Punishment and disclosure probabilities in an experimental deception game

Sascha Behnk[a]*

Iván Barreda-Tarrazona[b]

Aurora García-Gallego[b]

[a] *Department of Banking and Finance, University of Zurich, Switzerland*

[b] *LEE and Department of Economics, Universitat Jaume I, Castellón, Spain*

Abstract: Previous findings have shown that punishment does not necessarily reduce deception in principal agent-relationships. We shed further light on this issue by first identifying a punishment mechanism that substantially decreases deception in a sender-receiver game: the possibility of imposing severe sanctions that are cost-free for the enforcer. Keeping this effective combination of punishment costs and severity constant, we then investigate how a reduction in monitoring affects deception by comparing assured revelation of sender behavior ex post with a treatment in which it is disclosed with just 50% probability. We find a similarly strong deterrence effect in both treatments suggesting that punishment works in a particular way in the deception context: once it is a credible threat, it does not require complete monitoring to be effective. We also find that receivers show similar trust levels in senders' messages for both punishment treatments, which are significantly higher than in the corresponding baselines without sanctions, further supporting our conclusion.

Keywords: deception; punishment; disclosure probability, experiment

JEL Classifications: D03; D63; K42

* Corresponding author. Contact information: Department of Banking and Finance, University of Zurich, Plattenstrasse 14, CH-8032 Zurich, Switzerland. Phone: +41 44 634 19 03. Email: sascha.behnk@bf.uzh.ch.

1. INTRODUCTION

A large number of studies provide evidence that the possibility of being punished is a strong incentive to comply with social norms in human societies and that this mechanism to enhance cooperation is widely preferred to sanction-free settings (Gürerk et al., 2006). How strong the human appetite for penalizing unfair actions is can be best represented by the fact that people are willing to incur substantial costs in order to punish selfish individuals (e.g. Fehr and Gächter, 2002), even when punishers are not directly exposed to the negative consequences of the norm defection (e.g. Fehr and Fischbacher, 2004). Beside economic motivations in terms of avoiding a fine, the simple fact that victims or observers can take actions against the person breaking the norm seems to entail emotional discomfort that can keep individuals from defection (Fehr and Gächter, 2000, Hopfensitz and Reuben, 2009). Furthermore, the willingness to enforce punishment seems to be sensitive to the type of underlying social dilemma (Molenmaker et al., 2015).

We contribute to the discussion of punishment in different contexts by investigating its effect on deception in principal-agent relationships (Jensen and Meckling, 1976). Compared to the large experimental literature on punishment effects in cooperation problems, such as public goods, there exists only a reduced number of studies investigating the impact of sanctions on economic deception, although it is a prevailing problem with long-term repercussions in everyday life (Akerlof, 1970). Furthermore, the majority of these studies shows that punishment does not lead to less deception in this particular context (see for instance Sánchez-Pagés and Vorsatz, 2007, or Peeters et al., 2013).[1] Overall, the findings in the literature indicate that punishment is not a one-size-fits-all measure to successfully reduce deception, an issue that needs further exploration in the light of policy makers' attempts to tackle the prevailing problem of economic fraud.

We first identify a well-functioning punishment mechanism in a sender-receiver game identical in the other features to the design used in Behnk et al. (2014). In their article, the authors report experimental results on how ex post disclosure of conflicts of interest affects a principal-agent setting in which deception is possible, but they do not consider punishment. They showed that a 50% probability of disclosing sender behavior after the game was played significantly reduced deception through image concerns in case the sender obtained a small gain at the expense of a comparatively big loss for his counterpart. In our experiment, apart from replicating the basic design of Behnk et al. (2014), we allowed receivers to severely punish dishonest senders, without incurring monetary enforcement costs. We chose this particular calibration based on our literature review which indicates that costs and severity are not only crucial parameters for the general credibility of punishment with regard to deception but also for it to have an effect at all in this context.

Within this design, we address a factor of fundamental importance for the efficiency of a sanctioning system in environments with information asymmetries: the probability with which defection is revealed. A typical feature of economic cooperation problems as represented, for instance, in public goods games, is

[1] In general, punishment seems to be sensitive to the environment in which it is applied to foster cooperation. Contrary to the public goods setting, punishment can even reduce pro-social behavior in other social dilemmas, such as trust games (Fehr and Rockenbach, 2003).

the possibility to detect free-riders at the end of an interaction without bearing monitoring costs. In contrast, deception is linked to an exploitation of asymmetric information and, hence, victims are not always fully aware of the miserable consequences in the direct aftermath of an economic interaction. Famous examples are Akerlof's (1970) lemon market or the selling of sub-optimal products in the financial services sector (Angelova and Regner, 2013). The possibility to find out about the honesty or dishonesty of others is therefore particularly crucial in such principal-agent relationships. Since perfect monitoring of agent behavior can be costly, if not impossible, the question that arises is whether the deterrence effect of a punishment mechanism can be sufficiently severe to keep agents from deception even if their honesty or dishonesty is not disclosed to the principals in each and every case.

To shed further light on this issue, we used two treatments in which receivers were able to punish dishonest senders and across which we varied the disclosure probability. In one of the treatments, receivers always found out about their sender's honesty or dishonesty and each receiver had the possibility to monetarily punish her sender in case she had accepted a dishonest message. In the second treatment, a receiver could sanction a dishonest sender under the same conditions in case the payoff structure was revealed to her, which happened only with 50% probability. For each of the two treatments we used as a baseline the Behnk et al. (2014) data, obtained without punishment but with the same corresponding probabilities of sender behavior disclosure.[2] As a theoretical framework, we perform an expected utility analysis, which predicts that our punishment mechanism should lead to a lower rate of honest messages when the probability of revealing sender behavior ex post is halved.

Furthermore, Eisenkopf et al. (2011) have shown that the size of a lie matters for receivers in the sense that senders were punished harder the more they tried to gain from a lie. We address this issue from the sender's point of view regarding two types of lies, a lie that leads to deception and a lie that promotes an equalization of payoffs on a Pareto-dominated level, in the modified version of Gneezy's (2005) deception game that we are using. The application of a punishment mechanism to different payoff scenarios in a within-subject setting enables us to examine in detail how various monetary temptations and different consequences for their counterparts influence the senders' behavior when punishment is possible regarding the two types of lies.

In contrast to many previous results in the literature, we find that senders chose honest messages substantially more often in all scenarios of our punishment treatments compared to the baselines, confirming the efficacy of our punishment system. Strikingly, we do not observe a significant difference in the fractions of honest messages in any of our payoff scenarios comparing assured revelation with 50% disclosure probability. The stable deterrence effect of punishment, which we observe, implies that individual monitoring effort in such principal-agent relationships could be reduced while maintaining a similar level of honesty.

Our within-subject comparison of different payoff scenarios reveals that the deterrence effect of punishment is always significant but less strong when senders are able to gain a comparatively high

[2] In contrast to previous studies on punishment in environments with asymmetric information, such as deception games, our design provides an additional advantage since the ex post disclosure of norm defections, which is necessary for the enforcement of punishments, is also present in the baselines. This setting enables us to tell apart the pure deterrence effect of punishment from the image concerns caused by disclosure.

amount from deception. On the other hand, differences in the financial consequences for the receiver do not significantly affect sender decisions in our punishment setting. We also find that the second alternative to honesty, sending payoff-equalizing messages, even if frequently chosen in the baselines, nearly disappears with the possibility of punishment. Our analysis of individual beliefs reveals that punishment works as a positive selection screen in our design since it eliminates strategic sender actions in terms of falsely promoting an equal outcome in order to actually maximize profits.

In line with sender behavior, receivers show substantially higher trust levels when severe and cost-free punishment is possible. This is an important condition since an economic relationship can only be established if all involved parties actually agree to interact with each other. Furthermore, we do not find a significant difference in acceptance rates between the two punishment treatments, implying that receivers anticipate the similarly strong deterrence effect of punishment under both disclosure probabilities, independently of the low ex post enforcement rates that we observe. These similarly high trust levels further support our notion that the deterrence effect of our punishment mechanism allows for a reduction in the principals' agency costs.

Altogether, our findings indicate that punishment works in a particular way in the deception context, which should be taken into account by policy-makers in the development of efficient measures against fraud: once punishment is a credible threat, it does not require complete monitoring in order to be effective. By analyzing subjects' beliefs and punishment considerations, we can rule out a change in strategic components and suggest that the stable deterrence is observed due to anticipated psychological costs of being punished for deception.

The paper is structured as follows: In section 2 we review the related literature. Section 3 describes the experimental design and procedures. Hypotheses are derived from an expected utility analysis in section 4. In section 5 we present our results and section 6 concludes. The experimental instructions can be found in the appendix.

2. RELATED LITERATURE

The study of Brandts and Charness (2003) is one of the first experimental investigations that introduced punishment in a principal-agent-relationship in which deception is possible. The authors examined if intentions play a role when people decide whether to sanction their counterparts. They found that receivers were indeed more inclined to punish senders when the selfish act was preceded by a deceptive message. Their results indicate that a decision about whether to enforce a sanction is not only influenced by the financial loss caused by the unfair action but also by the way the counterpart intended to obtain his earnings (see also Eisenkopf et al., 2011).

The number of experimental studies that investigate deterrence in the deception context by comparing punishment settings with sanction-free baselines is limited and shows mixed results. In Appendix A, we summarize the parameters and findings of these selected studies. The first group of articles does not find a significant reduction in deception when agents face the possibility of being sanctioned for deceiving their principals. Sánchez-Pagés and Vorsatz (2007) used a repeated sender-receiver game with alternating roles in which both players' payoffs were reduced to zero in case the receiver punished her counterpart.

Although receivers showed more trust when punishment was possible, the authors did not find a significant difference in truth-telling between their baseline and the punishment treatment. Sánchez-Pagés and Vorsatz (2009) found the same pattern in a subsequent study in which senders had the possibility to remain silent as an alternative to sending a message to the receiver. Peeters et al. (2013) modified the design of Sánchez-Pagés and Vorsatz (2007) by randomly assigning subjects to settings with and without punishment and allowing players to choose between these institutions in later stages. Again, they found no overall difference in truth-telling with and without punishment. By analyzing subgroups of punishers and non-punishers, they showed that only in the self-selected punisher group the truth was told significantly more often when sanctioning was possible. Furthermore, Church and Kuang (2009) investigated how deception is affected by sanctions in combination with ex ante disclosure of conflicting interests, which can increase deceptive behavior through "moral licensing" (Cain et al., 2005). Receivers had to estimate an unknown value and received an advice from a better informed sender. In case a receiver punished a dishonest sender, she incurred a relatively low cost of 10% of her initial endowment and the sender's earnings were reduced to a comparably low extent, that is, by one quarter of his initial endowment. The possibility to get punished alone did not reduce the bias in the senders' advice.

On the other hand, the sanctioning system led to a significant reduction in the bias when receivers were informed about the senders' conflict of interest from the beginning. Their explanation for this effect is that "with common knowledge of incentives, sanctions provide a real threat that regulates behavior" (Church and Kuang, 2009: p. 512). Xiao (2013) also found a deterrence effect of punishment in a deception game with cost-free third-party punishment, which halved the sender's payoffs in case it was enforced. Deception rates decreased when receivers observed whether the enforcer decided to punish the sender before making their decisions. However, when enforcers were able to gain profits from punishing senders or when receivers were not aware of the punishment possibility, sanctions did not reduce deception compared to the baseline. With this design, Xiao provides evidence for the importance of the communication function of punishment regarding social norms.[3]

Kimbrough and Rubin (2013) found a positive effect of punishment in a dynamic environment. The authors used a repeated trust game with pre-play messages in order to compare the effects of costly in-group punishment depending on a jury decision and the provision of information about the other group members' behavior. In case of a positive jury decision, the deceiver had to pay the outstanding amount plus the cost of one fifth of the initial endowment that the victim had to disburse for initiating the jury process. They found that players deceived significantly less when the sanctioning system was applied and that both factors, punishment and information provision, worked in a mutually reinforcing way. While information sharing led to reputation building over time, punishment reduced deception already from the beginning, implying that its deterrence effect is not affected by repeated interactions. In another dynamic setting, Reuben and Stephenson (2013) showed that lying became unprofitable over time since players were inclined to report deceivers to a central authority in order to get them punished. However, when

[e] In a subsequent study, Xiao and Tan (2014) showed that the suboptimal effect of profitable punishment can be overcome when punishers have to justify their decisions.

members were able to select who else to include in their group, whistle-blowers were often avoided, leading to groups in which the sanctioning mechanism lost its functionality.[4]

Altogether, previous experiments led to mixed results regarding the deterrence effect of punishment in the deception setting. It is striking that in those studies which do not show an effect of sanctions on deception punishment either includes substantial costs for the enforcer (Sánchez-Pagés and Vorsatz, 2007, 2009, Peeters et al., 2013) or leads to comparably slight consequences for punished senders (Church and Kuang, 2009). Here, punishment only affected sender behavior positively in case the credibility of its threat was increased, e.g. through self-selection into a punishment institution or ex-ante disclosure of conflicts of interest. Although the results are not directly comparable, due to different design features that might have influenced the anticipation of being punished in various ways, we can conclude that effective sanctioning of deceptive behavior is related to a strong reduction in earnings and a comparatively low cost for the enforcer. This tendency is in line with Egas and Riedl (2008) and Nikiforakis and Normann (2008) who found that decreasing enforcement costs and, respectively, a higher severity of sanctions enhance cooperation. In order to reflect this tendency in our design, we use a punishment mechanism that reduces the final earnings of a dishonest sender considerably while, at the same time, its enforcement is cost-free for receivers.

Subsequently, we use this design to investigate how the probability with which a receiver learns about a sender's honesty or dishonesty affects deterrence. The economic analysis of detection probabilities goes back to Becker's (1968) seminal work about the temptation to enrich oneself through a criminal act and the threat of negative consequences in case of being caught. The author predicts that less people become criminals when the probability or the severity of punishment increases, based on economic rationality and depending on risk attitudes. A large body of theoretical and empirical studies has discussed this issue so far, leading to different, sometimes even contrasting results. Although increasing the profitability of alternatives to crime has in general a higher impact than certainty and severity (Carroll, 1978), a number of empirical studies provides evidence for a positive relationship between detection probability and crime reduction, for instance regarding free-riding in public transportation (Killias et al., 2009). Other studies find contradictory results in the sense that crime rates actually increase with a higher probability of detection (e.g. Myers, 1983). Paternoster (1987) provides an overview of early studies and discusses the used methodologies.

While certainty and severity of punishment exhibit interaction effects (Stafford et al., 1986), some experimental investigations confirm a relatively higher importance of detection probabilities, for instance in Nagin and Pogarsky (2006) while others show the opposite effect (Anderson and Stafford, 2003, Friesen, 2012). Schildberg-Hörisch and Strassmair (2012) compare severity and certainty of sanctions using an experimental stealing game and obtain results that contradict Becker's (1968) predictions, except for the presence of very strong incentives. Given these inconclusive findings and the particular role of punishment in the deception context, indicated by previous studies, we are interested in the effect of punishment on deception in a sender-receiver game with different probabilities of ex post disclosure.

[4] Punishment was also used in other contexts that include information transmission, for instance, to investigate the effect of confessions (Utikal, 2012) and apologies (Fischbacher and Utikal, 2013) on individual behavior.

3. EXPERIMENTAL DESIGN

We use data coming from the anonymous two-player sender-receiver game with ex post disclosure of Behnk et al. (2014) as our baselines and also from two treatments, with the same respective probabilities of sender behavior disclosure, run specifically for this study in which we allow receivers to punish dishonest senders in an otherwise identical setting.

3.1 Sender-receiver game

In this one-shot deception game, all subjects were randomly assigned the role of either a sender or a receiver, neutrally named Player 1 and Player 2. A sender and a receiver from the same session were randomly matched. The sender was then provided with information about the payoffs that both players could get from a set of options. The receiver's task was to implement one of the options without having any information about the corresponding payoffs and by doing so she was determining both player's payoffs. In this situation of information asymmetry, the sender transmitted a message to the receiver in which he recommended one of the options as the one that provided the highest payoff for the receiver. This message could either be honest or a lie and, since payoffs were misaligned, the sender had an economic incentive to be dishonest.

We used the strategy method and presented three different payoff scenarios to the sender. Each scenario included three options with a payoff for each player as shown in Table 1.[6] The scenarios were based on Gneezy's (2005) seminal design in combination with a Pareto-dominated third option similar to Rode (2010) and Angelova and Regner (2013). We added the dominated option to reduce the effect of "deception by telling the truth" (Sutter, 2009).

Scenario	Option	Payoff sender	Payoff receiver
1 (low+;low-)	A	5	6
	B	6	5
	C	3	3
2 (low+;high-)	A	5	15
	B	6	5
	C	3	3
3 (high+;high-)	A	5	15
	B	15	5
	C	3	3

Table 1
Sender and receiver payoffs by scenario and option (in euros).

While the dominated option C always provided an equal payoff of three euros to both players, their interests were misaligned between options A and B and the intensity of this misalignment varied across scenarios. In scenario 1, a successful deception led to a comparatively low additional gain of one euro for the sender compared to an equally low loss for the receiver. For this reason, we labeled this scenario

[6] The scenarios were presented on different screens and we controlled for order effects regarding their appearance as well as for the option order within a scenario. Receivers only saw the general structure of the scenarios and options without payoff information in the experimental instructions (see Appendix B).

(low+,low-).[7] In scenario 2, the sender obtained the same profit from implementing option B as in scenario 1 but now at the expense of a higher comparative loss of ten euros for his counterpart (low+,high-). Finally, in scenario 3, the sender was able to gain an additional profit of ten euros from option B, which is higher than in the other two scenarios, at the cost of his counterpart's loss which is equally high as in scenario 2 (high+,high-). In each of the payoff scenarios, the sender selected one of the following three messages to be sent to the receiver afterwards:

> Message 1: Option A will earn you more money than the other two options.
>
> Message 2: Option B will earn you more money than the other two options.
>
> Message 3: Option C will earn you more money than the other two options.

Since each of the three options provided a different payoff to the receiver, only one of the three messages was true (henceforth called "honest message"). The remaining messages were lies that can be characterized by their payoff structure as a "deceptive message" in case the sender recommended the option that provided him with the highest payoff within a scenario and as a "payoff-equalizing" message when referring to the option that provided equal payoffs on a low level for both players.

After the sender chose a message in each payoff scenario, the computer randomly selected one of the scenarios and sent the corresponding message to the receiver, who either accepted or rejected it. In case the receiver accepted the message, the recommended option was implemented and determined the payoffs for both players. In case of rejection, one of the two non-recommended options was randomly implemented by the computer.

Before informing them about the implemented option and presenting them their payoffs, we elicited the subject's beliefs.[8] We asked senders to estimate the percentage of receivers who would accept the received message in their session. In turn, receivers were asked about the percentage of senders who had transmitted honest messages in their session. These first-order beliefs are the basis for the subjective equilibrium analysis which we apply to the punishment treatments in section 3 of this study. Furthermore, we elicited a more individual form of these beliefs by asking senders if they expected their counterpart to accept the message and by asking receivers if they believed that they had received an honest message from the sender they were matched with.

We also elicited both player types' second-order beliefs about relative payoffs by asking how much they thought their counterpart expected to gain from the message relative to their own payoffs in five categories (“much less”, “less”, “equal”, “more” and “much more”). In order to elicit the players' peer group expectations, which could be interpreted as a kind of social norm, we presented the three scenarios again to the senders and asked them to estimate in four categories how likely they believed it was that other senders had chosen a message that favored themselves in the respective scenario ("very unlikely", "unlikely", "likely" and "very likely"). In turn, without knowing the payoff scenarios, receivers were asked

[7] The label indicates how much senders earn (+) and receivers lose (-) from option B compared to option A.

[8] A variety of studies have shown the importance of belief elicitation in this context, such as Charness and Dufwenberg (2006), Peeters et al. (2012) and López-Pérez and Spiegelman (2013).

about how likely they believed it was that other receivers in their session had accepted the message which was transmitted to them by their counterpart.

After the game was played, we asked senders in the punishment treatments to what extent they had taken into account the possibility that their counterpart would punish them when deciding which message to send (not at all, not so much, much, very much). Using the same scale, receivers were asked to what extent they considered how much money their counterpart made them lose when deciding whether to punish him.

3.2 Treatments

The experiment encompassed four treatments among which both the possibility to punish dishonest senders and the disclosure probability was manipulated, as summarized in Table 2. T-treatments and the respective data from Behnk et al. (2014) are used in this study as baselines without punishment. Both of them included ex-post disclosure in the sense that the receiver found out about the honesty or dishonesty of the sender with a certain probability after the game was played. Particularly, in T100 the receiver was always shown her own payoff on the final screen and, additionally, all payoffs of both players in each of the options in the scenario that had been selected. In T50, the probability of this additional disclosure was only 50%. In all treatments both player types were informed about the ex post transparency and the corresponding probability in the experimental instructions (See Appendix). This setup ensures that receivers did not only find out about the honesty or dishonesty of their counterpart after the game was played in the punishment settings but also in the sanction-free baselines. Therefore, we are able to tell apart the sanctioning system's deterrence effect from the image concerns caused by the disclosure, which alone can lead to less deception (Behnk et al., 2014).

Treatments	Nº of subjects	% of female subjects
T100 - subsequent disclosure Payoff revelation with 100% probability	144	56%
T50 - 50% subsequent disclosure Payoff revelation with 50% probability	168	61%
P100 - subsequent disclosure and punishment T100 with option to punish dishonesty	60	57%
P50 - 50% subsequent disclosure and punishment T50 with option to punish dishonesty	64	50%

Table 2
Treatments, number of subjects and percentages of female subjects.

P-treatments are the equivalents to the T-treatments with the receiver being able to cost-free punish her sender by reducing the sender's payoff to 2 euros.[9] In P100, where a receiver always found out about the sender's honesty or dishonesty, she was able to sanction the sender in case she accepted a dishonest

[9] An additional reason for the relatively strong financial punishment is that there would not have been a sanctioning threat for choosing the payoff-equalizing message if punishment would have reduced the payoff to an amount equal to or higher than 3 euros.

message. In the P50 treatment, a receiver could sanction a dishonest sender only in case the payoff structure was revealed to her, which happened with 50% probability. There was no possibility for retaliation or reputation building since the game was played only once.

A receiver was not able to punish dishonest behavior after rejecting a message in our experiment, similar to previous studies (e.g. Brandts and Charness, 2003). In this aspect, our design differs from the common view in many legal frameworks, that even an attempt to harm others can be penalized, without the need for it to be successful. However, this rule does not apply to lies per se in the sense that a person, who does not tell the truth, can only be held responsible in a legal way for causing damage in case others base their actual behavior on the lie, except for special cases like perjury. For instance, in a situation in which an advisor recommends a financial product to an investor, rejecting this recommendation can be interpreted as a cancellation of the economic relationship between both parties. Such a situation is not in the focus of the present experiment which investigates the effect of punishment on lies that lead to a reduction in financial rewards of other. Furthermore, studies like Sánchez-Pagés and Vorsatz (2007) and Peeters et al. (2013) have shown that the willingness to punish dishonesty does not only depend on the action of the defector but also on the behavior of the potential punisher. In their experiments, the sanction rate after the sequence lie-distrust was comparatively low, indicating that this action sequence was not worth punishing for the vast majority of their subjects.

3.3 Procedures

The experiment was conducted at the LEE - Laboratory of Experimental Economics at University Jaume I, Castellon, Spain. Altogether, 436 undergraduate students from different faculties were recruited with the Online Recruitment System for Economic Experiments ORSEE (Greiner, 2004). Ten approximately 45 minutes lasting sessions were run - three in each T-treatment and two in each P-treatment. This led to a total number of 144 and 168 subjects in the T-treatments as well as 60 and 64 subjects in the P-treatments. As presented in Table 2, the sample is almost balanced between men and women.

Upon arrival, subjects entered the laboratory one by one and chose a seat in front of a computer. The experiment was programmed in z-Tree (Fischbacher, 2007) and the player roles were randomly assigned to the seats. After the experimental instructions were read aloud the subjects answered a control question using the computers. In case some subjects did not answer correctly, one of the experimenters went to their seats and explained the instructions individually to ensure a full understanding of the game. After the experiment, we paid the subjects anonymously in cash. The average earnings were around €10.

4. THEORETICAL PREDICTIONS

With our experimental design, we seek to explore behavior in a sender-receiver game with punishment and different disclosure probabilities. We derive our hypotheses regarding sender behavior from an expected utility analysis. In each scenario $i \in \{1,2,3\}$, the sender selects a message m_i to be sent to his counterpart, promoting one option from the set of options Z = {*A(honest), B(deceptive), C(payoff-equalizing)*} which provide monetary payoffs $\pi_i(z) > 0$. Since senders gain the highest payoff from successful deception and both payoffs from deception and honesty are greater than the one from payoff-equalization in every scenario, we set $\pi_i(B) > \pi_i(A) > \pi_i(C)$ for the sender. The receiver, although not provided with information about the payoffs from accepting each of the possible messages, has a say in the outcome by either accepting or rejecting the message, an element that a rational sender takes into account. Therefore, we add the sender's first order beliefs to our analysis in terms of the subjective probability $p \in [0,1]$ with which he believes the receiver will accept the message. The sender's expected utility EU_T under risk neutrality in the T-treatments obtained from sending message $m_i = z$ is reproduced in equation (1).[10]

$$EU_T(m_i = z) = p\pi_i(z) + (1-p)\frac{\sum_{j \neq z} \pi_i(j)}{2} \tag{1}$$

According to this model, a sender would never send honest messages in any scenario of the T-treatments except for the indifference point $p = 0.33$ since

$$EU_T(A) \begin{cases} < EU_T(B) & if \ \ p > 0.33 \\ < EU_T(C) & if \ \ p < 0.33 \\ = EU_T(B) = EU_T(C) & if \ \ p = 0.33 \end{cases}$$

as shown in Appendix C and illustrated in Figure 1.

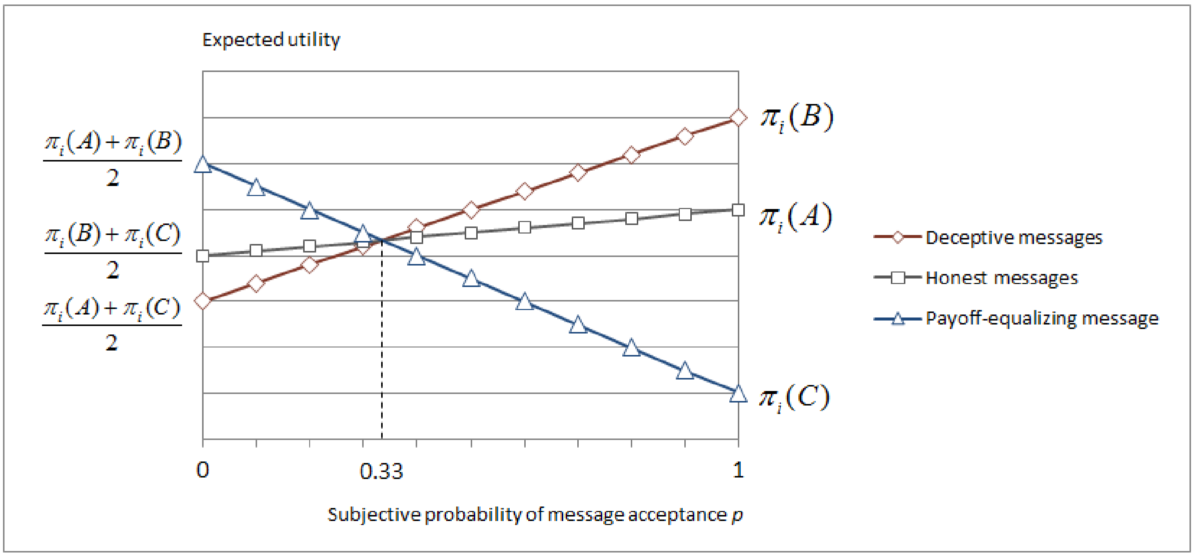

Figure 1
Sender's expected utility from sending messages in the scenarios of the T-treatments.

In the P-treatments, the receiver has the cost-free possibility to sanction her counterpart after accepting a dishonest message by reducing his payoff from the implemented option to $\pi_S < \pi_i(z)$. Since sending an honest message cannot be punished, the sender's expected utility from this action is the same as in the T-treatments. On the other hand, the expected utility from sending a deceptive or a payoff-equalizing

[10] In Behnk et al. (2014) psychological costs of lying L_i are included in the model. Since the authors did not find a significant difference in the message rates between the treatments with assured revelation and 50% disclosure probability, we assume here that L_i do not vary across treatments either. When L_i are assumed to be positive, we find a positive range of p-values in which honesty becomes a dominant strategy in the T-treatments and an identical range for the P-treatments if the subjective probability of punishment q is zero, since subsequent disclosure of sender behavior is present in all of our treatments. Therefore, we assume in the following for simplicity that $L_i = 0$.

message depends on the sender's subjective probability $q \in [0,1]$ of the receiver punishing him after sending one of these two message types in P100, which we include in equation (2).

$$EU_{P100}(m_i = z \neq A) = p(q\pi_S + (1-q)\pi_i(z)) + (1-p)\frac{\sum_{j \neq z} \pi_i(j)}{2} \tag{2}$$

In P50, the possibility to punish also depends on the 50% disclosure probability in the sense that a receiver can sanction a liar after acceptance of his message only if she becomes aware of his dishonesty, thanks to probabilistic disclosure, as included in equation (3).

$$EU_{P50}(m_i = z \neq A) = p(0.5\pi_i(z) + 0.5(q\pi_S + (1-q)\pi_i(z)) + (1-p)\frac{\sum_{j \neq z} \pi_i(j)}{2} \tag{3}$$

We show that sending honest messages becomes a dominant strategy for specific *p*-ranges in the punishment treatments, and that these ranges are wider in P100 than in P50, by comparing the intersections of the expected utility functions among the treatments.

Let us define two indifference values of *p*:

pAB : the value of *p* at which the expected utility from truth-telling *EUT(A)* = *EUT(B*), the expected utility from sending a deceptive message.

pAC : the value of *p* at which *EUT(A*) = *EUT(C*), the expected utility from sending a payoff-equalizing message.

In equation (4), we show that in the T-treatments the difference is zero between *pAC*, up to which sending payoff-equalizing messages dominates, and *pAB*, as of which sending deceptive messages dominates, leaving no value of *p* for which honesty provides the highest expected utility. The same is true for the P-treatments when *q* = 0, i.e., in case senders do not expect to be punished for dishonesty.

$$p_{AB,T} - p_{AC,T} = \frac{\pi_{\bar{A}} - \pi_{\bar{B}}}{\pi_{\bar{A}} - \pi_{\bar{B}} + \pi_i(B) - \pi_i(A)} - \frac{\pi_{\bar{C}} - \pi_{\bar{A}}}{\pi_{\bar{C}} - \pi_{\bar{A}} + \pi_i(A) - \pi_i(C)} = 0 \tag{4}$$

$$\text{with } \pi_{\bar{A}} = \frac{\pi_i(B) + \pi_i(C)}{2},\ \pi_{\bar{B}} = \frac{\pi_i(A) + \pi_i(C)}{2},\ \pi_{\bar{C}} = \frac{\pi_i(A) + \pi_i(B)}{2}, \text{where } \pi_{\bar{C}} > \pi_{\bar{A}} > \pi_{\bar{B}}$$

However, if a sender believes that *q* > 0, positive value ranges of *p* exist that increase with *q* and within which honesty is the dominant strategy in the P-treatments. This is shown in equations (5) and (6), in which we compare the values of *p* of the expected utility function intersections among the T-treatments and P100:

$$p_{AB,T} - p_{AB,P100} = \frac{\pi_{\bar{A}} - \pi_{\bar{B}}}{\pi_{\bar{A}} - \pi_{\bar{B}} + \pi_i(B) - \pi_i(A)} - \frac{\pi_{\bar{A}} - \pi_{\bar{B}}}{\pi_{\bar{A}} - \pi_{\bar{B}} + \pi_i(B) - \pi_i(A) - q(\pi_i(B) - \pi_S)} < 0 \tag{5}$$

$$p_{AC,T} - p_{AC,P100} = \frac{\pi_{\bar{C}} - \pi_{\bar{A}}}{\pi_{\bar{C}} - \pi_{\bar{A}} + \pi_i(A) - \pi_i(C)} - \frac{\pi_{\bar{C}} - \pi_{\bar{A}}}{\pi_{\bar{C}} - \pi_{\bar{A}} + \pi_i(A) - \pi_i(C) + q(\pi_i(C) - \pi_S)} > 0 \tag{6}$$

The difference $p_{AB,T}$ - $p_{AB,P100}$ is negative since $\pi_{\bar{A}} > \pi_{\bar{B}}$ and $\pi_i(B) > \pi_S$, which is equivalent to a rightward shift of the intersection point in P100 compared to the T-treatments, with punishment reducing the *p*-range within which sending deceptive messages is a dominant strategy. Correspondingly, the difference $p_{AC,T}$ - $p_{AC,P100}$ is positive since $\pi_{\bar{A}} < \pi_{\bar{C}}$ as well as $\pi_i(C) > \pi_S$ and reflects a leftward shift of the intersection point in P100 compared to the T-treatments, which leads to a comparatively smaller *p*-range within which sending payoff-equalizing messages dominates. As a consequence, truth-telling dominates between both intersection points $p_{AC,P100}$ and $p_{AB,P100}$ in P100. The same pattern appears regarding P50 as shown in equations (7) and (8):

$$p_{AB,T} - p_{AB,P50} = \frac{\pi_{\bar{A}} - \pi_{\bar{B}}}{\pi_{\bar{A}} - \pi_{\bar{B}} + \pi_i(B) - \pi_i(A)} - \frac{\pi_{\bar{A}} - \pi_{\bar{B}}}{\pi_{\bar{A}} - \pi_{\bar{B}} + \pi_i(B) - \pi_i(A) - 0.5q(\pi_i(B) - \pi_S)} < 0 \quad (7)$$

$$p_{AC,T} - p_{AC,P50} = \frac{\pi_{\bar{C}} - \pi_{\bar{A}}}{\pi_{\bar{C}} - \pi_{\bar{A}} + \pi_i(A) - \pi_i(C)} - \frac{\pi_{\bar{C}} - \pi_{\bar{A}}}{\pi_{\bar{C}} - \pi_{\bar{A}} + \pi_i(A) - \pi_i(C) + 0.5q(\pi_i(C) - \pi_S)} > 0 \quad (8)$$

Figure 2 illustrates this pattern in the punishment treatments for a selection of *q*-values. We show the pattern for representation reasons in scenario 3. While equalizing messages still dominate on the left side of the *p*-distributions, sending deceptive messages can even become a completely dominated action when the expected probability of punishment *q* is sufficiently high. Therefore, we expect to find a higher level of honesty in P100 compared to T100 and in P50 compared to T50.

H1: *Senders are more likely to send honest messages in the P-treatments compared to the respective baselines.*

It is straightforward to show that the *p*-range in which truth-telling dominates is comparatively smaller in scenario 3, where a sender can obtain a higher payoff from deceiving his counterpart than in the other two scenarios, for the same value of *q*. In Appendix C, we present the respective ranges of *p* applied to the scenario-specific payoffs in the P-treatments. This continues to be true if we allow for the possibility of the sender assigning different q_i to different payoff scenarios.[11] In such a case, it would be reasonable to assume that the sender assigns *q* a higher value in scenario 2 than in scenario 3, since the receiver's relative loss does not differ between the two scenarios, but the sender gains comparably less from deception in scenario 2, and could therefore assume that it is more likely to be punished for dishonesty in this 'mean' scenario. Hence, we hypothesize that the higher the potential payoff from deception, keeping the receivers' payoffs constant, the less inclined senders will be to send honest messages when punishment is possible.

H2*: In the P-treatments, senders are more likely to send honest messages in scenario 2 compared to scenario 3.*

Besides the different financial temptations among the scenarios, we are also interested in the sender's sensitivity to the harm he causes to his counterpart with successful deception under the punishment mechanism. We know from Gneezy (2005) that individuals take into account the negative consequences of

[11] In general, the influence of the different payoff scenarios on the senders' expected *q* should be limited since we do not provide receivers with information about the payoffs in the non-selected scenarios.

deception for others, which can be represented by higher costs of lying L_i in scenario 2 than in scenario 1. Hence, when we give up our simplicity assumption of zero costs of lying, this would lead to a wider p-range in which honesty becomes a dominant strategy in scenario 2 compared to scenario 1, as shown in Appendix C.

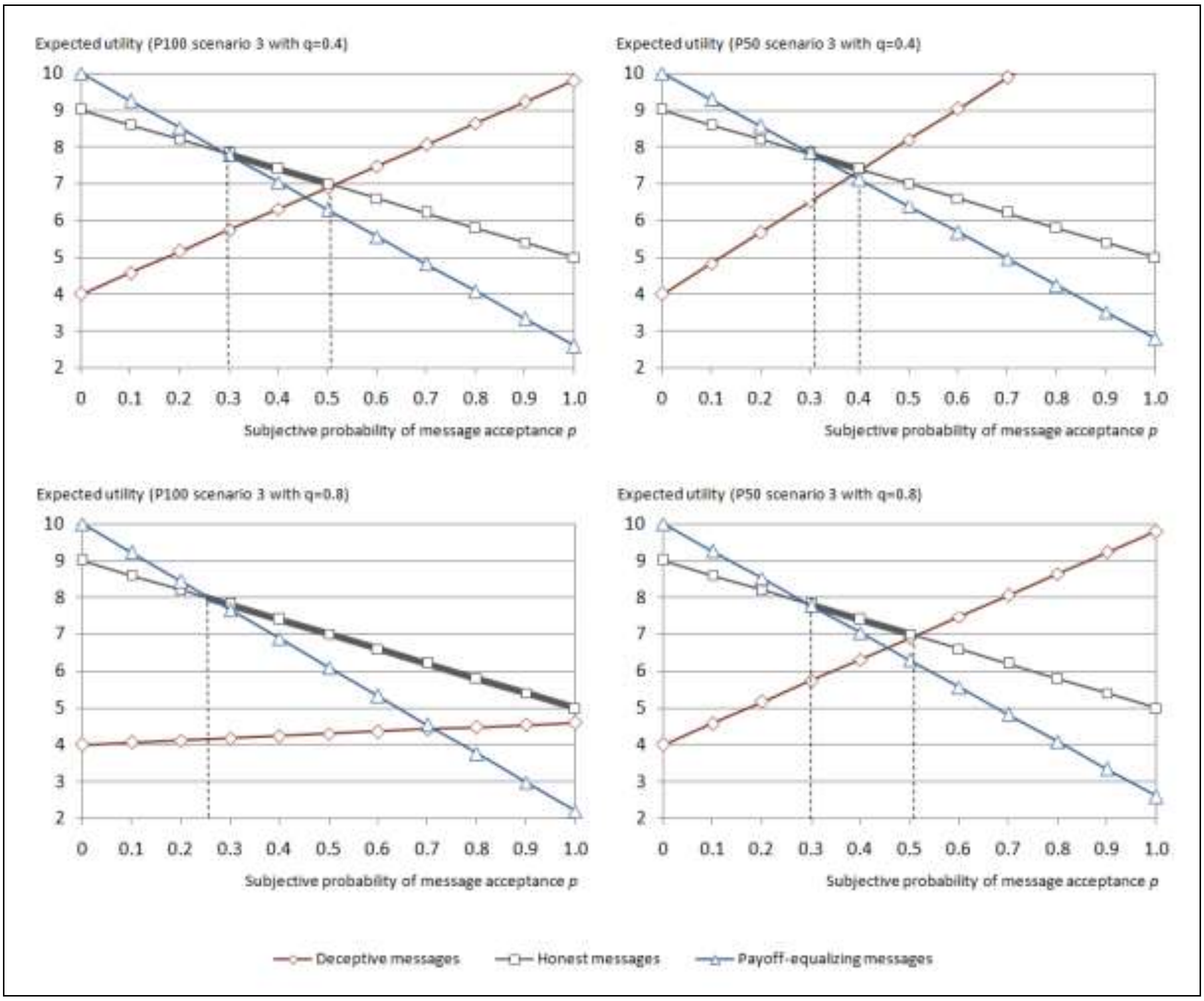

Figure 2
Sender's expected utility in P100 and P50 with low (q = 0.4) and high (q = 0.8) subjective punishment probability q in scenario 3.

Furthermore, if a sender adjusts q_i among the scenarios, it is reasonable to assume that he assigns q_i a higher value in scenario 2 than in scenario 1, due to the receiver's higher relative loss in scenario 2. The adjustment of q_i would lead to an even wider p-range in which honesty dominates in scenario 2 compared to scenario 1. Therefore, we hypothesize that a sender is less inclined to lie when deception causes a greater harm to his counterpart in the presence of punishment possibilities.[12]

H3: *In the P-treatments, senders are more likely to send honest messages in scenario 2 compared to scenario 1.*

[12] The comparison between scenarios 1 and 3 remains undefined since we do not know the relative strength of two possibly conflicting effects: On the one hand, senders have lower financial incentives to be dishonest in scenario 1 compared to scenario 3, while, at the same time, the reduced relative harm to the receiver might decrease the perceived punishment probability q and lead to more dishonest messages in scenario 1.

Finally, we can demonstrate that a decreasing probability of disclosing dishonesty reduces the *p*-range in which honesty is a dominant strategy in our setting by comparing the intersections of the expected utility functions between the P-treatments. We obtain that $P_{\mathrm{ABP,50}} < P_{\mathrm{AB,P100}}$ for all $q > 0$ as represented in equation (9) since $\pi_i(B) > \pi_i(A) > \pi_S$ and, hence, whenever the sender expects to be punished by the receiver with a positive probability, the *p*-range in which sending a deceptive message dominates is wider in P50 than in P100 as illustrated in Figure 2.

$$\begin{aligned} P_{AB,P50} &= \frac{\pi_{\bar{A}} - \pi_{\bar{B}}}{\pi_{\bar{A}} - \pi_{\bar{B}} + \pi_i(B) - \pi_i(A) - 0.5q(\pi_i(B) - \pi_S)} \\ < p_{AB,P100} &= \frac{\pi_{\bar{A}} - \pi_{\bar{B}}}{\pi_{\bar{A}} - \pi_{\bar{B}} + \pi_i(B) - \pi_i(A) - q(\pi_i(B) - \pi_S)} \end{aligned} \tag{9}$$

Accordingly, we obtain that $p_{\mathrm{AC,P50}} > p_{\mathrm{AC,P100}}$ for all $q > 0$ since $\pi_i(\mathrm{A}) > \pi_i(\mathrm{C}) > \pi_{\mathrm{S}}$ as shown in equation (10) and, hence, that the *p*-range in which sending a payoff-equalizing message dominates is also wider in P50 than in P100.

$$\begin{aligned} p_{AC,P50} &= \frac{\pi_C - \pi_{\bar{A}}}{\pi_{\bar{C}} - \pi_{\bar{A}} + \pi_i(A) - \pi_i(C) + 0.5q(\pi_i(C) - \pi_S)} \\ > p_{AC,P100} &= \frac{\pi_{\bar{C}} - \pi_{\bar{A}}}{\pi_{\bar{C}} - \pi_{\bar{A}} + \pi_i(A) - \pi_i(C) + q(\pi_i(C) - \pi_S)} \end{aligned} \tag{10}$$

Regarding the effectiveness of punishment with different disclosure probabilities, we conclude that senders are expected to be honest for smaller *p*-ranges in P50 compared to P100 if $q > 0$, independently of the payoff scenario and the amount to which the sender's earnings are reduced by punishment, as long as $\pi_{\mathrm{S}} < \pi_{\mathrm{i}}(z)$.

H4: *Senders are more likely to send honest messages in P100 than in P50.*

We turn now to the analysis of receiver behavior by asking whether these subjects trust their counterparts more often when they are able to punish them for dishonesty compared to a situation without the possibility of sanctioning. Receivers are blind regarding payoffs in our design and so we are not able to apply a similar subjective equilibrium analysis to this player type. The receiver's decision depends on her beliefs about the probability $r \in [0,1]$ with which her counterpart has sent an honest message. If the receiver expects r to be greater than 0.5, she maximizes her expected payoffs by accepting this message and vice versa in the T-treatments. A receiver is indifferent between accepting and rejecting a message in case $r = 0.5$. We hypothesize that receivers will anticipate the deterrence effect of potential punishment predicted by our analysis, i.e., that receivers will show more trust in the senders' message in P100 compared to T100 and in P50 compared to T50.

H5: *Receivers are more likely to accept the message in the P-treatments compared to the baselines.*

In line with our previous results, we further expect that receivers will anticipate a higher lying rate when the probability of disclosing dishonesty decreases and will trust their counterparts therefore more often in P100 compared to P50.

H6: *Receivers are more likely to accept the message in P100 than in P50.*

5. RESULTS AND DISCUSSION

5.1 Sender behavior

This section reports the senders' behavior in terms of their message choices in the three scenarios of each treatment as summarized in Figure 3. In contrast to many other punishment systems against deceptive behavior used in the experimental literature, we find that the presence of our severe and cost-free sanctions enhances pro-social behavior substantially, while we control for image effects by also using subsequent disclosure in our baselines. According to chi^2 tests, the rates of honest messages are higher in all scenarios of P100 compared to T100, and in P50 compared to T50, with significances at the 0.01 level. Hence, we find support for our hypothesis H1.

Result 1: *Senders select the honest message more often in the treatments with severe and cost-free sanctions compared to the baselines without punishment.*

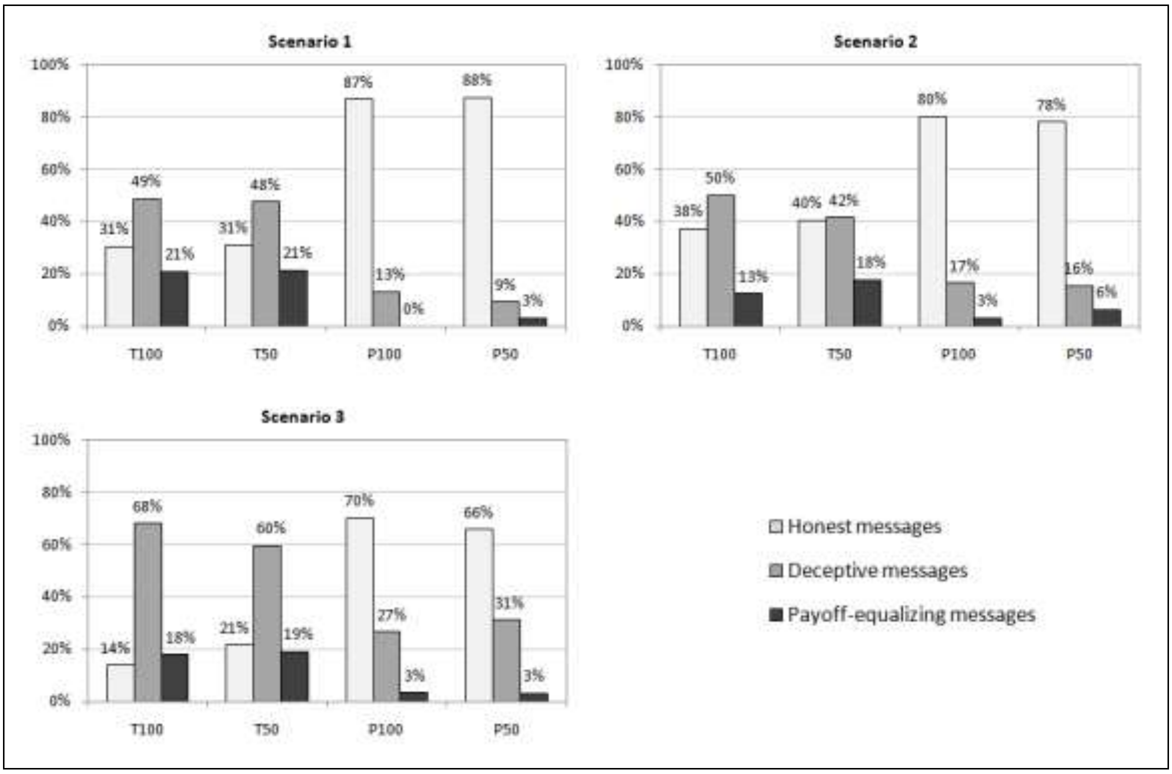

Figure 3
Fractions of messages chosen by senders per scenario and treatment.

In scenarios 1 and 2 of the P-treatments, the rates of honest messages vary between 78% and 88%. In P50, significantly fewer senders, 66%, act honestly in the third scenario, in which the sender can gain a comparatively higher amount from deception, according to a McNemar test that we used for this within-subject comparison (difference between scenarios 1 and 3: p= 0.035; difference between scenarios 2 and 3: p= 0.046). In P100, the difference between scenarios 1 and 3 is marginally significant (p=0.096) but not significant between scenarios 2 and 3. These results confirm our H2 partially in the sense that our punishment mechanism is less effective in reducing deception with high stakes in some cases, in line with the scenario comparison in Gneezy (2005). On the other hand, we do not find any significant differences between scenarios 1 and 2, in which the receiver suffers a comparatively higher loss. Therefore, we cannot confirm our H3. As the receiver is always entitled to sanction with the same severity after a disclosed successful lie, independently from the harm received, the sender seems to anticipate a similar frequency of punishment in both scenarios.

Result 2: *When punishment is possible, senders choose honest messages relatively less often when they can gain a comparatively higher amount from deception. However, their decisions are not significantly affected by a variation in the negative consequences for their counterparts.*

We turn now to the question of how the probability with which the payoff distribution is revealed to the receivers influences the deterrence effect. We compare the rates of honest messages within the two treatment families, which we define as punishment and no-punishment settings. Punishment induces a strong incentive for being honest and should therefore lead to a lower rate of honest messages when the probability of revealing sender behavior is halved in P50, as predicted by our expected utility analysis. Interestingly, we do not observe a significant difference in the fractions of honest messages in any of the scenarios between P50 and P100 and, hence, cannot confirm our H4. Senders are not affected by the difference in probabilities of their behavior being revealed and punishment becoming possible after sending dishonest messages. This finding implies that by using severe and cost-free punishment in such principal-agent relationships, monitoring effort could be reduced while maintaining the level of honesty almost as in the case of complete disclosure.

Result 3: *Independently of the different disclosure probability, we find the same pronounced deterrence effects in both P-treatments in each payoff scenario.*

Accordingly, senders deceive their counterparts less when punishment is possible.[13] The rate of deceptive messages is lower in all scenarios in the P-treatments compared to the respective baselines with significances at the 0.01 level. Altogether, between 9% and 30% of the senders choose a deceptive message in our punishment settings. In the case of 50% disclosure, it is possible that risk-loving subjects take into account the possibility of not being revealed in the end of the game and therefore choose to deceive their counterpart. However, this strategy would not explain the rates of deceptive messages in P100, with up to 27% of deception, where the revelation of deception is unavoidable. These players seem to believe that their counterparts will not punish them either because of a general lack of interest in sanctioning or a possible discomfort they experience when reducing the sender's payoff to a very low level. Let us recall that sanctioning does not entail any financial cost for the receiver but also no financial gain. As we describe section 5.2, a substantial fraction of the receivers is indeed not punishing senders although being in the position to enforce a sanction.

With regard to the payoff-equalizing messages, our subjective equilibrium analysis predicts that in treatments without punishment, sending this message type is a dominant strategy for risk-neutral senders who expect their counterparts to trust them with a subjective acceptance probability of $p < 0.33$. In line with these predictions, this alternative to honesty is frequently chosen, up to 21%, in the T-treatments. On the other hand, this message type nearly disappears with punishment. We find the highest rate, 6%, in scenario 2 of P50 while in scenario 1 of P100 none of the senders selected a payoff-equalizing message.

[13] This distinction is necessary since deception is not the only alternative to being honest in our design. Senders were also able to choose payoff-equalizing messages.

According to chi^2 tests, the respective fractions in the P-treatments are significantly lower than in treatments without sanctioning, except for scenario 2.

Let us recall from our theoretical analysis that the higher the disclosure probability, the smaller the dominance range of payoff-equalizing messages, i.e., fewer senders would maximize their expected utility by sending this message type. However, we find that varying the disclosure probabilities does again not influence sender behavior in our setting. The differences within each treatment family (comparing payoff-equalizing messages between T100 and T50, as well as between P100 and P50) are not significant.

Result 4: *Compared to the baselines, the emission of payoff-equalizing messages is significantly lower in both P-treatments.*

Sender beliefs	**Treatments**			
	T100	**T50**	**P100**	**P50**
First-order beliefs about receiver actions		Means		
Percentage of receivers accepting the message	45.18	41.88	61.00	53.97
Second-order beliefs about relative payoffs		Percentages		
Higher or much higher than sender's payoffs	29.17	22.62	46.67	21.88
Peer group beliefs		Percentages		
Other senders likely/very likely to deceive				
Scenario 1 (low+;low-)	76.39	76.19	26.67	53.13
Scenario 2 (low+;high-)	68.05	73.81	40.00	37.50
Scenario 3 (high+;high-)	68.06	71.42	66.67	78.13

Table 3
Sender beliefs across treatments.

Given the importance of beliefs on subjects' decisions in games focusing on information transmission (see e.g. Charness and Dufwenberg, 2010), we first examine their potential role as a driver of sender behavior in each treatment. The descriptive outcomes are summarized in Table 3. While senders expect receivers to accept a message with probabilities of 45% and 42% in the baselines, these first-order beliefs increase to 61% and 54% in the treatments with punishment. According to a Mann-Whitney rank-sum test, the difference between T100 and P100 is significant at the 0.01 level and, respectively, with p=0.014 between T50 and P50. The corresponding distributions are displayed in Figure 4 and are consistent with our finding that receivers actually show more trust when punishment is possible as presented in section 5.2. According to a chi^2-test, second-order beliefs do not vary significantly between the punishment treatments and the respective baselines. With regard to peer beliefs, we find a significant difference comparing T100 with P100 in scenario 1 (p<0.01) and a marginally significant one in scenario 2 (p=0.051). Between T50 and P50 peer beliefs are only significantly different in scenario 2 (p<0.01). In these cases, substantially fewer senders expect other players in their role to deceive their counterpart when there is a possibility for being punished. In scenario 3, in which stakes are high, subjects do not expect their peers to behave differently with and without punishment.

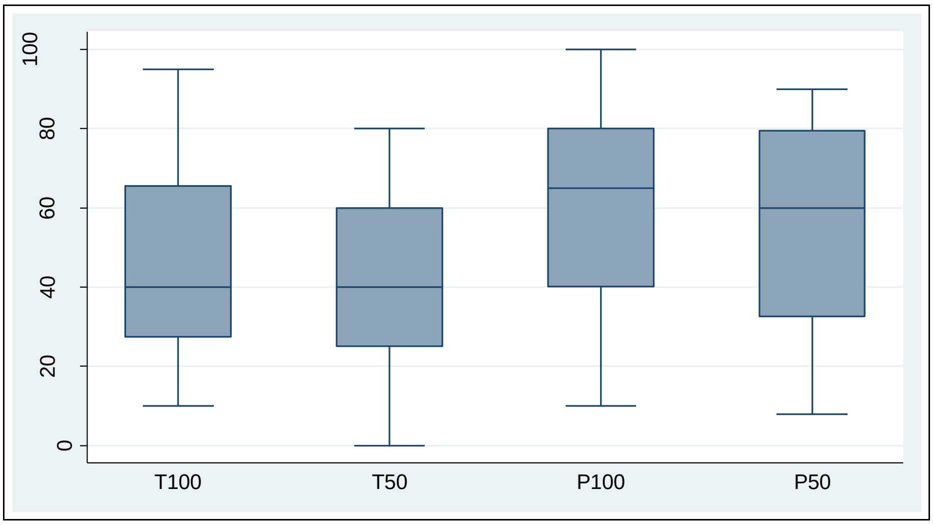

Figure 4
Box plots of sender first-order beliefs.

Result 5: *Senders anticipate an increased acceptance rate by the receivers when punishment is possible. Peer beliefs show situational differences.*

In order to shed further light on the reasons for the stable deterrence effect in the three payoff scenarios when disclosure probability is halved, we compare sender beliefs between P100 and P50. According to our expected utility analysis, payoff-maximizing senders should exhibit a higher willingness to deceive when the probability of disclosure is decreased. Since neither the experimental environment nor nominal payoffs differ between treatments, applying our theoretical framework, the stable deterrence effect could only be due to a change in the senders' subjective probabilities of message acceptance p or enforced punishment q, which could reduce the increased expected payoff in P50 to a level similar to the one in P100. Surprisingly, we do not find a significant difference in the senders' first-order beliefs or punishment considerations between P50 and P100 (Mann-Whitney rank-sum tests).

We turn now to assess a potential explanation for the stable deterrence effect, based on strategic sender behavior. The first one would be that in the P-treatments only those subjects deceive who do not expect to be punished by their counterparts anyway and, hence, would not be affected by the change in detection probabilities. However, we find that a substantial fraction of senders in our experiment, who deceived their counterpart in P100 and P50, actually considered in their decision the possibility of being punished. This fraction varies between 40% (scenario 2 in P50) and 100% (scenario 2 in P100). The second explanation would be that in the P-treatments only those subjects are honest who consider the possibility of being punished by their counterparts and, hence, abstain from deception independently of the change in detection probabilities. In our setting, we also find heterogeneity among the honest senders with regards to punishment considerations, since between 17% (scenario 2 in P50) and 33% (scenario 2 in P100) of the senders were honest in the P-treatments although they did not take into consideration the possibility of being punished. Therefore, we can also rule out that only those senders were honest (dishonest), who expected (did not expect) to be punished for deception and, hence, they are not responsible for the stable deterrence effect.

Since no strategic components change between P100 and P50 in a way that could explain the stable deterrence effect, we argue that the main driving force behind it can only be attributed to anticipated psychological costs of being punished for deceiving others. This explanation is in line with Fehr and Gächter (2000) as well as Hopfensitz and Reuben (2009) who find that punished subjects indeed suffer from negative emotions related to the experienced sanction. Furthermore, it seems as if such psychological costs are not related to guilt-aversion. Second-order beliefs vary significantly between the two punishment treatments according to a chi^2-test (p=0.039), indicating that, when the disclosure probability decreases, comparatively fewer senders believe that their counterparts expect relatively high payoffs from the message. Also, when comparing the social norm-related peer beliefs within each treatment family, we only find a significant difference between the punishment treatments in scenario 1 (p=0.034), implying that other senders are more often expected to deceive when the probability of disclosure is only 50%.

Result 6: *The stable deterrence effect cannot be attributed to a change in strategic components such as the senders' first-order beliefs or punishment considerations. We hypothesize that the effect is due to anticipated psychological costs of being punished for deception.*

In order to investigate to which extent beliefs affect sender behavior in our design, we use probit regressions for each payoff scenario with a dependent variable that takes the value 1 in case the sender chose to send an honest message.[15] We include the interactions of treatment effects and the senders' first-order beliefs as independent variables. Our expected utility analysis predicted that senders would most likely send honest messages within p-ranges located approximately on the right side of the T-treatment indifference point at p=0.33. Therefore, we leave the low p-tertile between 0 and 0.33 as the baseline. We also include second-order beliefs, peer beliefs and gender as additional independent variables. The second-order belief variable takes the value 1 whenever senders believe that their counterpart expects to achieve a comparatively higher payoff whereas the peer belief dummy takes the value 1 in case senders expect their peers to send a deceptive message in the respective scenario. The results of the probit regressions are presented in Table 4.

The models confirm the deterrence effect of punishment in our setting when individual beliefs and gender are taken into account. Senders who believe that receivers will accept their message with an intermediate or high probability are significantly more likely to send honest messages in all scenarios of the P-treatments. Additionally, in scenario 3, in which a sender can gain a comparatively higher amount from deception, expecting the receiver to follow with an intermediate or high probability leads to accordingly less honesty in the treatments without punishment. These models allow us to directly compare the coefficients of the respective variables with different disclosure probabilities. Although our theoretical model predicts a weaker punishment effect when the probability of revealing the sender's dishonesty decreases, the P50 coefficients are not significantly different from the P100 coefficients in any scenario.

[15] We do not use multinomial regressions to control for the three types of sender behavior (honesty, deception and payoff-equalization) since there are only few or even no observations regarding payoff-equalizing messages in the P-treatments.

These findings confirm that the effect of our punishment mechanism does not change between the two different disclosure probabilities even when beliefs and gender are taken into account.

Treatment, belief and gender effects		Scenario 1 (low+;low-)		Scenario 2 (low+;high-)		Scenario 3 (high+;high-)	
Honesty	T50 x first-order>0.33	-0.167	(0.248)	0.056	(0.237)	-0.497*	(0.272)
	T100 x first-order>0.33	-0.039	(0.257)	-0.200	(0.252)	-0.581**	(0.291)
	P50 x first-order>0.33	1.367***	(0.355)	0.877***	(0.331)	0.868***	(0.308)
	P100 x first-order>0.33	1.244***	(0.392)	0.883***	(0.338)	1.044***	(0.319)
	Second-order_more	0.544**	(0.213)	-0.074	(0.206)	0.219	(0.213)
	Peer_group_lying	-0.770***	(0.199)	-0.761***	(0.186)	-0.372	(0.237)
	Female	0.187	(0.193)	-0.121	(0.185)	0.380*	(0.199)
	Constant	-0.071	(0.218)	0.389*	(0.205)	-0.500*	(0.255)
N		218		218		218	
LR chi^2		64.48***		44.04***		47.10***	
Pseudo R^2		0.2140		0.1358		0.1721	

Note: *** p-value < 0.01; ** p-value < 0.05; * p-value < 0.1. Standard errors in parentheses.

Table 4
Probit regression models for sending honest messages in each scenario.

Second-order beliefs, on the other hand, are only significant in scenario 1, in the sense that senders are more likely to choose honest messages when they believe that their counterpart expects to gain a relatively higher profit in the scenario in which the payoff misalignment is relatively low. With regard to peer beliefs, the relative probability of choosing an honest message is lower when a sender expects the majority of the other players in their role to deceive in a specific scenario. This is true for scenarios 1 and 2, but not for scenario 3 in which senders face comparatively high financial incentives to deceive. In contrast to other findings in sender-receiver games like Dreber and Johannesson (2008) and Gneezy and Erat (2012), we find no significant gender effects regarding dishonesty at conventional levels.

We conclude from these results that punishment, in combination with first-order beliefs, is the overall driving factor for decisions between sending honest and dishonest messages in our design, while other beliefs seem to exhibit more situational effects. To explain this pattern it is important to make a distinction between the belief types. First-order beliefs have an actual relevance for strategic sender actions. Since receivers undertake the task of finally implementing the payoff-determining option, strategic senders have to incorporate in their decision the probability of message acceptance. On the one hand, given the substantial impact of punishment that we observed, strategic considerations become obsolete when a sender is forced to be honest in order to avoid a sanction. On the other hand, in case a sender actually considers deceiving his counterpart regardless of the possibility of being punished, which we observed for a particular fraction of subjects, first-order beliefs become relevant, regardless of the scenario. The guilt aversion-related second-order beliefs and beliefs about perceived social norms among peers are rather related to emotional motivations and have only limited relevance for strategic considerations resulting in scenario-specific effects.

Result 7: *Punishment and first-order beliefs are the driving factors for sending honest messages, while the influence of other belief types is comparatively more sensitive to the underlying payoff scenario.*

Finally, we aim at identifying an explanation with respect to why payoff-equalizing messages nearly disappear with punishment, and have a closer look at a specific form of the sender's first-order beliefs. This dummy variable takes the value 0 in case the sender expects the receiver he is matched with to reject the message, i.e., these senders choose payoff-equalizing messages strategically in order to get one of the remaining options implemented. As presented in Table 5, on average, over 90% of the senders who chose the payoff-equalizing message in the T-treatments believed that their counterpart would reject their message and, hence, have chosen this message for strategic reasons.

Payoff-equalizers per treatment and scenario	**Treatments**			
	T100	T50	P100	P50
Scenario 1				
Number of payoff-equalizers	18	15	1	0
Percentage of payoff-equalizers with FOB=0	83.3%	93.3%	0%	-
Scenario 2				
Number of payoff-equalizers	15	9	2	1
Percentage of payoff-equalizers with FOB=0	86.7%	100%	0%	0%
Scenario 3				
Number of payoff-equalizers	16	13	1	1
Percentage of payoff-equalizers with FOB=0	87.5%	92.3%	0%	0%

Table 5
Payoff-equalizing senders and their direct first-order beliefs (FOB) per treatment and scenario.

By contrast, all the remaining payoff-equalizers believed that their counterpart would actually implement the option mentioned in the message in the P-treatments. This indicates that the effectiveness of punishment does not only depend on the action per se but also on the intention with which the action is taken, even in a situation where the potential punisher does not receive information about the true intention of the defector and no communication between the two parties is possible to set the record straight. We conclude that punishment does not only successfully reduce the transmission of deceptive messages but also eliminates strategic sender actions in the sense of maximizing payoffs by promoting the Pareto-dominated outcome while expecting the receiver to reject the message.

Result 8: *Our punishment mechanism works in the sense of a positive selection screen, eliminating strategic sender actions.*

5.2 Receiver behavior

Receivers take on an active role in our design in terms of deciding whether the option mentioned in the message will actually be implemented or not. Since they do not get any information about the players' payoffs from each option, accepting a message can be considered as an act of trust and rejection as distrust in our setting. Therefore, a comparison of the message acceptance rates across treatments, as presented in Figure 5, allows us to make a statement about how punishment and disclosure probabilities affect the trust levels.

We find that only 33% and, respectively, 48% of the receivers accept the messages in the T-treatments, while trust levels are substantially higher in the punishment treatments, 70% in P100 and 78% in P50. The differences between T100 and P100 as well as T50 compared to P50 are significant at the 0.01 level according to a chi^2 test. Hence, we can confirm our H5 about receivers being more likely to accept messages in the P-treatments compared to the respective baselines.

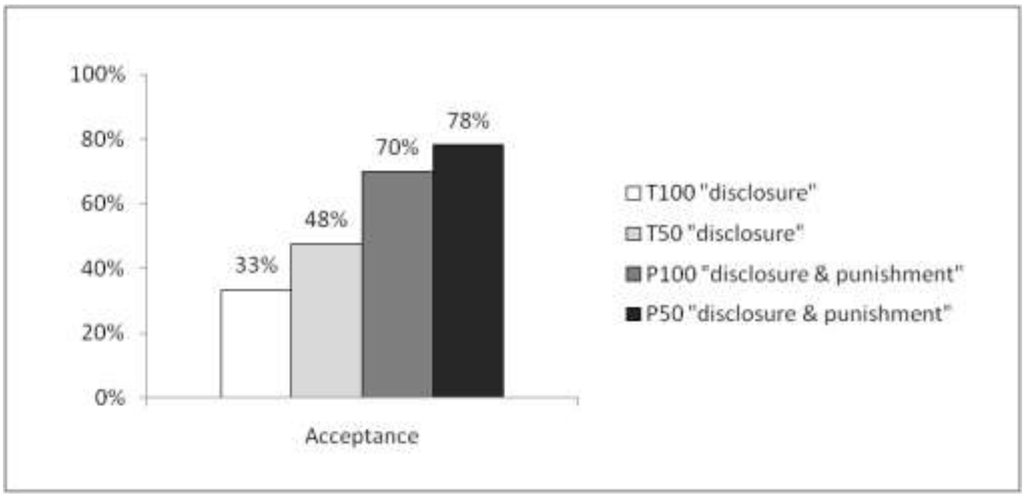

Figure 5
Acceptance rates per treatment.

Regarding the variation of disclosure probabilities, we find a higher acceptance rate in T50 compared to T100. This difference is marginally significant (p=0.071). However, the acceptance rates in the P-treatments are not significantly different and, therefore, we cannot confirm our H6. We conclude that receivers seem to anticipate that punishment provokes a substantial deterrence effect independently of the disclosure probability, a finding that supports our notion that monitoring effort could be reduced in our punishment setting while maintaining the levels of honesty.

Result 9: *Receivers trust their counterparts significantly more often when they are able to punish dishonest senders, showing similarly high trust levels in the punishment treatments independently of the disclosure probability.*

We now turn to the receivers' beliefs. In Table 6 we present the respective means and percentages of first-order, second-order and peer beliefs. In contrast to second-order expectations, first-order beliefs are significantly different, at the 0.01 level, between T100 and P100 according to a Mann-Whitney rank-sum test, in the sense that more receivers expect senders to send an honest message when punishment is possible. In line with our previous findings, comparatively more receivers believe that the other players in their role will accept the message in the P-treatments, independently of the disclosure probability. A chi^2 test shows significant differences at the 0.01 level when comparing T100 with P100 and, respectively, T50 with P50.

Receiver beliefs	**Treatments**			
	T100	**T50**	**P100**	**P50**
First-order beliefs about sender actions		Means		
Percentage of senders sending honest message	34.50	40.46	56.17	47.38
Second-order beliefs about relative payoffs		Percentages		
Higher or much higher than receiver's payoffs	29.17	28.57	30.00	34.38
Peer group beliefs		Percentages		
Other receivers likely/very likely to accept	43.06	42.86	80.00	68.75

Table 6
Receiver beliefs across treatments.

We use probit models, in which we control for treatment effects, individual beliefs and gender, to compare message acceptance among the punishment treatments and their respective baselines. The independent variable of this model measures if a receiver rejects (value 0) or accepts (value 1) the sender's message. We present two specifications of each model. In specification (1), we include dummy variables for the possibility of being punished and another for gender. As shown in Table 7, the relative probability of accepting a message increases when sanctioning is possible under both disclosure probabilities. The coefficient for punishment in P50 is not significantly different from the one in P100, confirming our earlier finding that the trust level is not higher with an increased probability of revealing sender behavior.

Treatment, belief and gender effects	**P50**		**P100**	
	(1)	(2)	(1)	(2)
Punishment	0.769*** (0.289)	1.074*** (0.391)	0.952*** (0.285)	0.413 (0.362)
First-order_honest		0.056*** (0.010)		0.031*** (0.009)
Second-order_more		-0.426 (0.366)		-0.789** (0.381)
Peer_group_accept		-0.186 (0.346)		0.700** (0.339)
Female	-0.283 (0.246)	-0.473 (0.342)	0.081 (0.263)	-0.066 (0.319)
Constant	-0.112 (0.203)	-1.799*** (0.427)	-0.478** (0.217)	-1.666*** (0.379)
N	116	116	102	102
LR chi^2	10.56***	75.65***	11.77***	49.04***
Pseudo R^2	0.0664	0.4754	0.0841	0.3504

Note: *** p-value < 0.01; ** p-value < 0.05; * p-value < 0.1. Standard errors in parentheses.

Table 7
Probit regression models for accepting messages comparing T50 with P50 (left) and T100 with P100 (right).

This pattern appears in a more pronounced form when we include the receivers' beliefs in specification (2). The possibility of punishing dishonest senders after accepting a message is still a main driving force for trusting senders in P50 while first-order beliefs also show a significant and consistent effect on acceptance rates in the sense that receivers are more likely to accept a message when they expect senders to be honest.

On the other hand, the pure treatment effect of P100 does not have a significant impact on receiver decisions when individual beliefs are taken into account. All three belief types show a significant effect on message acceptance and seem to overlay the punishment threat in P100, which is in line with our previous observation that first-order beliefs are only significantly different between T100 and P100. Beside the aforementioned first-order belief effect, receivers trust their counterparts less if they believe that senders have high relative payoff expectations according to the significantly negative effect of second-order beliefs on message acceptance. Furthermore, receivers are more likely to trust a sender when they expect their peers to accept the messages they received. Again, we do not find significant gender differences.

Result 10: *We find a strong effect of punishment on trust in P50 but not in P100, as it is mediated by the change in beliefs.*

We finally examine whether possible punishments are actually enforced by receivers. First of all, there are only few receivers who are able to sanction their counterpart, due to the fact that the vast majority of senders chooses honest messages in the P-treatments. The respective rates of punishment possibilities are 13.3% in P100 and 18.8% in P50, as presented in Table 8. Interestingly, we find that only two out of four possible punishments in P100 are enforced. The enforcement rate in P50 is even lower with one out of six receivers. It is surprising that so few receivers actually sanctioned their counterparts without having to bear any monetary costs in this anonymous setting.

Treatment	Possibilities to punish	Enforced punishments
P100	4 (13.3%)	2 (50%)
Scenario 1	0	0
Scenario 2	3	1
Scenario 3	1	1
P50	6 (18.8%)	1 (16.67%)
Scenario 1	1	0
Scenario 2	1	0
Scenario 3	4	1

Table 8
Punishment enforcement rates.

We expected more enforced punishments when the computer selected the 'mean' scenario 2, in which the sender can gain a comparatively low amount from deception at a high relative cost for the receiver. Altogether, there were four cases in which punishment was possible in scenario 2, out of which only one was finally enforced. In the other two cases, in which punishment was enforced, scenario 3 was previously selected by the computer, indicating that the punishment decision might have been taken rather independently of the underlying payoffs. These findings have to be interpreted with caution, given the overall low number of punishments, but show a potential for further research.

6. CONCLUSION

In this study, we first identified a potentially well-functioning punishment system against deception in principal-agent relationships. We further investigated how the possibility to punish dishonesty affects behavior in a one-shot deception game with different probabilities of disclosing dishonesty and compared this behavior to the case without punishment possibilities. Since receivers found out about their counterpart's behavior with the same probability after the game was played in the paired treatments, we were able to tell apart the deterrence effect of potential punishment from the image concerns caused by ex post disclosure, which alone can lead to less deception.

In contrast to many previous results, we showed that a substantially higher fraction of senders acts honestly when a cost-free and severe punishment mechanism is present. By using a within-subject comparison of different payoff scenarios, we also found that the deterrence effect of punishment is less pronounced when senders face the temptation of gaining a comparatively high amount from deception. On the other hand, senders seem to be unaffected by the financial consequences of deception in terms of the receiver's relative loss when facing the possibility of being sanctioned. Furthermore, our punishment mechanism works in the sense of a positive selection screen, eliminating the strategic sending of payoff-equalizing messages, while leaving a fraction of truly inequity-averse senders unaffected.

With regard to the different disclosure probabilities, a subjective equilibrium analysis predicts that our punishment mechanism should lead to a lower rate of honest messages when the probability of revealing sender behavior is halved. Interestingly, we did not observe a significant difference in the fractions of honest messages in any scenario comparing assured revelation with 50% disclosure probability. By analyzing subjects' beliefs and punishment considerations, we can rule out a change in strategic components and suggest that the stable deterrence is observed due to anticipated psychological costs of being punished for deception. Our conclusion from this stable deterrence effect of punishment is that monitoring effort could be reduced in similar settings while maintaining a level of honesty similar to the one of complete disclosure. Besides a potential reduction in agency costs, this implication is particularly important regarding the existence of thresholds above which monitoring is able to create crowding-out effects, in the sense that agents can even reduce their norm adherence when they are monitored excessively (Dickinson and Villeval, 2008, Ichino and Muehlheusser, 2008).

Receivers, on the other hand, show substantially higher trust levels when they are able to sanction dishonest senders. We conclude that punishment in principal-agent relationships in which deception is possible does not only lead to more honesty but enhances trust levels accordingly, even in a setting without repeated interactions. This is an important condition since both senders and receivers are active players and an intact economic relationship can only be established when all involved parties actually agree to interact with each other. Even a perfectly working deterrence mechanism is not an optimal tool if it does not convince principals to establish economic relationships in the first place. Furthermore, the difference between the high acceptance rates in the punishment treatments is not significant, implying that receivers anticipate that the punishment's deterrence effect is independent of our variation in the disclosure probabilities.

However, we find surprisingly low enforcement rates. Since sanctioning was cost-free and not-profitable to the enforcer in our design, we assume that the majority of receivers were reluctant to punish because of an emotional discomfort resulting from the severity of the sanction. Our results might be an indication for the act of punishing being not only linked to emotions like anger about the unethical act or guilt for otherwise forgone opportunities to punish (see Nelissen and Zeelenberg, 2009), but also to pity in terms of an anticipation of anger and guilt that the punished one experiences (Hopfensitz and Reuben, 2009). This might result in opposing motivations of the involved party with respect to the enforcement of punishment as in Whitson et al. (2015), based on the severity of the sanction in relation to the size of the norm defection.

Since we focus on cost-free punishment in terms of a sanction implemented by a central authority in our design, an open question for further research would be how different disclosure probabilities interact with a variety of punishment costs. Given that we fixed the disclosure probabilities exogenously, an additional potential for further investigation would be to allow for an endogenous monitoring level. As shown in the model of Buechel and Muehlheusser (2014), the deterrence effect of punishment can even decrease when the central authority delegates the monitoring task to an autonomous agent with own interests. Furthermore, the introduction of repeated interactions with sanctioning possibilities, as in Kimbrough and Rubin (2013), could cast some light on the stability of the deterrence effect of cost-free punishment and on the sanction enforcement rates under different disclosure probabilities in the long run.

ACKNOWLEDGEMENTS

We would like to thank Nikolaos Georgantzís, Nobuyuki Hanaki, Tibor Neugebauer, Stephan Meier and Dirk Sliwka for their helpful comments. Financial support by Universitat Jaume I (project P1.1B2015-48), the Spanish Ministry of Economics and Competitiveness (projects ECO2013-44409-P and ECO2015-68469-R) and the Bank of Spain Excellence Chair in Computational Economics (project 11I229.01/1) is gratefully acknowledged.

REFERENCES

Akerlof, G.A., 1970. The market for 'lemons': Quality uncertainty and the market mechanism. The Quarterly Journal of Economics 84(3), 488-500.

Anderson, L.R., Stafford, S.L., 2003. Punishment in a regulatory setting: Experimental evidence from the VCM. Journal of Regulatory Economics 24(1), 91-110.

Angelova, V., Regner, T., 2013. Do voluntary payments to advisors improve the quality of financial advice? An experimental sender-receiver game. Journal of Economic Behavior and Organization 93, 205-218.

Becker, G.S., 1968. Crime and punishment: An economic approach. Journal of Political Economy 76(2), 169-217.

Behnk, S., Barreda-Tarrazona, I., García-Gallego, A., 2014. The role of ex post transparency in information transmission - An experiment. Journal of Economic Behavior and Organization 101, 45-64.

Brandts, J., Charness, G., 2003. Truth or consequences: An experiment. Management Science 49, 116-130.

Buechel, B., Muehlheusser, G., 2014. Black sheep or scapegoats? Implementable monitoring policies with unobservable levels of misbehavior. CESifo Discussion Paper No. 4698 - revised and extended version.

Cain, D.M., Loewenstein, G., Moore, D.A., 2005. The dirt on coming clean: Perverse effects of disclosing conflicts of interest. Journal of Legal Studies 34(1), 1-25.

Carroll, J.S., 1978. A psychological approach to deterrence: The evaluation of crime opportunities. Journal of Personality and Social Psychology 36(12), 1512-1520.

Charness, G., Dufwenberg, M., 2006. Promises and partnership. Econometrica 74(6), 1579-1601.

Charness, G., Dufwenberg, M., 2010. Bare promises: An experiment. Economics Letters 107(2), 281-283.

Church, B.K., Kuang, X., 2009. Conflicts of interest, disclosure, and (costly) sanctions: Experimental evidence. The Journal of Legal Studies 38(2), 505-532.

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

Dreber, A., Johannesson, M., 2008. Gender differences in deception. Economics Letters 99(1), 197-199.

Egas, M., Riedl, A., 2008. The economics of altruistic punishment and the maintenance of cooperation. Proceedings of the Royal Society B: Biological Sciences 275(1637), 871-878.

Eisenkopf, G., Gurtoviy, R., Utikal, V., 2011. Size matters - When it comes to lies. Working Paper Series of the Department of Economics, 2011-14, University of Konstanz.

Fehr, E., Fischbacher, U., 2004. Third-party punishment and social norms. Evolution and Human Behavior 25, 63-87.

Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. American Economic Review 90(4), 980-994.

Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415(6868), 137-140.

Fehr, E., Rockenbach, B., 2003. Detrimental effects of sanctions on human altruism. Nature 422(6928), 137-140.

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

Fischbacher, U., Utikal, V., 2013. On the acceptance of apologies. Games and Economic Behavior 82, 592-608.

Friesen, L., 2012. Certainty of punishment versus severity of punishment: An experimental investigation. Southern Economic Journal 79(2), 399-421.

Gneezy, U., 2005. Deception: The role of consequences. American Economic Review 95(1), 384-394.

Gneezy, U., Erat, S., 2012. White lies. Management Science 58(4), 723-733.

Greiner, B., 2004. An online recruitment system for economic experiments. In: Kremer, K. and Macho, V. (Eds.). Forschung und wissenschaftliches Rechnen, GWDG Bericht 63, Gesellschaft für wissenschaftliche Datenverarbeitung, Göttingen, 79-93.

Gürerk, Ö., Irlenbusch, B., Rockenbach, B., 2006. The competitive advantage of sanctioning institutions. Science 312(5770), 108-111.

Hopfensitz, A., Reuben, E., 2009. The importance of emotions for the effectiveness of social punishment. Economic Journal 119, 1534-1559.

Ichino, A., Muehlheusser, G., 2008. How often should you open the door? Optimal monitoring to screen heterogeneous agents. Journal of Economic Behavior & Organization 67(3), 820-831.

Jensen, M.C., Meckling, W.H., 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. Journal of Financial Economics 3(4), 305-60.

Killias, M., Scheidegger, D., Nordenson, P., 2009. The effects of increasing the certainty of punishment: A field experiment on public transportation. European Journal of Criminology 6(5), 387-400.

Kimbrough, E.O., Rubin, J., 2013. Sustaining group reputation, Discussion Papers dp13-02, Department of Economics, Simon Fraser University.

López-Pérez, R., Spiegelman, E., 2013. Why do people tell the truth? Experimental evidence for pure lie aversion. Experimental Economics 16(3), 233-247.

Molenmaker, W.E., de Kwaadsteniet, E.W., van Dijk, E., 2014. On the willingness to costly reward cooperation and punish non-cooperation: The moderating role of type of social dilemma. Organizational Behavior and Human Decision Processes, 125(2), 175-183.

Myers, S.L., 1983. Estimating the economic model of crime: Employment versus punishment effects. Quarterly Journal of Economics 98, 157-166.

Nagin, D.S., Pogarsky, G., 2006. An experimental investigation of deterrence: Cheating, self-serving bias, and impulsivity. Criminology 41(1), 167-194.

Nelissen, R.M.A., Zeelenberg, M., 2009. Moral emotions as determinants of third party punishment: Anger, guilt, and the functions of altruistic sanctions. Judgment and Decision Making 4(7), 543-553.

Nikiforakis, N., Normann, H., 2008. A comparative statics analysis of punishment in public good experiments. Experimental Economics 11(4), 358-369.

Paternoster, R., 1987. The deterrent effect of the perceived certainty and severity of punishment: A review of the evidence and issues. Justice Quarterly 4(2), 173-217.

Peeters, R., Vorsatz, M., Walzl, M., 2012. Beliefs and truth-telling: A laboratory experiment. University of Innsbruck, Working Papers in Economics and Statistics 2012-17.

Peeters, R., Vorsatz, M., Walzl, M., 2013. Truth, trust, and sanctions: On institutional selection in sender-receiver games. Scandinavian Journal of Economics 115(2), 508-548.

Reuben, E., Stephenson, M., 2013. Nobody likes a rat: on the willingness to report lies and the consequences thereof. Journal of Economic Behavior and Organization 93, 384-391.

Rode, J., 2010. Truth and trust in communication: Experiments on the effect of a competitive context. Games and Economic Behavior 68(1), 325-338.

Sánchez-Pagés, S., Vorsatz, M., 2007. An experimental study of truth-telling in a sender-receiver game. Games and Economic Behavior 61(1), 86-112.

Sánchez-Pagés, S., Vorsatz, M., 2009. Enjoy the silence: an experiment on truth telling. Experimental Economics 12(2), 220-241.

Schildberg-Hörisch, H., Strassmair, C., 2012. An experimental test of the deterrence hypothesis. Journal of Law, Economics, and Organization 28(3), 447-459.

Stafford, M.C., Gray, L.N., Menke, B.A., Ward, D.A., 1986. Modeling the deterrent effects of punishment. Social Psychology Quarterly 49(4), 338-347.

Sutter, M., 2009. Deception through telling the truth?! Experimental evidence from individuals and teams. Economic Journal 119(534), 47-60.

Utikal, V., 2012. A fault confessed is half redressed - Confessions and punishment. Journal of Economic Behavior and Organization 81(1), 314-327.

Whitson, J.A., Wang, C.S., See, Y.H.M., Baker, W.E., Murnighan, J.K., 2015. How, when, and why recipients and observers reward good deeds and punish bad deeds. Organizational Behavior and Human Decision Processes 128, 84-95.

Xiao, E., 2013. Profit-seeking punishment corrupts norm obedience. Games and Economic Behavior 77(1), 321-344.

Xiao, E., Tan, F., 2014. Justification and legitimate punishment. Journal of Institutional and Theoretical Economics 170(1), 168-188.

APPENDIX A

	Sánchez-Pagés & Vorsatz (2007)	**Sánchez-Pagés & Vorsatz (2009)**	**Peeters et al. (2013)**
Game structure	2 options with misaligned payoffs	2 options with misaligned payoffs	2 options with misaligned payoffs
Ex ante transparency (conflicts of interest)	yes	yes	yes
Ex post transparency (conflicts of interest)	yes	yes	yes
Punishment severity	100% of the earnings	100% of the earnings	100% of the earnings
Punishment costs	100% of the earnings	100% of the earnings	100% of the earnings
Punishable actions	all	all	all
Number of rounds	50	50	60 + 40
Alternating roles	yes	yes	yes
Specific features		costly option to remain silent	institution choice in later stages
Findings regarding punishment	punishment does not reduce deception	punishment does not reduce deception	punishment does not reduce deception, except for self-selected punishment groups in the institution selection phase
	Church & Kuang (2009)	**Xiao (2013)**	**Kimbrough & Rubin (2013)**
Game structure	value assessment (coins)	2 options with misaligned payoffs	trust game with pre-play messages
Ex ante transparency (conflicts of interest)	no/yes	no	yes
Ex post transparency (conflicts of interest)	yes	enforcers: yes receivers: no	yes
Punishment severity	25 out of {100-140}	50% of the earnings	20% of initial endowment + outstanding amount
Punishment costs	10 out of {100-200}	no costs	20% of initial endowment if jury denies punishment
Punishable actions	dishonesty	all	dishonesty
Number of rounds	1	3 scenarios	10
Alternating roles	no	no	no

Specific features	outside option for receivers; punishment decision before final outcome is shown	third-party punishment;
		varying punishment profitability;
		varying receiver information
			jury system

Findings regarding punishment	punishment reduces deception only with ex ante disclosure of conflicts of interest	punishment reduces deception; but no effect when punishment is profitable or when receivers are not informed of punishments	punishment reduces deception from the beginning

Table A.1
Parameters of selected experimental studies examining punishment effects on deception.

APPENDIX B

Instructions for the experimental subjects (translated from Spanish)

Welcome to this experiment, we greatly appreciate your participation. From this moment on, please switch off your cell phone and do not talk or communicate in any way with the other participants. Read these instructions carefully and raise your hand if you have any questions during the session. One of the officials of the experiment will answer your questions individually.

Your decisions in this experiment will allow you to earn a certain amount of money that we will pay you in cash at the end of the session.

You will be a player in a two-player game. Your partner will be one of the participants in this session, randomly assigned by the computer. None of you will know the identity of the partner at any time. One of you will be assigned the role of "Player 1" and the other the role of "Player 2".

You will interact with your partner only once and this will take place through the computers. After this interaction, the experiment will end and you will be asked to fill in a short questionnaire.

Decision Making Player 1

During the experiment we will present three scenarios to Player 1, each one them contains three options. Each option consists of a payoff for Player 1 and a payoff for Player 2. This is the general structure of the options in each scenario that will be presented to Player 1:

Option A: Player 1 receives ... euros and Player 2 receives ... euros.

Option B: Player 1 receives ... euros and Player 2 receives ... euros.

Option C: Player 1 receives ... euros and Player 2 receives ... euros.

We will present to Player 1 the payoffs for both players of each option and in each scenario (the order of the options is at random). By contrast, Player 2 will not get this information. Player 1's task is to choose one of the following three messages that will be sent to Player 2 afterwards:

Message 1: Option A will earn you more money than the other two options.

Message 2: Option B will earn you more money than the other two options.

Message 3: Option C will earn you more money than the other two options.

Remember that there are three scenarios. That means, Player 1 has to decide, in each scenario, which message she wants to be sent to Player 2.

After Player 1 has chosen a message for each scenario, the computer will randomly select one of the scenarios. This scenario will then be implemented and the specific message that Player 1 chose for this scenario will be sent to Player 2. From this moment on, it depends on the decision of Player 2 which of the three corresponding options will be implemented and, according to this, which amount of money both players will earn.

Decision Making Player 2

Player 2 knows about the three options in the selected scenario but she knows nothing about the earnings associated with each option. The only information that Player 2 receives is the message that Player 1 chose for the implemented scenario.

After receiving player 1's message, Player 2 takes her decision, which is either to "accept" or "reject" the message. To "accept" the message means that Player 2 accepts the information of the message and that the option mentioned in the message determines the earnings of the two players. On the contrary, to "reject" the message means that Player 2 does not want the option mentioned in the message but another option to determine the earnings of both players. Therefore, if Player 2 accepts the message, the option in the message will be implemented and determines the payoffs of the players. In the case that Player 2 rejects the message, one of the remaining options of the selected scenario will be randomly implemented by the computer in order to determine the earnings of both players.

Earnings

Before your earnings are shown on the screen, you will answer some short questions. After that, Player 1 will receive information about the acceptance or rejection of her message, the implemented option corresponding to the scenario that was selected by the computer and the earnings of both players.

[***Treatment T100****:*

Player 2 will receive information about her own earnings corresponding to the implemented option. Furthermore, Player 2 will receive information about all potential payoffs for both players of each options in the scenario that has been implemented.]

[***Treatment T50****:*

In principle, Player 2 will only receive information about her own payoff corresponding to the implemented option. Furthermore, a possibility exists that the computer decides to provide additional information to player 2 about all potential payoffs for both players of each option in the implemented scenario. The probability of this happening is 50%.]

[***Treatment P100****:*

Player 2 will receive information about her own earnings corresponding to the implemented option. Furthermore, Player 2 will receive information about all potential payoffs for both players of each options in the scenario that has been implemented.

If player 1 sent a false message that was accepted, Player 2 will then have the possibility to punish Player 1 by reducing his profits to 2 euros.]

[***Treatment P50***:

In principle, Player 2 will only receive information about her own payoff corresponding to the implemented option. Furthermore, a possibility exists that the computer decides to provide additional information to player 2 about all potential payoffs for both players of each option in the implemented scenario. The probability of this happening is 50%.

If player 1 sent a false message that was accepted, Player 2 will then have the possibility to punish Player 1 by reducing his profits to 2 euros.]

The final earnings will be presented on the last screen.

After that we will pay you anonymously and in cash the amount that corresponds to your final earnings in the game.

Do you have any questions about these instructions? If so, please raise your hand. If you do not have any questions, remain silent until you get instructions from the experimenter.

APPENDIX C

Under standard economic theory, a sender would never send honest messages in the T-treatments since the expected utility of sending honest messages is lower than the one from sending deceptive messages for all $p > 0.33$:

$$EU_T(A) < EU_T(B)$$

$$\Leftrightarrow p\pi_i(A) + (1-p)\frac{\pi_i(B)+\pi_i(C)}{2} < p\pi_i(B) + (1-p)\frac{\pi_i(A)+\pi_i(C)}{2}$$

$$\Leftrightarrow p > \frac{\frac{\pi_i(B)+\pi_i(C)}{2} - \frac{\pi_i(A)+\pi_i(C)}{2}}{\frac{\pi_i(B)+\pi_i(C)}{2} - \frac{\pi_i(A)+\pi_i(C)}{2} + \pi_i(B) - \pi_i(A)} \Leftrightarrow p > \frac{1}{3}$$

The expected utility of sending honest messages is lower than the one from sending payoff-equalizing messages for all $p < 0.33$:

$$EU_T(A) < EU_T(C)$$

$$\Leftrightarrow p\pi_i(A) + (1-p)\frac{\pi_i(B)+\pi_i(C)}{2} < p\pi_i(C) + (1-p)\frac{\pi_i(A)+\pi_i(B)}{2}$$

$$\Leftrightarrow p < \frac{\frac{\pi_i(A)+\pi_i(B)}{2} - \frac{\pi_i(B)+\pi_i(C)}{2}}{\frac{\pi_i(A)+\pi_i(B)}{2} - \frac{\pi_i(B)+\pi_i(C)}{2} + \pi_i(A) - \pi_i(C)} \Leftrightarrow p < \frac{1}{3}$$

Accordingly, $EU_T(A) = EU_T(B) = EU_T(C)$ in case p = $\frac{1}{3}$.

Value ranges of p in which honesty is the dominant strategy exist in the punishment treatments with $\pi_i(A)$ = €5, $\pi_{1,2}(B)$ = €6, $\pi_3(B)$ = €15, $\pi_i(C)$ = €3 and a reduction to π_S = €2 in case of a sanction:

p-ranges	P100	P50
Scenarios 1 & 2	$\frac{1}{3+q} < p < \frac{1}{3-8q}$ in case $q < 0.375$	$\frac{1}{3+0.5q} < p < \frac{1}{3-4q}$ in case $q < 0.75$
	$\frac{1}{3+q} < p$ in case $q > 0.375$	$\frac{1}{3+0.5q} < p$ in case $q > 0.75$
Scenario 3	$\frac{1}{3+q} < p < \frac{1}{3-2.6q}$	$\frac{1}{3+0.5q} < p < \frac{1}{3-1.3q}$

Table C.1
Value ranges of p in which honesty is the dominant strategy in the punishment treatments applied to scenario-specific payoffs.

When we give up our simplicity assumption of zero costs of lying, we obtain the following p-ranges in which honesty is the dominant strategy, for instance, in scenario 1 and 2 in P100:

$$\frac{1-L_i}{3+q} < p < \frac{1+L_i}{3-8q} \quad \text{in case } q < 0.375; \qquad \frac{1-L_i}{3+q} < p \quad \text{in case } q > 0.375$$

Accordingly, we find the following p-ranges for scenario 1 and 2 in P50:

$$\frac{1-L_i}{3+0.5q} < p < \frac{1+L_i}{3-4q} \quad \text{in case } q < 0.75; \qquad \frac{1-L_i}{3+0.5q} < p \quad \text{in case } q > 0.75$$