monetary inventive structure, the treatments, and provide details concerning the implementation, respectively.

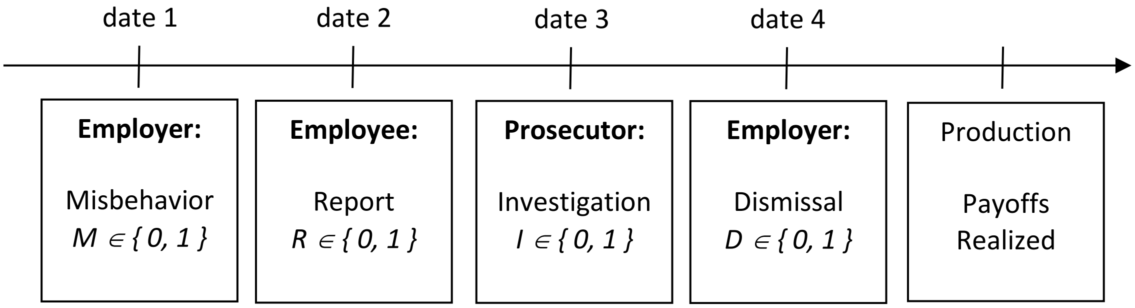
3.1 The Game Played in Each Period

Basic Setup In each of 30 periods per session, subjects were randomly (re-)matched into groups of four (stranger-design). They were assigned a role as either *employer*, *employee*, *prosecutor*, or *third party*, where the role assignments across periods are explained in more detail below. Employees are heterogenous with respect to their (exogenously given) productivity, which is either high ("H-employee") or low ("L-employee"), drawn randomly anew with equal probability at the beginning of each period. The third party is a purely passive player without any decisions to make, who suffers a loss from employer misbehavior. The third party is included in the experiment to make it more salient that misbehavior causes harm to others.

The remaining three players played the game illustrated in Figure 1: At date 1, the employer observes the productivity of her employee, and then chooses whether or not to misbehave. Misbehavior entails a gain, which is independent of her employee's productivity type. At date 2, we use the strategy method to elicit the employee's binary reporting decision in the case with and without employer misbehavior. Then, the employee observes the actual misbehavior decision of the employer. At date 3, the prosecutor observes whether or not a report is sent by the employee; but the prosecutor observes neither the underlying misbehavior decision of the employer nor the employee's productivity type.[18] The prosecutor then decides on trigger-

[18] Hence, we consider reports that are external in the sense of being directed towards the (outside) prosecutor.

Figure 1: The Sequence of Events in Each Period



ing an investigation. An investigation implies a private cost for the prosecutor and perfectly reveals whether or not the employer has misbehaved. The assumption that the prosecutor has discretion whether to initiate an investigation is in line with both the related literature (see, e.g., Chassang and Padró i Miquel, 2019; Givati, 2016; Heyes and Kapur, 2009) and legal practice (e.g., under SOX). If misbehavior is uncovered this benefits both the prosecutor and the third party, while the employer must pay an exogenously given fine. Finally, at date 4, the employer decides whether or not to dismiss the employee. If dismissed, the employee receives a payoff of zero and is replaced by a computerized outsider, who is more (less) productive than an L-employee (H-employee). However, dismissal is only feasible as long as the employee is not shielded by whistleblower protection. The observability of the employee's reporting decision to the employer is discussed below when we introduce the various treatments. At the end of each period, subjects learn their individual payoffs from the current period, and the decisions leading to these payoffs.

Monetary Incentives and Parameter Values In the experiment, the players' monetary payoff components, which were common knowledge ex ante, had the following properties: Unless detected, an employer's monetary payoff is higher upon misbehavior. Moreover, the difference between the productivity and the wage of the L-employee (H-employee) is smaller (larger) compared to employing the replacement outsider. Hence, the employer's monetary payoff is higher when dismissing (retaining) the L-employee (H-employee). By contrast, the monetary payoff of each employee type is always higher when retained. The monetary payoff of the third party is highest under no misbehavior, followed by detected, and then undetected misbehavior.

Some whistleblower laws stipulate that firms must establish internal reporting systems, and that whistleblowers must use these internal channels first, before resorting to outsiders. Incorporating this issue would require a richer framework, which might be an interesting topic for future research.

The motivation for this payoff ranking is that detecting misbehavior might allow to (at least partly) curb the associated harm. Finally, despite the investigation costs, when there actually is misbehavior, the prosecutor's monetary payoff is higher when he investigates.[19] In contrast, without misbehavior, the prosecutor's monetary payoff is higher when he does not investigate.

We used the following parameter values, where the numbers indicate experimental points: The productivities of H-employees, L-employees, and the outside replacement are given by 80, 30, and 70, respectively.[20] Employees who are not dismissed receive a fixed wage of 40. The employer's payoff from misbehavior is 50, and, in case of detection, she faces a fine of 60. When there is no misbehavior, the prosecutor's payoff is -20 (0) if he investigates (does not investigate). When there is misbehavior, his payoff is -10 (-20) if he investigates (does not investigate). The fine does not accrue to the prosecutor. Finally, the third party suffers a loss of 50 (70) from detected (undetected) misbehavior. In order to avoid negative payoffs at the end of the experiment, only prosecutors and third parties (who otherwise would face only negative payoff consequences) received per-period endowments of 60 and 40, respectively.

3.2 Treatments

We consider a total of six treatments. Treatments *NoP* and *P-R* form our baseline comparison of settings without and with whistleblower protection, respectively. Protection means that a whistleblower cannot be dismissed at date 4, which is the only difference between *NoP* and *P-R*. In both treatments, each role – employer, employee, prosecutor, and third party – is played by a real subject. Treatment *NoP* corresponds to a benchmark setting in which employment protection is not available. In treatment *P-R*, protection is obtained by sending a report (i.e., when $R = 1$). As discussed above, this treatment is meant to capture real-world legal regimes, where protection is granted when the employee can demonstrate a reasonable belief with respect to the presence of misbehavior. Reports are assumed to satisfy this reasonable-belief criterion, i.e., from the perspective of the prosecutor, reports are not obviously unsubstantiated. Moreover, because of the binary reporting decision, by design any report $R = 1$ looks the same to the prosecutor, independent of underlying misbehavior. In treatments *NoP* and *P-R*, both the prosecutor and the employer learn the reporting decision. This design choice was motivated by the fact that we wanted to rule out the possibility of erroneous updating by the employer

[19] Hence, given that there is misbehavior, an investigation is not only beneficial to the third party, but also to the prosecutor (which, in practice, might for example come in the form of a reputation gain).

[20] In Section 6, we discuss treatments where these payoffs are modified.

as a potential driver for dismissal decisions.[21] Details about the additional treatments will be introduced in Sections 5 (*P-RI* and *P-RIM*) and 6 (*NoP-Low* and *P-R-Low*).

3.3 Implementation of the Experiment

Summary Information The experiment was conducted in the experimental laboratory of the University of Hamburg and programmed in z-Tree (Fischbacher, 2007). We employed a between-subjects design, so that each subject participated in one treatment only. Sessions lasted for approximately 140 minutes, and participants earned 21 Euro on average (including a show-up fee of 12 Euro). For the recruitment of a total of 648 subjects, we used the software tool *hroot* (Bock, Baetge, and Nicklisch, 2014). The number of sessions and subjects per treatment is as follows: *NoP* (5;120), *P-R* (5;120), *P-RI* (4;88), *P-RIM* (4;96), *NoP-Low* (5;112), and *P-R-Low* (6;120). Virtually all subjects were undergraduate or master students at the University of Hamburg from a variety of fields (40% majors or minors in economics, business, or a related field), and 51% were female.

Session Design, Instructions, and Payments In each session, the design of the experiment was common knowledge, and all subjects received the same instructions (a translation is provided in Appendix B). Sessions consisted of 30 periods. In addition to the random rematching of groups in each period, also the role assignments varied across periods as follows: Each subject who was assigned the role of employer in the first period retained this role throughout all 30 periods. All other subjects randomly switched roles across periods, either between employees and third parties or between prosecutors and third parties. This was communicated in the instructions, where we also stated that role assignments were independent of subjects' behavior. The aim of this re-shuffling was to make the negative consequences of misbehavior more salient; in particular to the employee and the prosecutor, whose decisions might curb the harm inflicted by the employer on the third party. In addition, in order to ensure that subjects indeed understood the game, after going through the instructions, subjects had to answer a series of control questions, and we discussed any wrong answers with them in private before launching the experiment. At the end of the experiment we asked subjects to complete a post-experimental questionnaire. Finally, to determine each subject's payment, three periods were

[21] While many whistleblower protection laws require firms to establish anonymous reporting channels, Chassang and Padró i Miquel (2019) argue that the protection offered by a formal requirement of anonymity might be limited in practice as in many cases the set of people informed about misbehavior will be small to begin with (and hence, the identity of the whistleblower can be conjectured).

randomly selected, and the subject's total points earned in these periods were converted at the rate of 1 Euro per 15 points. Together with the show-up fee, this was paid out (in private) in cash at the end of the session.

Framing In the experiment, we framed the game as an employer-employee relationship, where the employee could file a report to a prosecutor.[22] However, in the instructions (see Appendix B), we avoided the use of strongly judgemental terms such as "misbehavior", "illegal" or "whistleblowing". For example, in the experiment, we referred to an employer's misbehavior decision as a choice between two alternatives labelled "circle" (i.e., no misbehavior) and "triangle" (i.e., misbehavior). However, all subjects were informed that a (fictitious) law for the protection of the third party says that "triangle" should not be chosen as it harms the third party. Moreover, the employee's reporting decision was not referred to as whistleblowing, but as "asking the prosecutor to trigger an investigation".

4 Basic Comparisons: Treatments *NoP* and *P-R*

We start with the basic comparison between treatment *P-R*, where protection is granted for all whistleblowers who send a report, and the benchmark *NoP*, where protection is not available. We first present theoretical predictions and then discuss the experimental results.

4.1 Theoretical Predictions

In this section, we present theoretical predictions for treatments *NoP* and *P-R*. The underlying model is formally spelled out and analyzed in Appendix A.The structure of the game and the *monetary* payoffs of the players have already been described in Section 3.1 above. In addition, our model incorporates three *behavioral* motives, that are relevant in the context of whistleblowing, but which are not incentivized in the experiment.[23] First, with respect to whistleblowers, "conscience cleansing" has been shown to be an important driver of their reporting behavior.[24] Consequently, in our model employees suffer an idiosyncratic disutility

[22] In experimental economics, there is a discussion about the conditions under which a neutral or a loaded framing is more appropriate, see e.g., Eckel and Grossman (1996) and Alekseev, Charness, and Gneezy (2017). Framing is also discussed in other contexts involving misbehavior, e.g., in experiments on corruption (Abbink and Hennig-Schmidt, 2006; Barr and Serra, 2009) and tort litigation (Loewenstein, Issacharoff, Camerer, and Babcock, 1993; Babcock, Loewenstein, Issacharoff, and Camerer, 1995).

[23] Otherwise, players are assumed to have standard preferences. For a summary of the payoff functions of the players, see Table 2 in Appendix A.

[24] For example, see Jos, Tompkins, and Hays (1989), Miceli and Near (1992) and Alford (2001). Conscience cleansing is also a crucial motivation of whistleblowers in the model of Heyes and Kapur (2009).

from undetected misbehavior, which they can potentially avoid by sending a report. Second, as discussed in the Introduction, there is ample evidence that employers retaliate against unprotected whistleblowers. We incorporate this in the model by introducing an idiosyncratic disutility for employers that they incur when retaining a known whistleblower.[25] As a result, employers might want to retaliate against whistleblowers by dismissing them. Third, while misbehavior yields a fixed monetary gain to employers, we assume that they differ with respect to their "net benefit" from misbehavior. This allows us to capture moral concerns on the side of employers, and it gives rise to a distribution of net benefits from misbehavior as in the theoretical literature on law enforcement in the tradition of Becker (1968).

The theoretical predictions for treatments *NoP* and *P-R* are derived from the pure-strategy Perfect Bayesian Equilibria of the model (see Appendix A, in particular Propositions 1 and 2). We focus on *informative equilibria* where the prosecutor triggers an investigation if and only if the employee sends a report. This directly leads to

Prediction I (Investigation): *In both treatments, prosecutors trigger (do not trigger) an investigation upon receiving (not receiving) a report by the employee.*

Hence, every misbehavior that is reported is detected in the subsequent investigation. Turning to dismissals, the employer prefers to dismiss an L-employee whenever this is feasible, because the productivity of the outside replacement is higher. By contrast, an H-employee will only be dismissed upon reporting, and only if the employer's dislike of employing a whistleblower exceeds the H-employee's productivity advantage.[26] This leads to:

Prediction D (Dismissal): *In both treatments: (i) L-employees are dismissed, unless they are protected. (ii) H-employees are retained when sending no report, while they are dismissed with positive probability when sending a report and absent protection.*

We now turn to the reporting decision: As employees are assumed to suffer a disutility from undetected misbehavior, either productivity type is more willing to report when misbehavior

[25] Such heterogeneity is consistent with empirical findings. For example, Near and Miceli (1996, pp.517ff) find retaliation rates ranging from 6% to 38%, suggesting that employers do differ with respect to their attitude towards whistleblowing (see also the National Business Ethics Survey of 2013 available at https://www.ibe.org.uk/userassets/surveys/nbes2013.pdf).

[26] As can be shown, for none of the treatments the theoretical predictions depend on whether the reporting decision of the employee is observed by the prosecutor only or by both the prosecutor and the employer. Intuitively, this is driven by the fact that the employer can observe the investigation decision and by our focus on informative equilibria where the prosecutor investigates if and only if a report occurs.

actually has occurred. However, in anticipation of the subsequent investigation and dismissal decisions, the reporting behavior differs across types. The reason is that L-employees expect to be dismissed whenever feasible, while H-employees are less vulnerable due to their higher productivity. This gives L-employees a higher incentive to report irrespective of the underlying misbehavior decision. In fact, in treatment *P-R*, L-employees always have an incentive to report even when there is no misbehavior, as this protects them from dismissal. By contrast, H-employees whose employer misbehaves face a trade-off between the disutility from undetected misbehavior when not sending a report and the higher risk of dismissal when sending a report. This leads to:

Prediction R (Reporting): *Reporting behavior is summarized in the following table, where the entries indicate the fraction of reports in the respective condition, and where* $[0, 1]$ *indicates some value within this interval:*

Treatment	***NoP***		***P-R***	
Employee Type	L	H	L	H
Misbehavior	1	$[0, 1]$	1	1
No Misbehavior	0	0	1	0

In particular: (i) Reports are triggered by misbehavior, but there are also non-meritorious claims by L-employees in treatment P-R (i.e., a report is sent although there is no misbehavior). (ii) L-employees exhibit a (weakly) higher willingness to report than H-employees. (iii) In P-R, both employee types exhibit a (weakly) higher willingness to report relative to NoP.

Finally, we turn to the employer's misbehavior decision. When the employer faces an L-employee, the incentive to misbehave is identical in both treatments. The reason is twofold: First, in both treatments the L-employee would report any misbehavior (see *Prediction R*). Second, in neither of the two treatments the misbehavior decision affects the subsequent dismissal decision: The L-employee is always dismissed in treatment *NoP*, while she is always shielded from dismissal in treatment *P-R* (because she always sends a report). When the employer faces an H-employee, the incentive to misbehave is lower (and hence deterrence is higher) in treatment *P-R* than in *NoP*. This treatment difference is driven by the H-employee's higher willingness to report misbehavior in *P-R* (see again *Prediction R*).

Prediction M (Misbehavior): *Relative to NoP, there is less misbehavior in treatment P-R.*

This is driven by employers matched with H-employees, while there is no treatment difference for employers matched with L-employees.

To summarize, our model predicts that the introduction of whistleblower protection increases the willingness to report (both truthfully and non-meritoriously), and leads to more detection and deterrence of misbehavior.

4.2 Experimental Results for Treatments *NoP* and *P-R*

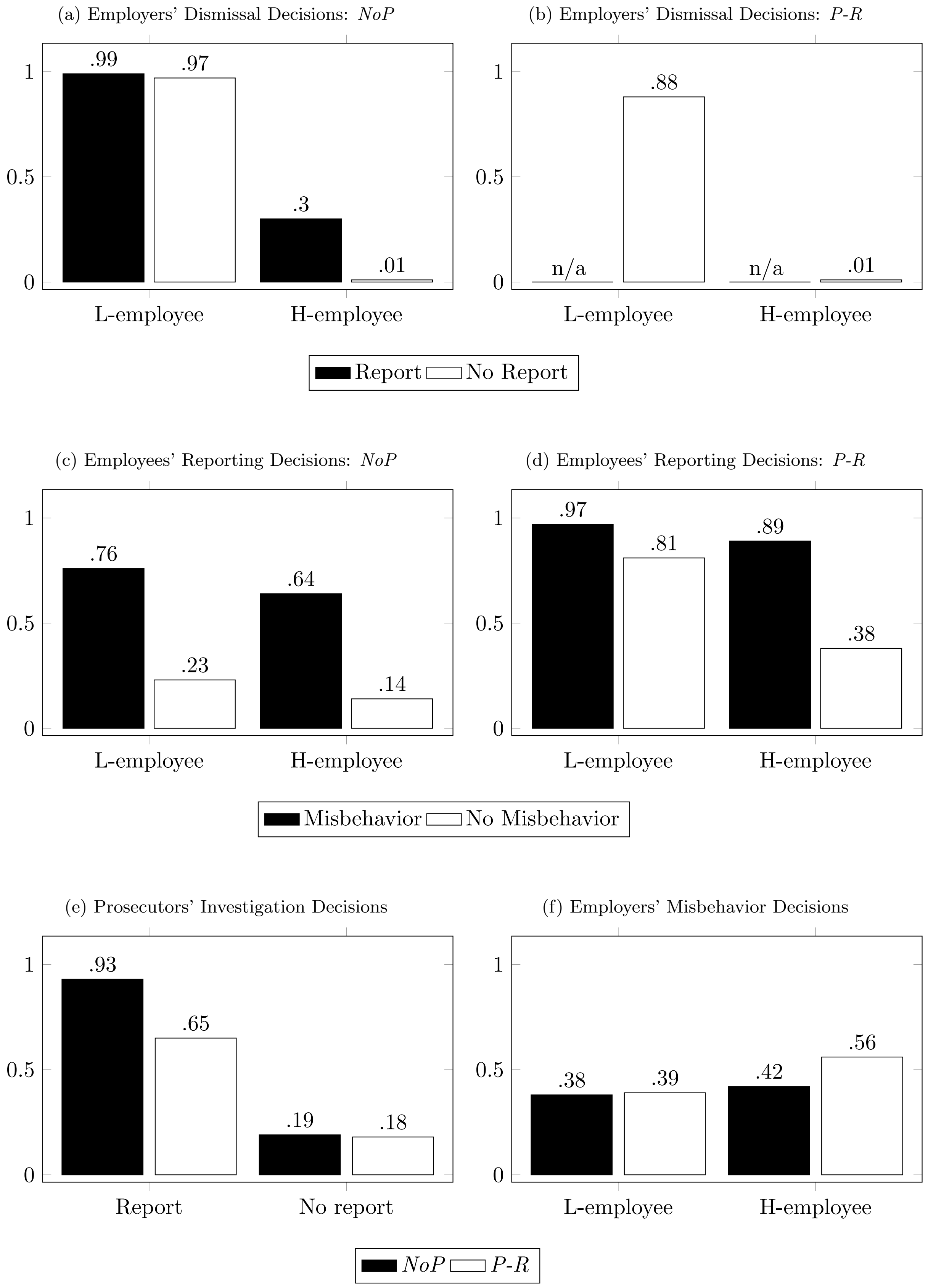
Based on the predictions, our experimental findings are illustrated in Figure 2. Systematic statistical testing is done using regression analysis, and the results are reported in Table 1. There, we estimate linear probability models with the respective underlying decision as the dependent variable. The unit of observation are individual decisions with standard errors clustered at the session level. For the sake of comparability, all subsequent regression tables have the same basic structure. In Table 1, we test all within- and between-treatment predictions of Section 4.1, but in the following discussion we focus on the key insights.[27] In a nutshell, we find that whistleblower protection indeed increases the willingness to report, but that the predicted beneficial effects on detection and deterrence do not materialize.

Employers' Dismissal Decisions: Testing Prediction D According to *Prediction D*, the dismissal decision depends on the respective employee's productivity type and reporting decision. The results are depicted in panels (a) and (b) of Figure 2, and the regression results are shown in columns (1)–(3) of Table 1.

All findings are fully supportive of *Prediction D*. First, in both treatments L-employees are virtually always dismissed when feasible. Second, also in both treatments, H-employees who do not report are almost always retained. Third, 30% of H-employee whistleblowers are indeed dismissed in treatment *NoP*, which is significantly more compared to the dismissal of non-reporting H-employees, and significantly less compared to the dismissal of L-employees who do report. The dismissal of H-employee whistleblowers supports our (behavioral) model feature that some employers retaliate even if this is costly to them due to a productivity loss.

[27] As illustrated in Figure 2, *Prediction I* leads to a total of four hypotheses (two within- and between-treatment comparisons, respectively). *Prediction D* leads to seven hypotheses (four within-treatment comparisons in *NoP*, one in *P-R*, and two between-treatment comparisons). *Prediction R* leads to 12 hypotheses (four within-treatment comparisons in each of *NoP* and *P-R*, and four between-treatment comparisons). Finally, *Prediction M* leads to two between-treatment hypotheses (note that there no within treatment-predictions comparing employee types). This leads a total of 25 hypotheses emerging from our predictions.

Figure 2: Experimental Behavior in Treatments *NoP* and *P-R* (in Fractions)



(a) Employers' Dismissal Decisions: *NoP*

(b) Employers' Dismissal Decisions: *P-R*

(c) Employees' Reporting Decisions: *NoP*

(d) Employees' Reporting Decisions: *P-R*

(e) Prosecutors' Investigation Decisions

(f) Employers' Misbehavior Decisions

Table 1: Regressions Results for Treatments *NoP* and *P-R*. Testing Predictions D, R, I, and M

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Dismiss	Dismiss	Dismiss	Report	Report	Report	Report	Investigate	Misbehave
	(*NoP*)	(*P-R*, No Rep.)	(No Rep.)	(*NoP*)	(*P-R*)	(H-emp.)	(L-emp.)		
H-employee	-0.963*** (0.000)	-0.871*** (0.000)	-0.963*** (0.000)	-0.0953* (0.097)	-0.431*** (0.001)				0.0406 (0.311)
Report	0.0171 (0.257)							0.745*** (0.000)	
Report x H-emp.	0.274** (0.040)								
P-R			-0.0959 (0.186)			0.248*** (0.003)	0.584*** (0.000)	-0.00403 (0.950)	0.00973 (0.854)
P-R x H-emp.			0.0913 (0.201)						0.126 (0.167)
Misbehavior				0.526*** (0.001)	0.154** (0.024)	0.500*** (0.000)	0.526*** (0.000)		
Misb. x H-emp.				-0.0259 (0.450)	0.357** (0.016)				
Misb. x P-R						0.0112 (0.918)	-0.372*** (0.000)		
P-R x Report								-0.276** (0.013)	
Observations	900	227	774	1800	1800	1854	1746	1800	1800
Adjusted R^2	0.805	0.819	0.908	0.278	0.290	0.319	0.361	0.349	0.020
F-Test1	0.001		0.000	0.049	0.311	0.004	0.007	0.000	0.053
F-Test2	0.022		0.610	0.000	0.008	0.001	0.005	0.001	0.122

Notes: Each column refers to a linear probability model with the respective underlying decision (dismissal, reporting, investigation, misbehavior) as the dependent variable. All regressions use individual observations with standard errors clustered at the session level, p-values are reported in parentheses, and *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. When a qualifier is stated in parenthesis in the header of a regression, it refers to the subset of observations used; for example, "H-emp." indicates that only H-employees are considered, and "No Rep." indicates that only observations where no report was sent are considered. When no such qualifier is stated, the regression uses all observations for the respective decision from both treatments. The entries at the bottom of each column show the p-values of the F-tests for the joint significance of the interaction term with the first regressor (*F-Test1*) and of the interaction term with the second regressor (*F-Test2*).

Employees' Reporting Decisions: Testing Prediction R According to *Prediction R*, the reporting decision of employees depends on their productivity type and the misbehavior decision of their employer. Our findings are illustrated in panels (c) and (d) of Figure 2, and the regression results are shown in columns (4)–(7) of Table 1. First, in both treatments, the reporting rates are significantly higher when there is misbehavior, which is consistent with the "conscience cleansing" motive as included in our theoretical framework. This motive emerges most cleanly for H-employees in *NoP*, where sending a report is costly in the sense of significantly increasing the probability of dismissal. Second, L-employees send more reports than H-employees. These findings are broadly in line with *Predictions R(i) and (ii)*.[28]

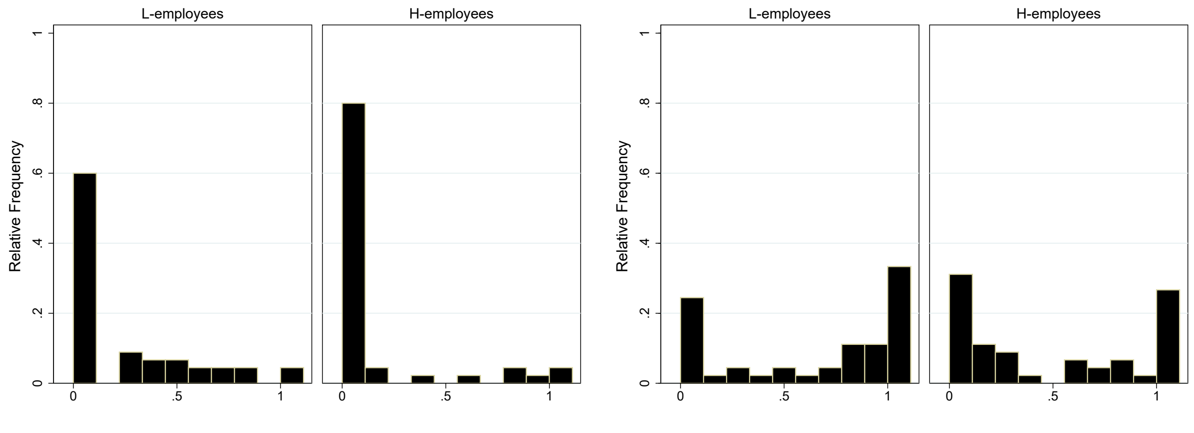
Prediction R(iii) captures treatment effects. Comparing the reporting behavior across treatments, we find that reporting is significantly higher in treatment *P-R* for all four combinations

[28] One exception is the behavior of L-employees in *P-R*, which was not predicted to depend on whether or not there is misbehavior. Another exception is the reporting of misbehavior in *P-R*, where misbehavior is predicted to always be reported by either productivity type.

of productivity type and employer misbehavior (see columns (6) and (7) of Table 1). This is in line with the predictions for truthful reports of H-employees and non-meritorious reports of L-employees, while for the other two combinations no treatment difference was predicted. Overall, it can be seen that a whistleblower protection scheme such as *P-R* leads to a pronounced increase of reporting of misbehavior.

The downside is that also the fraction of non-meritorious claims rises in treatment *P-R*, in particular by L-employees, which is in line with *Prediction R(i).* We also observe an unpredicted and non-negligible amount of non-meritorious claims by both employee types in *NoP*, and by H-employees in *P-R*. To investigate this issue further, we consider the subject-specific fractions of non-meritorious claims: In particular, over the course of a given session, each subject observed in the employee role makes multiple reporting decisions. Consequently, for each subject we determine the fraction of non-meritorious claims. Thereby, a value of 0 (1) indicates a subject who never (always) sends a non-meritorious claim. Figure 3 shows the histograms of these subject-specific fractions split by treatment and productivity type. As can be seen, the predicted reporting behavior is always the modal behavior in the experiment. In *NoP*, this is true by a large margin. However, in *P-R*, there is, in addition, a certain number of subjects who always act against the prediction, i.e., in the role of an L-employee (H-employee), they never (always) send a non-meritorious claim.

Figure 3: Histograms of Subject-Specific Fractions of Non-Meritorious Claims


(a) *NoP*

(b) *P-R*

Notes: In a given session, each subject observed in the employee role takes multiple reporting decisions. This leads to subject-specific fractions of non-meritorious claims (i.e., the cases where the given subject sends a report although there is no misbehavior). These subject-specific fractions are shown on the horizontal axis, where a value of 0 (1) indicates subjects who never (always) send a non-meritorious claim.

Detection of Misbehavior by Prosecutors: Testing Prediction I Recall that when deciding on whether or not to investigate, the prosecutor only observes whether or not a report was sent by the employee, but neither the employee's productivity type nor the underlying misbehavior decision. Figure 2(e) illustrates the experimental findings, and the regression results are shown in Table 1, column (8). Overall, prosecutors indeed seem to perceive reports as informative with respect to the presence of misbehavior, since in both treatments the number of investigations is significantly higher when a report occurs. In *NoP*, *Prediction I* is broadly confirmed: The fraction of investigations following a report (no report) is 0.93 (0.19), and hence close to the predicted values of 1 (0).

However, in *P-R*, we observe an interesting deviation in the sense that protection seems to reduce prosecutors' responsiveness to reports. In particular, with a value of 0.65 the fraction of investigations conditional on a report is well below the predicted value of 1, and significantly lower compared to *NoP*. We will revisit prosecutor behavior in more detail in Section 5 below.

With respect to the detection of misbehavior, recall that in an informative equilibrium, every report triggers an investigation. Therefore, our theory predicts that whenever misbehavior is more likely to be reported, it is also more likely to be detected in the course of an investigation. However, while protection indeed improves the willingness of employees to report misbehavior in the experiment, the predicted higher detection rate of misbehavior fails to materialize due to the prosecutors' lower responsiveness to reports. In particular, we find that in *P-R*, 38 percent of all actual misbehavior remains undetected, as opposed to 29 percent in *NoP*.[29]

Employers' Decisions to Misbehave: Testing Prediction M According to *Prediction M*, protection leads to less misbehavior, and this effect is driven by employers matched with H-employees. Figure 2(f) displays the fractions of employers who chose to misbehave, and the respective regression results are shown in Table 1, column (9). The experimental results strongly support the prediction of no treatment effect for employers with L-employees. However, the predicted reduction of misbehavior for employers with H-employees fails to materialize, as there is no significant treatment difference. If anything, there seems to be more (rather than less) misbehavior in *P-R*.

[29] The respective numbers are 18 and 12 percent when relating undetected actual misbehavior to all *potential* cases of misbehavior (i.e., the total number of employer decisions for or against misbehavior).

Summary and Robustness To summarize our experimental findings for the baseline comparison of treatments *NoP* and *P-R*, whistleblower protection indeed achieves the goal of inducing a higher willingness to report misbehavior. However, because of the lower responsiveness of prosecutors to reports, the predicted improvement in terms of detection of misbehavior does not materialize, and there is no effect on deterrence. As shown in Appendix C, these findings are robust (i) under alternative statistical specifications (i.e., alternative unit of observation, non-parametric testing, controlling for multiple-hypothesis testing), (ii) when focussing on the last ten periods of each session, and (iii) when including personal characteristics as additional controls. In Appendix C, we also show that the elicited personal characteristics do not seem to have systematic effects on behavior.

5 A Closer Look at Prosecutor Behavior: *P-RI* and *P-RIM*

One question arising from the findings of Section 4 is whether the lower responsiveness of prosecutors to reports in *P-R* is solely driven by the presence of non-meritorious claims. Such reports do seem to play a role due to their higher frequency in *P-R*, which makes reports less informative about underlying misbehavior compared to *NoP*. In fact, if there is a report, the empirical frequency of underlying misbehavior is 0.71 in *NoP*, but only 0.58 in *P-R*.[30] However, this does not seem to be the full explanation for the lower responsiveness of prosecutors, as suggested by the analysis of two additional treatments *P-RI* and *P-RIM*.

The treatment comparison between *P-RI* and *P-RIM* allows to gauge the effect of non-meritorious claims on prosecutor behavior, because in *P-RIM* all incentives for non-meritorious claims are removed. In particular, in *P-RI*, whistleblower protection requires not only a report, but also a subsequent investigation by the prosecutor. Treatment *P-RIM* has the additional requirement for protection that the investigation indeed needs to uncover misbehavior. Otherwise, *P-RI* and *P-RIM* are identical.[31] In both of these treatments, the employer learns the reporting decision if there is an investigation. For this reason, *P-RI* and *P-R* (where the employer always learns the reporting decision) differ along a second dimension. However, the theoretical predictions for *P-RI* and *P-R* coincide (because in informative equilibrium, every

[30] When there is no report, the empirical frequencies of underlying misbehavior are 0.20 (*NoP*) and 0.19 (*P-R*), and hence there is no treatment difference.

[31] We also ran two additional treatments as variants of *P-RI* and *P-RIM*, which introduce a cost for employers whenever an investigation occurs, and investigation errors, respectively. For these two treatments, the interested reader is referred to an earlier version of the paper (Mechtenberg, Muehlheusser, and Roider, 2017).

report triggers an investigation), and this is also borne out in the experiment.[32] For *P-RIM*, theory predicts that all misbehavior is truthfully reported (as in *P-RI*), while all incentives for non-meritorious claims are removed (which is in contrast to *P-RI*).[33] The reason is that in *P-RIM* protection can only be obtained in the presence of misbehavior.

The experimental results are shown in Figure 4 below, and in Table 9 in Appendix D. Comparing panels (a) and (b) of Figure 4 reveals that the predictions on reporting behavior are borne out in the experiment. There are no treatment differences for truthful reporting which occurs with high frequency in both treatments. However, in *P-RIM* there are sharp and significant drops in the number of non-meritorious claims. Importantly, as shown in panels (c) and (c) of Figure 4, while the lower number of non-meritorious claims strongly improves the informativeness of reports in *P-RIM* relative to *P-RI*, this does not lead to significantly more investigations by prosecutors upon a report.[34] As a result, the detection rate of misbehavior does not improve in *P-RIM*. In fact, while in *P-RIM* the informativeness of reports is even higher than in the no-protection treatment *NoP*, the responsiveness of prosecutors to reports is significantly lower.[35]

To summarize, non-meritorious claims cannot fully explain the lower responsiveness of prosecutors to reports in the presence of whistleblower protection. It could be that prosecutors underestimate the information content of reports when whistleblower protection is in place, even when non-meritorious claims only play a minor role.

Note that treatment *P-RIM* itself is not meant as a policy recommendation, because it abstracts from a number of important real-world aspects. For example, in this treatment the merit of a claim is revealed immediately in the course of an investigation. In practice, however, in constituencies where protection is only granted after the presence of misbehavior has been

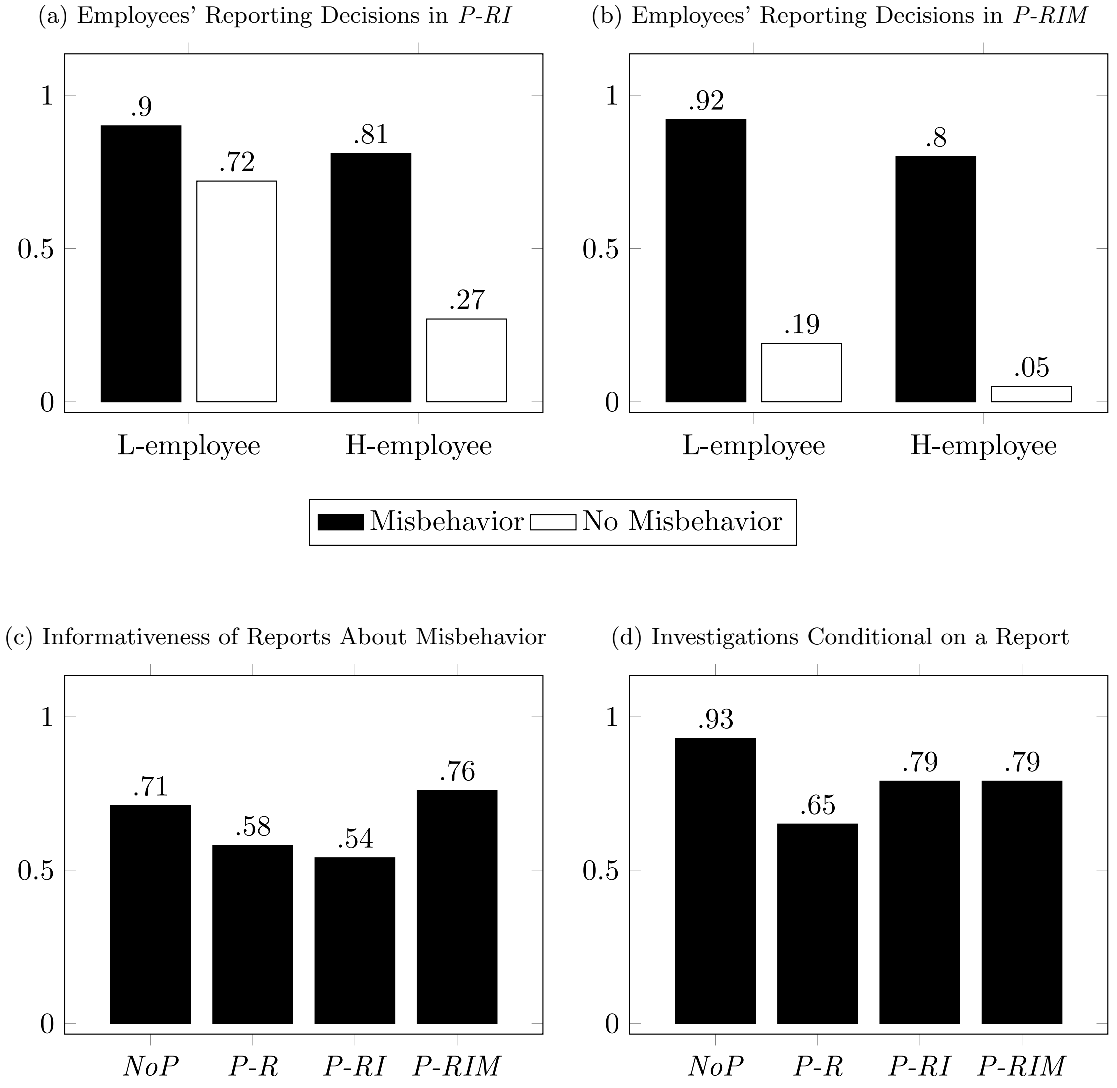
[32] Note that there is only one case in which treatments *P-R* and *P-RI* would have different implications: the employee sends a report, but the prosecutor does not investigate, which would lead to protection in *P-R*, but not in *P-RI*. However, this case does not occur on the equilibrium path. Experimental behavior in these two treatments is indeed very similar (see Table 8 in Appendix D): With one exception there are no statistically significant treatment effects. The exception is the frequency of investigations conditional on a report, which is significantly higher in *P-RI*, but still well below one (see Figure 4(d)).

[33] The theoretical predictions for *P-RIM* are derived in a completely analogous way to those for *P-R*. Details are available upon request.

[34] Proceeding analogously as outlined in Appendix C.2, we have also checked whether this effect differs between early and late periods. This is not the case.

[35] The regression results for the comparison of *NoP* and *P-RIM* are shown in Table 10, where in column (8) the relevant "F-Test2" reveals significance at the 1%-level. Note that dismissals in *P-RIM* follow the same pattern as in the other treatments. In particular, unprotected L-employees are virtually always dismissed, while non-reporting H-employees are basically always retained. The incentive to misbehave is smaller in *P-RIM* compared to *P-RI*, which is also borne out in the experiment (see Table 9, column (9) in the Appendix).

Figure 4: Reporting, Informativeness of Reports About Misbehavior, and Investigations


(a) Employees' Reporting Decisions in *P-RI*

(b) Employees' Reporting Decisions in *P-RIM*

(c) Informativeness of Reports About Misbehavior

(d) Investigations Conditional on a Report

established in court, this often leads to long waiting times for whistleblowers.[36]

6 A Richer Incentive Structure: ***NoP-Low*** and ***P-R-Low***

In this section, we analyze the robustness of our findings for the basic treatments *NoP* and *P-R* as discussed in Section 4. Recall that in *NoP* and *P-R*, employers had a strong incentive to

[36] For example, this is highlighted by the *Heinisch v. Germany* case, where several German courts had refused to reverse the dismissal of a whistleblower (a geriatric nurse who had truthfully reported misbehavior by her employer) before protection was eventually affirmed by the European Court of Human Rights (see for example the discussion in Thuesing and Forst, 2016, pp. 12).

dismiss L-employees, irrespective of whether or not they are whistleblowers, because they are less productive compared to the outside replacement. In treatment *P-R*, this created a strong incentive for L-employees to seek employment protection by lodging a report, irrespective of whether or not there was actually misbehavior by the employer.

We now remove the productivity-based rationale for the dismissal of L-employees. This leads to a richer incentive structure for the dismissal and reporting decisions. In particular, we have run two additional treatments *NoP-Low* and *P-R-Low*, in which the productivity of the outside replacement is reduced to match that of the L-employee.[37] Otherwise, *NoP-Low* and *P-R-Low* are identical to *NoP* and *P-R*, respectively. In a nutshell, the main effects of whistleblower protection on reporting behavior and the detection and deterrence of misbehavior remain robust, but there are also additional findings.

From a theoretical point of view, the employer still suffers a disutility when retaining a whistleblower. However, an L-employee who does not report now generates exactly the same payoff for the employer as the outside replacement. Ceteris paribus, this should reduce employers' incentive to dismiss L-employees. In turn, this also affects the reporting incentives of L-employees (while the prediction for the H-employee case is unaffected). First, in *NoP-Low*, L-employees have a lower incentive to report misbehavior than in *NoP* (as remaining silent now increases their chance of retention). Second, in *P-R-Low*, L-employees have a weaker incentive to lodge a non-meritorious claim than in *P-R*.[38]

The experimental results for the additional treatments *NoP-Low* and *P-R-Low* are illustrated in Figure 5 and hence can be compared to those for *NoP* and *P-R* (see Figure 2 above). The full sets of regression results for the three pair-wise treatment comparisons (*NoP-Low* and *P-R-Low*, *NoP-Low* and *NoP*, *P-R-Low* and *P-R*) are relegated to Tables 11 – 13 in Appendix E. In the following, we focus on a discussion of the key findings.

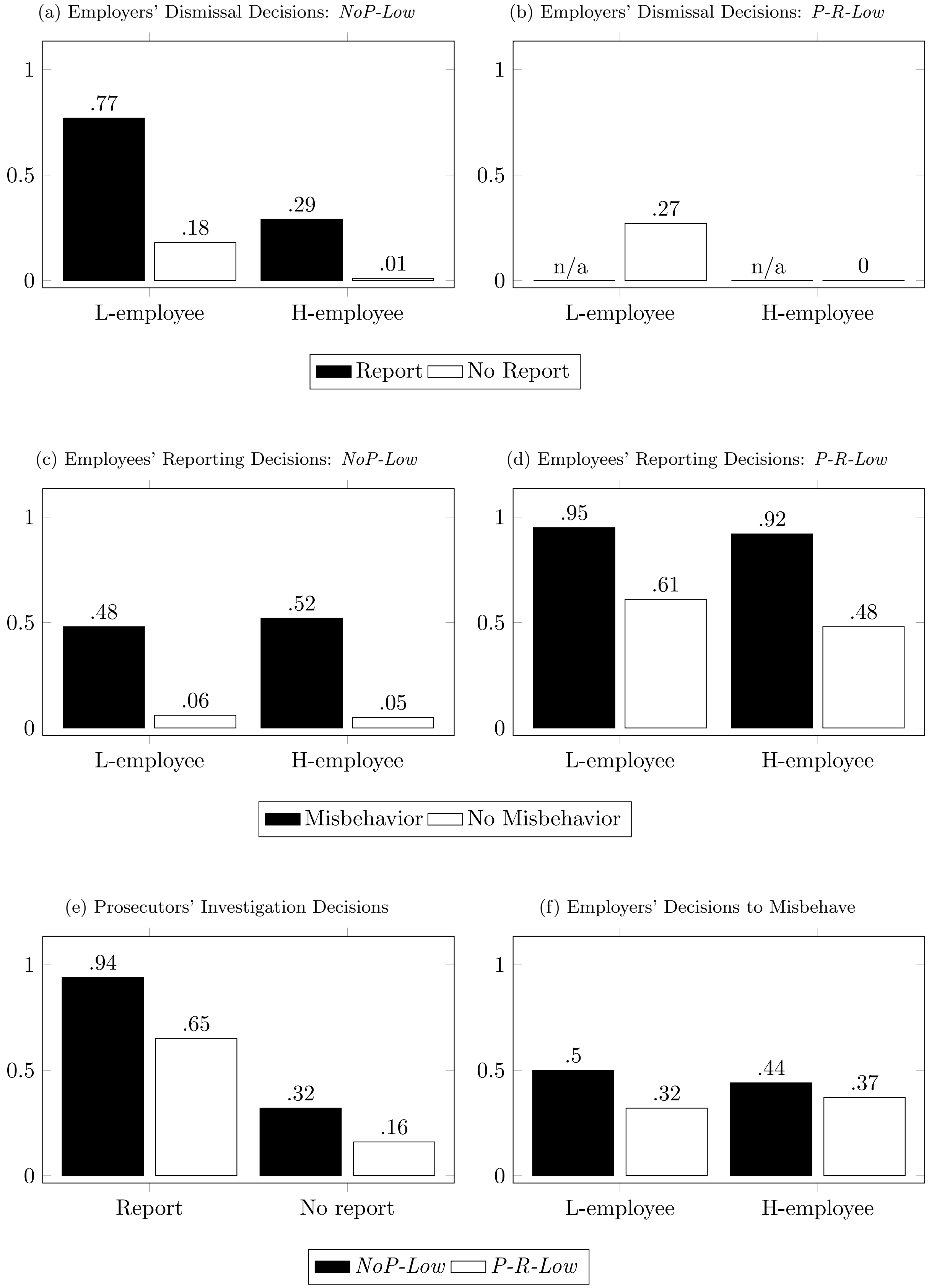
First, consider dismissal decisions (see panels (a) and (b) of Figures 5 and 2). As expected, L-employees are indeed substantially and significantly less often dismissed in *NoP-Low* and *P-R-Low*, compared to *NoP* and *P-R*,[39] and their probability of retention now strongly depends on

[37] That is, in the experiment we reduce the productivity of the outside replacement from 70 to 30.

[38] Deriving the informative-equilibrium predictions for *NoP-Low* and *P-R-Low* is analogous to treatments *NoP* and *P-R*. Details are available upon request.

[39] In particular, the dismissal rate of non-reporting L-employees is 80 percentage points lower in *NoP-Low* compared to *NoP*, and even for whistleblowers we find a sizeable reduction by 20 percentage points. Similarly, the dismissal rate of L-employees is 60 percentage points lower in *P-R-Low* compared *P-R*. Moreover, and also as expected, the fractions of dismissed H-employees are comparable in both labor-market settings.

Figure 5: Results for Treatments *NoP-Low* and *P-R-Low*



(a) Employers' Dismissal Decisions: *NoP-Low*

(b) Employers' Dismissal Decisions: *P-R-Low*

(c) Employees' Reporting Decisions: *NoP-Low*

(d) Employees' Reporting Decisions: *P-R-Low*

(e) Prosecutors' Investigation Decisions

(f) Employers' Decisions to Misbehave

the reporting decision: In the absence of protection, remaining silent increases an L-employee's likelihood of retention by 59 percentage points in *NoP-Low*, as compared to just 2 percentage points in *NoP*.

Next, consider reporting behavior (see panels (c) and (d) of Figures 5 and 2). Compared to *NoP*, in *NoP-Low* we find a substantial and significant reduction in the number of truthful and non-meritorious claims, which are now virtually eliminated. We observe a similar pattern for H-employees.[40] As for the effect of whistleblower protection, in *P-R-Low* we observe a doubling of the reporting rate of misbehavior relative to *NoP-Low*. This is a stronger effect than for the baseline comparison between *NoP* and *P-R*. The introduction of protection has a comparable effect on non-meritorious claims in both labor-market settings.

As for investigations, comparing Figures 5(e) and 2(e) reveals that prosecutor behavior is very similar in both labor-market settings. In particular, the introduction of whistleblower protection again leads to a significant (and virtually identical) reduction in prosecutors' responsiveness to reports. Finally, as for deterrence, in the new labor-market setting the introduction of whistleblower protection has a (mild) positive effect on deterrence (see Figures 5(f) and 2(f), where only the effect for employers with L-employees is statistically significant). This is in contrast to the baseline comparisons, where protection had no significant effect on deterrence.

To summarize, while this alternative labor-market setting provides a richer incentive structure with respect to the dismissal and reporting decisions, the effects of whistleblower protection are qualitatively very similar to those in the baseline comparison of treatments *NoP* and *P-R*. In particular, protection has a pronounced effect on the reporting of misbehavior. However, under protection, prosecutors are again less responsive to reports, which hampers the detection of misbehavior, while this time there is some positive effect on deterrence. These findings re-iterate the importance of taking into account behavioral responses of prosecutors when introducing whistleblower protection.

A common feature of Sections 4 and 6 is the rather high level of non-meritorious claims when protection is available. This could result from our design choice that an employee's payoff difference between the case of guaranteed retention and dismissal is positive (where the respective payoffs are the employee's wage and zero, respectively). Of course, in practice, for

[40] By contrast, for the case of protection, there are no significant treatment differences between *P-R* and *P-R-Low*, except that in *P-R-Low*, non-meritorious claims by L-employees decrease by 20 percentage points. Note that there is no statistically significant effect across treatments for non-meritorious claims by H-employees (see Table 13, column (6)).

many employees sending a non-meritorious claim will impose costs that might exceed the benefit from a claim (e.g., due to features not captured in our setup such as lower future employment prospects or social sanctioning). Hence, in real-world settings the number of non-meritorious claims is likely to be lower than in the experiment. Nevertheless, based on the discussion in the Introduction, there seem to exist employees for which obtaining employment protection is the dominating motive (e.g., due to a lack of outside employment opportunities). Of course, how prevalent such employees are is ultimately an empirical question (which would be, however, very difficult to answer with field data where non-meritorious claims are typically unobserved). We have not considered treatments where an employee's net gain from sending a non-meritorious claim is negative (i.e., settings where any benefits are outweighed by potential costs). However, in treatment *P-RIM* of Section 5, this net gain is zero, and there the number of non-meritorious claims is indeed low (see Figure 4). While this drives up the informativeness of reports, the responsiveness of prosecutors to reports when protection is available is still relatively low.

7 Conclusion

The fight against corporate fraud looms high on the policy agenda in many countries and international bodies. Because of their access to crucial information, employee whistleblowers potentially play an important role in uncovering corporate fraud. However, whistleblowers often face retaliation which deters employees from coming forward. As a consequence, whistleblower protection laws have been enacted in recent years to protect whistleblowers from retaliation. In addition, these laws intend to foster employees' willingness to report fraud, thereby also improving the detection and deterrence of illegal activities. This paper contributes to a growing empirical and experimental literature on corporate fraud and employee whistleblowing.

To this end, we conduct a theory-guided laboratory experiment that allows us to observe information (such as undetected misbehavior) usually not observed in the field. In our setup, employees (as potential whistleblowers) interact with employers (as potential wrong-doers) and prosecutors (who may investigate the allegations of whistleblowers against their employers). We consider various treatments with and without protection.

Our main experimental finding is that whistleblower protection indeed accomplishes the aim of enhancing the reporting of misbehavior, while the desired (and predicted) positive effects on detection and deterrence of misbehavior do not materialize. One crucial factor for these latter observations is a lower responsiveness of prosecutors to reports when protection is available. Interestingly, this finding (which is robust across all four treatments with protection) does not seem to be driven by non-meritorious claims alone. One potential explanation for this phenomenon is that prosecutors underestimate the informativeness of reports.

Given the empirical importance of corporate fraud and in order to make the best use of reports, designing effective whistleblower-protection policies is a crucial, but still under-researched topic. One conclusion from our analysis is that in this endeavour the behavioral responses of all parties involved should to be taken into account. For example, our findings suggest that whistleblower-protection policies could be even more effective if reports induced more investigations. However, in our view, more research is needed before robust policy implications can be drawn from this observation. Future research should also explicitly consider other potentially important policy variables, such as the standard of proof for triggering an investigation or (financial) capacity constraints of investigation agencies. To the best of our knowledge, these issues have so far not been analyzed systematically.

References

Abbink, K. and H. Hennig-Schmidt (2006): "Neutral Versus Loaded Instructions in a Bribery Experiment," *Experimental Economics*, 9, 103–121.

Abeler, J., D. Nosenzo, and C. Raymond (2019): "Preferences for Truth-Telling," *Econometrica*, 87, 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.

Anechiarico, F. and J. B. Jacobs (1996): *The Pursuit of Absolute Integrity*, University of Chicago Press, Chicago IL.

Apesteguia, J., M. Dufwenberg, and R. Selten (2007): "Blowing the Whistle," *Economic Theory*, 31, 143–166.

Association of Certified Fraud Examiners (2014): *Report to the Nations on Occupational Fraud and Abuse: 2014 Global Fraud Study*, http://www.acfe.com/rttn/docs/2014-report-to-nations.pdf.

Babcock, L., G. Loewenstein, S. Issacharoff, and C. Camerer (1995): "Biased Judgments of Fairness in Bargaining," *American Economic Review*, 85, 1337–1343.

Barr, A. and D. Serra (2009): "The Effects of Externalities and Framing on Bribery in a Petty Corruption Experiment," *Experimental Economics*, 12, 488–503.

Bartuli, J., B. Djawadi, and R. Fahr (2016): "Business Ethics in Organizations: An Experimental Examination of Whistleblowing and Personality," *IZA Discussion Paper No. 10190.*

Becker, G. (1968): "Crime and Punishment: An Economic Approach," *Journal of Political Economy*, 76, 169–217.

Benabou, R. and J. Tirole (2003): "Intrinsic and Extrinsic Motivation," *Review of Economic Studies*, 70, 489–520.

Bénabou, R. and J. Tirole (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96, 1652–1678.

Benoît, J. and J. Dubra (2004): "Why Do Good Cops Defend Bad Cops?" *International Economic Review*, 45, 787–809.

Berth, H. and S. Goldschmidt (2006): "NEO-PI-R. NEO-Persönlichkeitsinventar nach Costa und McCrae," *Diagnostica*, 52, 95–99.

Blount, J. and S. Markel (2012): "The End of the Internal Compliance World as we Know it, or an Enhancement of the Effectiveness of Securities Law Enforcement-Bounty Hunting under the Dodd-Frank Act's Whistleblower Provisions," *Fordham Journal of Corporate & Financial Law*, 17, 1023–1061.

Bock, O., I. Baetge, and A. Nicklisch (2014): "hroot: Hamburg Registration and Organization Online Tool," *European Economic Review*, 71, 117–120.

Bowen, R., A. Call, and S. Rajgopal (2010): "Whistle-Blowing: Target Firm Characteristics and Economic Consequences," *Accounting Review*, 85, 1239–1271.

Butler, J., D. Serra, and G. Spagnolo (2019): "Motivating Whistleblowers," *Management Science, forthcoming.*

Callahan, E. and T. Dworkin (1992): "Do Good and Get Rich: Financial Incentives for Whistleblowing and the False Claims Act," *Villanova Law Review*, 37, 273.

Casey, A. J. and A. Niblett (2014): "Noise Reduction: The Screening Value of Qui Tam," *Washington University Law Review*, 91, 1169–1217.

Chassang, S. and G. Padró i Miquel (2019): "Crime, Intimidation and Whistleblowing: A Theory of Inference from Unverifiable Reports," *Review of Economic Studies*, forthcoming.

Costa, P. and R. McCrae (1992): *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory. Professional Manual*, Psychological Assessment Resources, Odessa, FL.

Cotten, S. and R. Santore (2016): "Whistleblowers, Amnesty, and Managerial Fraud: An Experimental Investigation," *University of Tennessee, mimeo.*

Council of Europe (2014): "Recommendation CM/Rec(2014)7 of the Committee of Ministers to Member States on the Protection of Whistleblowers (Adopted by the Committee of Ministers on 30 April 2014)," *https://rm.coe.int/16807096c7.*

Crawford, V. and J. Sobel (1982): "Strategic Information Transmission," *Econometrica*, 50, 1431–1451.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. Wagner (2011): "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences," *Journal of the European Economic Association*, 9, 522–550.

Dyck, A., A. Morse, and L. Zingales (2010): "Who Blows the Whistle on Corporate Fraud?" *Journal of Finance*, 65, 2213–2253.

——— (2014): "How Pervasive is Corporate Fraud?" *Chicago Booth School of Business, mimeo.*

Ebersole, D. (2011): "Blowing the Whistle on the Dodd-Frank Whistleblower Provisions," *Ohio State Entrepreneurial Business Law Journal*, 6, 123–174.

Eckel, C. and P. Grossman (1996): "Altruism in Anonymous Dictator Games," *Games and Economic Behavior*, 16, 181–191.

Farrar, J., C. Hausserman, and M. Rennie (2019): "The Influence of Revenge and Financial Rewards on Tax Fraud Reporting Intentions," *Journal of Economic Psychology*, 71, 102–116.

Feltovich, N. and Y. Hamaguchi (2018): "The Effect of Whistle-Blowing Incentives on Collusion: An Experimental Study of Leniency Programs," *Southern Economic Journal*, 84, 1024–1049.

Fiorin, S. (2019): "Reporting Peers' Wrongdoing: Experimental Evidence on the Effect of Financial Incentives on Morally Controversial Behavior," *University of California San Diego, mimeo.*

Fischbacher, U. (2007): "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments," *Experimental Economics*, 10, 171–178.

Fleischer, H. and K. U. Schmolke (2012): "Financial Incentives for Whistleblowers in European Capital Markets Law," *European Company Law*, 9, 250–259.

Frederick, S. (2005): "Cognitive Reflection and Decision Making," *Journal of Economic Perspectives*, 19, 25–42.

Givati, Y. (2016): "A Theory of Whistleblower Rewards," *The Journal of Legal Studies*, 45, 43–72.

Gneezy, U., S. Meier, and P. Rey-Biel (2011): "When and Why Incentives (Don't) Work to Modify Behavior," *Journal of Economic Perspectives*, 191–209.

Gobert, J. and M. Punch (2000): "Whistleblowers, the Public Interest, and the Public Interest Disclosure Act 1998," *The Modern Law Review*, 63, 25–54.

Hansberry, H. L. (2012): "In Spite of its Good Intentions, the Dodd-Frank Act has Created an FCPA Monster," *The Journal of Criminal Law and Criminology (1973-)*, 102, 195–226.

Hartmann, L. M. (2011): "Whistle While You Work: The Fairytale-Like Whistleblower Provisions of the Dodd-Frank Act and the Emergence of Greedy, the Eighth Dwarf," *Mercer Law Review*, 62, 1279–1313.

Healy, P. and K. Palepu (2003): "The Fall of Enron," *Journal of Economic Perspectives*, 17, 3–26.

Heyes, A. and S. Kapur (2009): "An Economic Model of Whistle-Blower Policy," *Journal of Law, Economics & Organization*, 25, 157–182.

Hinloopen, J. and A. Soetevent (2008): "Laboratory Evidence on the Effectiveness of Corporate Leniency Programs," *RAND Journal of Economics*, 39, 607–616.

Holm, S. (1979): "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 65–70.

Howse, R. and R. Daniels (1995): "Rewarding Whistleblowers: The Costs and Benefits of an Incentive-Based Compliance Strategy," in *Corporate Decisionmaking in Canada*, ed. by R. Daniels and R. Morck, Calgery: University of Calgary Press.

Jos, P., M. Tompkins, and S. Hays (1989): "In Praise of Difficult People: A Portrait of the Committed Whistleblower," *Public Administration Review*, 49, 552–61.

Kohn, S., M. Kohn, and D. Colapinto (2004): *Whistleblower Law: A Guide to Legal Protections for Corporate Employees*, Praeger Publishers.

Kroll (2016): *Global Fraud Report: Vulnerability on the Rise*, http://www.kroll.com/en-us/global-fraud-report.

List, J. A., A. M. Shaikh, and Y. Xu (2016): "Multiple Hypothesis Testing in Experimental Economics," *Experimental Economics*, 1–21.

Loewenstein, G., S. Issacharoff, C. Camerer, and L. Babcock (1993): "Self-Serving Assessments of Fairness and Pretrial Bargaining," *The Journal of Legal Studies*, 22, 135–159.

Mechtenberg, L., G. Muehlheusser, and A. Roider (2017): "Whistle-Blower Protection: Theory and Experimental Evidence," *CEPR Working Paper No. 11898.*

Mesmer-Magnus, J. and C. Viswesvaran (2005): "Whistleblowing in Organizations: An Examination of Correlates of Whistleblowing Intentions, Actions, and Retaliation," *Journal of Business Ethics*, 62, 277–297.

Miceli, M., T. Dworkin, and J. Near (2008): *Whistle-Blowing in Organizations*, Routledge.

Miceli, M. and J. Near (1992): *Blowing the Whistle: The Organizational and Legal Implications for Companies and Employees*, Lexington Books.

Miceli, M. P., J. P. Near, and T. M. Dworkin (2009): "A Word to the Wise: How Managers and Policy-Makers can Encourage Employees to Report Wrongdoing," *Journal of Business Ethics*, 86, 379–396.

Muehlheusser, G. and A. Roider (2008): "Black Sheep and Walls of Silence," *Journal of Economic Behavior & Organization*, 65, 387–408.

Near, J. and M. Miceli (1986): "Retaliation Against Whistle Blowers: Predictors and Effects." *Journal of Applied Psychology*, 71, 137.

——— (1996): "Whistle-blowing: Myth and Reality," *Journal of Management*, 22, 507–526.

OECD (2011): "G20 Anti-Corruption Action Plan: Protection of Whistleblowers," *https://www.oecd.org/g20/topics/anti-corruption/48972967.pdf.*

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

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.

Romano, J. P., A. M. Shaikh, and M. Wolf (2010): "Hypothesis Testing in Econometrics," *Annual Review Economics*, 2, 75–104.

Rose, A. M. (2014): "Better Bounty Hunting: How the SEC's New Whistleblower Program Changes the Securities Fraud Class Action Debate," *Northwestern University Law Review*, 108, 1235–121302.

Schmidt, M. (2005): "Whistle-Blowing Regulation and Accounting Standards Enforcement in Germany and Europe: An Economic Perspective," *International Review of Law and Economics*, 25, 143–168.

Schmolke, K. and V. Utikal (2016): "Whistleblowing: Incentives and Situational Determinants," *FAU Discussion Papers in Economics No. 9/16.*

Thüsing, G. and G. Forst (eds.) (2016): *Whistleblowing - A Comparative Study*, Springer.

USA Today (2004): "Whistleblower Complaints Are Up, But Why?" *November 21 Issue*, http://usat.ly/1LitrYG.

Vadera, A., R. Aguilera, and B. Caza (2009): "Making Sense of Whistle-Blowing's Antecedents: Learning From Research on Identity and Ethics Programs," *Business Ethics Quarterly*, 19, 553–586.

Wallmeier, N. (2019): "The Hidden Costs of Whistleblower Protection," *University of Hamburg, mimeo.*

Zingales, L. (2004): "Want to Stop Corporate Fraud? Pay Off Those Whistle-Blowers," *Washington Post*, January 18 Issue.

Appendix

A Theory

This Appendix is structured as follows: In Section A.1, the model is presented, and in Section A.2, we derive the equilibrium outcomes for treatments *NoP* and *P-R*. We focus on pure-strategy Perfect Bayesian Equilibria that are *informative equilibria* in the sense that the prosecutor triggers an investigation if and only if the employee sends a report. Hence, we do not consider babbling equilibria throughout. The theoretical predictions of Section 4.1 follow immediately from Propositions 1 and 2. The comparisons of the fractions of employers who misbehave (as stated in *Prediction M*) are derived at the end of Section A.2.

A.1 Model

The Game Played We consider a game played by three players, an employer, an employee, and a prosecutor (see also Figure 1 in the main text).[41] The employer (she) is matched with an employee of type θ whose productivity x_θ the employer appropriates. In addition, the employer decides whether or not to misbehave, denoted by $M \in \{0, 1\}$ (where $M = 0$ indicates no misbehavior), which is observed by the employee, but not by the prosecutor.

The employee has productivity x_θ, $\theta = L, H$, which is either high ($\theta = H$: H-employee) or low ($\theta = L$: L-employee, where $x_H > x_L$). The employee's productivity is known to the employer but not to the prosecutor who only knows that there is a share $h \in (0, 1)$ of H-employees in the population. The employee's only choice is whether or not to send a report to the prosecutor indicating that the employer engaged in misbehavior, i.e., $R \in \{0, 1\}$, where $R = 1$ indicates that the employee sends a report. As a tie-breaking rule, we assume that employees refrain from reporting when being indifferent between reporting and not reporting.[42] The prosecutor always observes whether or not a report is sent.

After the employee's reporting decision, the prosecutor decides on initiating an investigation, $I \in \{0, 1\}$, where $I = 1$ indicates an investigation. Upon investigating the prosecutor learns whether or not the employer indeed has misbehaved. Whether or not an investigation is initiated and whether or not the employer is found to be guilty is publicly observable.

[41] As discussed in Section 3, in the experiment we have added a "third party", which is a purely passive player without any decisions to take. In the experiment, it is only included to make it more salient that misbehavior causes harm to others.

[42] Our results would also hold in a model where employees face some (small) reporting costs.

Table 2: Monetary Payoffs (in Bold) and Non-Monetary Payoffs (in Non-Bold) Conditional on the Misbehavior (M), Investigation (I), and Dismissal (D) Decisions

(a) Payoffs When There is No Protection: *NoP* and *P-R* (if R=0)

M	I	D	Employee	Prosecutor	Employer
0	0	0	$\boldsymbol{\omega}$	$\boldsymbol{0}$	$\boldsymbol{x_i - \omega} - \beta \cdot \tau$
0	0	1	$\boldsymbol{0}$	$\boldsymbol{0}$	$\boldsymbol{\overline{x} - \omega}$
0	1	0	$\boldsymbol{\omega}$	$\boldsymbol{-K_1}$	$\boldsymbol{x_i - \omega} - \beta \cdot \tau$
0	1	1	$\boldsymbol{0}$	$\boldsymbol{-K_1}$	$\boldsymbol{\overline{x} - \omega}$
1	1	1	$\boldsymbol{0}$	$\boldsymbol{-K_1 - K_2}$	$\boldsymbol{\overline{x} - \omega - f} + y$
1	1	0	$\boldsymbol{\omega}$	$\boldsymbol{-K_1 - K_2}$	$\boldsymbol{x_i - \omega - f + z} - \zeta - \beta \cdot \tau$
1	0	1	$\boldsymbol{0} - \delta$	$\boldsymbol{-K_2 - K_3}$	$\boldsymbol{\overline{x} - \omega + z} - \zeta$
1	0	0	$\boldsymbol{\omega} - \delta$	$\boldsymbol{-K_2 - K_3}$	$\boldsymbol{x_i - \omega + z} - \zeta - \beta \cdot \tau$

(b) Payoffs When There is Protection: *P-R* (if R=1)

M	I	D	Employee	Prosecutor	Employer
0	0	n/a	$\boldsymbol{\omega}$	$\boldsymbol{0}$	$\boldsymbol{x_i - \omega} - \beta \cdot \tau$
0	1	n/a	$\boldsymbol{\omega}$	$\boldsymbol{-K_1}$	$\boldsymbol{x_i - \omega} - \beta \cdot \tau$
1	1	n/a	$\boldsymbol{\omega}$	$\boldsymbol{-K_1 - K_2}$	$\boldsymbol{x_i - \omega - f + z} - \zeta - \beta \cdot \tau$
1	0	n/a	$\boldsymbol{\omega} - \delta$	$\boldsymbol{-K_2 - K_3}$	$\boldsymbol{x_i - \omega + z} - \zeta - \beta \cdot \tau$

Notes: Monetary payoffs as implemented in the experiment (see Section 3.1) are indicated in bold letters, where we use the following parameterization: $\omega = 40$, $x_L = 30$, $x_H = 80$, $\overline{x} = 70$, $z = 50$, $f = 60$, $K_1 = 20$, $K_2 = -10$, $K_3 = 30$. Non-monetary payoffs due to behavioral motives (see Section 4.1) are indicated in non-bold letters (and they are not incentivized in the experiment). As the employee's reporting decision has no *direct* effect on the payoff of neither player, we omit a separate column for the sake of readability.

Finally, before production eventually takes place, the employer decides whether or not to dismiss the employee, $D \in \{0, 1\}$, where $D = 1$ indicates a dismissal. A dismissed employee is replaced by an outsider of some intermediate productivity $\overline{x}$, with $x_L \leq \overline{x} < x_H$. In this case, the employer appropriates the outsider's productivity. In treatment *NoP*, the employer is free to dismiss the employee. In treatment *P-R*, a dismissal is prohibited if and only if $R = 1$.

Payoffs All payoffs (monetary and non-monetary) are summarized in Table 2. First, the payoff of the employer depends on whether or not she misbehaves, whether or not there is an investigation, and whether or not she employs a whistleblower. The employer's potential net gain y from misbehavior consists of a monetary payoff z minus some disutility from misbehavior ζ (which might reflect moral reservations of the employer). We assume that ζ is randomly distributed (and the realization is private information of the employer), and hence this is also the case for y. In particular, we assume that y is distributed according to $H(\cdot)$, with full support, and mean $\overline{y}$. If the prosecutor investigates and there is misbehavior, the employer faces an (exogenously given) fine $f > 0$. The employer receives the employee's or the outside replacement's productivity (i.e., x_L, x_H, or $\overline{x}$) and pays a fixed wage ω. The employer dislikes employing a whistleblower, and the respective disutility is denoted by $\tau > 0$. It is drawn from a distribution $G(.)$, and it is the employer's private information. The employer forms a belief $\beta \in [0, 1]$ that her employee has sent a report.

Second, the employee gets a fixed wage ω if he is not dismissed by the employer, and zero otherwise. In addition, misbehavior that remains undetected by the prosecutor imposes a disutility $\delta > 0$ on the employee, which reflects a preference for conscience cleaning as discussed in the main text, and which is the employee's private information. From the viewpoint of the other players, δ is drawn from a distribution $F(\delta)$. We assume $F(\omega) < 1$ which ensures that there exist values of δ for which the respective disutility outweighs the (H-employee's) fear of dismissal. Moreover, in case of undetected misbehavior, δ accrues to the employee independently of whether or not he is dismissed.

Finally, the payoff of the prosecutor depends on whether there is misbehavior and whether an investigation takes place. When there is no misbehavior, the prosecutor's payoff is $-K_1$ (0) if he investigates (does not investigate). Hence, $K_1 > 0$ can be considered as investigation costs. When there is misbehavior, his payoff is $-K_1 - K_2$ if he investigates and $-K_2 - K_3$ if he does not investigate, where we assume $K_3 > K_1$. Hence, when there is (no) misbehavior, the prosecutor's payoff is higher if he conducts (does not conduct) an investigation.

A.2 Equilibrium Analysis

A.2.1 Preliminaries

When deriving our predictions, we focus on 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), that are informative in the sense that the prosecutor's investigation decision varies with the employee's report:

Definition 1. *A Perfect Bayesian Equilibrium is called* ***informative equilibrium*** *if the prosecutor's equilibrium strategy is given by $I^*(R) = R$ for all $R \in \{0, 1\}$.*

In the following, we provide conditions for the existence of an informative equilibrium under each treatment, and we assume that it is always played given that it exists. To derive our predictions, we proceed as follows: First, under the assumption that the prosecutor plays his equilibrium strategy $I^*(R) = R$, we characterize optimal behavior with respect to misbehavior, reporting, and dismissal, denoted by $M^*(\cdot)$, $R^*(\cdot)$, and $D^*(\cdot)$, respectively. Note that in informative equilibrium, the employer's belief that the employee has sent a report satisfies $\beta^* \in \{0, 1\}$. Second, we derive conditions under which $I^*(R) = R$ is in fact optimal for the prosecutor (i.e., for each treatment, we provide conditions that ensure existence of informative equilibrium). Third, this leads to the equilibrium outcome, which depends on the realizations of the random variables δ, τ, and y (where these realizations are unknown to the experimenter). Taking into account the prior distributions of these random variables, the predictions of Section 4.1 are then based on the *expected equilibrium outcomes* (see Propositions 1 and 2).

A.2.2 Treatment *NoP*: Equilibrium Outcome

In the following, we assume that the report is observed by both the prosecutor and the employer (as in the experiment), and we solve the game backwards, starting with the employer's dismissal decision at date 4 (see Figure 1 in the main text). In doing so, we write $D^*(\cdot)$ as a function of I rather than R, because $I = R$ for all $R \in \{0, 1\}$ in the informative equilibrium:

Lemma 1 (***NoP*: Dismissal**). *In the informative equilibrium, the following holds: The L-employee is always dismissed. The H-employee is dismissed only if both a report occurs and the employer's disutility from retaining a known whistle-blower is sufficiently large. That is,*

$$D^*(x_\theta, I, \tau) = \begin{cases} 1 & \textit{if } x_\theta = x_L, \\ 1 & \textit{if } x_\theta = x_H, \textit{ and } R = 1 \textit{ and } \tau > \overline{\tau}, \textit{ and} \\ 0 & \textit{else.} \end{cases}$$

where $\overline{\tau} := x_H - \overline{x}$.

Proof. First, when $R = I = 0$, the employer gets x_θ if retaining the employee, and $\overline{x}$ if dismissing him. Since $x_L < \overline{x} < x_H$, in this case, the L-employee (H-employee) is dismissed (retained). Second, when $R = I = 1$, the employer gets $x_\theta - \tau - M \cdot f$ if retaining the employee and $\overline{x} - M \cdot f$ if dismissing him. Hence, the L-employee is again dismissed, while the H-employee is dismissed only if τ is sufficiently large, i.e., for $\tau > \overline{\tau} := x_H - \overline{x}$. □

In the informative equilibrium, the employee's optimal reporting behavior at date 2 can be characterized as follows:

Lemma 2 (***NoP*: Reporting**). *In the informative equilibrium, the following holds: The L-employee reports if and only if the employer misbehaves. The H-employee reports if and only if there is both misbehavior and his disutility δ from undetected misbehavior is sufficiently large. That is,*

$$R^*(x_\theta, M, \delta) = \begin{cases} 1 & \textit{if } M = 1 \textit{ and } x_\theta = x_L, \\ 1 & \textit{if } M = 1,\ x_\theta = x_H \textit{ and } \delta > \overline{\delta}, \textit{ and} \\ 0 & \textit{else,} \end{cases}$$

where $\overline{\delta} := (1 - G(\overline{\tau})) \cdot \omega$.

Proof. The L-employee is always dismissed independent of his reporting decision (see Lemma 1). Hence, the L-employee's payoff is $-\delta \cdot M$ if he does not report and 0 if he reports. Again from Lemma 1, when not reporting, the H-employee is not dismissed, and hence gets $\omega - \delta \cdot M$. Upon reporting, he is retained with probability $G(\overline{\tau})$, and hence his payoff is $G(\overline{\tau}) \cdot \omega$. □

Next, consider the employer's misbehavior decision at date 1:

Lemma 3 (***NoP*: Misbehavior**). *In the informative equilibrium, the employer's misbehavior decision is given by:*

$$M^*(x_\theta, y, \tau) = \begin{cases} 1 & \textit{if } x_\theta = x_L \textit{ and } y > f, \\ 1 & \textit{if } x_\theta = x_H \textit{ and } \tau < \overline{\tau} \textit{ and } y > y_1, \\ 1 & \textit{if } x_\theta = x_H \textit{ and } \tau > \overline{\tau} \textit{ and } y > y_2, \textit{ and} \\ 0 & \textit{else,} \end{cases}$$

where $y_1 := (1 - F(\overline{\delta}))(f + \tau)$ *and* $y_2 := (1 - F(\overline{\delta}))(x_H - \overline{x} + f)$.

Proof. First, suppose the employer faces an L-employee. In this case, Lemmas 1 and 2 imply that the employer's payoff is $\overline{x} + y - \omega - f$ if she misbehaves, and $\overline{x} - \omega$ if she does not misbehave. Hence, misbehavior is optimal if $y > f$. Second, consider the situation where the employer is facing an H-employee. When the employer chooses $M = 0$, then Lemmas 1 and 2 imply that her payoff is $x_H - \omega$. When choosing $M = 1$ instead, then the employer's payoff also depends on the subsequent dismissal decision, and hence it also depends on τ. Case (i): $\tau < \overline{\tau}$ (no subsequent dismissal). From Lemma 2, it follows that the employer's expected payoff when choosing $M = 1$ is $x_H + y - \omega - \left(1 - F(\overline{\delta})\right)(f + \tau)$. In this case, the employer optimally misbehaves if $y > y_1 := (1 - F(\overline{\delta}))(f + \tau)$. Case (ii): $\tau > \overline{\tau}$ (subsequent dismissal). Here, the expected payoff from choosing $M = 1$ is $y - \omega + F(\overline{\delta})x_H + \left(1 - F(\overline{\delta})\right)(\overline{x} - f)$. In this case, the employer optimally misbehaves if $y > y_2 := (1 - F(\overline{\delta}))(x_H - \overline{x} + f)$. □

Finally, consider the prosecutor's investigation decision, and recall that the prosecutor does not observe the employee's productivity. Define the prosecutor's equilibrium belief with respect to misbehavior conditional on R as $B_0 := \Pr\{M = 1 \mid R = 0\}$ and $B_1 := \Pr\{M = 1 \mid R = 1\}$. Given Lemmas 1 - 3, in equilibrium this leads to $B_1 = 1$ (as there are no non-meritorious claims) and $B_0 < 1$ (as misbehavior is not always reported). In particular,

$$B_0 = \frac{h \cdot p_H^0 \cdot F\left(\overline{\delta}\right)}{h \cdot \left(p_H^0 \cdot F\left(\overline{\delta}\right) + 1 - p_H^0\right) + (1 - h) \cdot H(f)}, \tag{1}$$

where

$$p_H^0 := G\left(\overline{\tau}\right) E_\tau \left[1 - H\left(y_1\right) \mid \tau < \overline{\tau}\right] + \left(1 - G\left(\overline{\tau}\right)\right)\left(1 - H\left(y_2\right)\right) \tag{2}$$

and where in (2) expectations are formed over τ (as y_1 is a function of τ). Intuitively, in (1) the numerator states the probability of unreported misbehavior (recall that this occurs with H-employees only), and the denominator states the overall probability that no report is sent.

Lemma 4 (*NoP*: Investigation). *Given the behavior of the other players as described in Lemmas 1 - 3, if $B_0 \leq \frac{K_1}{K_3}$ holds, then choosing $I^*(R) = R$ is optimal for the prosecutor.*

Proof. First, if $R = 0$, upon choosing $I = 0$, the prosecutor's expected payoff is $-B_0 \cdot (K_3 + K_2)$. When choosing $I = 1$ instead, he gets $-K_1 - B_0 \cdot K_2$. Hence, given $R = 0$, $I = 0$ is optimal iff $B_0 \leq \frac{K_1}{K_3}$. Second, if $R = 1$, when choosing $I = 0$, the prosecutor's expected payoff is $-B_1 \cdot (K_3 + K_2)$. When choosing $I = 1$ instead, he gets $-K_1 - B_1 \cdot K_2$. Hence, given $R = 1$, $I = 1$ is optimal iff $B_1 > \frac{K_1}{K_3}$. Since in equilibrium $B_1 = 1$, this is always satisfied (recall that $K_1 < K_3$ by assumption). □

Lemmas 1 to 4 characterize behavior in informative equilibrium. As this also depends on the random variables τ, δ and y (which are unobservable to the experimenter), we now state the *expected* equilibrium outcome given the prior distributions of these random variables. This expected equilibrium outcome is the basis for the predictions in Section 4.1:

Proposition 1 (*NoP*: Expected Equilibrium Outcome). *The informative equilibrium in treatment* NoP *has the following expected equilibrium outcome: (i) L-employees always (never) report if there is (no) misbehavior. (ii) L-employees are always dismissed. (iii) Given misbehavior, the probability of observing a report by an H-employee is $E_\delta[R^*(x_H, 1, \delta)] = 1 - F(\overline{\delta})$, and,* *in the absence of misbehavior, H-employees never send a report. (iv) Given that an H-employee sends a report, the probability of observing his dismissal is $E_\tau[D^*(x_H, 1, \tau)] = 1 - G(\overline{\tau})$, while when sending no report, he is never dismissed. (v) The probability of observing misbehavior by the employer when matched with an L-employee is $m_L^{No} := E_{y,\tau}[M^*(x_L, y, \tau)] = 1 - H(f)$. (vi) The probability of observing misbehavior by the employer when matched with an H-employee is $m_H^{No} := E_{y,\tau}[M^*(x_H, y, \tau)] = p_H^0$ as defined in (2). (vii) When (not) receiving a report, prosecutors always (never) trigger an investigation.*

A.2.3 Equilibrium Outcome in Treatment *P-R*

Again, we assume that the report is observed by both the prosecutor and the employer (as in the experiment), and we solve the game backwards:

Lemma 5 (***P-R*: Dismissal**). *In the informative equilibrium, the following holds: The L-employee is dismissed whenever this is feasible (i.e., if $R = 0$). The H-employee is never dismissed. That is,*

$$D^*(x_\theta, I, \tau) = \begin{cases} 1 & \text{if } x_\theta = x_L \text{ and } R = 0, \text{ and} \\ 0 & \text{else.} \end{cases}$$

Proof. In treatment *P-R*, a dismissal is only feasible when $R = 0$. Analogously to Lemma 1, the L-employee is always dismissed (when feasible). Moreover, the employer might only want to dismiss the H-employee, if the latter sends a report (in which case a dismissal is, however, not feasible). □

In informative equilibrium, the employee's optimal reporting behavior at date 2 can be characterized as follows:

Lemma 6 (***P-R*: Reporting**). *In the informative equilibrium, the following holds: The L-employee always sends a report, irrespective of whether or not there is misbehavior. In contrast, the H-employee sends a report if and only if there is misbehavior. That is,*

$$R^*(x_\theta, M, \delta) = \begin{cases} 1 & \text{if } x_\theta = x_L, \\ 1 & \text{if } x_\theta = x_H \text{ and } M = 1, \text{ and} \\ 0 & \text{else.} \end{cases}$$

Proof. From Lemma 5, the L-employee anticipates that he will be dismissed unless sending a report (thereby obtaining protection). For $M = 1$, his payoff upon choosing $R = 1$ is ω (since the report triggers an investigation), while he would get only $-\delta$ when choosing $R = 0$ instead. For $M = 0$, the L-employee still gets ω upon choosing $R = 1$, but would get zero upon choosing $R = 0$. Hence, always sending a report is optimal for the L-employee. An H-employee who observes $M = 1$ gets ω when choosing $R = 1$, and $\omega - \delta$ when choosing $R = 0$. If $M = 0$, he gets ω regardless of his reporting decision. Since we assume no reporting in case of indifference, the optimal response to $M = 0$ is $R = 0$. □

Next, consider the employer's misbehavior decision at date 1.

Lemma 7 (***P-R*** **: Misbehavior).** *In the informative equilibrium, the employer's misbehavior decision is given by:*

$$M^*(x_\theta, y, \tau) = \begin{cases} 1 & \text{if } x_\theta = x_L \text{ and } y > f, \\ 1 & \text{if } x_\theta = x_H \text{ and } y > f + \tau, \text{ and} \\ 0 & \text{else.} \end{cases}$$

Proof. Given Lemmas 5 and 6, when matched with an L-employee, the employer anticipates that the employee always reports, and hence always triggers an investigation. Therefore, when choosing $M = 1$, the employer gets $x_L - \omega + y - f - \tau$. Upon choosing $M = 0$, she gets $x_L - \omega - \tau$. By contrast, when matched with an H-employee, the employer anticipates that a report is sent if and only if $M = 1$ is chosen. Hence, upon choosing $M = 1$ she gets $x_H - \omega + y - f - \tau$, and $x_H - \omega$ upon choosing $M = 0$. □

Finally, consider the prosecutor's investigation decision. Given Lemmas 5 - 7, his equilibrium beliefs with respect to misbehavior conditional on R are given by $B_0 = 0$ (in equilibrium, any misbehavior is reported) and

$$B_1 = \frac{h \cdot p_H^1 + (1-h) \cdot (1 - H(f))}{h \cdot p_H^1 + (1-h)} \in (0,1)\,, \tag{3}$$

where

$$p_H^1 := E_\tau \left[1 - H\left(f + \tau\right)\right]. \tag{4}$$

Lemma 8 (***P-R*** **: Investigation).** *Given the behavior of the other players as described in Lemmas 5 - 7, if $\frac{K_1}{K_3} \leq B_1$ holds, then choosing $I^*(R) = R$ is optimal for the prosecutor.*

Proof. First, if $R = 0$, then, when choosing $I = 0$, the prosecutor's expected payoff is $-B_0 \cdot (K_3 + K_2) = 0$ due to $B_0 = 0$. When choosing $I = 1$ instead, the prosecutor gets $-K_1 - B_0 \cdot K_2 < 0$, which is strictly worse. Second, if $R = 1$, when choosing $I = 0$, the prosecutor's expected payoff is $-B_1 \cdot (K_3 + K_2)$. When choosing $I = 1$ instead, he gets $-K_1 - B_1 \cdot K_2$. Hence, given $R = 1$, $I = 1$ is optimal iff $\frac{K_1}{K_3} \leq B_1$. □

Note that the condition $\frac{K_1}{K_3} \leq B_1$ in Lemma 8 can be reformulated as follows (inserting for B_1): $\frac{K_1}{K_3} \leq \frac{1}{1+\frac{1}{1-H(f)}}$ Furthermore, note that $\frac{1}{1-H(f)}$ represents the fraction of the reporting frequency of L-employees (which is 1) and the frequency of misbehavior of employers matched with L-employees (which is $1 - H(f)$). This fraction is nothing else than the inverse of the measure of informativeness of the reports sent by L-employees: If it is 1, these reports are perfectly informative; but for higher values, they are less informative. If the percentage of non-meritorious claims among reports sent by L-employees converges to 100%, i.e., if the frequency of misbehavior of employers matched with L-employees converges to zero, the measure of informativeness converges to zero, too; its inverse converges to infinity, and thus, B_1 converges to zero. Hence, for $\frac{K_1}{K_3}$ bounded away from zero, the condition for existence of the informative equilibrium is violated for a sufficiently high percentage of non-meritorious claims among reports sent by L-employees. For now, we assume that the condition is not violated. We come back to the possibility of violation below.

Lemmas 5 - 8 characterize behavior in informative equilibrium. As this also depends on the random variables τ, δ and y (which are unobservable to the experimenter), we now state the *expected* equilibrium outcome given the prior distributions of these random variables. This expected equilibrium outcome is the basis for the predictions in Section 4.1:

Proposition 2 (*P-R*: Expected Equilibrium Outcome). *The informative equilibrium in treatment* P-R *has the following expected equilibrium outcome: (i) L-employees send a report regardless of whether or not there is misbehavior. (ii) L-employees are never dismissed. (iii) H-employees always (never) report if there is (no) misbehavior. (iv) H-employees are never dismissed. (v) The probability of observing misbehavior by the employer when matched with an L-employee is $m_L^R := E_{y,\tau}[M^*(x_L, y, \tau)] = 1 - H(f)$. (vi) The probability of observing misbehavior by the employer when matched with an H-employee is $m_H^R := E_{y,\tau}[M^*(x_H, y, \tau)] = p_H^1$ as defined in (4). (vii) When (not) receiving a report, prosecutors always (never) trigger an investigation.*

A.2.4 Comparing Employer Misbehavior

Propositions 1 - 2 directly lead to the predictions concerning investigations, dismissals, and reporting as presented in Section 4.1. The comparison of employer misbehavior across treatments and employee productivity types (see *Prediction M*) requires some further elaboration: From Lemmas 3 and 7, for a given productivity type of the employee, the employer misbehaves if y exceeds a certain threshold. First, when the employer is matched with an L-employee, then in both treatments, the employer misbehaves if $y > f$. Hence, $m_L^{No} = m_L^R$. Second, when the employer is matched with an H-employee, then $m_H^{No} > m_H^R$ holds: From Lemma 7, the threshold for y that determines m_H^R is $f + \tau$. From Lemma 3, the threshold for y that determines m_H^{No} depends on τ: First, if $\tau < \overline{\tau}$, the threshold is $(1 - F(\overline{\delta}))(f + \tau) < (f + \tau)$. Second, if $\tau > \overline{\tau}$, the threshold is $(1 - F(\overline{\delta}))(x_H - \overline{x} + f) = (1 - F(\overline{\delta}))(f + \overline{\tau}) < (f + \tau)$ because $\overline{\tau} = x_H - \overline{x}$ and we are in the case $\tau > \overline{\tau}$.

B Instructions

Note: We report here a translation of the instructions (originally in German) for treatments *NoP* and *P-R*, where all changes in *P-R* are indicated in square brackets as follow: [In *P-R* only: ...]. The respective modifications for the other treatments were made accordingly and are available upon request.

Welcome to today's experiment!

You are taking part in a decision situation, where you can earn some money. How much you will earn depends on your decisions and on the decisions of the other participants that are allocated to you. Moreover, your earnings depend on the role that is randomly assigned to you. The experiment consists of **two parts**. You now receive the instructions for the first part. After having finished the first part, you will get the instructions for the second part. What happens in the first part of the experiment will not have any influence on the amount of money that you might earn in the second part of the experiment. And vice versa. After having completed both parts, you will also have to answer a short questionnaire.

Please note that from now on until the end of the experiment it is **not allowed to communicate!** If you have any questions, please raise your hand out of your cubicle. One of the experimenters will come to you. Throughout the experiment, it is forbidden to use mobile phones, smartphones, tablets, or alike. Participants intentionally violating the rules may be asked to leave the experiment and may not be paid. All decisions are made anonymously, i.e., none of the participants will learn about the identity of the others. The payment for both parts of the experiment will also be made anonymously at the end of the experiment.

Instructions for the first part of the experiment

Please notice that if subsequently we refer to the "experiment", this relates to the **first part** of the experiment.

1. What it is about - A short overview

This experiment is about making decisions in a **group of four people** that consists of an **employer**, an **employee**, a **third party**, and a **prosecutor**, where these decisions may affect the payoffs of all members of the group. All decisions are made by the employer, the employee, and the prosecutor; the affected person cannot make any decisions. The employer chooses between two alternatives, **CIRCLE** and **TRIANGLE**. A (fictitious) **law for the protection of the third party** says that TRIANGLE should not be chosen as it harms the third party. Nevertheless, if an employer chooses TRIANGLE, he goes **completely unpunished** and even earns a higher profit - **provided that the prosecutor does not initiate an investigation**. The employer's decision between the two alternatives can only be observed by the employee. **The employee - and only him - can (but does not have to) ask the prosecutor to initiate an investigation.** The prosecutor may initiate an investigation even if the employee has not asked him to do so. The employer learns whether an investigation is initiated or not. He also learns whether the employee asked the prosecutor to initiate an investigation or not. At the end of a given round (of which there will be several) **the employer decides on whether the employee is dismissed or not**. [In *P-R* only: If, however, the employee has asked the prosecutor to conduct an investigation, **a dismissal of the employee is not possible**. This applies regardless of whether the employer chose CIRCLE or TRIANGLE and regardless of whether the prosecutor initiated an investigation or not.] In the following, the experiment will be explained more in detail.

2. The assignment of roles

At the beginning of the experiment, the computer randomly assigns every participant a role either as employer, employee, third party or prosecutor. **Employers will stay employers throughout the whole experiment**. However, over the course of the experiment, prosecutors and employees will sometimes also take the role of third party; and third parties will sometimes take the role of either employee or prosecutor. **Prosecutors will never take the role of employer, and employees will never take the role of prosecutor.** The change of roles occurs randomly, and is consequently not affected by current or prior decisions. The change of roles only takes place between rounds. During a given round of the experiment, each member of the group remains in his or her role. In each round, the computer randomly matches the participants into groups of four consisting of an employer, an employee, a third party, and a prosecutor. The employee is also randomly assigned **a productivity level (high or low)**.

Both productivity levels are equally likely, and the productivity level is **independent across rounds**, i.e., the productivity level of an employee might change from round to round. In the following, the course of events in a given round will be described. The experiment consists of **30 rounds**.

3. The sequence of events in a given round

3.1. The sequence of events in a given round from the perspective of the employer

The employer **does not receive an initial endowment**; i.e., his earnings depend exclusively on his decisions and the decisions of the other group members. First, the employer learns whether the **productivity level of his employee is high or low**. A **high-productivity employee**, who does not get dismissed, will earn the employer **80 points** for the current round; a **low-productivity employee**, who does not get dismissed, is worth **30 points**. If the employer dismisses his employee at the end of the round [In *P-R* only: (which is only possible if the employee did **not** ask the prosecutor to conduct an investigation)], he will get a **new employee** whose productivity will earn him **70 points**. Each employee who is **not dismissed** (and also any new employee replacing a dismissed employee) **earns a wage of 40 points**. An employee who got dismissed does not earn a wage in the current round.

Before the employer decides on whether to dismiss the employee or not, he has to take another decision: He has to choose between two alternatives, **CIRCLE and TRIANGLE**. This decision is observed by the employee only.

<u>If CIRCLE is chosen</u>

If the employer chooses **CIRCLE, he will not receive any extra earnings**, and **he will not cause any financial loss for the third party**. In this case, his earnings in the current round only result from the productivity of the employee (80, 30, or 70 points, depending on the productivity of the initial employee and depending on whether the initial employee is replaced by a new one) minus the employee's salary (40 points).

- An **employer with a high-productivity employee**, who chooses **CIRCLE**, gets 80 - 40 = **40 points** if he keeps the employee. If the employee gets replaced by a new one, the employer receives 70 - 40 = **30 points**.

- An **employer with a low-productivity employee** who chooses option **CIRCLE** gets 30 - 40 = **-10 points** if he keeps the employee. If the initial employee is replaced by a new one, the employer receives 70 - 40 = **30 points**.
- These payments are **irrespective of the prosecutor's decision for conducting an investigation or not**.

<u>If TRIANGLE is chosen</u>

If the employer chooses **TRIANGLE**, there are two [In *P-R*: four] distinct cases, depending on [In *P-R* only: whether the employee asked the prosecutor to investigate or not, and on] whether the prosecutor conducts an investigation or not.
In any of these cases if the employer chooses TRIANGLE, then he receives **an extra payment of 50 points in addition to the productivity of his employee**. In the case of **no investigation**, the employer goes unpunished and does not have to pay a fine, while in the case of an investigation, he has to pay a **fine of 60 points**, which, hence, exceeds the extra payment resulting from the choice of TRIANGLE. [In *P-R* only: Furthermore, the employee can **only** be dismissed if he did not ask the prosecutor to conduct an investigation, i.e., if he **kept silent**.]

- If the prosecutor does **not conduct an investigation**, and the employer consequently remains unpunished, the following holds:
  - An **employer with a high-productivity employee** who chooses **TRIANGLE** gets 80 + 50 - 40 = **90 points** if he keeps the employee. If the employee is replaced by a new one [In *P-R* only: (which is only possible if the employee kept silent)], the employer receives 70 + 50 - 40 = **80 points**.
  - An **employer with a low-productivity employee** who chooses **TRIANGLE** gets 30 + 50 - 40 = **40 points** if he keeps the employee. If the old employee is replaced by a new one [In *P-R* only: (which is only possible if the employee kept silent)], the employer receives 70 + 50 - 40 = **80 points**.
- If the prosecutor **conducts an investigation**, the following holds:

- An **employer with a highproductivity employee** who chooses **TRIANGLE** gets 80 + 50 - 60 - 40 = **30 points** if he keeps the employee. If the employee gets replaced by a new one [In *P-R* only: (which is only possible if the employee kept silent)], the employer receives 70 + 50 - 60 - 40 = **20 points**.
- An **employer with a low-productivity employee** who chooses **TRIANGLE** gets 30 + 50 - 60 - 40 = **-20 points** if he keeps the employee. If the old employee is replaced by a new one [In *P-R* only: (which is only possible if the employee kept silent)], the employer receives 70 + 50 - 60 - 40 = **20 points**.

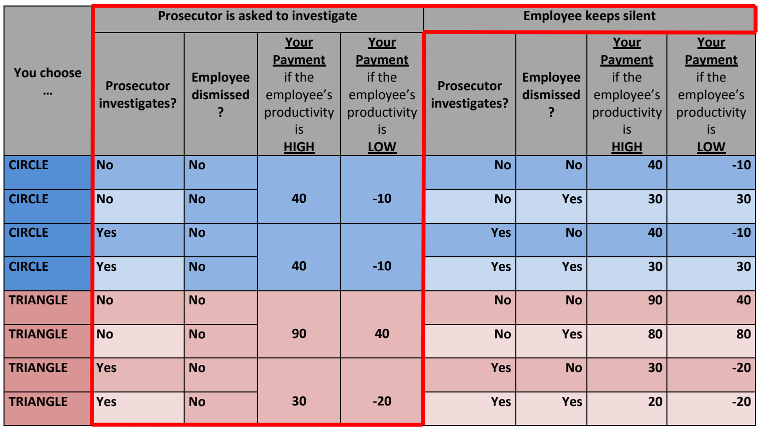
The potential fine is higher than the extra payment the employer receives when choosing TRIANGLE. Thus, it depends on the prosecutor's decision to conduct an investigation or not whether the employer earns more when choosing TRIANGLE or when choosing CIRCLE.

However, the employer choosing TRIANGLE implies a **loss of 70 points for the third party**. As the third party has an initial endowment of **40 points**, if the employer chooses TRIANGLE, the third party **loses 30 points** in the current round. However, this only applies if the prosecutor does not conduct an investigation, because choosing TRIANGLE violates the (fictitious) **law for the protection of the third party**. If the prosecutor conducts an investigation (potentially because he was asked to do so by the employee), the third party receives a partial refund of his damage in the form of a **compensation of 20 points**. In the role of third party, it is thus possible to complete the first part of the experiment with a loss. However, no participant will finish the entire experiment with a loss.

The total payoff (for the current round) of the employer (depending on the productivity of his employee as well as on his own decisions and the decision of the prosecutor) is summarized in the below table. In the experiment, this table is shown on the employer's decision screen. [In treatment *P-R*, the part of the table marked by the red bold frame is displayed in addition to the remainder of the table.]

The employer should keep in mind that the employee observes his choice between the two alternatives and may ask the prosecutor to initiate an investigation. [In *P-R* only: In this case, a dismissal of the employee is not possible.]

3.2 The sequence of events in a given round from the perspective of the employee

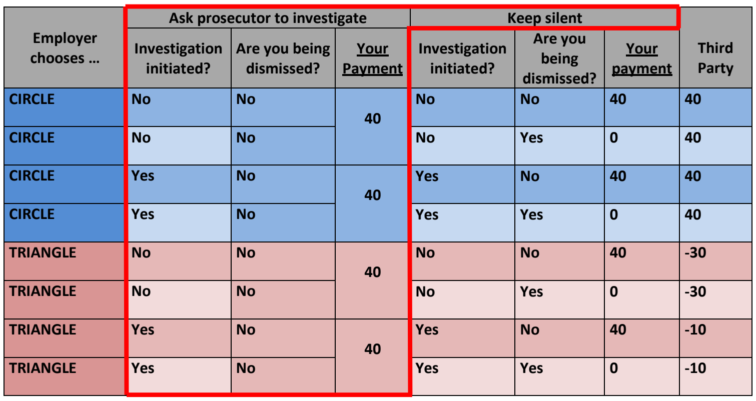


You choose …	Prosecutor is asked to investigate				Employee keeps silent			
	Prosecutor investigates?	Employee dismissed ?	**Your Payment** if the employee's productivity is **HIGH**	**Your Payment** if the employee's productivity is **LOW**	Prosecutor investigates?	Employee dismissed ?	**Your Payment** if the employee's productivity is **HIGH**	**Your Payment** if the employee's productivity is **LOW**
CIRCLE	No	No	40	-10	No	No	40	-10
CIRCLE	No	No			No	Yes	30	30
CIRCLE	Yes	No	40	-10	Yes	No	40	-10
CIRCLE	Yes	No			Yes	Yes	30	30
TRIANGLE	No	No	90	40	No	No	90	40
TRIANGLE	No	No			No	Yes	80	80
TRIANGLE	Yes	No	30	-20	Yes	No	30	-20
TRIANGLE	Yes	No			Yes	Yes	20	-20

The employee does **not receive an initial endowment**, i.e., his earnings depend exclusively on his decisions and the decisions of the others. First, the employee is informed about whether his **productivity level is high or low**. Both productivity levels are equally likely. At the end of the round, the employer can dismiss the employee. [In *P-R* only: However, a dismissal is only possible, if the employee did **not** ask the prosecutor to conduct an investigation, i.e., if he kept silent.] If the employee gets **dismissed**, he earns **0 points** in the current round. If the employee **does not get dismissed**, he receives a **wage of 40 points** from the employer.

The employee observes whether the employer chose **CIRCLE** or **TRIANGLE**. He then decides on whether to ask the prosecutor to conduct an investigation. This decision is taken as follows: The employee indicates both whether he wants to ask the prosecutor to conduct an investigation in case that the employer chose CIRCLE and also whether he wants to ask the prosecutor to conduct an investigation in case that the employer chose TRIANGLE. The computer then effectuates the decision (depending on the actual decision of the employer). Also the **employer** observes whether or not the employee decides to ask the prosecutor to conduct an investigation. If the **prosecutor conducts an investigation**, the following applies: If the employer chose CIRCLE, nothing happens. If, however, the employer chose TRIANGLE, the employer has to **pay a fine of 60 points**, while the **third party receives a compensation payment of 20 points**.

The total payoff (for the current round) of the **employee and the third party**, respectively, (depending on his own decision as well as on the decisions of the employer and the prosecutor) are summarized in the below table. In the experiment, this table is shown on the employee's decision screen. [In treatment *P-R*, the part of the table marked by the red bold frame is displayed in addition to the remainder of the table.]



Employer chooses ...	Ask prosecutor to investigate			Keep silent			Third Party
	Investigation initiated?	Are you being dismissed?	Your Payment	Investigation initiated?	Are you being dismissed?	Your payment	
CIRCLE	No	No	40	No	No	40	40
CIRCLE	No	No		No	Yes	0	40
CIRCLE	Yes	No	40	Yes	No	40	40
CIRCLE	Yes	No		Yes	Yes	0	40
TRIANGLE	No	No	40	No	No	40	-30
TRIANGLE	No	No		No	Yes	0	-30
TRIANGLE	Yes	No	40	Yes	No	40	-10
TRIANGLE	Yes	No		Yes	Yes	0	-10

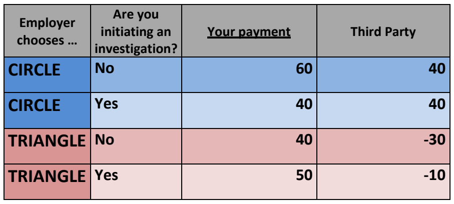
The employee should keep in mind two things. Firstly, if the employer chooses TRIANGLE, the employee may ask the prosecutor to conduct an investigation, and, if the prosecutor acts on his request, thereby reduce the loss of the affected person. Secondly, the employer can observe whether the employee asks the prosecutor to conduct an investigation or not.

3.3 The sequence of events in a given round from the perspective of the prosecutor

The prosecutor receives an **initial endowment of 60 points** at the beginning of each round. His task is to decide on whether to investigate the employer or not. If he conducts an **investigation**, he has **costs of 20 points**. If he does **not conduct an investigation** and the employer chose **CIRCLE**, the prosecutor keeps his initial endowments.

If the employer chose **TRIANGLE**, the **prosecutor loses 20 points** if he does not conduct an investigation. If he investigates (and in spite of the investigation cost of 20 points), he only has to bear a (smaller) loss of **10 points**. When deciding on whether to investigate or not, the prosecutor can observe whether the employee asked him to investigate or not.

The total payoff (for the current round) of the **prosecutor and the third party**, respectively, (depending on his own decision and the decisions of the employer and employee) are summarized in the below table. In the experiment, this table is shown on the prosecutor's decision screen.

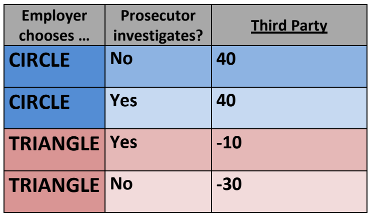


Employer chooses ...	Are you initiating an investigation?	Your payment	Third Party
CIRCLE	No	60	40
CIRCLE	Yes	40	40
TRIANGLE	No	40	-30
TRIANGLE	Yes	50	-10

The prosecutor should keep in mind two things: If the employer chose TRIANGLE, the prosecutor is the only one who can reduce both his own loss and the loss faced by the third party. If the employer chose CIRCLE, an investigation only leads to expenses. Thus, it is important for the prosecutor to think about how informative the employee's request (or lack of a request) to conduct an investigation is.

3.4 The sequence of events in a given round from the perspective of the third party

The third party gets an **initial endowment of 40 points** and does not have any own decisions to make. If the employer chooses **CIRCLE**, the third party can **keep its initial endowment**, irrespective of what the employee and the prosecutor do. If the employer chooses **TRIANGLE** and the prosecutor does **not conduct an investigation**, the third party **loses 70 points**, so that its payoff in the current round is **-30 points**. If the employer chooses **TRIANGLE** and the prosecutor **does conduct an investigation**, the third party again **loses 70 points**. However, in this case the third party also receives a **compensation payment of 20 points** so that its earnings in the current round are **-10 points**. In the experiment, this table is shown on the third party's decision screen.



Employer chooses ...	Prosecutor investigates?	Third Party
CIRCLE	No	40
CIRCLE	Yes	40
TRIANGLE	Yes	-10
TRIANGLE	No	-30

4. Summary of the sequence of events in a given round

- Each participant learns his or her role.

- The employer and the employee learn the productivity level of the employee (high or low).

- The employer chooses between two alternatives: CIRCLE and TRIANGLE

- The employee decides whether he wants to ask the prosecutor to conduct an investigation in case that the employer chooses CIRCLE, and also whether he wants to ask the prosecutor to conduct an investigation in case that the employer chooses TRIANGLE.

- The prosecutor learns whether the employee asks him to conduct an investigation or not. The prosecutor then decides on whether to conduct an investigation or not.

- The employer learns whether the employee asked the prosecutor to conduct an investigation or not. The employer decides whether he dismisses the employee or not. [In *P-R* only: However, dismissal is only possible in case that the employee did not ask the prosecutor to conduct an investigation.]

- All participants learn their individual payoffs from the current round, and the decisions leading to these payoffs.

- Behavior in a given round does not affect earnings in upcoming rounds.

5. Total earnings for the first part of the experiment

At the end of both parts of the experiment, three rounds out of the total of 30 rounds will be selected randomly and independently from each other. The points that you have earned in these three rounds will be summed up and exchanged into EURO. The exchange rate is 1 EURO = 15 points. The resulting payoff plus the show-up fee of 12 EURO plus your earnings from the second part of the experiment will then constitute your overall payoff from the experiment.

C Robustness Checks for Treatments *NoP* and *P-R*

In this Appendix, we document that our results for treatments *NoP* and *P-R* as presented in Table 1 are robust when (i) considering alternative statistical specifications, (ii) comparing behavior in late and early periods, and (iii) including personal characteristics as additional controls.

C.1 Alternative Statistical Specifications

Unit of Observation In the main regressions, we have used individual decisions as the unit of observation (with clustering at the session level). Recall that in each experimental session, each subject played 30 periods in a given treatment. Hence, a subject might be observed more than once in a given role, for example as an L-employee. Consequently, we have also performed the regression analysis with the unit of observation formed from the subjects' average behavior in the respective role. As can be seen from Table 3, this leads to virtually identical results.

Table 3: Robustness of Results for Treatments *NoP* and *P-R*: Using Observations at the Aggregate Level

	(1) Dismiss (*NoP*)	(2) Dismiss (*P-R*, No Rep.)	(3) Dismiss (No Rep.)	(4) Report (*NoP*)	(5) Report (*P-R*)	(6) Report (H-emp.)	(7) Report (L-emp.)	(8) Investigate	(9) Misbehave
H-employee	-0.960***	-0.902***	-0.960***	-0.0946*	-0.429***				0.0518
	(0.000)	(0.000)	(0.000)	(0.075)	(0.001)				(0.310)
Report	0.0129							0.756***	
	(0.513)							(0.000)	
Report x H-emp.	0.241*								
	(0.081)								
P-R			-0.0620			0.246***	0.580***	0.0325	0.00463
			(0.116)			(0.002)	(0.000)	(0.615)	(0.921)
P-R x H-emp.			0.0584						0.123
			(0.125)						(0.185)
Misbehavior				0.524***	0.158**	0.471***	0.524***		
				(0.001)	(0.025)	(0.000)	(0.000)		
Misb. x H-emp.				-0.0530*	0.349**				
				(0.085)	(0.014)				
Misb. x P-R						0.0360	-0.366***		
						(0.733)	(0.000)		
P-R x Report								-0.317***	
								(0.007)	
Observations	119	55	115	180	180	180	180	180	120
Adjusted R^2	0.860	0.859	0.924	0.375	0.412	0.397	0.544	0.607	0.022
F-Test1	0.002		0.000	0.042	0.269	0.002	0.007	0.001	0.036
F-Test2	0.042		0.766	0.000	0.007	0.000	0.005	0.001	0.162

Notes: The explanation below Table 1 applies. The only difference is that the analysis is based on aggregate data instead of individual data.

Non-Parametric Approach We have also verified that the findings of Table 1 are robust when conducting non-parametric testing instead, and indeed the results are virtually identical (see Table 4). For these tests, the unit of observation is again the respective average at the subject-role level (see the previous paragraph).

Table 4: Robustness of Results for Treatments *NoP* and *P-R*: Non-Parametric Tests

	Dismissal	***NoP***	***P-R***		**Reporting**	***NoP***	***P-R***	
		(A)	(B)			(C)	(D)	
				Columns (A) & (B)				Columns (C) & (D)
	L-employee				**L-employee**			
(1)	Report	.99	n/a	n/a	Misbehavior	.76	.97	***
(2)	No Report	.97	.88	n.s.	No Misbehavior	.23	.81	***
	Rows (1) & (2)	n.s.	n/a		Rows (1) & (2)	***	***	
	H-employee				**H-employee**			
(3)	Report	.3	n/a	n/a	Misbehavior	.64	.89	***
(4)	No Report	.01	.01	n.s.	No Misbehavior	.14	.38	***
	Rows (3) & (4)	***	n/a		Rows (3) & (4)	***	***	
	Rows (1) & (3)	***	n/a		Rows (1) & (3)	**	**	
	Rows (2) & (4)	***	***		Rows (2) & (4)	**	***	

	Investigation	***NoP***	***P-R***		**Misbehavior**	***NoP***	***P-R***	
		(A)	(B)			(C)	(D)	
				Columns (A) & (B)				Columns (C) & (D)
(5)	Report	.93	.65	***	L-employee	.38	.39	n.s.
(6)	No Report	.19	.18	n.s.	H-employee	.42	.56	n.s.
	Rows (5) & (6)	n.s.	***		Rows (5) & (6)	n.s.	***	

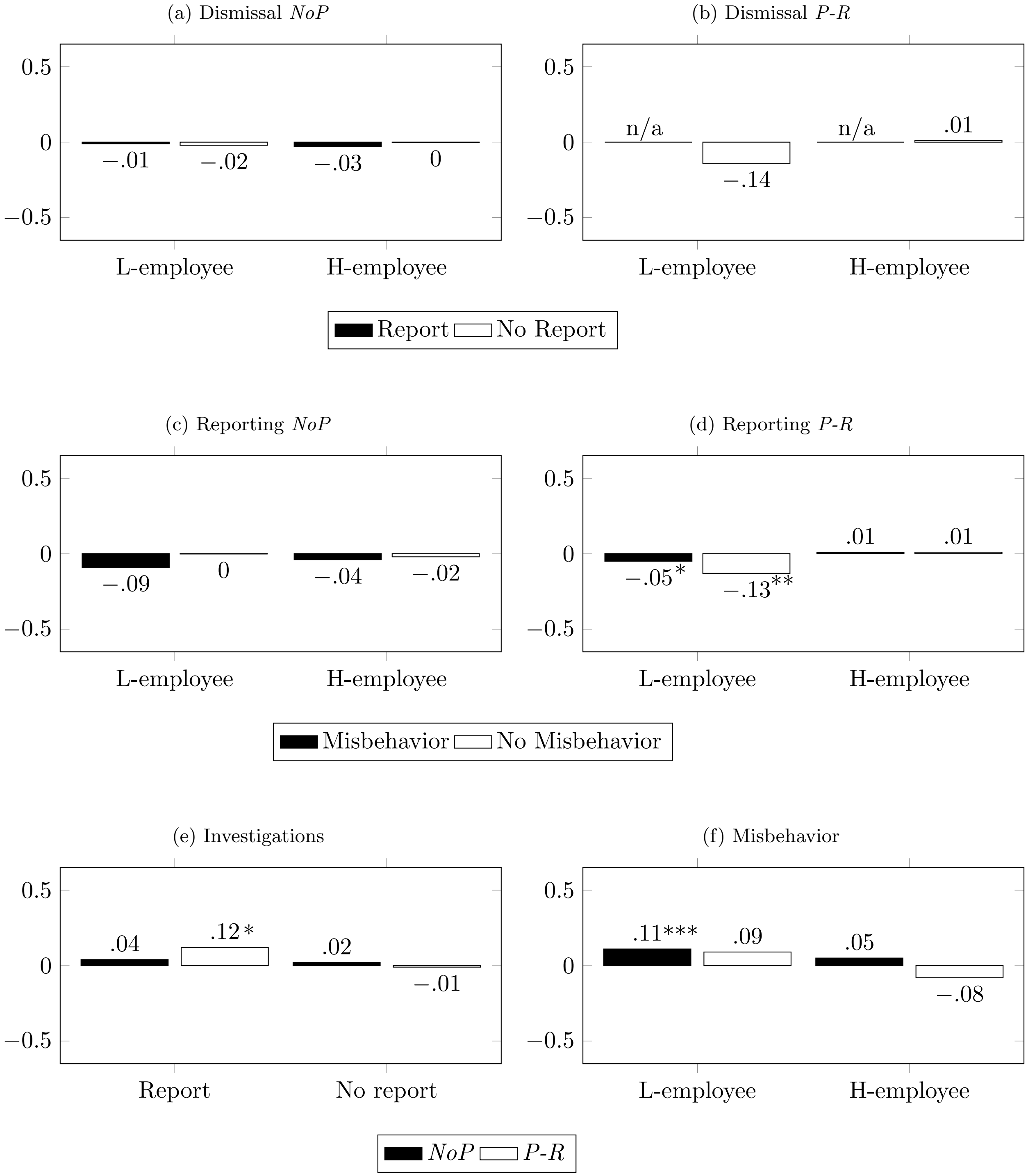
Notes: The values in the table are those reported in Figure 2. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively, and "n.s." indicates a lack of statistical significance. The entries with "n/a" are due to the fact that a dismissal is not feasible in treatment *P-R* when there is a report. For within-treatment comparisons (i.e., when comparing rows), we employ the Wilcoxon Signed-Rank test. For between-treatment comparisons (i.e., when comparing columns), we employ the Mann-Whitney-U test. In all tests, the unit of observation is a subject's average behavior in the respective role and condition: For the example of dismissals, a subject in the role of employer may be observed in four conditions: with either an L- or an H-employee, who either has or has not reported. For each of these four conditions, the respective employer's average behavior is one observation. We proceed analogously for employees and prosecutors. We have also conducted all non-parametric tests with averages taken on the session-role level under the different conditions as the unit of observation (rather than on the subject-role level). While this substantially reduces the number of observations, our results are remarkably robust: All previously statistically significant between-treatment comparisons remain significant at the 1% level (investigations, reporting behavior of both employee types given no misbehavior, and of L-employees given misbehavior) or the 4% level (reporting of H-employees given misbehavior). Also, the previously statistically significant within-treatment comparisons remain significant at the 7% level.

Controlling for Multiple Hypotheses The hypotheses underlying the empirical tests in Table 1 are derived from an explicit theoretical model. Nevertheless, we have also considered the issue of "multiple hypothesis testing" by controlling for the family-wise error rate (i.e., the probability of one or more false discoveries). In particular, we apply the procedure by Holm (1979), which assumes a worst-case dependence structure of the test statistics (for overviews, see Romano, Shaikh, and Wolf, 2010, and List, Shaikh, and Xu, 2016). In main Table 1, we tested a total of 25 hypotheses with respect to within- and between treatment difference, or the lack thereof (see Footnote 27). Out of these, 19 were confirmed in the empirical analysis. When applying the Holm (1979) procedure, the number of confirmed hypotheses in fact remains the same: More precisely, there are two cases where a difference is predicted, which vanishes after applying the Holm (1979) procedure. At the same, there are also two cases where no difference is predicted, and where the previously obtained significant effect vanishes after applying the Holm (1979) procedure. Details are available upon request.

C.2 Comparing Behavior in Early and Late Periods

As discussed in Section 3, we have strived to ensure that subjects understand the underlying game and the incentive structure before actual play started. Given the complexity of the game, it might still be the case that subjects learn over the course of the 30 periods of a given session. As a result, behavior in later periods might, in principle, differ systematically from behavior in earlier periods. However, Figure 6 illustrates that there do not seem to exist systematic time effects. In particular, this figure has the same structure as Figure 2 above, but displays the differences in average behavior between the first 20 periods and the last 10 periods. As can be seen, experimental behavior exhibits only minor and unsystematic changes over time: Panel (d) shows a slight increase in reporting over time by L-employees in *P-R*, thereby moving somewhat closer towards the theoretical prediction. Also in *P-R*, panel (e) shows that the responsiveness of prosecutors to reports declines over time, thereby reinforcing this deviation from the prediction as discussed in the main text. Finally, panel (f) shows that in *NoP*, there is less misbehavior by employers matched with L-employees in later periods. The respective regression results are reported in Table 5. It shows the adapted baseline regressions, where time effects are captured by a dummy variable *Early* that takes the value 1 for periods 1-20 and zero otherwise. We have also run regressions where we capture time effects by including a linear time trend (and the respective interactions). This does not affect the results qualitatively.

Figure 6: Robustness of Results for Treatments *NoP* and *P-R*: Early Versus Late Periods



(a) Dismissal *NoP*

(b) Dismissal *P-R*

(c) Reporting *NoP*

(d) Reporting *P-R*

(e) Investigations

(f) Misbehavior

Notes: The entries in the figure show the difference between average behavior over time (periods 1-20 minus periods 21-30). The stars indicate significant differences between behavior in periods 1-20 (early) and period 21-30 (late), where *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. The statistical tests are based on the regression results reported in Table 5. For example, consider panel (f) of the present figure: The difference of 0.11 between the behavior in early and late periods is statistically significant because the coefficient for the dummy variable *Early* in Table 5, column (9), is significant ($p = 0.001$, t-test). On the other hand, 0.05 is insignificant, because *Early* and the interaction term *Early* × *H-emp.* in the same column are not jointly significant ($p = 0.460$, F-Test). All other tests are performed in an analogous way.

Table 5: Robustness of Results for Treatments *NoP* and *P-R*: Early Versus Late Periods

	(1) Dismiss (*NoP*)	(2) Dismiss (*P-R*, No Rep.)	(3) Dismiss (No Rep.)	(4) Report (*NoP*)	(5) Report (*P-R*)	(6) Report (H-emp.)	(7) Report (L-emp.)	(8) Investigate	(9) Misbehave
H-employee	-0.978*** (0.000)	-1.000 (.)	-0.978*** (0.000)	-0.0825 (0.171)	-0.527*** (0.001)				0.0820 (0.250)
Report	0.0115 (0.374)							0.731*** (0.000)	
Report x H-emp.	0.299*** (0.007)								
P-R			0.0115 (0.317)			0.231*** (0.002)	0.676*** (0.000)	0.0145 (0.893)	0.0193 (0.812)
P-R x H-emp.			-0.0217 (0.110)						0.196 (0.150)
Misbehavior				0.591*** (0.003)	0.0960* (0.068)	0.510*** (0.000)	0.591*** (0.000)		
Misb. x H-emp.				-0.0807 (0.270)	0.413*** (0.010)				
Misb. x P-R						-0.00136 (0.991)	-0.495*** (0.001)		
P-R x Report								-0.347** (0.017)	
Early	-0.0232 (0.201)	-0.137 (0.173)	-0.0232 (0.140)	0.00356 (0.928)	-0.129** (0.011)	-0.0159 (0.474)	0.00356 (0.921)	0.0176 (0.657)	0.107*** (0.001)
Early x H-emp.	0.0236 (0.343)	0.146 (0.145)	0.0236 (0.284)	-0.0194 (0.528)	0.138*** (0.004)				-0.0593 (0.301)
Early x Report	0.00910 (0.631)							0.0192 (0.610)	
Early x Report x H-emp.	-0.0386 (0.649)								
Early x P-R			-0.114 (0.184)			0.0248 (0.550)	-0.133** (0.015)	-0.0272 (0.707)	-0.0161 (0.873)
Early x P-R x H-emp.			0.123 (0.154)						-0.106 (0.451)
Early x Misb.				-0.0954 (0.188)	0.0836* (0.078)	-0.0152 (0.676)	-0.0954 (0.126)		
Early x Misb. x H-emp.				0.0802 (0.280)	-0.0795** (0.046)				
Early x Misb. x P-R						0.0193 (0.665)	0.179** (0.024)		
Early x P-R x Report								0.109 (0.169)	
Observations	900	227	774	1800	1800	1854	1746	1800	1800
Adjusted R^2	0.804	0.820	0.909	0.279	0.294	0.318	0.367	0.353	0.024
F-Test1	0.000		.	0.105	0.130	0.020	0.099	0.008	0.027
F-Test2	0.005		0.306	0.001	0.008	0.001	0.027	0.003	0.090
F-Test-Early1	0.983	0.380	0.981	0.520	0.816	0.798	0.001	0.470	0.460
F-Test-Early2	0.193		0.112	0.286	0.066	0.466	0.225	0.874	0.368
F-Test-Early3	0.704		0.322	0.512	0.391	0.334	0.026	0.069	0.234

Notes: The explanation below Table 1 applies. In addition, the dummy variable *Early* takes the value 0 for periods 21 to 30, and the value 1 for periods 1 to 20. The entries in row "F-Test-Early1" ("F-Test-Early2") show the p-values for the joint significance of *Early* and the first (second) interaction term below *Early*. The entries in row "F-Test-Early3" show the p-values for the joint significance of *Early*, the two interaction terms, and the corresponding triple interaction term.

C.3 Controlling for Personal Characteristics

In Table 6, we check the robustness of our main findings when controlling for personal characteristics, as elicited in a post-experimental questionnaire (available upon request). These are: (i) socio-demographic information: *Age* (in years), *Male* (a dummy), and *Econ* (a dummy indicating whether the subject majors in economics or a related field), (ii) *Risk Aversion* (measured on a 10-point scale through the “100.000 Euro question” of Dohmen, Falk, Huffman, Sunde, Schupp, and Wagner, 2011), (iii) *CRT* (measured on a 4-point scale through the “Cognitive Reflection Test” of Frederick, 2005), (iv) *WB Attitude* (measured on a 5-point scale through multiple questions to infer a subject’s attitude towards revealing misbehavior), (v) *Dutifulness* (measured on a 5-point scale through the respective sub-factor of the Big Five personality trait “Conscientiousness” in the“NEO Personality Inventory”, see Costa and McCrae, 1992; Berth and Goldschmidt, 2006). As to make these questions not too salient, they were interspersed with some unrelated questions. Finally, *Offer* measures a subject’s social preferences (on a scale of 0-100), which were elicited through an incentivized dictator game. As Table 6 shows, the main findings of Table 1 are remarkably robust.

Table 6: Robustness of Results for Treatments *NoP* and *P-R*: Personal Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Dismiss	Dismiss	Dismiss	Report	Report	Report	Report	Investigate	Misbehave
	(*NoP*)	(*P-R*, No Rep.)	(No Rep.)	(*NoP*)	(*P-R*)	(H-emp.)	(L-emp.)		
H-employee	-0.959***	-0.871***	-0.965***	-0.0923	-0.428***				0.0405
	(0.000)	(0.000)	(0.000)	(0.116)	(0.001)				(0.317)
Report	0.0193							0.758***	
	(0.256)							(0.000)	
Report x H-emp.	0.275**								
	(0.043)								
P-R			-0.110			0.215***	0.551***	0.00118	0.00164
			(0.131)			(0.008)	(0.000)	(0.987)	(0.979)
P-R x H-emp.			0.0948						0.125
			(0.169)						(0.152)
Misbehavior				0.526***	0.154**	0.500***	0.526***		
				(0.001)	(0.025)	(0.000)	(0.000)		
Misb. x H-emp.				-0.0259	0.357**				
				(0.451)	(0.016)				
Misb. x P-R						0.0112	-0.372***		
						(0.918)	(0.000)		
P-R x Report								-0.288**	
								(0.014)	
Observations	900	227	774	1800	1800	1854	1746	1800	1800
Adjusted R^2	0.809	0.824	0.909	0.327	0.322	0.353	0.392	0.381	0.099
F-Test1	0.001		0.000	0.071	0.326	0.008	0.005	0.001	0.046
F-Test2	0.023		0.141	0.000	0.008	0.001	0.005	0.000	0.179
Add. Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The explanation below Table 1 applies. The only difference is that personal characteristics as explained above are included as additional controls.

Next, we investigate the effect of personal characteristics on the various decisions in more detail. In particular, a subject's personal characteristics might affect behavior differentially depending on what happened earlier in the game (as observed by the subject). Therefore, for each decision, we have run separate regressions for each possible history of prior decisions. The results are shown in Table 7. For example, when the employer takes the dismissal decision, she is aware of the prior misbehavior and reporting decisions. Note that when there is a report, a dismissal is only possible in *NoP*, which explains why Table 7, columns (1) and (2), only consider this treatment. All in all, personal characteristics do not seem to have clear and systematic effects.

Table 7: The Role of Personal Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Dismissal	Dismissal	Dismissal	Dismissal	Report	Report	Investigate	Investigate	Misbehave
	(*NoP*, Rep.,	(*NoP*, Rep.,	(No Rep.,	(No Rep.,	(Misb.)	(No Misb.)	(Rep.)	(No Rep.)	
	Misb.)	No Misb.)	Misb.)	No Misb.)					
H-employee	-0.590**	-0.976***	-0.893***	-0.977***	-0.119**	-0.0850*			0.0405
	(0.013)	(0.000)	(0.000)	(0.000)	(0.026)	(0.055)			(0.317)
P-R			0.0105	-0.124	0.198***	0.529***	-0.268***	-0.0189	0.00164
			(0.919)	(0.121)	(0.010)	(0.000)	(0.000)	(0.745)	(0.979)
P-R x H-emp.			0.00626	0.102	0.0464	-0.339***			0.125
			(0.948)	(0.195)	(0.547)	(0.000)			(0.152)
Age	-0.00234	-0.00404	0.000156	-0.00643*	0.00625	0.00383	0.00616	0.0107	-0.00612
	(0.781)	(0.386)	(0.985)	(0.052)	(0.140)	(0.555)	(0.421)	(0.113)	(0.681)
Male	-0.00270	-0.0193	0.0160	0.0177	0.00650	-0.192*	-0.0501	-0.0313	0.104
	(0.977)	(0.381)	(0.730)	(0.155)	(0.872)	(0.059)	(0.531)	(0.689)	(0.133)
Econ	0.168**	-0.0117	-0.0478*	0.0300**	-0.0662	-0.0385	0.000106	0.0110	-0.0270
	(0.032)	(0.394)	(0.061)	(0.038)	(0.203)	(0.577)	(0.999)	(0.856)	(0.713)
Risk Aversion	-0.0135	-0.00391	0.0164***	-0.00591*	-0.00331	-0.0218	0.0149**	-0.0155	-0.00772
	(0.728)	(0.431)	(0.001)	(0.066)	(0.712)	(0.180)	(0.017)	(0.147)	(0.515)
CRT	0.0649	0.0144	0.00104	-0.0122	-0.00804	-0.0189	0.0310	-0.0483*	-0.0406*
	(0.119)	(0.263)	(0.939)	(0.188)	(0.703)	(0.538)	(0.191)	(0.080)	(0.087)
WB Attitude	-0.0418	-0.00417	0.0149	0.00427	-0.0573	-0.110**	0.0263	0.0412	-0.0592
	(0.631)	(0.527)	(0.642)	(0.720)	(0.115)	(0.026)	(0.557)	(0.439)	(0.428)
Dutifulness	-0.0456	-0.0435	-0.00430	-0.0209	0.00138	-0.0975	-0.0103	0.0417	0.124
	(0.657)	(0.361)	(0.899)	(0.251)	(0.984)	(0.129)	(0.865)	(0.727)	(0.154)
Offer	0.00174	0.000375	-0.00146	-0.000334	-0.000201	-0.000348	0.00408**	0.00216*	-0.00405**
	(0.435)	(0.245)	(0.139)	(0.282)	(0.657)	(0.751)	(0.032)	(0.079)	(0.034)
Observations	251	102	152	622	1800	1800	1026	774	1800
Adjusted R^2	0.498	0.955	0.828	0.928	0.120	0.328	0.163	0.053	0.099

Notes: Each column refers to a linear probability model with the respective underlying decision (dismissal, reporting, investigation, misbehavior) as the dependent variable. All regressions use individual observations with standard errors clustered at the session level, p-values are reported in parentheses, and *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. When a qualifier is stated in parenthesis in the header of a regression, it refers to the subset of observations used; in particular, "Rep." ("No Rep.") indicates that only observations are used where a report (no report) is sent. Analogously, "Misb." ("No Misb.") indicates that only observations are used where there is misbehavior (no misbehavior) by employers. When no such qualifier is stated, the regression uses all observations for the respective decision from both treatments *NoP* and *P-R*. As for columns (1) and (2), note that an employee who sends a report can only be dismissed in *NoP*, which explains the restriction to this treatment.

D Regression Results for Section 5

This Appendix contains the regression tables for the treatment comparisons discussed in Section 5. For the sake of comparability, they all have the same basic structure as main Table 1.

Table 8: Regression Results for Treatments *P-RI* and *P-R*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Dismiss	Dismiss	Dismiss	Report	Report	Report	Report	Investigate	Misbehave
	(*P-R*)	(*P-RI*)		(*P-R*)	(*P-RI*)	(H-emp.)	(L-emp.)		
H-employee	-0.871***	-0.882***	-0.871***	-0.431***	-0.452**				0.167*
	(0.000)	(0.000)	(0.000)	(0.001)	(0.033)				(0.058)
Report		0.118**						0.469***	
		(0.011)						(0.001)	
Report x H-emp.		-0.118**							
		(0.011)							
P-RI			0.0519			-0.118	-0.0973	-0.0417	-0.0840
			(0.491)			(0.182)	(0.265)	(0.485)	(0.155)
P-RI x H-emp.			-0.0578						0.0899
			(0.437)						(0.365)
Misbehavior				0.154**	0.184**	0.511***	0.154***		
				(0.024)	(0.037)	(0.001)	(0.006)		
Misb. x H-emp.				0.357**	0.365**				
				(0.016)	(0.016)				
Misb. x P-RI						0.0376	0.0297		
						(0.727)	(0.649)		
P-RI x Report								0.181*	
								(0.092)	
Observations	227	310	537	1800	1320	1638	1482	1560	1560
Adjusted R^2	0.819	0.890	0.865	0.290	0.271	0.297	0.068	0.256	0.044
F-Test1		.	0.000	0.311	0.249	0.377	0.154	0.000	0.002
F-Test2		.	0.319	0.008	0.000	0.000	0.005	0.041	0.930

Notes: The explanation below Table 1 applies. In columns (1)–(3), we consider observations where a dismissal is feasible. Note that no F-tests are reported in column (2). The reason is that in treatment *P-RI*, whenever a dismissal is possible, L-employees are virtually always dismissed and H-employees are always retained.

Table 9: Regression Results for Treatments *P-RI* and *P-RIM*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Dismiss	Dismiss	Dismiss	Report	Report	Report	Report	Investigate	Misbehave
	(*P-RI*)	(*P-RIM*)		(*P-RI*)	(*P-RIM*)	(H-emp.)	(L-emp.)		
H-employee	-0.882***	-0.951***	-0.929***	-0.452**	-0.136**				0.257***
	(0.000)	(0.000)	(0.000)	(0.033)	(0.046)				(0.003)
Report	0.118**	0.0114						0.649***	
	(0.011)	(0.716)						(0.000)	
Report x H-emp.	-0.118**	-0.0339							
	(0.011)	(0.431)							
P-RIM			0.0474			-0.214**	-0.530***	-0.0106	-0.112**
			(0.154)			(0.012)	(0.000)	(0.779)	(0.050)
P-RIM x H-emp.			-0.0278						0.0102
			(0.422)						(0.875)
Misbehavior				0.184**	0.733***	0.549***	0.184***		
				(0.037)	(0.001)	(0.000)	(0.006)		
Misb. x H-emp.				0.365**	0.0199				
				(0.016)	(0.835)				
Misb. x P-RIM						0.204*	0.549***		
						(0.069)	(0.000)		
P-RIM x Report								0.0171	
								(0.742)	
Observations	310	554	864	1320	1440	1400	1360	1380	1380
Adjusted R^2	0.890	0.915	0.906	0.271	0.567	0.454	0.407	0.434	0.082
F-Test1	.	0.000	0.000	0.249	0.232	0.918	0.697	0.000	0.000
F-Test2	.	0.096	0.040	0.000	0.004	0.000	0.000	0.872	0.073

Notes: The explanation below Table 1 applies. In columns (1)–(3), we consider observations where a dismissal is feasible. Note that no F-tests are reported in column (1). The reason is that in treatment *P-RI*, whenever a dismissal is possible, L-employees are virtually always dismissed and H-employees are always retained.

Table 10: Regression Results for Treatments *NoP* and *P-RIM*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Dismiss	Dismiss	Dismiss	Report	Report	Report	Report	Investigate	Misbehave
	(*NoP*)	(*P-RIM*)		(*NoP*)	(*P-RIM*)	(H-emp.)	(L-emp.)		
H-employee	-0.963***	-0.951***	-0.871***	-0.0953*	-0.136**				0.0406
	(0.000)	(0.000)	(0.000)	(0.097)	(0.046)				(0.317)
Report	0.0171	0.0114						0.745***	
	(0.257)	(0.716)						(0.000)	
Report x H-emp.	0.274**	-0.0339							
	(0.040)	(0.431)							
P-RIM			-0.00409			-0.0842**	-0.0438	-0.0563	-0.186***
			(0.776)			(0.040)	(0.439)	(0.244)	(0.003)
P-RIM x H-emp.			-0.0865**						0.226***
			(0.021)						(0.001)
Misbehavior				0.526***	0.733***	0.500***	0.526***		
				(0.001)	(0.001)	(0.000)	(0.000)		
Misb. x H-emp.				-0.0259	0.0199				
				(0.450)	(0.835)				
Misb. x P-RIM						0.253**	0.207**		
						(0.032)	(0.024)		
P-RIM x Report								-0.0781*	
								(0.081)	
Observations	900	554	1454	1800	1440	1616	1624	1620	1620
Adjusted R^2	0.805	0.915	0.824	0.278	0.567	0.411	0.391	0.496	0.038
F-Test1	0.001	0.000	0.000	0.049	0.232	0.067	0.033	0.000	0.000
F-Test2	0.022	0.096	0.008	0.000	0.004	0.000	0.000	0.008	0.560

Notes: The explanation below Table 1 applies. In columns (1)–(3), we consider observations where a dismissal is feasible.

E Regression Results for Section 6

This Appendix contains the regression tables for the treatment comparisons discussed in Section 6. For the sake of comparability, they all have the same basic structure as main Table 1.

Table 11: Regression Results for Treatments *NoP-Low* and *P-R-Low*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Dismiss	Dismiss	Dismiss	Report	Report	Report	Report	Investigate	Misbehave
	(*NoP-Low*)	(*P-R-Low*, No Rep.)	(No Rep.)	(*NoP-Low*)	(*P-R-Low*)	(H-emp.)	(L-emp.)		
H-employee	-0.168**	-0.266***	-0.168**	-0.00216	-0.134**				-0.0633
	(0.043)	(0.001)	(0.011)	(0.830)	(0.038)				(0.167)
Report	0.586***							0.613***	
	(0.000)							(0.000)	
Report x H-emp.	-0.305***								
	(0.009)								
P-R-Low			0.0844			0.427***	0.559***	-0.158*	-0.184*
			(0.210)			(0.000)	(0.000)	(0.087)	(0.051)
P-R-Low x H-emp.			-0.0973						0.110*
			(0.166)						(0.097)
Misbehavior				0.423***	0.333**	0.466***	0.423***		
				(0.002)	(0.011)	(0.000)	(0.000)		
Misb. x H-emp.				0.0426	0.104*				
				(0.368)	(0.097)				
Misb. x P-R-Low						-0.0282	-0.0899		
						(0.773)	(0.384)		
P-R-Low x Report								-0.126	
								(0.332)	
Observations	840	292	931	1680	1800	1702	1778	1740	1740
Adjusted R^2	0.320	0.166	0.107	0.246	0.203	0.373	0.408	0.239	0.019
F-Test1	0.010		0.000	0.377	0.058	0.000	0.000	0.000	0.296
F-Test2	0.023		0.043	0.002	0.003	0.000	0.002	0.003	0.410

Notes: The explanation below Table 1 applies.

Table 12: Regression Results for Treatments *NoP* and *NoP-Low*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Dismiss	Dismiss	Dismiss	Dismiss	Report	Report	Report	Report	Investigate	Misbehave
	(*NoP*)	(*NoP-Low*)	(No Rep.)	(Rep.)	(*NoP*)	(*NoP-Low*)	(H-emp.)	(L-emp.)		
H-employee	-0.963***	-0.168**	-0.963***	-0.688***	-0.0953*	-0.00216				0.0406
	(0.000)	(0.043)	(0.000)	(0.000)	(0.097)	(0.830)				(0.311)
Report	0.0171	0.586***							0.745***	
	(0.257)	(0.000)							(0.000)	
Report x H-emp.	0.274**	-0.305***								
	(0.040)	(0.009)								
NoP-Low			-0.792***	-0.223***			-0.0817**	-0.175***	0.138	0.121
			(0.000)	(0.002)			(0.044)	(0.008)	(0.124)	(0.149)
NoP-Low x H-emp.			0.794***	0.215						-0.104
			(0.000)	(0.124)						(0.102)
Misbehavior					0.526***	0.423***	0.500***	0.526***		
					(0.001)	(0.002)	(0.000)	(0.000)		
Misb. x H-emp.					-0.0259	0.0426				
					(0.450)	(0.368)				
Misb. x NoP-Low							-0.0341	-0.103		
							(0.631)	(0.232)		
NoP-Low x Report									-0.132	
									(0.213)	
Observations	900	840	1186	554	1800	1680	1692	1788	1740	1740
Adjusted R^2	0.805	0.320	0.730	0.433	0.278	0.246	0.271	0.295	0.406	0.006
F-Test1	0.001	0.010	0.012	0.001	0.049	0.377	0.083	0.004	0.000	0.172
F-Test2	0.022	0.023	0.778	0.942	0.000	0.002	0.000	0.000	0.889	0.817

Notes: The explanation below Table 1 applies. Since protection is available in neither treatment, the employee can always be dismissed in both treatments.

Table 13: Regression Results for Treatments *P-R* and *P-R-Low*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Dismiss	Dismiss	Dismiss	Report	Report	Report	Report	Investigate	Misbehave
	(*P-R*, No Rep.)	(*P-R-Low*, No Rep.)	(No Report)	(*P-R*)	(*P-R-Low*)	(H-emp.)	(L-emp.)		
H-employee	-0.871***	-0.266***	-0.871***	-0.431***	-0.134**				0.167**
	(0.000)	(0.001)	(0.000)	(0.001)	(0.038)				(0.049)
P-R-Low			-0.612***			0.0978	-0.200**	-0.0162	-0.0731
			(0.000)			(0.266)	(0.046)	(0.809)	(0.259)
P-R-Low x H-emp.			0.606***						-0.120
			(0.000)						(0.194)
Misbehavior				0.154**	0.333**	0.511***	0.154***		
				(0.024)	(0.011)	(0.000)	(0.004)		
Misb. x H-emp.				0.357**	0.104*				
				(0.016)	(0.097)				
Misb. x P-R-Low						-0.0736	0.179*		
						(0.567)	(0.076)		
Report								0.469***	
								(0.000)	
P-R-Low x Report								0.0182	
								(0.879)	
Observations	227	292	519	1800	1800	1864	1736	1800	1800
Adjusted R^2	0.819	0.166	0.546	0.290	0.203	0.260	0.146	0.187	0.034
F-Test1			0.000	0.311	0.058	0.684	0.317	0.000	0.296
F-Test2			0.307	0.008	0.003	0.000	0.002	0.981	0.065

Notes: The explanation below Table 1 applies.