DISCUSSION PAPER SERIES

DP11898

WHISTLE-BLOWER PROTECTION: THEORY AND EXPERIMENTAL EVIDENCE

Lydia Mechtenberg, Gerd Muehlheusser and Andreas Roider

FINANCIAL ECONOMICS, INDUSTRIAL ORGANIZATION, LABOUR ECONOMICS and PUBLIC ECONOMICS

ISSN 0265-8003

WHISTLE-BLOWER PROTECTION: THEORY AND EXPERIMENTAL EVIDENCE

Lydia Mechtenberg, Gerd Muehlheusser and Andreas Roider

Discussion Paper DP11898
Published 10 March 2017
Submitted 10 March 2017

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

WHISTLE-BLOWER PROTECTION: THEORY AND EXPERIMENTAL EVIDENCE

Abstract

Whistle-blowing by employees plays a major role in uncovering corporate fraud. Various recent laws aim at improving protection of whistle-blowers and enhancing their willingness to report. Evidence on the effectiveness of such legislation is, however, scarce. Moreover, critics have raised worries about fraudulent claims by low-productivity employees. We study these issues in a theory-guided lab experiment. Easily attainable ("belief-based") protection indeed leads to more reports, both truthful and fraudulent. Fraudulent claims dilute prosecutors' incentives to investigate, and thereby hamper deterrence. These effects are ameliorated under more stringent ("fact-based") protection.

JEL Classification: C91, D83, D73, K42, M59

Keywords: Corporate Fraud, Corruption, Whistle-Blowing, Business Ethics, Cheap-Talk Games, Lab Experiment

Lydia Mechtenberg - lydia.mechtenberg@uni-hamburg.de
University of Hamburg

Gerd Muehlheusser - gerd.muehlheusser@uni-hamburg.de
University of Hamburg, IZA, CESifo

Andreas Roider - andreas.roider@ur.de
University of Regensburg, IZA, CESifo and CEPR

Acknowledgements
We gratefully acknowledge financial support by the Fritz Thyssen Foundation (Grant 10.13.2.097). We also thank Ralph Bayer, Matthew Ellman, Eberhard Feess, Leonie Gerhards, Luise Goerges, Johann Graf Lambsdorff, Igor Legkiy, Niklas Wallmeier, and seminar participants at Aarhus, Bonn, Fribourg, Hamburg, Innsbruck, Konstanz, Marburg, Paris X (Nanterre), Passau, Portsmouth, and Regensburg for their comments and suggestions. Christos Litsios and Pamela Mertens provided excellent research assistance.

Whistle-Blower Protection: Theory and Experimental Evidence*

Lydia Mechtenberg† Gerd Muehlheusser‡ Andreas Roider§

March 6, 2017

Abstract

Whistle-blowing by employees plays a major role in uncovering corporate fraud. Various recent laws aim at improving protection of whistle-blowers and enhancing their willingness to report. Evidence on the effectiveness of such legislation is, however, scarce. Moreover, critics have raised worries about fraudulent claims by low-productivity employees. We study these issues in a theory-guided lab experiment. Easily attainable ("belief-based") protection indeed leads to more reports, both truthful and fraudulent. Fraudulent claims dilute prosecutors' incentives to investigate, and thereby hamper deterrence. These effects are ameliorated under more stringent ("fact-based") protection.

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

Keywords: Corporate Fraud, Corruption, Whistle-Blowing, Business Ethics, Cheap-Talk Games, Lab Experiment

*We gratefully acknowledge financial support by the Fritz Thyssen Foundation (Grant 10.13.2.097). We also thank Ralph Bayer, Matthew Ellman, Eberhard Feess, Leonie Gerhards, Luise Görges, Johann Graf Lambsdorff, Igor Legkiy, Niklas Wallmeier, and seminar participants at Aarhus, Bonn, Fribourg, Hamburg, Innsbruck, Konstanz, Marburg, Paris X (Nanterre), Passau, Portsmouth, and Regensburg 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

‡Corresponding author: 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 whistle-blowers play an important role in uncovering it. In response to recent scandals, and in order to encourage employee whistle-blowing and to increase deterrence, some countries have enacted comprehensive whistle-blower protection laws that grant rather easy access to protection. In contrast, in other countries, the requirements for obtaining protection are more stringent. In this paper, we conduct a theory-guided lab experiment in which we analyze the impact of introducing whistle-blower protection. In particular, we compare different legal regimes ("belief-based" versus "fact-based") with respect to their effects on employers' misbehavior, employees' truthful and fraudulent reports, prosecutors' investigations, and employers' retaliation. Belief-based regimes have less stringent requirements for granting protection to whistle-blowers than fact-based regimes. Our results suggest that the latter lead to better outcomes in terms of reporting behavior and deterrence.

The topicality of corporate fraud is well-documented. At an anecdotal level, this is exemplified by recent scandals at Volkswagen, Enron, or Worldcom. More systematic evidence is presented in a recent study by the Association of Certified Fraud Examiners (2014), according to which 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 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. Moreover, in 81% of those 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. This evidence suggests that policies curbing corporate fraud should be of first order importance for legislators and policy makers.

In recent years, employee whistle-blowers (who are not participating in the misbehavior) have been recognized as a powerful source of uncovering fraud involving company insiders, primarily because of their access to crucial information. In fact, in a number of high-profile corporate fraud scandals (such as Enron or Worldcom) the misbehavior was uncovered by

employee whistle-blowers.[1] There is also systematic evidence for the importance of employee whistle-blowers. For example, Dyck, Morse, and Zingales (2010) consider all reported cases of fraud in large U.S. companies between 1996 and 2004. They find that in 17% of the 216 cases they study, the fraud was uncovered by employee whistle-blowers; thereby outnumbering other players such as the SEC, auditors, non-financial market regulators, or the media. According to the Association of Certified Fraud Examiners (2014), employees were the source in 49% of tips leading to the detection of fraud.[2]

Previous research has identified a number of factors such as "conscience cleansing" that are deemed crucial for whistle-blowers' decision to come forward (see e.g., Jos et al., 1989; Miceli and Near, 1992; Alford, 2001). However, there are also strong countervailing factors, in particular the fear of retaliation from co-workers or management (see e.g., Near and Miceli, 1986; Alford, 2001; Near et al., 2004; Rehg et al., 2008). For this reason, the overall willingness of employees to report misbehavior is often perceived as rather low. As a consequence, whistle-blowers might be encouraged to come forward by legally protecting them from retaliation. As documented in the extensive comparative law study by Thüsing and Forst (2016), many countries have enacted legislation to that end, where prominent examples include the influential Sarbanes-Oxley Act (SOX) and Dodd-Frank Act in the U.S., or the Public Interest Disclosure Act (PIDA) in the UK. For example, under SOX (enacted in 2002, in the wake of the Enron and Worldcom scandals) a variety of adverse actions against whistle-blowers, ranging from "tangible employment actions" (such as dismissal or demotion) to softer forms of retaliation are explicitly prohibited (see e.g., Kohn, Kohn and Colapinto, 2004, pp. 97).

The potential gains from more employee whistle-blowing – more deterrence and earlier detection of fraud – are widely acknowledged. However, academics and practitioners alike have expressed concerns that better protection of whistle-blowers might also increase the number of fraudulent claims. In particular, there are worries that low-performing individuals might knowingly lodge false reports with the sole motivation of being sheltered from unfavorable actions such as dismissal (see e.g., Kohn, Kohn, and Colapinto, 2004; Schmidt, 2005; Bowen, Call, and Rajgopal, 2010).[3] For example, it has been argued that, under SOX, the incentive

[1] The rise and fall of Enron are documented in Healy and Palepu (2003). At an anecdotal level, as of January 2017 the Enron and Worldcom cases rank second and third on a top-ten list of the worst accounting scandals of all time, see http://www.accounting-degree.org/scandals/.

[2] Miceli, Near, and Dworkin (2009) survey fraud cases unveiled by whistle-blowers in over 20 countries.

[3] Other concerns that have been raised are potential adverse effects on productive efficiency, for example due to lower effort incentives (Friebel and Guriev, 2012), inefficient hiring decisions (Friebel and Raith, 2004), or by creating an atmosphere of distrust at the workplace (Dworkin and Near, 1997).

for filing fraudulent claims are non-negligible since the requirements to qualify for protection are relatively mild.[4] In particular, a SOX whistle-blower is not required to provide proof of the allegations made, but only needs to hold a "reasonable belief". Under such a "belief-based" legal regime, protection is granted immediately after a claim is lodged, and hence before its actual validity is established in the course of an investigation.[5] Moreover, such protection remains intact even if, in the end, the allegations turn out to be incorrect.[6] Furthermore, sanctions for such (ex post) incorrect claims are typically either rather mild or even ruled out altogether.[7] Also, under SOX, allegations can be made with respect to a wide range of fraudulent conduct as well as against a wide range of individuals in the organization (Kohn, Kohn, and Colapinto, 2004, pp. 76 and pp. 92), which offers plenty of opportunities for lodging (potentially fraudulent) claims.

In other jurisdictions, such as France or Germany, whistle-blower protection is, in general, more difficult to obtain (as a stronger weight is put on employees' legal duty of loyalty towards their employers). Under such "fact-based" legal regimes, protection is typically only granted ex post, i.e., after the validity of the whistle-blower's claim has been established in the course of an investigation or in court. For example, the particular challenge of obtaining protection in Germany is highlighted by the seminal case *Heinisch v. Germany*: In this case, several German courts had refused to reverse the dismissal of a whistle-blower (a geriatric nurse who had (correctly) reported fraudulent behavior by her employer) before protection was eventually affirmed by the European Court of Human Rights (see e.g., Thüsing and Forst, 2016, pp. 12). Hence, legal approaches to whistle-blower protection vary substantially across jurisdictions.

1.2 Research Question, Framework, and Results

The main goal of this paper is to study, in a unified framework, the effects of belief- versus fact-based legal whistle-blower protection on corporate misbehavior, employee whistle-blowing,

[4]For example, at an anecdotal level, a USA Today (2004) article quotes legal practitioners with statements such as "...a genuine explosion of whistle-blower claims", "...the allegations are the invention of an employee who knew he was on thin ice for poor performance", "...many of the complaints are bogus and are largely efforts by marginal employees to squeeze settlements out of their companies", and "some of the more difficult problems I've had is whistle-blowers who will raise issues in which we find some merit, but where they will raise them to gain personal protection for marginal performance".

[5]For example, see the discussion by the Eighth Circuit Court of Appeals in Beacom v. Oracle America, http://media.ca8.uscourts.gov/opndir/16/06/151729P.pdf.

[6]For example, many of the 23 countries studied in Thüsing and Forst (2016) do uphold protection of whistle-blowers whose allegations turn out to be without merit ex post.

[7]For example, for all claims administered by the US Department of Labor, sanctions are not permitted. Similarly, under SEC and IRS whistle-blower programs, while claims that turn out to be fraudulent will not lead to awards, sanctions are not explicitly specified (see also Givati, 2016).

investigations, and retaliation by employers against whistle-blowers. 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 whistle-blower in the form of dismissal. We allow employees to be heterogenous with respect to their productivity: The incentive structure is such that, whenever feasible, the employer prefers to dismiss low-productivity employees, while their high-productivity counterparts might face retaliation only if they blow the whistle. Hence, in line with the evidence discussed above, this might give low-productivity employees an incentive to file fraudulent claims in order to gain employment protection (i.e., to avoid dismissal). In this paper, we focus on whistle-blower protection in the form of employment protection.[8]

We implement six experimental treatments: Four main treatments that capture different legal regimes and two robustness checks. First, in a benchmark treatment *NoWBP*, whistle-blower protection (i.e., employment protection) is not available. Second, in treatment *WBP1* protection is easily obtained by just filing a report. Hence, this treatment is meant to capture a "belief-based regime" where the concept of a *reasonable belief* is interpreted in the most lenient way. Third, in treatment *WBP2* protection is obtained if a report is lodged and, in addition, if it indeed triggers an investigation by the prosecutor. Hence, *WBP2* is also belief-based, but the requirements for protection are more stringent. Fourth, in treatment *WBP3*, protection is granted if and only if a report is filed, the prosecutor investigates, and there is indeed misbehavior. Hence, *WBP3* is meant to capture a "fact-based regime".

The theoretical predictions can be summarized as follows: First, in the benchmark treatment *NoWBP*, fraudulent claims by employees are not an issue, but not all employer misbehavior is reported. Second, in treatment *WBP1*, all misbehavior is reported, but low-productivity employees also lodge fraudulent claims. Still, it turns out that there is less misbehavior in *WBP1* compared to the benchmark *NoWBP*. Third, the predictions for *WBP1* and *WBP2* coincide.[9] Fourth, in treatment *WBP3*, again all misbehavior is reported, while (in contrast to

[8] In both the experiment and the model, retaliation takes the form of dismissal, which we implement by assuming that, unless the employee is protected, the employer can reduce the employee's payoff (down to zero). In practice, this payoff reduction could also represent other forms of retaliation such as demotions or reduced career perspectives. Whistle-blower programs (such as SOX) that offer strong employment protection typically also stipulate a wide range of other banned actions, see e.g., Kohn, Kohn, and Colapinto (2004, pp. 97).

[9] This is driven by our focus on informative equilibria in which the prosecutor triggers an investigation if and only if there is a report by the employee.

WBP1) fraudulent claims do not arise. As a consequence, our theory suggests that deterrence is strongest in treatment *WBP3*.

The main experimental findings are as follows: First, most of the theoretical predictions (with respect to dismissal, misbehavior, and the reporting behavior of the different productivity types) are broadly supported by the experimental data, but, second, there are also interesting deviations. In treatment *WBP1*, we find that fraudulent claims are indeed an issue, and even more so than predicted by theory. Moreover, these fraudulent claims do not only affect "productive efficiency" (in the sense that low-productivity employees cannot be replaced by more productive ones). Fraudulent claims also reduce prosecutors' responsiveness to reports, as these are now less informative about underlying misbehavior. As a consequence, the predicted reduction of misbehavior in *WBP1* relative to *NoWBP* does not materialize. Third, as predicted, the behavior in *WBP1* and *WBP2* is very similar. Fourth, in *WBP3*, there are substantially fewer fraudulent claims than in *WBP1*. Moreover, prosecutors make better decisions in terms of less undetected misbehavior and unnecessary investigations, and employer misbehavior is lower, too. These findings point to potential shortcomings of a belief-based approach.

From a methodological point of view, we would like to argue that our lab experiment complements empirical research with field data on whistle-blowing. With field data one typically observes only those cases of misbehavior that come to light, but not the degree of undetected misbehavior. Also, observing an unaltered number of reports after the introduction of whistle-blower protection might have at least two possible explanations. First, the whistle-blower protection scheme might simply be ineffective. Alternatively, it might indeed increase the willingness to report as intended which, in turn, deters misbehavior to such a degree, that the number of observed reports remains constant. With field data, it is usually difficult to distinguish between these two explanations. By contrast, a lab experiment allows to observe crucial variables such as the underlying (und potentially undetected) level of misbehavior, the willingness to send both truthful and fraudulent reports, and the prosecutors' response to them.[10] Moreover, one can run "policy experiments"; thereby (pre-)testing various features of whistle-blower protection programs, which would be difficult to run outside the lab.

The remainder of the paper is structured as follows: Section 2 discusses the related literature. Section 3 introduces the game played and the design of the experiment, while Section

[10]These methodological advantages are similar to those advanced in the related experimental literature on leniency programs in antitrust (see e.g., Apesteguia, Dufwenberg, and Selten, 2007; Hinloopen and Soetevent, 2008; Feltovich and Hamaguchi, 2016) which is discussed below.

4 presents the theoretical predictions and the underlying intuition. The experimental results are discussed in Section 5. Section 6 concludes. Appendix A contains a full-fledged analysis of the model from which the theoretical predictions of Section 4 are derived. Appendix B contains translations of the experimental instructions. Appendix C provides an overview over the number of observations across decisions and treatments.

2 Related Literature

To the best of our knowledge, our paper is the first to analyze (both theoretically and experimentally) the effects of various legal whistle-blower protection schemes. In doing so, we complement existing research that has focussed on alternative channels to achieve this aim. First, there is a theoretical literature on whistle-blowing that analyzes the optimal responsiveness of prosecutors to reports. In particular, in Chassang and Padró i Miquel (2016) whistle-blowing is fostered through investigation policies that generate "garbled" information. In particular they show that, to shield a whistle-blower from retaliation by his employer, the optimal investigation policy (to which the investigator can commit ex ante) must not be too responsive to reports.[11] The reason is that a relatively responsive policy would reveal that whistle-blowing has in fact occurred, which would then trigger retaliation. In turn, this would undermine the incentive to report in the first place. Like the present paper, Chassang and Padró i Miquel (2016) analyze a cheap-talk game in which the decisions to misbehave, to report, and to investigate are endogenous.[12] 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 the impact of different legal protection regimes on equilibrium behavior, and allow for heterogeneity of workers with respect to productivity. Moreover, we focus on pure-strategy equilibria where the investigator 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 lab experiment.

Second, there is a literature that analyzes the role of monetary rewards in fostering whistle-blowing, as for example implemented in the False Claims Act and the Dodd-Frank Act.[13] Dyck,

[11] Benoît and Dubra (2004) and Muehlheusser and Roider (2008) consider whistle-blowing and "walls of silence" in the context of work teams, and they 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.

[12] Using a different modeling approach, Heyes and Kapur (2009) analyze how the optimal responsiveness of investigations depends on different behavioral motives for whistle-blowing such as conscience cleansing, social welfare considerations, or disgruntlement. Our model captures the first of these motives by assuming that potential whistle-blowers suffer a disutility from undetected misbehavior.

[13] However, as shown by Thüsing and Forst (2016), financial rewards are not too widespread across jurisdictions.

Morse, and Zingales (2010) and Zingales (2004) stress the beneficial role of such rewards in uncovering fraud, while others discuss potentially adverse effects such as fostering fraudulent claims or even the fabrication of cases (see e.g., Givati, 2016; Howse and Daniels, 1995; Callahan and Dworkin, 1992). Our paper focusses on analyzing the impact of different requirements for obtaining (employment) protection, and hence we do not consider financial rewards.[14]

Third, there is a literature on leniency programs in anti-trust, which studies the self-reporting of cartel members (see e.g., the surveys by Spagnolo, 2008, and Marvão and Spagnolo, 2014 and the recent experimental studies by Apesteguia, Dufwenberg, and Selten, 2007, Hinloopen and Soetevent, 2008, and Feltovich and Hamaguchi, 2016).[15] This body of (theoretical, empirical and experimental) research also analyzes how to foster the reporting of illegal activities. However, it considers settings of oligopolistic competition in which every party is involved in the illegal behavior, while in our setup the whistle-blowers are innocent bystanders.

Finally, our paper relates to an empirical literature (in fields such as psychology, sociology, organizational behavior, and business ethics) analyzing the impact of situational and personal factors on the reporting decision of whistle-blowers. For example, such factors are the threat of retaliation, whether or not co-workers were harmed, the type of misbehavior and its severity, whether individuals are rather high-performers or low-performers, and the strength of behavioral motivations such as conscience cleansing, see e.g., the overviews by Miceli and Near (1992), Miceli et al. (2008), Mesmer-Magnus and Viswesvaran (2005) and Vadera et al. (2009), and the recent incentivized lab experiment by Bartuli, Djawadi, and Fahr (2016). In our paper, we focus on how the reporting decisions of employees are affected by their productivity and the underlying legal regime. In addition, we also elicit a number of personal characteristics and situational factors in the post-experimental questionnaire.

3 Experimental Design

This section explains the setup of the experiment. That is, we provide summary information and describe in detail the game played, the incentive structure, the session design and payments, the various treatments and their framing, as well as the post-experimental procedures.

Beyond such legal provisions, according to a recent report by the Association of Certified Fraud Examiners (2014) only 11% of all organizations considered world-wide had a reward scheme in place.

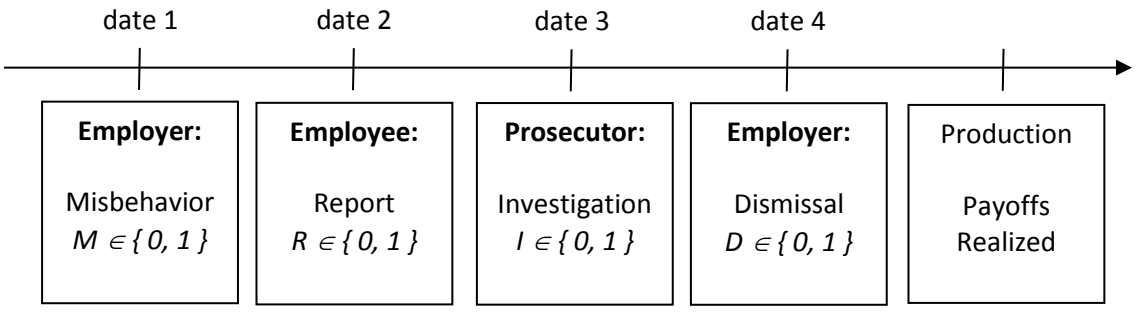
[14] In an experimental study, Schmolke and Utikal (2016) compare financial rewards for whistle-blowers and fines for non-reporting. They find that the latter are more effective in increasing reporting rates than the former.

[15] Moreover, also in an experimental setup, Cotten and Santore (2016) analyze the impact of transparency and amnesty rules in the context of corporate fraud by criminal teams.

Summary Information The experiment was conducted in the experimental lab of the University of Hamburg between April 2015 and March 2016 and programmed in z-Tree (Fischbacher, 2007). In total, we ran four main treatments (which differ with respect to the requirements for obtaining employment protection) and two further treatments as robustness checks (see Table 1 below). We employed a between-subjects design, so that each subject participated in one treatment (and hence, one session) 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 600 subjects, we used the software tool *hroot* (Bock, Baetge, and Nicklisch, 2014). 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.

The Game Played in Each Period 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 (exogenous) 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 losses from employer misbehavior. It is included in the experiment to make it more salient that misbehavior causes harm to others. The remaining three players played the following game (summarized in Figure 1, which is an implementation of the game analyzed in the theory part as laid out in Appendix A):

Figure 1: The Game Played in Each Period



At date 1, the employer observes the productivity of his employee. She then chooses whether or not to misbehave, which entails a gain (which is independent of her employee's productivity

type), but which is costly to others. At date 2, we used the strategy method to elicit the employee's binary reporting decision for both cases 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 neither the underlying employer misbehavior decision nor the employee's productivity type). The prosecutor then decides on triggering an investigation (which incurs a private costs for the prosecutor). An investigation perfectly reveals whether or not the employer has misbehaved.[16] Moreover, the uncovering of the misbehavior in the course of an investigation benefits both the prosecutor and the third party, while the employer must pay an (exogenously) fine.

Finally, at date 4, the employer decides whether or not to dismiss the employee. In case of dismissal, the employee is replaced by a (computerized) outsider, who is more (less) productive than an L-employee (H-employee). However, a dismissal is only feasible as long as the employee is not shielded by whistle-blower 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.

Incentive Structure: Monetary Incentives and Parameter Values In the experiment, the players' monetary payoff components (which were common knowledge ex ante) had the following properties:[17] Unless detected, an employer's monetary payoff is higher with misbehavior. Moreover, the difference between the productivity and the wage of an L-employee (H-employee) is smaller (larger) compared to employing the replacement outsider. Hence, the employer's monetary payoff is higher when dismissing (retaining) an L-employee (H-employee). In 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.[18] Finally, despite its costs, when there actually is misbehavior, the prosecutor's monetary payoff is higher when he investigates. In contrast, in the absence of

[16]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, 2016; Givati, 2016; Heyes and Kapur, 2009) and legal practice (e.g., under SOX). The case where investigations do not perfectly reveal underlying misbehavior is discussed in Section 5.3 below.

[17]The payoff structure of the underlying model is summarized in Table 5 in Appendix A.

[18]The underlying idea for this payoff ranking is that detecting the misbehavior might allow to at least partly curb the associated harm. For example, for misbehavior in the form of illegally dumping hazardous waste, when detected in the course of an investigation, the environmental damage will typically be lower compared to the case without an investigation.

misbehavior, the prosecutor's monetary payoff is higher when he does not investigate.[19]

We used the following parameter values throughout (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. Employees receive a fixed wage of 40. The employer's (gross) payoff from misbehavior is 50, and, in case of detection, he 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). Note that 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.

Incentive Structure: Potential Behavioral Motivations As discussed below, our experimental design is not fully neutral, and we do provide subjects with some information about the context in which they operate. Consequently, in addition to the monetary payoff components, there might also exist behavioral motivations that shape subjects' behavior (which were not incentivized in the experiment): First, employees do not receive a direct monetary reward when reporting misbehavior, and we rely on their potential behavioral motivations to report misbehavior instead. As discussed above, the literature has identified conscience cleansing as a main motive of whistle-blowers to come forward. Second, it is well documented that whistle-blowers are often no longer well-liked at their workplace. That is, employers might feel tempted to retaliate in the form of dismissal, even though this might reduce the employers' profit due to a loss of productivity (i.e., when matched with an H-employee). Third, employers might have moral reservations such that their "net benefit" from misbehavior is smaller than their pure monetary gain. In the theoretical analysis in Appendix A, on which the predictions of Section 4 are based, we allow for heterogeneity with respect to the intensity of these behavioral preferences (i.e., employees' dislike of undetected misbehavior, employers' dislike of employing whistle-blowers,[20] and employers' aversion against misbehavior). Apart from that,

[19]Hence, apart from lowering the loss for the third party, this also gives the prosecutor a monetary incentive to discover any misbehavior. In reality, this might for example come in the form of a reputation gain.

[20]The existence of such heterogeneity on the employers' side is consistent with empirical findings on the relevance of retaliation. For example, Near and Miceli (1996, pp. 517) find retaliation rates ranging from 6% to 38%, suggesting that employers do differ with respect to their attitude towards whistle-blowing (see also the National Business Ethics Survey of 2013).

the theoretical predictions rely on the assumption that subjects have standard preferences.

Session Design and Payments In each session, the design of the experiment was common knowledge and all subjects received the same instructions. Sessions consisted of 30 periods and usually had 24 participants (6 groups).[21] In addition to the (random) re-matching of groups in each period, also the role assignments varied as follows across periods: Each subject who was assigned the role of employer in the first period retained this role throughout all 30 periods. In contrast, all other subjects randomly switch roles across periods, either between employees and third partys or between prosecutors and third partys. 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 (directly or indirectly) curb the harm inflicted by the employer on the third party.

In addition, ensure that subjects indeed understood the game, after going through the instructions they had to answer a series of control questions, and we discussed any wrong answers with them in private before finally launching the experiment. Finally, to determine each subject's payment, three out of the 30 periods were randomly selected, and the subject's total points earned in these three 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.

Treatments We consider four main treatments (see Table 1). Treatment *NoWBP* corresponds to a benchmark setting in which employment protection is not available at all. In addition, there are three treatments where protection is available (to the effect that a protected employee cannot be dismissed at date 4), but which differ with respect to the requirements under which it is obtained. We consider two treatments that are meant to capture "belief-based" legal regimes in the sense that protection is granted on the basis of sending a (convincing) report: In particular, in treatment *WBP1*, protection is (easily) obtained by just sending a report. As discussed in the Introduction, this is meant to capture the lenient approach of the U.S. Sarbanes-Oxley-Act (SOX), where, to gain protection, the whistle-blower only needs to hold a *reasonable belief* that misbehavior has actually occurred.[22] In treatment *WBP2*, the

[21] In three out of a total of 26 sessions, the number of participants was 16 because of no-shows.

[22] Hence, in the experiment we assume that it is not possible for the prosecutor to prove that the employee does not hold a reasonable belief when sending a report.

whistle-blower gains protection if his report leads the prosecutor to investigate the case. Finally, treatment *WBP3* captures a "fact-based" regime, in which protection is only granted if, in addition to the requirements of *WBP2*, there is actually misbehavior by the employer. This treatment is meant to capture the legal situation in countries such as Germany or France as discussed above. As discussed in more detail in Section 5.3 below, as robustness checks we ran two further treatments *R1* and *R2*, in which employers face a (reputation) loss whenever an investigation occurs and in which there are investigation errors, respectively.

Table 1: Treatments

Conditions for Protection	Never	Report Only	Report + Investigation	Report + Investigation + Detected Misbehavior
Main Treatments	*NoWBP*	*WBP1*	*WBP2*	*WBP3*
Robustness Checks			*R1* (reputation loss)	*R2* (investigation errors)

As can be shown, the theoretical predictions for none of these four main treatments depend on whether the reporting decision of the employee is observed by the prosecutor only or by both the prosecutor and the employer:[23] 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. As, in the experiment, the prosecutor's behavior might deviate from this prediction, it might be more difficult for the employer to correctly infer the reporting decision from observing the investigation decision only. Hence, as we wanted to rule out the possibility of erroneous updating as a potential driver for dismissal decisions (rather than any potential dislike of whistle-blowing), in treatments *NoWBP* and *WBP1*, both the prosecutor and the employer learn the reporting decision. In treatments *WBP2* and *WBP3* (where a report alone is not sufficient for obtaining protection) the employer learns the reporting decision in the course of an investigation.[24]

[23]Hence, both types of report are "external" in the sense of being directed towards the (outside) prosecutor. Some whistle-blower laws also stipulate that firms must establish internal reporting systems, and that whistle-blowers 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. In general, while whistle-blower protection laws such as SOX do require firms to establish anonymous reporting channels, recent evidence casts doubt that anonymity can be upheld in practice, see e.g., Kaplan and Schultz (2007) and the discussion in Chassang and Padró i Miquel (2016).

[24]Hence, the comparisons of treatments *NoWBP* and *WBP1*, and *WBP2* and *WBP3*, respectively, follow a

Framing In experimental economics, there is a discussion on the conditions under which a neutral or a loaded framing is more appropriate.[25] In this respect, we have followed a middle course. That is, we do give subjects some information about the context in which their behavior is placed (e.g., we frame the game as a employer-employee relationship, where the employee can file a report to a prosecutor). However, in the experimental instructions (see Appendix B), we avoided the use of strongly judgemental terms such as "misbehavior", "illegal" or "whistle-blowing". For example, in the experiment, we refer to a employer's misbehavior decision as a choice between two alternatives 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" (see Appendix B). Moreover, the employee's reporting decision is not referred to as "whistle-blowing", but as "asking the prosecutor to trigger an investigation".

Post-Experimental Procedures At the end of the respective session, subjects completed a (non-incentivized) questionnaire in which we elicited socio-demographic information (e.g., age, gender, and field of study), risk preferences (via the "100.000 Euro question" of Dohmen, Falk, Huffman, Sunde, Schupp, and Wagner, 2011), and cognitive abilities (via the "Cognitive Reflection Test" of Frederick, 2005), and their attitudes towards revealing misbehavior (measured on a five-level Likert scale).[26] In addition, we elicited subjects' "Dutifulness" (i.e., their sense of duty and obligation) as a sub-factor of the Big Five personality trait "Conscientiousness" (where the respective questions were taken from the "NEO Personality Inventory", see Costa and McCrae, 1992; Berth and Goldschmidt, 2006). As to make these issues not too salient, the above questions were interspersed with some unrelated questions. We also elicited information about subjects' social preferences by letting them play an incentivized standard one-shot dictator game in which they had to decide on how to split 100 points between themselves and a "receiver". We used the strategy method so that subjects had to make their choice before

one-change-at-a-time principle. While this is not the case for the comparison between treatments *WBP1* and *WBP2* (where both the requirements for obtaining protection and the observability of reports change), the theoretical prediction for these two treatments is identical (see *Prediction WBP2* below), and this is also borne out in the experiment (see Section 5.3).

[25] For general discussions, see e.g., Eckel and Grossman (1996) and Alekseev, Charness, and Gneezy (2017). In experimental studies on whistle-blowing in organizations, a loaded framing is used in Bartuli, Djawadi, and Fahr (2016) and Cotten and Santore (2016), while Schmolke and Utikal (2016) choose a neutral design. 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 et al., 1993; Babcock et al., 1995).

[26] The post-experimental questionnaire is available upon request.

they knew whether they were actually (randomly) assigned the role of dictator or receiver. We then converted their resulting points at the rate of 1 Euro per 20 points and added this to the monetary payoff they received at the end of the experiment.

4 Theoretical Predictions

Our theoretical predictions are derived from the pure-strategy Perfect Bayesian Equilibria of the game described in Section 3, and formally spelled out and analyzed in Appendix A (see Propositions 1 - 4). We focus on *informative equilibria* in the sense that the prosecutor triggers an investigation if and only if the employee sends a report. This directly leads to

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

Also, the prediction for the employer's dismissal decision is straightforward. Intuitively, the employer prefers to dismiss an L-employee whenever this is feasible because the (expected) productivity of the outside replacement is higher. In contrast, an H-employee will only be dismissed upon reporting, and only if the employer's dislike of employing a whistle-blower exceeds the H-employee's productivity advantage. This leads to

Prediction D (Dismissal): *In all treatments: (i) unless protected, L-employees are dismissed. (ii) H-employees are retained when sending no report, while they are dismissed with positive probability when sending a report and not being protected.*

The predictions for the reporting and misbehavior decisions are treatment-specific: We start with the comparison of treatments *NoWBP* and *WBP1*, and then discuss treatments *WBP2* and *WBP3*.

Prediction R (Reporting): *The reporting behavior in treatments* NoWBP *and* WBP1 *is summarized in Table 2. In particular: (i) In both treatments, misbehavior leads to a (weakly) higher willingness to report for either productivity type. (ii) For either misbehavior decision, L-employees exhibit a (weakly) higher willingness to report than H-employees. (iii) For either misbehavior decision, both productivity types exhibit a (weakly) higher willingness to report in treatment* WBP1. *(iv) Fraudulent claims are sent by L-employees only, and they occur in treatment* WBP1 *only.*

Table 2: Theoretical Prediction: Fraction of Employees Sending a Report

Treatment	***NoWBP***		***WBP1***		***WBP3***	
Employee Type	Low	High	Low	High	Low	High
Misbehavior	1	$\in [0,1]$	1	1	1	1
No Misbehavior	0	0	1	0	0	0

Note: The prediction for treatment *WBP2* is the same as for treatment *WBP1*.

Hence, in treatment *NoWBP*, but not in *WBP1*, misbehavior by the employer is a necessary, but not sufficient, condition for reporting to occur (see Table 2). Intuitively, recall that in our model employees are assumed to suffer a disutility from undetected misbehavior, so that either productivity type tends to be more willing to report when misbehavior actually occurs. However, in anticipation of the subsequent investigation and dismissal decisions, the reporting behavior differs across types as L-employees expect to be dismissed whenever feasible, while H-employees are less vulnerable due to their higher productivity. This gives the former a higher incentive to send both truthful and fraudulent reports: When misbehavior actually occurs, H-employees are facing a trade-off between any disutility from undetected misbehavior under no reporting and the higher risk of dismissal when doing so. Moreover, in treatment *WBP1*, L-employees have an incentive to report even when there is no misbehavior, as this protects them from dismissal. Finally, we have:

Prediction M (Misbehavior): *Misbehavior in treatments* NoWBP *and* WBP1 *is summarized in Table 3. In particular: (i) When the employer is matched with an L-employee, the frequency of misbehavior is the same in* NoWBP *and* WBP1*. (ii) When the employer is matched with an H-employee, the frequency of misbehavior is strictly lower in* WBP1 *than in* NoWBP*.*

Intuitively, misbehavior of employers with L-employees does not vary across the two treatments as the decision whether or not to misbehave has no effect on the dismissal of L-employees: This productivity type is always dismissed in treatment *NoWBP* (irrespective of any earlier decisions), and he is always shielded from dismissal in treatment *WBP1* (as L-employees always report, again irrespective of their employer's earlier misbehavior decision). Because H-employees report any misbehavior in treatment *WBP1*, the incentive to misbehave is smaller in this treatment.

Table 3: Theoretical Prediction: Fraction of Employers Misbehaving

Treatment	***NoWBP***		***WBP1***		***WBP3***
L-employee	m_L^{no}	$=$	m_L^1	$>$	m_L^3
			$\vee$		$\wedge$
H-employee	m_H^{no}	$>$	m_H^1	$=$	m_H^3

Notes: The prediction for treatment *WBP2* is the same as for treatment *WBP1*. m_θ^j denotes the frequency of misbehavior by an employer matched with an employee of productivity $\theta = L, H$ in treatment $j = \{no, 1, 3\}$.

With respect to treatment *WBP2*, our focus on informative equilibria implies that the theoretical predictions coincide with those of treatment *WBP1*. The reason is that the only case in which the two treatments would have different implications does not occur on the equilibrium path (i.e., the case that the employee sends a report, but the prosecutor does not investigate, which would lead to protection in *WBP1*, but not in *WBP2*). This implies

Prediction WBP2: *The predictions for treatment* WBP2 *coincide with those for* WBP1*.*

Predictions change, however, in treatment *WBP3*, in which protection is only granted upon a report, followed by an investigation and actual misbehavior by the employer:

Prediction WBP3: In treatment *WBP3*, the following holds: *(i) Misbehavior is always reported. (ii) Fraudulent claims do not occur. (iii) The frequency of misbehavior in* WBP3 *is strictly lower than in* NoWBP *(for either employee type) and weakly lower than in* WBP1*.*

The results for treatment *WBP3* are also displayed in Tables 2 and 3. Intuitively, in *WBP3*, by conditioning protection on actual misbehavior, all incentives for fraudulent claims are removed. With respect to truthful claims, the incentives for L-employees (H-employees) are as in treatment *NoWBP* (*WBP1*), and hence any misbehavior is reported. The incentive to misbehave is (weakly) decreasing from *NoWBP* to *WBP1* to *WBP3*. The reason is that in treatment *WBP3*, through her misbehavior decision, the employer can directly affect the employee's access to protection. As protection is potentially costly for the employer (because of constraining her dismissal decision), this makes the employer more reluctant to misbehave.

To summarize, compared to a benchmark without protection, we predict that belief-based whistle-blower protection regimes increase both truthful and fraudulent reports, and decrease

misbehavior of employers with H-employees. In contrast, a fact-based whistle-blower protection regime increases truthful reports without triggering fraudulent claims, and reduces misbehavior by *all* employers.

5 Experimental Results

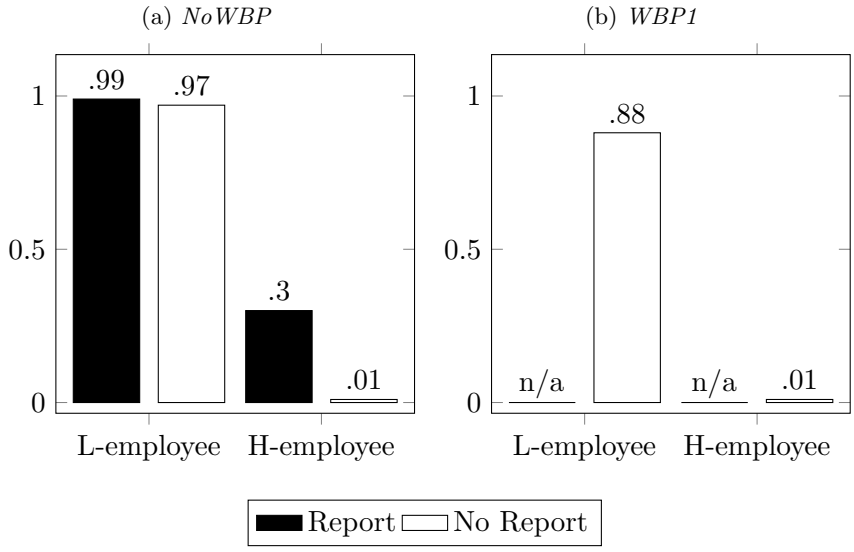
As for our empirical strategy, recall that in each session of the experiment, each subject played 30 periods in a given treatment, but possibly in different roles. Hence, as we observe each subject more than once, in non-parametric tests, our unit of observation are averages on the subject-level (which is explained in more detail in Footnote 27 below), and the resulting numbers of observations are reported in Appendix C. For within-treatment (across-treatment) comparisons, we use Wilcoxon Signed-Rank (Mann-Whitney-U) tests. Moreover, in regression analysis, standard errors are clustered at the subject-level. We start with a discussion of the results for treatments *NoWBP* and *WBP1*.

5.1 Comparing Treatments *NoWBP* and *WBP1*

Employers' Dismissal Decisions: Testing Prediction D Figure 2 displays the fractions of dismissed workers in treatments *NoWBP* and *WBP1*, depending on their productivity type and reporting behavior (where the non-feasibility of dismissal upon reporting in treatment *WBP1* is indicated by "n/a"). The results are fully supportive of *Prediction D*. In particular, in treatment *NoWBP*, virtually all L-employees are dismissed. In *WBP1* (where dismissal is only feasible, when there is no report) this fraction is somewhat lower but still at 0.88, and the difference is not statistically significant according to a MWU test).[27] Moreover in both treatments, H-employees who do not report are almost always retained. In contrast, and again in line with *Prediction D*, around 30% of H-employees who do report are fired in treatment *NoWBP*. This is significantly more compared to non-reporting H-employees (0.30 versus 0.01, Wilcoxon Signed-Rank test, $p < 0.001$) and significantly less compared to L-employees who do report (0.99 versus 0.30, Wilcoxon Signed-Rank test, $p < 0.001$). The considerable fraction of dismissed H-employees is consistent with the prediction of our model, which is based on a

[27] As discussed at the end of Section 3, in each of the two treatments shown in Figure 2 there are four conditions under which employers are (repeatedly) observed: with either an L- or an H-employee, who either reports or does not report. For each of these four conditions, we aggregate a given subject's behavior into an average, and these averages then form the units of observation in the reported non-parametric tests. In all other reported non-parametric tests, when analyzing the behavior of employers, employees, and prosecutors, we proceed analogously.

Figure 2: Fraction of Employers Dismissing Their Employee



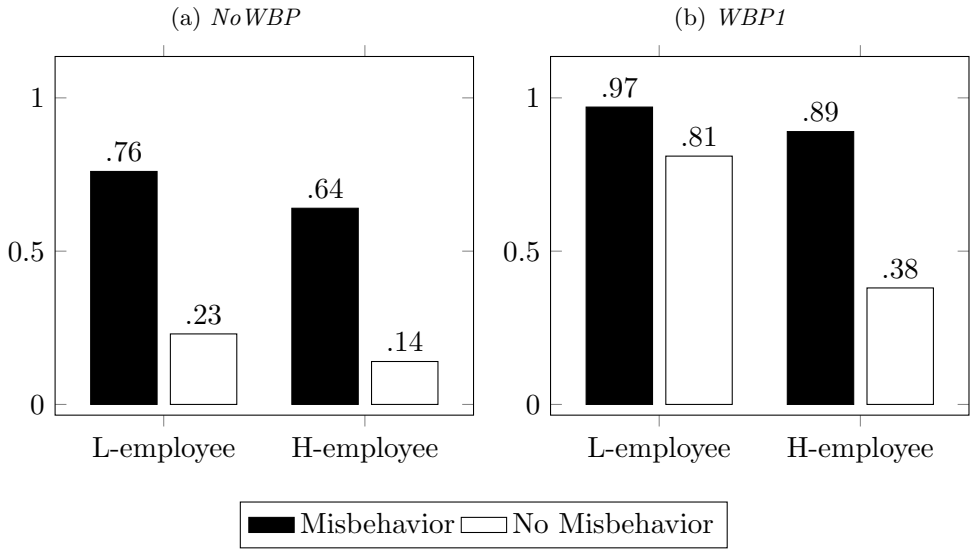
utility loss associated with retaining whistle-blowers, such that employers might be willing to forgo a higher productivity in order to avoid that loss.

Employees' Reporting Decisions: Testing Prediction R Figure 3 illustrates our results concerning the reporting behavior of employees in treatments *NoWBP* and *WBP1*. It turns out that *Prediction R* as summarized in Table 2 is broadly supported. In particular, as for Prediction *Prediction R(i)*, the reporting rates of both types are higher when there is misbehavior (i.e., in Figure 3, compare pairwise the black bars and the white bars). These differences are all statistically significant (all with $p < 0.001$, Wilcoxon Signed-Rank tests) and hence, with the exception of the case of L-employees in treatment *WBP1*, also in line with the prediction.

As for *Prediction R(ii)*, in both treatments also the reporting rates of L-employees are generally higher than those of H-employees, irrespective of whether or not misbehavior actually occurred (i.e., in Figure 3, compare for each treatment the two black and the two white bars, respectively). Again, all of these four differences are statistically significant, which is in line with *Prediction R(ii)*, except for the reporting of misbehavior in treatment *WBP1*, which should be reported by either productivity type (see Table 2). For this latter case (i.e., comparing 0.97 with 0.89), a Wilcoxon Signed-Rank test yields $p < 0.028$. For the other three cases, we have 0.81 versus 0.38 ($p < 0.001$), 0.76 versus 0.64 ($p < 0.019$), and 0.23 versus 0.14 ($p < 0.013$).

Also *Prediction R(iii)* is broadly supported: Comparing the reporting behavior across treat-

Figure 3: Fraction of Employees Sending a Report

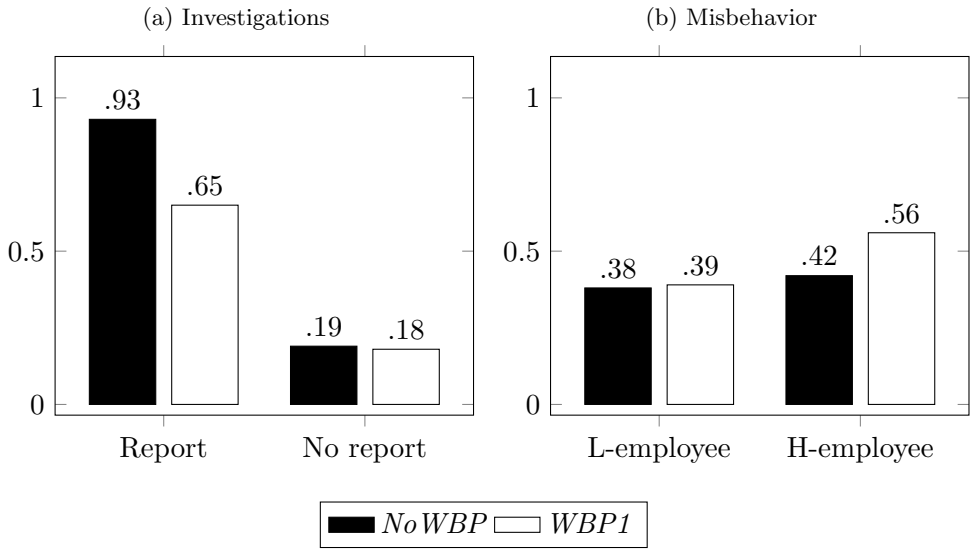


ments (i.e., comparing each of the bars in the left panel of Figure 3 with the respective counterpart in the right panel) reveals that it is generally higher in treatment *WBP1*. Again, all of these treatment differences are statistically significant (all with $p < 0.0001$, MWU test), although for the two pairs $(0.76, 0.97)$ and $(0.14, 0.38)$ they are predicted to be weak only (see Table 2). Overall, it can be seen that a lenient whistle-blower protection scheme such as whistle-blower protection leads to high reporting rates close to one for both productivity types in the case of misbehavior (as predicted by the model).

The downside is that also the fraction of fraudulent claims rises sharply in treatment *WBP1*, in particular by L-employees, which is fully in line with *Prediction R(iv)*. However, we also observe an (unpredicted) increase in the fraction of fraudulent claim by H-employees. This issue and its potential implications are discussed in Section 5.2 below.

Prosecutors' Investigation Decisions: Testing Prediction I Figure 4(a) illustrates the experimental results for the investigation decisions in treatments *NoWBP* and *WBP1*. First, in both treatments, prosecutors indeed seem to perceive employee's report as an informative signal about the presence of misbehavior, and the number of investigations is significantly higher when a report occurs (0.93 versus 0.19, and 0.65 versus 0.18, both with $p < 0.001$, Wilcoxon Signed-Rank tests).

Figure 4: Fractions of Investigations and Misbehavior



Moreover, the point predictions of *Prediction I* are broadly confirmed in treatment *NoWBP*, where the fraction of investigations following a report is 0.93 (and hence, indeed close to one as predicted). When there is no report, the fraction of investigations is 0.19 (and hence, somewhat further away from the predicted level of zero).

In treatment *WBP1*, we find very similar results for the case of no report (0.18). However, in contrast to treatment *NoWBP*, the willingness to investigate conditional on a report is significantly lower in *WBP1* (0.65 versus 0.93, $p < 0.001$, MWU test). This difference, which was not predicted by the model, might be driven by the observed reporting behavior, in particular the high number of fraudulent claims by H-employees in treatment *WBP1*. This reduces the informativeness of reports, and hence dilutes the incentive of prosecutors to trigger costly investigations. This issue will be discussed in more detail in Section 5.2 below.[28]

Employers' Decisions to Misbehave: Testing Prediction M Figure 4(b) displays the fractions of employers who chose to misbehave in treatments *NoWBP* and *WBP1*. As can be seen, *Prediction M(i)* is strongly supported: The fractions of misbehaving employers with L-employees are basically identical in the two treatments (0.38 versus 0.39, where the difference

[28] We have also checked how the behavior of prosecutors varies across periods. In particular, there is no period effect in treatment *NoWBP*, while the frequency of investigations conditional on a report decreases slightly over time in *WBP1*.

is not statistically significant). *Prediction M(ii)* is not borne out by the data: For employers matched with H-employees, there is no statistically significant difference in misbehavior between treatments *NoWBP* and *WBP1* (0.42 and 0.56, respectively). If anything, there is more (rather than less) misbehavior upon the introduction of whistle-blower protection in *WBP1*.

To summarize, many of the experimental results for treatments *NoWBP* and *WBP1* strongly support the theoretical predictions. However, there are also some deviations, in particular in treatment *WBP1*, which are discussed next.

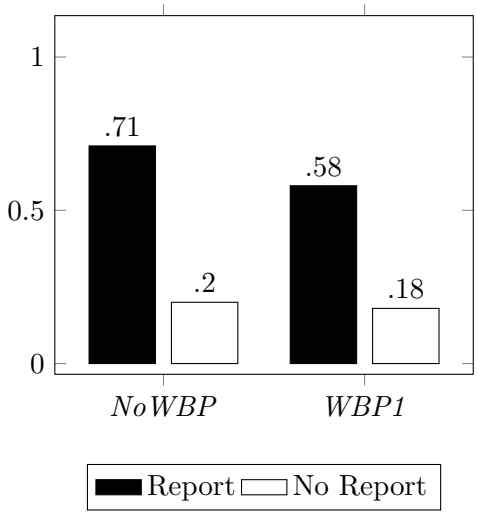
5.2 Treatment *WBP1*: A Closer Look at Deviations From the Predictions

Relative to the theoretical predictions for treatment *WBP1*, we observe a considerable number of fraudulent claims by H-employees, a lower responsiveness of prosecutors to reports, and no reduction in the level of misbehavior. In this subsection, we first show that the latter two findings can be rationalized under the assumption that prosecutors and employers correctly anticipate the *actual* behavior in the experiment (rather than the theoretically predicted one). Furthermore, as for the first finding, we investigate potential drivers for sending fraudulent claims, using the post-experimental questionnaire.

Investigation Decisions and the Informativeness of Reports Recall from *Prediction M* that both employee types should exhibit a higher overall willingness to report in treatment *WBP1* compared to *NoWBP*, but that there should also be an incentive for L-employees to send fraudulent claims. Hence, even from a theoretical point of view, the treatment comparison for the number of *truthful reports* (i.e., reports that are sent if and only if there is misbehavior) is ambiguous. Actual play in the experiment reveals that the effect on the number of fraudulent claims dominates, and the fraction of truthful reports is lower in treatment *WBP1* than in *NoWBP* (0.66 versus 0.76). As a result, in *WBP1* reports are less informative for prosecutors about underlying misbehavior. In particular, as illustrated in Figure 5, the (empirical) frequency of misbehavior conditional on receiving a report drops from 0.71 to 0.58, while it remains basically unchanged when no report is sent.

In a next step, using the information of Figure 5, we derive the optimal investigation decision of prosecutors under the assumption that they correctly take into account the actual empirical relationship between reporting and underlying misbehavior. In addition, we also allow for the possibility that prosecutors internalize the harm from misbehavior inflicted on the third party

Figure 5: Fraction of Underlying Misbehavior Conditional on a Report



(with weight $\alpha \in [0,1]$). Hence, for $\alpha = 0$, the prosecutor only cares about his own payoff, while for $\alpha = 1$, he fully internalizes the third party's harm.[29]

Under these assumptions, it would be optimal for the prosecutor (i) not to investigate when no report is sent both in treatment *NoWBP* and *WBP1* (irrespective of α), (ii) to investigate in treatment *NoWBP* whenever a report is sent (irrespective of α), and (iii) to investigate in treatment *WBP1* when a report is sent and at the same time $\alpha > 0.22$.[30] Hence, under these assumptions, in treatment *WBP1* it would not be necessarily optimal for the prosecutor to trigger an investigation upon receiving a report. All in all, this modified prediction is very well in line with the findings for the prosecutors' investigation decisions as reported in Figure 4(a).

Moreover, also the effect of α on the investigation decision is in line with the above prediction. In particular, we proxy α by the offer in the (incentivized) dictator game, which was played at the end of the experiment (where the offer to the other party was an integer between 0 and 100). Table 4 reports the result of a linear probability model for treatment *WBP1*. In

[29] Note that the theoretical predictions of Section 4 are based on $\alpha = 0$.

[30] Recall that prosecutors receive an endowment of 60 points, their cost of an investigation is 20 points, and their payoff is reduced by 20 (10) points in the case of undetected (detected) misbehavior. Moreover, third parties receive an endowment of 40, which is reduced by 50 (70) points in case of detected (undetected) misbehavior. For example, based on Figure 5 in treatment *NoWBP* and conditional on receiving a report, the prosecutor's expected payoff when choosing $I = 1$ is $0.71 \cdot (50 - 10\alpha) + 0.29 \cdot (40 + 40\alpha)$. Analogously, when choosing $I = 0$ instead, the prosecutor expects $0.71 \cdot (40 - 30\alpha) + 0.29 \cdot (60 + 40\alpha)$, which leads to a payoff difference of $1.3 + 14.2\alpha > 0$, and hence the prosecutor would optimally trigger an investigation independent of α. For all the other cases, the calculations are analogous.

Table 4: Regression Analysis: Investigation Decisions in Treatment *WBP1*

	Investigate
Report	0.344***
	(0.000)
Offer	-0.00236
	(0.177)
Report x Offer	0.00780**
	(0.003)
Constant	0.292
	(0.481)
Observations	860
Adjusted R^2	0.242

Notes: The table reports the results from a linear probability model with the investigation decision as the dependent variable. p-values are reported in parentheses. Standard errors are clustered at the subject level, where *, **, and *** indicate statistical significance at the 5%, 1%, and .1% level, respectively. Further controls included are: age, gender, proxies for (i) risk aversion, (ii) cognitive reflection, (iii) attitude towards revealing misbehavior, (iv) dutifulness, and (v) a dummy for a major or minor in a field related to economics or business. The coefficients of these controls are all insignificant, and hence are not reported.

line with Figure 4(a), a crucial driver for the investigation decision is indeed whether or not a report arrives.[31] Moreover, a higher offer (i.e., a higher α) increases the probability of investigation only in the case in which a report is sent.[32] Finally, the effect of social preferences is considerable: For example, for a prosecutor with preferences for an equitable outcome (which corresponds to an offer of 50, and which was chosen by around 20% of subjects), this increases the likelihood of an investigation by 27 percentage points (compared to the case where $\alpha = 0$).

Employer Misbehavior According to *Prediction M*, there should be no treatment effect on misbehavior for L-employees (which is supported by the data), while, for H-employees, it should be lower in treatment *WBP1* than in *NoWBP* (which is not borne out in the data). We find that, similar to above, the observed relative frequencies of misbehavior across treatments and employee types (see Figure 4(b)) can be rationalized under the assumption that employers correctly anticipate the *actual* payoff consequences from their misbehavior decision. To show this, in a first step we determine the difference of the employer's average payoff when choosing

[31] We do not report the regression results for treatment *NoWBP* as the likelihood of investigations is strongly determined by whether or not a report is sent (which is fully in line with Figure 4(a)), and the behavior in the dictator game has no effect.

[32] The coefficients for *Offer* and the interaction term are also jointly significant (F-test, $p < 0.001$).

$M = 1$ and $M = 0$, respectively, from the experimental data, and we do this separately for each treatment and for each productivity type of the employee with whom the employer might be matched. For treatment *NoWBP*, these payoff differences are 4.32 when the employer is matched with an L-employee, and 7.37 when the employer matched with an H-employee. Proceeding analogously for treatment *WBP1*, we get 7.66 and 13.18. Note that the ranking of these four payoff differences (4.32, 7.37, 7.66, and 13.18) is the same as the ranking of the corresponding misbehavior frequencies observed in the experiment (0.38, 0.39, 0.42, and 0.56) and as reported in Figure 4(b). Hence, large (small) payoff differences correspond to high (low) levels of misbehavior. While these monetary payoff differences are all positive, an employer will prefer not to misbehave when his (moral) aversion towards misbehavior is sufficiently large. This is more likely to occur the smaller the monetary payoff is in the first place.[33]

Fraudulent Claims by H-Employees As argued above, in treatment *WBP1* the higher than predicted number of fraudulent claims by H-employees can rationalize the decisions to investigate and, in turn, also to misbehave. That is, prosecutors and employers seem to understand that in treatment *WBP1* (where protection is easy to obtain) fraudulent claims are a crucial issue which they (directly or indirectly) seem to take into account.

We now study in more detail potential drivers for such claims by looking at the characteristics of these whistle-blowers. In total, there are 44 distinct subjects in treatment *WBP1* who played the role of an H-employee and whose employer did not misbehave. Out of these 44 subjects, 16 behaved exactly in line with *Prediction R(iv)*, i.e., they never sent a fraudulent claim. This is also the modal behavior. However, there is also a substantial fraction of 10 subjects who always send fraudulent claims. It turns out that on average, these subjects exhibit a degree of risk aversion (which, recall, we elicited in the post-experimental questionnaire) that is 0.5 standard deviations higher compared to those 16 subjects who never report. While, in the experiment, the frequency of dismissal of non-reporting H-employees is negligible (see Figure 2), these risk-averse subjects might nevertheless prefer to insure themselves against this risk. This is straightforward to achieve in treatment *WBP1* where protection is easy to obtain. Below, we will contrast this with treatments *WBP2* and *WBP3* where a mere report no longer suffices to obtain protection.

[33] A (unreported) regression akin to the one of Table 4 reveals that the propensity to misbehave is negatively related to employers' risk aversion (which is line with a similar finding by Minor, 2015) and to the intensity of social preferences (again proxied by the amount offered in the dictator game), where $p = 0.013$ and $p = 0.067$, respectively.

Note also that there is no time trend in the frequency of fraudulent claims across periods. In particular, it is not the case that H-employees (erroneously) file such claims in early periods, while after experiencing that they are not dismissed when remaining silent, they refrain from reporting in later periods (thereby behaving in accordance with theory).[34]

5.3 Treatments With More Stringent Requirements for Protection

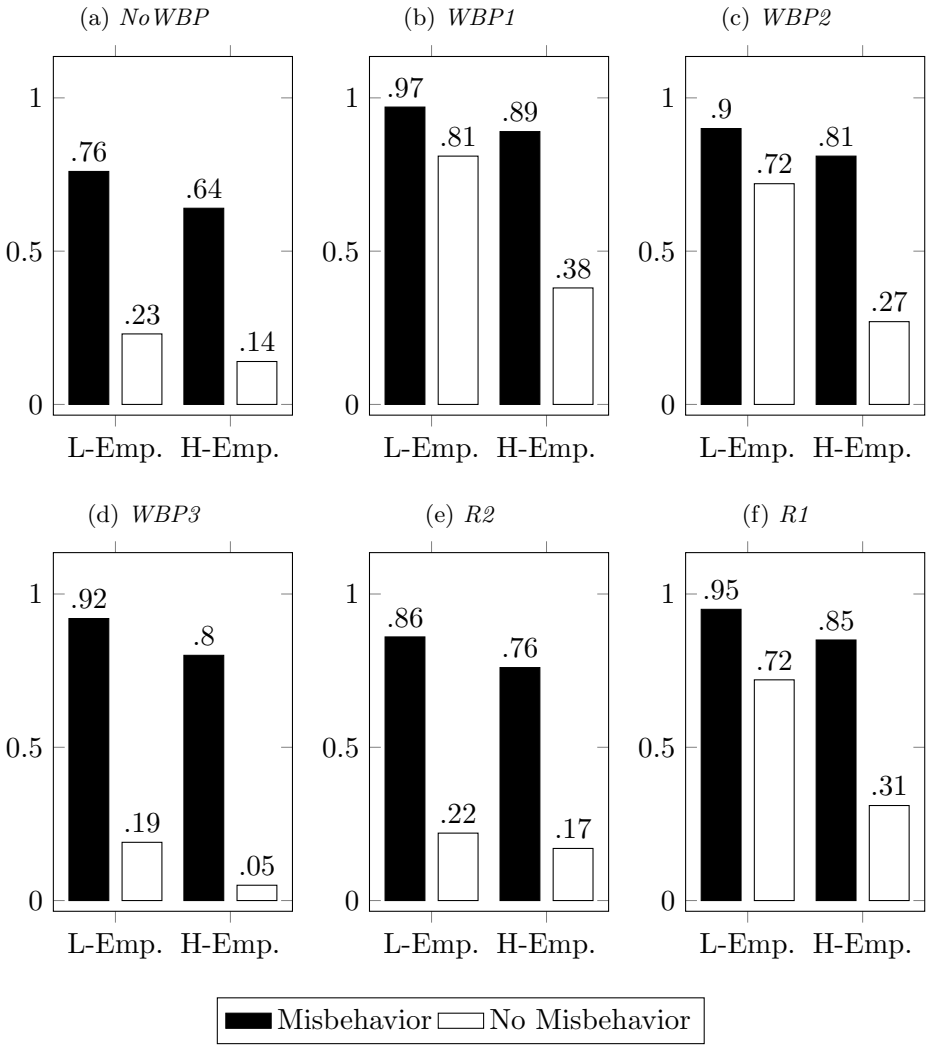
The results above suggest that, when introducing employment protection for whistle-blowers, fraudulent claims might be a serious issue, not only because of the high number of such claims, but also because they dilute the responsiveness of prosecutors to reports (as reports are no longer good indicators for underlying misbehavior). In turn, this seems to dilute the deterrence effect of whistle-blower protection. In a next step, we inquire whether these findings depend on the fact that protection is relatively easy to obtain in treatment *WBP1* (which was meant to capture the legal situation under SOX). Consequently, we now discuss two additional treatments where the requirements for obtaining protection are more stringent (see Table 1). In what follows, we focus on the reporting, investigation, and misbehavior decisions.[35]

Treatment *WBP2* Recall that the only difference between treatments *WBP1* and *WBP2* is that in the latter, the employee only obtains protection when his report also triggers an investigation. However, as discussed in Section 4, the theoretical predictions for both of these treatments coincide. In fact, this is also borne out in the experiment, and the results for treatment *WBP2* are illustrated in Figures 6 and 7 (where for the sake of comparability, we repeat the results for treatments *NoWBP* and *WBP1*). It turns out that there are no statistically significant differences compared to treatment *WBP1* with respect to either reporting or misbehavior, and, in particular, the number of fraudulent claims does not drop significantly compared to *WBP1*. That is, none of the pair-wise tests (MWU) for differences in reporting behavior between treatment *WBP1* $(0.97, 0.81, 0.89, 0.38)$ and *WBP2* $(0.9, 0.72, 0.81, 0.27)$ is statistically significant (see Figure 6(b) and (c)). The tests for treatment differences with respect to misbehavior and investigations (see Figure 7) are performed analogously. Here, the only significant difference occurs with respect to the frequency of investigations conditional on

[34] Decomposing the fraction of fraudulent claims by H-employee (0.38) across periods yields 0.37, 0.40, and 0.38 for periods 1-10, 11-20, and 21-30, respectively.

[35] As for dismissals, the results remain strongly in line with *Prediction D* and hence are not reported here. In particular, conditional on dismissal being feasible, the fraction of dismissed L-employees (H-employees) is 0.93 (0.00) in treatment *WBP2* and 0.98 (0.02) in treatment *WBP3*.

Figure 6: Varying the Requirements for Protection: Fraction of Employees Sending a Report



a report, which is higher in treatment *WBP2* (0.79 versus 0.65, $p = 0.02$, MWU).

Robustness Check I (Treatment *R1*): Reputation Cost of Investigations One might suspect that the high number of fraudulent claims in treatments *WBP1* and *WBP2* is partially driven by the fact that filing a fraudulent claim does not impose any cost on the (innocent) employer (which, in practice, might come in the form of a reputation loss in the course of a subsequent investigation). At the same time, recent findings from the experimental literature on lying (see e.g., Gneezy, 2005; Gneezy, Rockenbach, and Serra-Garcia, 2013; Fischbacher and Föllmi-Heusi, 2013) suggest that many individuals are subject to lying aversion. Moreover, as

shown in Gneezy (2005), this aversion seems to be the stronger the bigger the harm imposed on others through a lie. Hence, as a robustness check, we ran treatment *R1*, which differs from *WBP2* only by the fact that the employer's payoff is reduced by 10 points whenever an investigation occurs. We find, however, that this does not affect behavior. In particular, all pair-wise tests for treatment differences between *R1* and *WBP2* (with respect to reporting, investigations, and misbehavior) are not significant (for an illustration, see Figures 6 and 7).

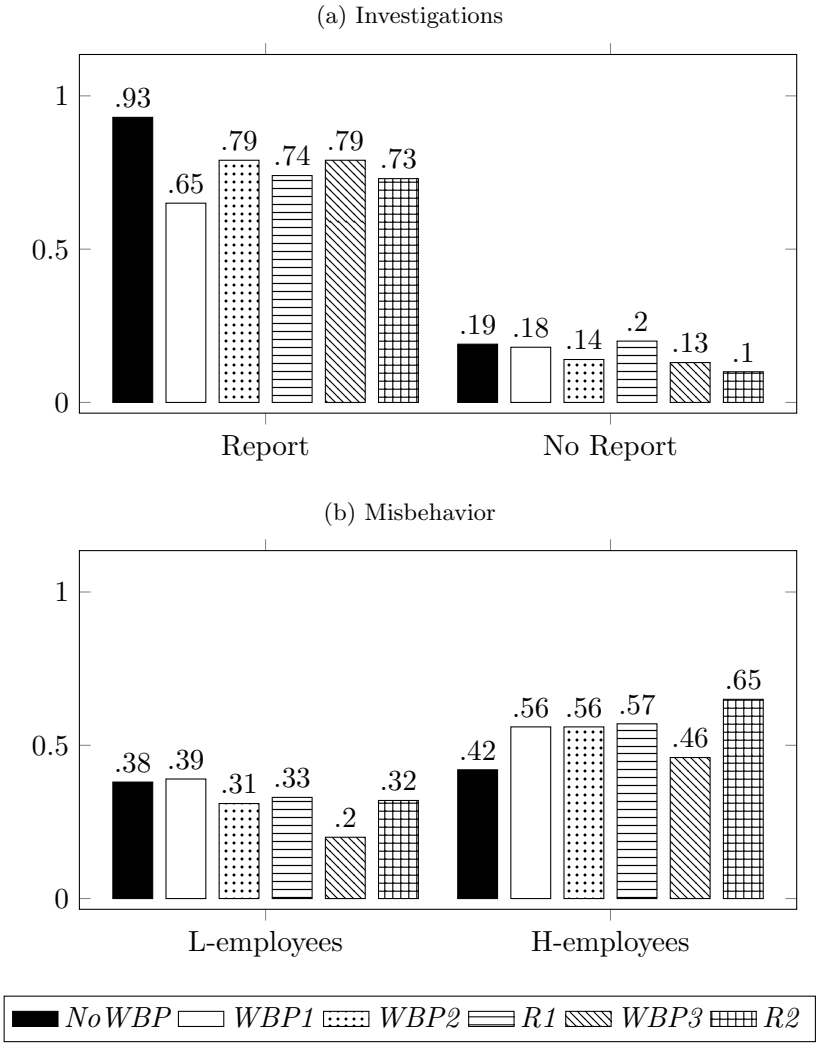
Treatment *WBP3* In treatment *WBP3* (which is meant to capture more stringent, fact-based legal regimes such as the ones in Germany or France), it is more difficult for the employee to obtain protection as, in addition to a report and an investigation (as in *WBP2*), actual misbehavior is also required. According to *Predictions WBP3*(i) and (ii), relative to *WBP1*, this does not affect truthful reports, but all fraudulent claims should be eliminated. Indeed, this prediction is strongly supported by the experimental data (see Figure 6(d)): Note first that the share of truthful claims (black bars) remains at high levels and there is no statistically significant difference to either treatment *WBP1* or *WBP2* (again using pair-wise comparisons). In fact, in all treatments with whistle-blower protection, the willingness to report misbehavior is significantly higher (at the 1% or 5% level, MWU tests) compared to *NoWBP*.

Moreover, in treatment *WBP3* fraudulent claims (white bars) indeed go down strongly for both productivity types, and they are significantly lower than in any other treatment with whistle-blower protection. For example, compared to *WBP2*, they are virtually fully eliminated for H-employees (a drop from 27% to 5%, MWU). For L-employees we also observe a substantial decrease of more than 50 percentage points (both effects with $p < 0.001$, MWU). The shares of fraudulent claims are also lower in *WBP3* compared to treatment *NoWBP*, but these differences (0.19 versus 0.23 , and 0.05 versus 0.14) are not statistically significant (MWU).

Furthermore, with respect to the responsiveness of prosecutors to reports, as can be seen in Figure 7(a), in treatment *WBP3* the frequency of investigations is significantly higher ($p < 0.01$, MWU) compared to *WBP1* (where, as discussed above, the number of investigations was lower than predicted). This finding is in line with the above reasoning that the low number of fraudulent claims in *WBP3* seems to lead to a higher responsiveness of prosecutors to reports.

Overall, in all treatments with whistle-blower protection, the level of investigations conditional on receiving a report is significantly lower compared to *NoWBP* (in all cases, $p < 0.05$, MWU). In contrast, when reporting does not occur, we find no significant differences in any

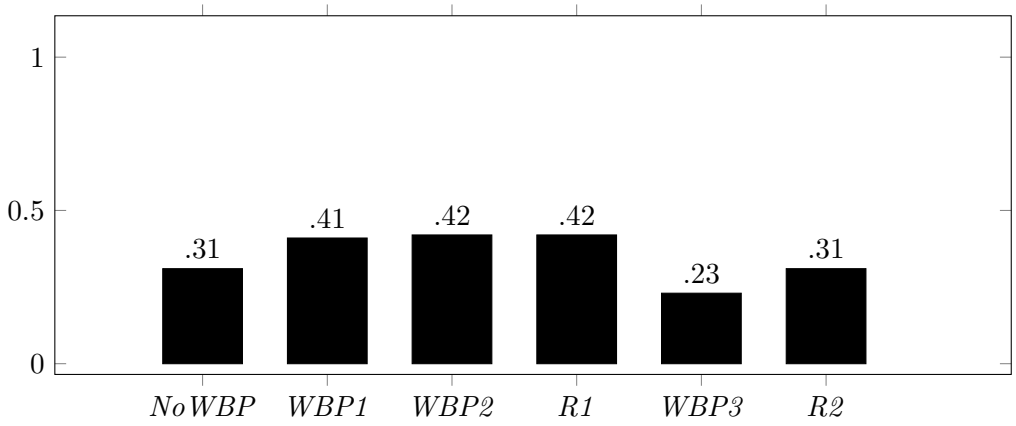
Figure 7: Varying the Requirements for Protection: Fraction of Investigations and Misbehavior



pair-wise comparison. This suggests that, in general, prosecutors seem to have more doubts about the truthfulness of reports whenever a whistle-blower protection scheme is in place.

Consider next the frequency of employer misbehavior in *WBP3* (see *Prediction WPB3* (iii), and Table 3). As displayed in Figure 7(b), for employers matched with L-employees, the frequency of misbehavior in *WBP3* (0.2) is significantly lower ($p < 0.01$, MWU) than in *WBP1* (0.39), which is in line with predictions, but the difference to *NoWBP* (0.38) turns out not to be significant ($p = 0.16$, MWU). For employers matched with H-employees, as predicted there is no statistically significant treatment difference between *WBP3* and *WBP1*. As shown above, there was no treatment difference for H-employees between *WBP1* and *NoWBP*. It is therefore

Figure 8: Fraction of Incorrect Investigation Decisions (Undetected Misbehavior and Unnecessary Investigations)



not surprising that the same holds true for the comparison of *WBP3* and *NoWBP*.

Finally, in contrast to field data, our experimental design also allows us to compare the quality of the investigation decisions across treatments. For example, with field data, it will be very difficult to assess the extent of undetected misbehavior. In particular, we can check whether a prosecutor's decision is "incorrect" in the sense of leading to either undetected misbehavior (if $I = 0$, but $M = 1$) or unnecessary investigations (if $I = 1$, but $M = 0$). To this end, we create an indicator variable that takes the value 1 if the respective prosecutor's decision is incorrect. As before, we then aggregate on the subject level, so that the unit of observation is the average number of incorrect investigation decisions by a subject who is observed in the role of prosecutor. The results are shown in Figure 8, and they provide further evidence in favor of a fact-based regime such as *WBP3*: In particular, the fraction of incorrect investigations is significantly lower compared to *NoWBP*, *WBP1*, and *WBP2* (all $p < 0.02$, MWU). Breaking this result further down, we find that the fraction of unnecessary investigations is significantly lower in treatment *WBP3* compared to *NoWBP*, *WBP1*, and *WBP2*. Moreover, also the amount of undetected misbehavior is lowest in treatment *WBP3* (e.g., it is nine percentage points lower than in *WBP1*), but these treatment differences are not statistically significant.

In summary, the fact-based whistle-blower protection regime implemented in treatment *WBP3* seems quite effective in bringing down fraudulent claims without hampering truthful reports, and it also leads to relatively low levels of misbehavior.

Robustness Check II: Erroneous Investigations (Treatment *R2*) The beneficial features of treatment *WBP3* might be driven by the assumption that investigations are very reliable in the sense of perfectly verifying the employer's actual misbehavior decision. When this assumption is relaxed, there might be a negative countervailing effect on the overall willingness to report. Presumably, when investigations are prone to type-1 error (i.e., misbehavior is not always detected), then even a truthful report might not lead to protection in case of an investigation, while the employee would be uncovered as a whistle-blower. Arguably, this leads to lower incentives for sending truthful reports and also decreases deterrence.

Consequently, we conducted a further treatment that differs from *WBP3* only in that an investigation uncovers misbehavior only with some probability. In the experiment, we have set this probability to 0.7; thereby implementing a sizeable error rate of 0.3. As can be seen in Figure 6, compared to *WBP3* this has no significant effect on the fraction of truthful claims (neither for L-employees nor for H-employees). As for fraudulent claims, there is no treatment difference for L-employees, while for H-employees the increase slightly (from 0.05 to 0.17, $p = 0.029$, MWU), but there are not significantly more fraudulent claims than in the benchmark treatment *NoWBP*.

Moreover, as can be seen from Figure 7(b), the frequency of misbehavior goes up in treatment *R2* compared to *WBP3*, but these effects are only marginally significant ($p = 0.058$ and $p = 0.045$ for employers matched with L-employees and H-employees, respectively, MWU). In addition, comparing fact-based versus belief-based regimes, there is no statistically significant treatment effect compared to *WBP1*.

6 Conclusion

In this paper, we have studied employee whistle-blowing as a means for fighting corporate fraud. To this end, we have considered, theoretically and experimentally, a setting where employees (as potential whistle-blowers) interact with employers (as potential wrong-doers) and prosecutors (who may investigate the allegations of whistle-blowers against their employers). Our main goal was to compare legal forms of whistle-blower protection, that differ with respect to the requirements under which protection can be obtained (belief-based versus fact-based). In doing so, we aimed at capturing different legal frameworks in countries such as the U.S., the U.K., and Germany. Our findings suggest that when protection is relatively easy to (obtain as under

belief-based regimes), fraudulent claims indeed become a prevalent issue. This reduces the informativeness of reports to which prosecutors respond with a lower propensity to investigate. As a consequence, the introduction of such whistle-blower protection schemes might not lead to the intended reduction of misbehavior. In contrast, these effects are mitigated under a fact-based regime where the requirements for protection are more stringent.

There are a number of issues in the context of whistle-blowing that could be analyzed in further research, possibly by enriching our framework. First, it has been argued in the business ethics literature that fostering whistle-blowing might create an "atmosphere of distrust" at the workplace, which could also be harmful for efficiency. For example, according to Dworkin and Near (1997, p. 10) "encouraging snitching can have significant organizational consequences. Such a system can nourish a climate of suspicion, hostility, and defensiveness, which can result in a loss of group identity, loyalty, and morale, with a consequent loss of efficiency." To the best of our knowledge, this issue has not yet been analyzed in economic research. Presumably, such effects might be particularly pronounced in environments where the incentive to file fraudulent claims is high. In the light of our results (in particular, the high number of fraudulent claims in treatment *WBP1*), this might constitute a further caveat against whistle-blower schemes where protection is relatively easy to obtain.

Second, while we have focussed on protection as a means of fostering whistle-blowing, an additional instrument to achieve that might be monetary rewards. Many whistle-blower laws do not allow for such rewards, but there are important exceptions, such as the U.S. False Claims Act and the Dodd-Frank-Act. As discussed in Section 2, Dyck, Morse, and Zingales (2010) and Zingales (2004) have argued in favor of using monetary rewards more intensively. In our view, it would be interesting to complement the existing research by analysing how financial rewards affect the incentive to file fraudulent claims. Furthermore, the literature on motivational crowding-out has pointed out that (modest) monetary rewards may lead to dysfunctional behavioral responses as they potentially crowd-out intrinsic motivation (see e.g. the survey by Gneezy, Meier, and Rey-Biel, 2011). Presumably, such behavioral phenomena might also come into play in the context of whistle-blowing, at least of only moderate financial rewards for whistle-blowers are introduced.

References

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

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.

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*.

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.

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.

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.

Chassang, S. and G. Padró i Miquel (2016): "Corruption, Intimidation and Whistleblowing: A Theory of Inference from Unverifiable Reports," *mimeo, New York University.*

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," *mimeo, University of Tennessee.*

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.

Dworkin, T. and J. Near (1997): "A Better Statutory Approach to Whistle-Blowing," *Business Ethics Quarterly*, 7, 1–16.

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

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

Feltovich, N. and Y. Hamaguchi (2016): "The Effect of Whistle-Blowing Incentives on Collusion: An Experimental Study of Leniency Programmes," *mimeo, Monash University.*

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

Fischbacher, U. and F. Föllmi-Heusi (2013): "Lies in Disguise - An Experimental Study on Cheating," *Journal of the European Economic Association*, 11, 525–547.

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

Friebel, G. and S. Guriev (2012): "Whistle-Blowing and Incentives in Firms," *Journal of Economics & Management Strategy*, 21, 1007–1027.

Friebel, G. and M. Raith (2004): "Abuse of Authority and Hierarchical Communication," *RAND Journal of Economics*, 35, 224–244.

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

Gneezy, U. (2005): "Deception: The Role of Consequences," *American Economic Review*, 95, 384–394.

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.

Gneezy, U., B. Rockenbach, and M. Serra-Garcia (2013): "Measuring Lying Aversion," *Journal of Economic Behavior & Organization*, 93, 293–300.

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.

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.

Kaplan, S. E. and J. J. Schultz (2007): "Intentions to Report Questionable Acts: An Examination of The Influence of Anonymous Reporting Channel, Internal Audit Quality, And Setting," *Journal of Business Ethics*, 71, 109–124.

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.

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.

Marvão, C. and G. Spagnolo (2014): "What Do We Know About the Effectiveness of Leniency Policies? A Survey of the Empirical and Experimental Evidence," *University of Stockholm, SITE Working Paper No. 28*.

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.

Minor, D. (2015): "Risk Preferences and Misconduct: Evidence from Politicians," *mimeo, Harvard Business School*.

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.

Near, J., M. Rehg, J. Van Scotter, and M. Miceli (2004): "Does Type of Wrongdoing Affect the Whistle-Blowing Process?" *Business Ethics Quarterly*, 219–242.

Rehg, M., M. Miceli, J. Near, and J. Van Scotter (2008): "Antecedents and Outcomes of Retaliation Against Whistleblowers: Gender Differences and Power Relationships," *Organization Science*, 19, 221–240.

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*.

Spagnolo, G. (2008): "Leniency and Whistleblowers in Antitrust," in *Handbook of Antitrust Economics*, ed. by P. Buccirossi, MIT Press, 259–304.

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.

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 outcome for each treatment. There, 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. The theoretical predictions of Section 4 then follow immediately from Propositions 1 - 4. The comparisons of the fractions of employers who misbehave (as stated in Table 3) 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).[36] 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 $x_H > x_L$). This 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 $R \in \{0, 1\}$ to the prosecutor indicating that the employer engaged in misbehavior, 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.[37] The prosecutor always observes whether or not a report is sent. In treatments *NoWBP* and *WBP1*, this is also observed by the employer. In treatments *WBP2* and *WBP3*, the employer learns the reporting decision in the course of an investigation (see also the discussion in Section 3 above).

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

[36] As discussed in Section 5.2, 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.

[37] For example, this could be motivated by assuming that employees face some small reporting cost.

initiated and whether or not the employer is found to be guilty is publicly observable.

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 $x_L < \overline{x} < x_H$. In this case, the employee appropriates the outsider's productivity.

Treatments We capture the following four legal regimes, which correspond to the four main treatments in the experiment (see Table 1 in the main text): In treatment *NoWBP*, the employer is free to dismiss the employee. In treatment *WBP1*, a dismissal is prohibited if and only if $R = 1$. In treatment *WBP2*, a dismissal is prohibited if and only if $R = I = 1$. Finally, in treatment *WBP3*, a dismissal is prohibited if and only if $R = I = M = 1$.

Payoffs All payoffs (monetary and non-monetary) are summarized in Table 5. 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 whistle-blower. 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. Last, but not least, the employer dislikes employing a whistle-blower, 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 which remains undetected by the prosecutor imposes a disutility $\delta > 0$ on the employee, which could reflect 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

Table 5: Payoffs

(a) *NoWBP*, *WBP1*, *WBP2*, *WBP3*: Payoffs When There is No Protection

Misbehavior	Investigation	Dismissal	Employee	Prosecutor	Employer
0	0	0	ω	0	$(x_i - \omega) - \beta \cdot \tau$
0	0	1	0	0	$(\overline{x} - \omega)$
0	1	0	ω	$-K_1$	$(x_i - \omega) - \beta \cdot \tau$
0	1	1	0	$-K_1$	$(\overline{x} - \omega)$
1	1	1	0	$-K_1 - K_2$	$(\overline{x} - \omega) + y - f$
1	1	0	ω	$-K_1 - K_2$	$(x_i - \omega) + y - f - \beta \cdot \tau$
1	0	1	$-\delta$	$-K_2 - K_3$	$(\overline{x} - \omega) + y$
1	0	0	$\omega - \delta$	$-K_2 - K_3$	$(x_i - \omega) + y - \beta \cdot \tau$

(b) *WBP1*: Payoffs When There is Protection

Misbehavior	Investigation	Dismissal	Employee	Prosecutor	Employer
0	0	n/a	ω	0	$(x_i - \omega) - \beta \cdot \tau$
0	1	n/a	ω	$-K_1$	$(x_i - \omega) - \beta \cdot \tau$
1	1	n/a	ω	$-K_1 - K_2$	$(x_i - \omega) + y - f - \beta \cdot \tau$
1	0	n/a	$\omega - \delta$	$-K_2 - K_3$	$(x_i - \omega) + y - \beta \cdot \tau$

(c) *WBP2*: Payoffs When There is Protection

Misbehavior	Investigation	Dismissal	Employee	Prosecutor	Employer
0	1	n/a	ω	$-K_1$	$(x_i - \omega) - \beta \cdot \tau$
1	1	n/a	ω	$-K_1 - K_2$	$(x_i - \omega) + y - f - \beta \cdot \tau$

(d) *WBP3*: Payoffs When there is Protection

Misbehavior	Investigation	Dismissal	Employee	Prosecutor	Employer
1	1	n/a	ω	$-K_1 - K_2$	$(x_i - \omega) + y - f - \beta \cdot \tau$

Notes: The table depicts the players' payoffs as a function of the employer's misbehavior and dismissal decisions and the prosecutor's investigation decision. As the only directly payoff-relevant effect of the employee's reporting decision occurs through the employer's belief β for facing a whistle-blower, we omit a separate column for the employee's reporting decision R for the sake of readability. Note that the belief β is endogenous and depends on the (equilibrium) behavior of the other players. Moreover, when dismissing the employee ($D = 1$) the employer never incurs disutility τ, independent of β. The payoffs in Panel (a) apply whenever the employee is *not* shielded from dismissal, and hence (i) always in treatment *NoWBP*, (ii) in treatment *WBP1* if $R = 0$ holds, (iii) in treatment *WBP2* either if $R = 0$ holds or if both $R = 1$ and $I = 0$ hold, and (iv) in treatment *WBP3* if $R = I = M = 1$ does not hold. The payoffs in Panel (b) apply in treatment *WBP1* if the employee is protected from dismissal (i.e., if $R = 1$). The payoffs in Panel (c) apply in treatment *WBP2* if the employee is protected from dismissal (i.e., if $R = I = 1$). The payoffs in Panel (c) apply in treatment *WBP3* if the employee is protected from dismissal (i.e., if $R = I = M = 1$).

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 $K_3 > K_1$.[38] 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 following sense:

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

Assumption 1. *An informative equilibrium exists and is always played.*

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, and which are unknown to the experimenter. Taking into account the prior distributions of these random variables, the predictions of Section 4 are then based on the *expected equilibrium outcomes* (see Propositions 1 - 4).

[38] The experimental payoffs of the prosecutor as reported in the main text are obtained when setting $K_2 = -10$ and $K_3 = 30$.

A.2.2 Treatment *NoWBP*: 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 $I \equiv R$ in informative equilibrium:

Lemma 1 (***NoWBP*: 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,*

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

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 (***NoWBP*: 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,*

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

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 next the employer's misbehavior decision at date 1:

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

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

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 fraudulent 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 (***NoWBP*: 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 the equilibrium outcome in informative equilibrium. As this outcome 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, which is the basis for the predictions in Section 4:

Proposition 1 (***NoWBP*: Expected Equilibrium Outcome**). *The informative equilibrium in treatment* NoWBP *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) When matched with an L-employee, the probability of observing misbehavior by the employer is* $m_L^{no} := E_{y,\tau}[M^*(x_L, y, \tau)] = 1 - H(f)$*. (vi) When matched with an H-employee, the probability of observing misbehavior by the employer 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 *WBP1*

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 (***WBP1*: 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,

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

Proof. In treatment *WBP1*, 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 he sends a report (which, however, is not feasible). □

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

Lemma 6 (***WBP1*: 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,*

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

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 (***WBP1*: Misbehavior**). *In the informative equilibrium, the employer's misbehavior decision is given by:*

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

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 (***WBP1*: 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 $\frac{K_1}{K_3} \leq B_1$. □

Lemmas 5 - 8 characterize the equilibrium outcome in informative equilibrium. As this outcome 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, which is the basis for the predictions in Section 4:

Proposition 2 (***WBP1*: Expected Equilibrium Outcome**). *The informative equilibrium in treatment* WBP1 *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) When matched with an L-employee, the probability of observing misbehavior by the employer is $m_L^1 := E_{y,\tau}[M^*(x_L, y, \tau)] = 1 - H\left(f\right)$. (vi) When matched with an H-employee, the probability of observing misbehavior by the employer is $m_H^1 := E_{y,\tau}[M^*(x_H, y, \tau)] = p_H^1$ as*

defined in (4). (vii) When (not) receiving a report, prosecutors always (never) trigger an investigation.

A.2.4 Equilibrium Outcome in Treatment *WBP2*

Note that the only difference between treatments *WBP1* and *WBP2* is that in the latter, an investigation must be triggered if the employee is to obtain protection after sending a report. Since in an informative equilibrium we have $I = R$, reporting (no reporting) always results in (no) protection, as in treatment *WBP1*. It follows that the respective equilibrium outcomes are the same in both treatments:

Proposition 3 (***WBP2*: Expected Equilibrium Outcome**). *In treatments* WBP2 *and* WBP1*, the expected equilibrium outcomes coincide.*

A.2.5 Equilibrium Outcome in Treatment *WBP3*

In treatment *WBP3*, the reporting decision is not directly observed by the employer, but since $I = R$ holds in an informative equilibrium, the employer can perfectly infer the reporting decision from observing whether or not an investigation occurs. We solve the game backwards starting with the employer's equilibrium dismissal decision at date 4:

Lemma 9 (***WBP3*: Dismissal**). *In the informative equilibrium, the following holds: The L-employee is always dismissed whenever this is feasible. The H-employee is dismissed if $I = 1$, $M = 0$, and τ sufficiently large. That is,*

$$D^*(x_\theta, I, \tau) = \begin{cases} 1 & \text{if } x_\theta = x_L \text{ holds and } R = I = M = 1 \text{ does not hold,} \\ 1 & \text{if } x_\theta = x_H,\ M = 0,\ I = 1, \text{ and } \tau > \overline{\tau}, \\ 0 & \text{else.} \end{cases}$$

where $\overline{\tau} = x_H - \overline{x}$ as defined in Lemma 1.

Proof. In treatment *WBP3*, the employee is protected from dismissal when $R = I = M = 1$. Since $I = R$ in the informative equilibrium, dismissal is, hence, feasible if either $I = 0$ holds (irrespective of M) or if both $I = 1$ and $M = 0$. If $I = 0$, the employer always dismisses the L-employee and always retains the H-employee (because $x_L < \overline{x} < x_H$). If $I = 1$ and $M = 0$, the L-employee is dismissed (again because $x_L - \tau < \overline{x}$), while the H-employee is dismissed if $x_H - \tau < \overline{x}$, i.e., for $\tau > \overline{\tau}$. □

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

Lemma 10 (***WBP3*: Reporting**). *In the informative equilibrium, both the L- and the H-employee send a report if and only if there is misbehavior. That is,*

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

Proof. The L-employee anticipates that he will be dismissed unless obtaining protection. For $M = 1$, his payoff upon choosing $R = 1$ is ω, while he would get only $-\delta$ when choosing $R = 0$. Hence, he reports. For $M = 0$, the L-employee gets zero in any case, and thus no reporting is a best response. to summarize, for the L-employee we have $R \equiv M$ for all M. Next, consider the H-employee: For $M = 1$, when choosing $R = 1$, the H-employee gets ω and $\omega - \delta$ otherwise. Hence, he reports. If $M = 0$, when choosing $R = 1$, he is retained with probability $G(\overline{\tau})$ and hence gets $G(\overline{\tau})\,\omega$. When choosing $R = 0$, he gets ω, which is strictly larger. To summarize, also for the H-employee, we have $R \equiv M$ for all M. □

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

Lemma 11 (***WBP3*: Misbehavior**). *In the informative equilibrium, the employer's misbehavior decision is given by:*

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

Proof. Given Lemmas 9 and 10, the employer anticipates that both employee types will report if and only if $M = 1$, which then leads to protection. When matched with an L-employee, when choosing $M = 1$ the employer gets $x_L - \omega + y - f - \tau$. When choosing $M = 0$, he gets $\overline{x} - \omega$. Hence, $M = 1$ is preferred if $y > f + \tau - x_L + \overline{x}$. When matched with an H-employee, when choosing $M = 1$ the employer get $x_H - \omega + y - f - \tau$. When choosing $M = 0$, he gets $x_H - \omega$. Hence, $M = 1$ is preferred if $y > f + \tau$. □

Finally, consider the investigation decision of the prosecutor. It follows from Lemma 10 that $B_0 = 0$ and $B_1 = 1$. Since $K_1 < K_3$, the condition $B_0 \leq \frac{K_1}{K_3} < B_1$ (see the proofs of Lemmas 4 and 8) is always satisfied.

Lemma 12 (***WBP3*: Investigation**). *Given the behavior of the other players as described in Lemmas 9 - 11, choosing $I^*(R) = R$ is optimal for the prosecutor.*

Lemmas 9 - 12 characterize the equilibrium outcome in informative equilibrium. As this outcome 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, which is the basis for the predictions in Section 4:

Proposition 4 (***WBP3***). *The informative equilibrium in treatment* WBP3 *has the following expected equilibrium outcome: (i) Employees of either productivity type send a report if and only if there is misbehavior. (ii) L-employees are always dismissed whenever this is feasible. (iii) H-employees are never dismissed. (iv) When matched with an L-employee, the probability of observing misbehavior by the employer is $m_L^3 := E_{y,\tau}[M^*(x_L, y, \tau)] = p_L^3$, where $p_L^3 = E_\tau[1 - H(f + \tau - x_L + \overline{x})]$. (v) When matched with an H-employee, the probability of observing misbehavior by the employer is $m_H^3 := E_{y,\tau}[M^*(x_H, y, \tau)] = p_H^1$ as defined in (4). (vi) When (not) receiving a report, prosecutors always (never) trigger an investigation.*

A.2.6 Comparing Employer Misbehavior

Propositions 1 - 4 directly lead to the predictions concerning investigations, dismissals, and reporting as presented in Section 4. The comparison of employer misbehavior across treatments and employee productivity types (see Table 3) requires some further elaboration: From Lemmas 3, 7, and 11, 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, Lemmas 3 and 7 imply that both in treatment *NoWBP* and *WBP1*, the employer misbehaves if $y > f$, while Lemma 11 implies that in treatment *WBP3* the employer misbehaves if $y > f + \tau - x_L + \overline{x}$, where $\tau - x_L + \overline{x} > 0$. Hence, $m_L^{no} = m_L^1 > m_L^3$. Second, when the employer is matched with an H-employee, Lemmas 7 and 11 imply that both in treatment *WBP1* and *WBP3*, the employer misbehaves if $y > f + \tau$, and hence $m_H^1 = m_H^3$. Moreover, the discussion of the threshold levels above immediately implies $m_L^1 > m_H^1 = m_H^3 > m_L^3$. It remains to show that $m_H^{no} > m_H^1$ holds. From Lemma 7, the threshold for y that determines m_H^1 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 *NoWBP* and *WBP1*, where all changes in *WBP1* are indicated in square brackets as follow: [In *WBP1* 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 *WBP1* 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 *WBP1* 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**.

If TRIANGLE is chosen

If the employer chooses **TRIANGLE**, there are two [In *WBP1*: four] distinct cases, depending on [In *WBP1* 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 *WBP1* 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 *WBP1* 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 *WBP1* 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 *WBP1* 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 *WBP1* 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 *WBP1*, 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 *WBP1* 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 *WBP1* 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 *WBP1*, 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 *WBP1* 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 Overview: Number of Observations

Table 6: Number of Observations Across Treatments and Conditions

(a) Number of Observations in Figure 2 (Dismissal)

	NoWBP	*WBP1*	*WBP2*	*WBP3*	*R1*	*R2*
L-employee + Report	30	30	22	24	22	22
L-employee + No Report	30	26	20	24	19	21
H-employee + Report	29	30	22	23	22	21
H-employee + No Report	30	29	21	23	22	21

(b) Number of Observations in Figure 6 (Reporting)

	NoWBP	*WBP1*	*WBP2*	*WBP3*	*R1*	*R2*
L-employee	45	45	33	36	33	33
H-employee	45	45	33	36	33	33

(c) Number of Observations in Figure 7(a) (Investigations)

	NoWBP	*WBP1*	*WBP2*	*WBP3*	*R1*	*R2*
Report	45	45	33	36	33	33
No Report	45	45	33	36	31	33

(d) Number of Observations in Figure 7(b) (Misbehavior)

	NoWBP	*WBP1*	*WBP2*	*WBP3*	*R1*	*R2*
L-employee	30	30	22	24	22	22
H-employee	30	30	22	24	22	22

(e) Number of Observations in Figure 8 (Investigations)

NoWBP	*WBP1*	*WBP2*	*WBP3*	*R1*	*R2*
45	45	33	36	33	33

Notes: As discussed in Section 5 above, in each session of the experiment, each subject played 30 periods in a given treatment, but possibly in different roles. Hence, our unit of observation are averages on the subject-level. Therefore, the number of observations in each treatment also depends on role assignments. As for panel (a), Figure 2 only exhibits treatments *NoWBP* and *WBP1*. For the other treatments, obtaining protection no longer depends on the reporting decision only. As discussed in Section 5.3, the results remain strongly in line with *Prediction D* and hence are not reported in the main text. In panel (b), since we used the strategy method to elicit the employees' reporting decision, the number of observations in Figure 6 does not vary across misbehavior decisions. The number of observations for Figures 3 and 4 are not reported here separately, as these two figures re-appear in Figures 6 and 7, and hence are already included in panels (b) to (d).