Games and Economic Behavior 77 (2013) 321–344

Contents lists available at SciVerse ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Profit-seeking punishment corrupts norm obedience

Erte Xiao *

Department of Social and Decision Sciences, 208 Porter Hall, Carnegie Mellon University, Pittsburgh, PA 15213, USA

ARTICLE INFO

Article history:
Received 22 July 2010
Available online 30 October 2012

JEL classification:
C91
C72
D82
D03

Keywords:
Punishment
Norms
Corruption
Sender–receiver game
Experiment

ABSTRACT

Punishment typically involves depriving violators of resources they own such as money or labor. These resources can become revenue for authorities and thus motivate profit-seeking punishment. In this paper, we design a novel experiment to provide direct evidence on the role punishment plays in communicating norms. Importantly, this allows us to provide experimental evidence indicating that if people know that enforcers can benefit monetarily by punishing, they no longer view punishment as signaling a norm violation. The result is a substantial degradation of punishment's ability to influence behavior. Our findings draw attention to the detrimental effect of profit-seeking enforcement on the efficacy of punishment.

1. Introduction

Punishment plays an important role in maintaining social order in human society. One important feature of punishment is that it often deprives the punishee of some resources he/she owns (e.g., money or labor) and the enforcer often profits by capturing these resources. In principle, profitable punishment could motivate prosecutors to put more effort into catching offenders. Likewise, increasing the severity of punishment could reduce the frequency of violations. On the other hand, this opportunity for punishers to profit may lead them to choose to further their own revenue interests rather than pursuing the goals of norm conformity and promoting socially desirable behavior. In this paper, we draw attention to the norm expression function of punishment. We offer both theory and evidence from a laboratory experiment that punishment signals a norm violation. However, this norm communication function of punishment is greatly diminished if people know that enforcers can *potentially* earn benefit by punishing. The result is that punishment loses its ability to influence behavior.

Punishment promotes compliance not only by changing incentives but also by communicating social norms. Research in law and economics has argued that norm expression is an important function of punishment (Kahan, 1998; Cooter, 1998; Sunstein, 1996; Masclet et al., 2003; Tyran and Feld, 2006; Weibull and Villa, 2005; Funk, 2007; Galbiati and Vertova, 2008; Kube and Traxler, 2011; Galbiati et al., 2009). Punishment is used to inform violators and the public that the targeted behavior is not approved, and that it violates a social norm. By nature, the social norms that punishment enforces are often inconsistent with people's self-interest (Bicchieri, 2006); yet, norms can substantially affect people's decisions (Elster, 1989). It follows that the norm expression function of punishment can have a significant effect on behavior.

For example, to enforce the honor code, many organizations (e.g., West Point and Kellogg Graduate School of Management) announce to the community when an honor-code violation occurs and is punished. Xiao and Houser (2011) provide

* Fax: +1 412 268 6938.
E-mail address: exiao@andrew.cmu.edu.

0899-8256/$ – see front matter
http://dx.doi.org/10.1016/j.geb.2012.10.010

experimental evidence suggesting that publicly implemented punishment is more effective in promoting cooperation than privately implemented punishment. Their experiment design ensures that the difference cannot be attributed to shame or differences in information. Rather, their evidence suggests it is the ability *to observe* the punishment of free riders that promotes the effectiveness of punishment.

The effectiveness of the norm expressing function of punishment is particularly important in a society where good norms do not exist and bad ones do exist. In these situations, laws can help to reconstruct existing norms and change the social meaning of the targeted behavior (Sunstein, 1996). For example, to promote recycling norms, some cities, such as Philadelphia, have enacted laws that punish the households who fail to recycle. Enforcing such a law not only increases the cost of disobedience but also signals to the public that recycling is the appropriate behavior.

In light of the importance of punishment in expressing social norms, it is relevant to ask what factors may interfere with the communication function of punishment. In this paper we argue that if enforcers can profit from imposing punishment, a pervasive feature of corrupt societies, punishment may fail to communicate to the public that the punished action is not appropriate. To test our hypothesis, we design a lab experiment based on a sender–receiver game used to study cheating behavior (Gneezy, 2005). Thus, the norm violation we studied in this environment is the violation of the truth-telling norm. We revise the sender–receiver game so that it can test the norm expression function of punishment. Our game includes a third person who knows whether the sender has sent a true or false message to the receiver about which of the two options will give the receiver a higher payoff. The sender earns more (less) and the receiver earns less (more) if the sender sends a dishonest message and the receiver follows (does not follow) it.

The experiment consists of five treatments: two non-profitable punishment treatments (NPP and NPP′); two profitable punishment treatments (PP and PP′) and a no-punishment treatment (NP). In the non-profitable punishment treatments, the third person can decide whether to assign a payoff-cut to the sender after observing the message, and earns a fixed amount of money independent of his/her payoff-cut decision. NPP and NPP′ differ in that in NPP treatment the receiver sees the third person's punishment decision before deciding whether to follow the sender's message; in NPP′ treatment, the receiver is not aware of the third-party's punishment opportunity throughout the experiment and thus the punishment cannot signal anything to the receiver.

The two profitable punishment treatments, PP and PP′ are identical to the non-profitable punishment treatments, NPP and NPP′, respectively, except that in both PP and PP′ the third person's earnings increase upon assigning the payoff-cut to the sender. Finally, in the no-punishment treatment, the third person does not have the option of punishment but only observes the sender's message.

We find that in the two non-profitable punishment treatments, most third parties punish only false messages. In contrast, in the two profitable punishment treatments, third parties are significantly more likely to punish true messages. More importantly, compared with NPP′ treatment where receivers do not see the third-party's punishment decision, receivers in the NPP treatment are less likely to follow the message when they see the sender is punished. Senders are also less likely to lie in the NPP treatment than the NPP′ treatment. This difference provides direct evidence supporting the norm communication function of punishment.

Moreover, when the punishment is not visible to the receiver, the sender's deception rate does not differ significantly between the NPP′ treatment and the PP′ treatment. In contrast, when punishment is visible, receivers in the NPP treatment are much less likely to follow the message after observing senders being punished than those in the PP treatment, and senders in the PP treatment are also significantly more likely to lie. Indeed, the PP treatment maintains as high a level of lying as found in the NP treatment where punishment is not available. These results are strong evidence that, in our setting, diminished norm communication under profitable punishment causes punishment to be ineffective in deterring the violation of truth-telling norms.

2. Background

2.1. Expressive function of punishment

In past decades, theoretical and empirical analyses of punishment have focused largely on punishment's ability to increase the cost of undesired behavior. For example, the standard theory of optimal deterrence argues that when the probability of detecting norm violators is low, the severity of punishment should be high (Becker, 1968; Shavell, 2004).[1]

Recently, researchers have conducted experiments designed to explore the effect of the norm expression function of punishment. For instance, Tyran and Feld (2006) showed that in a public goods environment compliance improves greatly when punishment is imposed by group members rather than exogenously. The reason is that voting for punishment expresses support for a cooperation norm. Exogenously imposed punishment, on the other hand, lacks such a signaling function. In related work, Galbiati and Vertova (2008) reported data from a public goods experiment supporting the idea that punishment

[1] Experimental research has investigated people's willingness to incur costs to punish non-cooperators, showing that this peer punishment can be severe enough to enforce cooperation (Andreoni et al., 2003; Benabou and Tirole, 2003; Dickinson, 2001; Fehr and Gächter, 2000; Ostrom et al., 1992; Sefton et al., 2007; Yamagishi, 1986). On the other hand, recent research on incentives also reveals that, in some circumstances, incentives can have a detrimental effect on cooperation (Ariely et al., 2009; Falk and Kosfeld, 2006; Fehr and Falk, 2002; Fehr and Rockenbach, 2003; Fehr and List, 2004; Frey and Oberholzer-Gee, 1997; Fuster and Meier, 2010; Herrmann et al., 2008; Houser et al., 2008; Gneezy and Rustichini, 2000).

informs people of what they should or should not do, and that this established obligation has a significant effect on cooperation. The experiment showed that the expressive power of punishment can influence behavior independent of the incentive mechanism. Finally, Xiao and Houser (2011) found that when punishment is implemented publicly but anonymously, the norm becomes salient to both the recipient and the observers. This enhanced norm salience can promote cooperation.[2]

The above research indicates that punishment's ability to express disapproval directly affects its efficacy in enforcing norms. Thus, it is important to understand what factors might affect the expressive power of punishment. In this paper, we point out that when punishers can make profit by punishing, neither those who observe the punishment nor the punishment recipients view punishment as expressing social norms.

2.2. Sender–receiver game

Our experiment is based on a sender–receiver game. Gneezy (2005) was the first to design experiments using this game to study the nature of people's aversions to lying in a cheap talk environment. A sender and a receiver are paired. The receiver can choose one of two options which specify the payoffs of the receiver and the sender. The receiver must make a decision without knowing the payoffs associated with each option. The sender, however, knows this information. The sender sends a message to the receiver that purports to reveal the option that earns the receiver more money.

The payoffs of the two alternatives are designed so that the option that gives the receiver a higher payoff also leads to a lower payoff for the sender. Gneezy (2005) defined deception as the sender sending the false message to the receiver about which option will earn the receiver a higher payoff. Based on this definition, he found that senders are more likely to deceive receivers when the possible gains from deception are high and the associated losses for the receiver are low. Since then, a series of studies have used Gneezy's (2005) definition to investigate truth-telling (Hurkens and Kartik, 2009; Rode, 2010; Sánchez-Pagés and Vorsatz, 2007, 2009).

Nevertheless, as Sutter (2009) points out, this definition of deception focuses on behavior and thus fails to take account of the intention to deceive. Sutter broadens the notion of deception to include intentions, defining deception as any decision that the sender expects to lead the receiver to choose the low-payoff option. Using this alternative definition, Sutter's results track Gneezy (2005) closely with respect to the effect of costs and benefits on the frequency of deception.[3] In this paper, we use Sutter's definition to investigate how profitable punishments affect senders' intentions regarding deception.

The reason we choose this game is that the norm of truth-telling is clear in this setting. More importantly, a receiver's desire to follow the true message allows us to draw a clean inference regarding the receiver's perception of punishment from his/her behavior: whether to follow the sender's message if the sender is punished. As we discuss in more detail below, when we introduce third-party punishment in this game, norm-communicating punishment and norm-irrelevant punishment can lead to different predictions for the behavior of both senders and receivers. Consequently, our data provide direct evidence on how incentives for profit-seeking punishment affect norm communication.

3. Experiment design

Our experiment consists of five treatments: two non-profitable punishment treatments (NPP and NPP′), two profitable punishment treatment (PP and PP′), and one no-punishment treatment (NP).

3.1. Non-profitable punishment treatments (NPP and NPP′)

In these two treatments three subjects form a group. One person acts as a sender, one person acts as a receiver, and one person acts as an enforcer. The receiver chooses between two options: A and B. Each option specifies the payoffs to the sender and receiver if that option is chosen. Only the sender knows the payoffs to each option. The enforcer only knows whether the sent message is true.

Before the receiver chooses an option, the sender sends her one of the following messages:

Message A: Option A earns you more money than Option B.

Message B: Option B earns you more money than Option A.

The enforcer sees the sent message and decides whether to impose a payoff-cut to the sender. If a payoff-cut is imposed, the sender's payoff, which is decided later according to the option chosen by the receiver, is cut by 50%.

The only difference between the NPP and NPP′ treatments is the information provided to the receiver. In the NPP treatment, the receiver sees both the message and the enforcer's payoff-cut decision before choosing an option. In the NPP′ treatment, the receiver is not aware of the enforcer's punishment opportunity. The receiver is only informed about the presence of the enforcer and that the enforcer will see the sender's message. This is common knowledge: In particular, both the

[2] Researchers have also investigated other signaling functions of punishment. For example, it is argued that the principal's choice of incentives reveals his/her beliefs about the trustworthiness of the target group (Benabou and Tirole, 2003; Sliwka, 2007; Ellingsen and Johannesson, 2008; Van Der Weele, 2012).

[3] To avoid this potential confound regarding intentions to deceive, a recent paper by Erat and Gneezy (2012) develops a sender–receiver game with a richer message space that reduces the likelihood that senders will send a true message in order to deceive.

Table 1
The payoffs in each game.

Game	Option	Sender's payoff	Receiver's payoff
1	A	10	5
	B	0	6
2	A	4	4
	B	6	2
3	A	4	8
	B	8	4

sender and the enforcer are told that the receiver will not know about the enforcer's punishment decision or even the opportunity to punish (see instructions in Appendix A for details). This feature of the NPP′ treatment ensures that punishment cannot have signaling value for the receiver.

Thus, the difference in receiver behavior between NPP and NPP′ provides direct evidence supporting the norm expression function of punishment when punishment is not profitable. Moreover, the difference in sender behavior between these two treatments can also inform the extent to which the pure norm communication function of non-profitable punishment can discourage deception.

It is well-established that decisions can be influenced by relative payoff comparisons (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Dawes et al., 2007). We designed our experiment to minimize such confounds. In particular, enforcers are blind to each option's payoff. One of the four payoffs in Option A and Option B is randomly selected as the enforcer's payoff. Thus, the enforcer's payoff is independent of the payoff of either the sender or the receiver. Moreover, the enforcer does not know her randomly assigned payoff until the end of the experiment. That is, the enforcer does not know her payoff when making the payoff-cut decision in each round. Neither the sender nor the receiver knows the enforcer's payoff. Likewise, the enforcer is kept ignorant of the sender's and receiver's payoffs as well. All these conditions are common knowledge.

To allow within-subject analysis of the decisions of receivers and enforcers (e.g., how the enforcer imposes the payoff-cut when seeing truthful as opposed to false messages sent), we had subjects play three sender–receiver games with different payoff structures. Subjects were not informed of the outcome of each game until the end of the experiment. Thus, the design did not allow for learning.

In each game, each participant was randomly and anonymously paired with two other participants. The payoffs of the options in each game are listed in Table 1. The payoffs in each game imply different incentives to lie. For example, in Game 1, the sender can earn \$10 more, and the receiver earns only \$1 less, if the sender succeeds in deceiving the receiver so that the receiver chooses Option A. In contrast, in Game 2, successful deception earns the sender only \$2 more, and also leaves the receiver earning \$2 less. In previous studies of deception, subjects were found to be averse to lying so the frequency of lies increased along with the potential gains from deception (Gneezy, 2005; Sutter, 2009).[4] This suggests that even though the sender is better off lying in all three games, the sender is most likely to lie in Game 1 and least likely to lie in Game 2.

Since we define deception in terms of the intention to lie, it is necessary to obtain data on senders' beliefs regarding whether receivers follow their messages. For each game, we elicited senders' beliefs regarding: (1) which option receivers chose, and (2) whether the enforcers imposed the punishment. We did this after all three games concluded and before anyone was informed of the outcome of any game. (For each game the senders were reminded of the two options' details as well as their own choices.) Similarly, the enforcers were reminded which message was true and which message was sent by the sender in each game and their choices. Then they were asked to guess which option the receivers chose. In both the sender's and the enforcer's surveys, one question was randomly chosen and subjects earned an additional \$2 for giving the correct answer.

We also asked the receivers their beliefs regarding how the senders and the enforcers made their decisions. In order to keep the receiver ignorant of the complete nature of the options in each game, we were unable to reward receivers for answering the survey questions accurately.[5] All survey questions are in Appendix B.

3.2. *Profitable punishment treatments (PP and PP′)*

The only difference among PP and NPP, PP′ and NPP′ is that in both PP and PP′ treatments the enforcer's randomly assigned payoff is increased by 50% if and only if the enforcer imposes a 50% payoff-cut to the sender. In PP, both the sender and the receiver are aware of this. In PP′, only the sender is aware of this and the receiver, as in NPP′, is not aware of the punishment option. Again, the enforcer does not know her own payoff, or the payoffs of the sender or receiver when making payoff-cut decisions. This feature of the experiment minimizes the possibility that the enforcer's decision might be affected by concerns related to efficiency or inequality.

[4] See also Boles et al. (2000), Brandts and Charness (2003), Charness and Dufwenberg (2006), Crawford (2003), Crawford and Sobel (1982), Croson et al. (2003), Dreber and Johannesson (2008), Hurkens and Kartik (2009), Ellingsen and Johannesson (2004), Ellingsen et al. (2009), Lundquist et al. (2009), Mazar et al. (2008), Sánchez-Pagés and Vorsatz (2007).

[5] It appears lack of incentives led the receivers to answer survey questions inattentively. For example, some receivers said that enforcers never punish senders even when they experienced punishment in all the three games. In view of this, we did not draw conclusions from the receiver's survey answers.

Comparing behaviors between NPP and PP treatments, we can learn whether the profitable punishment can signal norm violations to the receiver as well as when punishment is not profitable. Unlike receiver behavior, which is affected only by the communication function of punishment, the sender's decisions are potentially influenced by two functions of punishment: (1) the communication function that influences receiver behavior and thus indirectly impacts the sender's payoff; and (2) the direct incentive effect of punishment on the sender's payoff. Thus, we can first compare the sender's behavior in PP′ and NPP′ treatments to learn how the profit opportunity of punishment may affect sender behavior due to the incentive. Then we can compare the difference between PP′ and NPP′ to that between PP and NPP treatments in order to infer to what extent the weakened signaling function of punishment diminishes the effectiveness of punishment in promoting honest behavior.

3.3. No-punishment treatment (NP)

The setup of the NP treatment is identical to the NPP treatment with the exception that the "enforcer" does not have the option to punish. The senders' or receivers' decisions might be affected by the knowledge that someone can observe the sender's decisions. In view of this, and to minimize all other possible treatment differences other than punishment opportunities, we keep this treatment the same as the previous one in that an enforcer remains involved in the game but he/she can only observe the sender's behavior. As in the other four treatments, the payoffs of the enforcers are also randomly assigned. All of this is common knowledge among the game's players. Further, the receivers in the NP treatment receive the same instructions as those in the NPP′ and PP′ treatments. This treatment enables inferences regarding how senders and receivers behave when there is no punishment available to communicate norms.

3.4. Procedures

The experiment was conducted in the Pittsburgh Experimental Economics Laboratory lab using z-tree (Fischbacher, 2007). Subjects were randomly and anonymously assigned a role and the role was fixed in all three rounds. One round was randomly chosen as the payoff round. Each subject was paid according to the outcome in that round. Subjects were paid privately. Each subject participated in exactly one treatment.

4. Theoretical analysis and hypothesis

In the sender–receiver game, the receiver receives only: (1) a message from the sender claiming which option will give the receiver a higher payoff; and (2) in NPP and PP treatments, the enforcer's punishment decision. The receiver does not have any information regarding the payoff table. In particular, the receiver does not know how the sender's payoff is decided in each option. Nor does the receiver receive any feedback after each round. Thus, it is reasonable to assume that the receiver will make her decisions based on her own subjective assessment of the decision environment. In view of this feature, we apply Subjective Equilibrium Analysis (Kalai and Lehrer, 1995) to derive the predictions in each treatment.[6] Our subjective equilibrium analysis of the NP treatment (a sender–receiver game without punishment) is similar to Rode (2010). The details of our theory analysis are in Appendix C.

We assume it is common knowledge that the enforcer's utility function consists of two parts: (1) the monetary payoff from the implementation of the duty; and (2) the disutility of how poorly she fulfills the duty:

$$U_e = \pi_e(d_e) - \sigma_e(d_e)$$

where $d_e = 1$ if the enforcer punishes the sender; $= 0$ if the enforcer does not punish the sender;

$$\sigma_e(d_e) = \begin{cases} \alpha, & \text{if } d_e \neq v_s \\ 0, & \text{if } d_e = v_s \end{cases}$$

where $\alpha > 0$ denotes the enforcer's dissatisfaction of her performance of her duty. $v_s = 1$ if the sender sends a false message; $= 0$ if the sender sends a true message. Thus, the enforcer incurs a disutility of α if she fails to punish the norm violator or punishes the norm follower. The enforcer decides whether to punish the sender according to:

$$\max_{d_e} U_e = \pi_e(d_e) - \sigma_e(d_e)$$

When punishment is non-profitable, $\pi_e(1) = \pi_e(0) = 0$. In this case, the optimal strategy for the enforcer is to punish the sender if and only if the sender sends a false message. We define norm-communicating punishment as punishment that credibly and perfectly signals that a norm violation occurred. Thus, under the preference structure posited above, punishment is norm communicating when it does not include a pecuniary benefit.

If punishment is profitable, then $\pi_e(1) > \pi_e(0) = 0$. In this case, if the sender sends a false message, the utility-maximizing strategy of the enforcer is to punish the sender. If the sender sends a true message, the optimal strategy for the

[6] For theoretical work on the sender–receiver game when the receiver is given information regarding the alignment of incentives, see Blume et al. (2001), Crawford (1998, 2003).

enforcer is to punish the sender if and only if $\pi_e(1) > \alpha$. If $\pi_e(1) = \alpha$, the enforcer is indifferent between punishing and not punishing the sender (see Appendix C for details). The intuition is that when punishment is profitable, the enforcer will punish a truthful sender if the pecuniary value of doing so is large enough to overcome the dissatisfaction of performing her duty poorly (i.e., punishing a norm follower). When a sender sends a false message, punishment increases the monetary payoff and also satisfies the enforcer's duty. Thus, the enforcer is always better off punishing the sender.

We define "norm-irrelevant" punishment as punishment that does not convey any information regarding whether a norm violation occurred. Thus, when punishment is profitable and subjects believe $\pi_e(1) > \alpha$, punishment is norm irrelevant.

As we mentioned earlier, one of the reasons we designed our experiment based on the sender–receiver game is that norm-communicating punishment and norm-irrelevant punishment can lead to different predictions regarding the behavior of both senders and receivers. Thus, by comparing the choices of senders and receivers among our treatments, we are able to draw inferences regarding people's perceptions of the norm-communicating function of punishment in the two punishment environments (see Appendix C for details).

Hypothesis 1 *(Enforcer's decisions).* Our first hypothesis is that there is more (less) norm-communicating (norm-irrelevant) punishment in NPP treatment than in the PP treatment:

$$freq(\textit{norm-communicating punishment}):\ PP < NPP$$
$$freq(\textit{norm-irrelevant punishment}):\ NPP < PP$$

Hypothesis 2 *(Receiver's decisions).* Our second hypothesis is that receivers interpret punishment as signaling norm violations when punishment is not profitable, but this is diminished if punishment is profitable for the enforcer. (i.e., a positive proportion of receivers believe $\pi_e(1) > \alpha$ in PP treatment). We test this hypothesis by comparing between treatments the frequency with which receivers follow senders' messages:

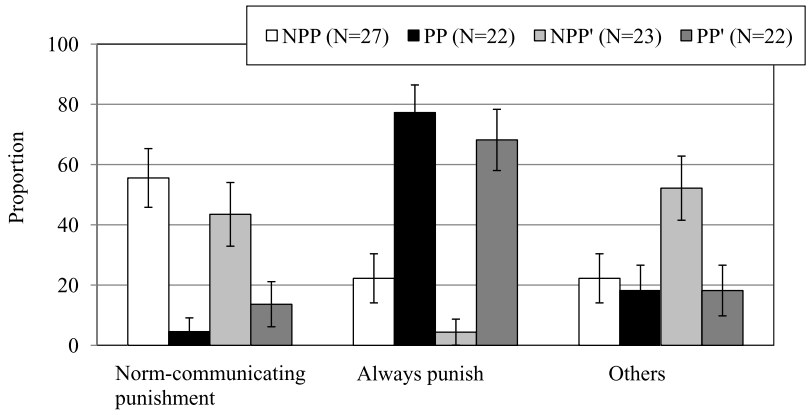
$$freq^{NPP}(\textit{follow message}|\textit{sender is punished}) < freq^{PP}(\textit{follow message}|\textit{sender is punished}) \leqslant freq^{NP}(\textit{follow message}) = freq^{PP'}(\textit{follow message}) = freq^{NPP'}(\textit{follow message})$$

Hypothesis 3 *(Sender's decisions).* Our third hypothesis is that senders are less likely to expect punishment to communicate norm violations when punishment is profitable than when punishment is not profitable (i.e., a positive proportion of senders believe $\pi_e(1) > \alpha$ when punishment is profitable). We test this hypothesis by comparing sender behavior across treatments. We denote $f_s \in \{0, 1\}$ as the sender's belief regarding whether the receiver will believe the message is true and will follow the message. Since a sender's decision depends on the value of f_s, we divide the sender's decisions as follows[7]: (1) T_F: send a true message and expect the receiver to follow ($f_s = 1$); (2) T_NF: send a true message and expect the receiver not to follow ($f_s = 0$); (3) F_F: send a false message and expect the receiver to follow ($f_s = 1$); and (4) F_NF: send a false message and expect the receiver not to follow ($f_s = 0$). As discussed above, lies include both T_NF and F_F.

In the case where receivers observe enforcers' punishment decisions (NPP and PP treatments), if punishment is norm communicating, then senders send true messages and expect receivers to follow the messages (T_F) (i.e., senders always behave honestly). If punishment is norm irrelevant (e.g., senders believe $\pi_e(1) > \alpha$ when punishment is profitable), then the sender's profit-maximizing strategy becomes the same as that in the NP treatment (i.e., always lie). This means that the sender sends a false message if she believes the receiver will follow the message (i.e., $f_s = 1$) (F_F); otherwise, she sends a true message (T_NF).

In the cases where receivers are not informed about the observer's punishment option (NPP′ and PP′ treatments), punishment cannot signal norm violations to the receivers. If punishment is not profitable, a sender expects to be punished if and only if she sends a false message. It is straightforward to see that in Game 1, the sender's profit-maximizing strategy is always to lie. In Game 2, the sender will lie if he believes the receiver will not follow his message ($f_s = 0$) and tell the truth if he believes the receiver will follow his message ($f_s = 1$). In Game 3, the sender will send a true message if she believes the receiver will not follow her (i.e., lie) and the sender is indifferent between sending a true or a false message if she believes the receiver will follow her message. If punishment is profitable, (1) when the sender expects to be punished regardless of the message (i.e., the sender believes $\pi_e(1) > \alpha$), the sender's profit-maximizing strategy becomes the same as that in the NP treatment (i.e., always lie); (2) when the sender believes that the enforcer will never punish sending a true message (i.e., the sender believes $\pi_e(1) \leqslant \alpha$), the sender's decision problem becomes the same as the case when punishment is not profitable (NPP′ treatment).

[7] We obtained data on senders' expectations regarding whether the enforcer would impose punishment and whether the receiver would follow the message. In principle, these data can inform the differences in senders' expectations regarding the norm communication functions of PP and NPP. While sample sizes are small, the evidence from the expectations data seems consistent with our findings drawn from the behavioral data. For example, among those who sent at least one false and one true message, compared with the NPP treatment, significantly fewer senders in the PP treatment expect enforcers to punish if and only if the message is false (2 out of 16 vs. 14 out of 21, Z-test, $p < 0.01$). Since we are interested in the effect of the norm communication function of punishment on behavior, the current paper focuses on the senders' behavioral data. Expectations data are available on request.



Note: The data include only those who experienced both false message and true message scenarios.
NPP: Non-profitable punishment and punishment is visible to receivers.
PP: Profitable punishment and punishment is visible to receivers.
NPP′: Non-profitable punishment and punishment is not visible to receivers.
PP′: Profitable punishment and punishment is not visible to receivers.

Fig. 1. Enforcer's behavior by treatment.

Taken together, our hypothesis that a positive proportion of senders believe $\pi_e(1) > \alpha$ when punishment is profitable has the following implications for treatment difference in the rates of deception and truth-telling (see Appendix C for details)[8]:

$$freq(deception)\text{: } NPP < NPP'; NPP < PP \leqslant NP; NPP' \leqslant PP' \leqslant NP$$

$$freq(send\ a\ true\ message\ and\ believe\ the\ message\ to\ be\ followed)\text{: } NPP > NPP'; NPP > PP \geqslant NP; NPP' \geqslant PP' \geqslant NP$$

Since the only difference between NPP and NPP′ treatments is that the receivers are not aware of the enforcer's punishment option, the increase of deception in NPP′ treatments as compared to NPP treatment can inform the extent to which the pure norm communication function of non-profitable punishment discourages deception in this setting.

Our predictions suggest that, in addition to the signaling effect of punishment, non-profitable punishment may decrease the sender's tendency to lie due to pure incentive effect (i.e., the deception rate in NPP′ treatment is lower than that in NP treatment). Such incentive effect can be weakened when the punishment is profitable, which is another possible explanation for the higher deception rate in the profitable punishment condition compared to the non-profitable punishment condition. Thus we first compare the deception rate in NPP′ to NP and PP′ to investigate the extent to which the profit opportunity of punishment can mitigate the direct incentive effect on the sender's behavior. We then compare the difference in sender behavior between NPP′ and PP′ with the difference between NPP and PP in order to discover to what extent the profit opportunity of punishment reduces the signaling effect of punishment, thus diminishing the effectiveness of punishment on enforcing truth-telling.

5. Results

We obtained observations on 432 subjects: 29 groups (of three) in NPP; 29 groups in NPP′; 30 groups in PP; 31 groups in PP′; and 25 groups in NP.

To avoid order effects we randomly switched the order of Game 2 and Game 3. In all sessions we ran Game 1 first. This enabled us to test whether our findings were robust to inexperienced subjects. We were able to draw similar conclusions from the data of Game 1.[9] Thus, here we report only our findings from the analysis using data from all three games. We first compare the enforcers' behavior between the two treatments in which they make decisions. Then we analyze the behavior of receivers and senders across all treatments.

5.1. Enforcer's behavior

To examine the punishment decisions of the enforcers, we calculate the proportion of enforcers who: (1) punish if and only if the sender sends a false message; (2) always punish regardless of the truthfulness of the message; and (3) exhibit punishment behavior that does not fall into either one of the aforementioned categories. Fig. 1 plots the distribution of

[8] The difference in sender behavior between PP and PP′ treatments, if any, can potentially inform to what extent the senders believe the profitable punishment can still communicate norm violations to the receivers. We compare sender behavior in PP and PP′ treatments in the results section.

[9] This of course does not include those findings reported below from within-subject analysis using all the data from the three games. The data are available on request.

Table 2
Receivers' decisions by treatment.

Treatment	Sender	Follow	Not follow
NPP (# of obs = 87)	Punished	7	42
	Not punished	36	2
PP (# of obs = 90)	Punished	41	40
	Not punished	7	2
NP (# of obs = 75)	–	51	24
NPP′ (# of obs = 87)	–	49	38
PP′ (# of obs = 93)	–	54	39

NPP: Non-profitable punishment and punishment is visible to receivers.
PP: Profitable punishment and punishment is visible to receivers.
NP: No punishment treatment.
NPP′: Non-profitable punishment and punishment is not visible to receivers.
PP′: Profitable punishment and punishment is not visible to receivers.

different types of enforcers in the two punishment treatments. The data includes only those enforcers who have seen both true messages and false messages. As shown in Fig. 1, while a majority (15 out of 27) of enforcers imposed norm-communicating punishment in the NPP treatment, only one out of 22 enforcers did so in the PP treatment. The difference is significant between the two treatments (Z-test, $p < 0.01$). On the other hand, most enforcers (17 out of 22) always punished the sender regardless of the message sent in the PP treatment, but only 6 out of 27 did so in the NPP treatment (Z-test, $p < 0.01$). This result supports Hypothesis 1. The enforcers display similar behavioral pattern in NPP′ and PP′. More enforcers in NPP′ punish if and only if the message is false than do in PP′ (10 out of 23 vs. 3 out of 22, Z-test, $p = 0.03$) and fewer enforcers always punish compared with PP′ (1 out of 23 vs. 15 out of 22; Z-test, $p < 0.01$).

We also obtained data on enforcer's beliefs regarding whether they think receivers will follow senders' messages when senders receive the payoff-cut. The data suggest treatment differences in beliefs. When receivers are aware of the enforcer's punishment option (NPP and PP), on average, enforcers believe the receiver will follow the sender's message about 12% of the time when a non-profitable punishment is imposed, about 26% of the time when a profitable punishment is imposed, and about 41% of the time when there is no punishment available (12% vs. 26%, Mann–Whitney test, $p = 0.01$; 26% vs. 41%, Mann–Whitney test, $p = 0.055$; 12% vs. 41%, Mann–Whitney test, $p < 0.01$). Our data suggest that enforcers recognize that while punishment can leave receivers less likely to follow senders' messages, the profit opportunity of punishment can weaken this effect.

5.2. Receivers' choices

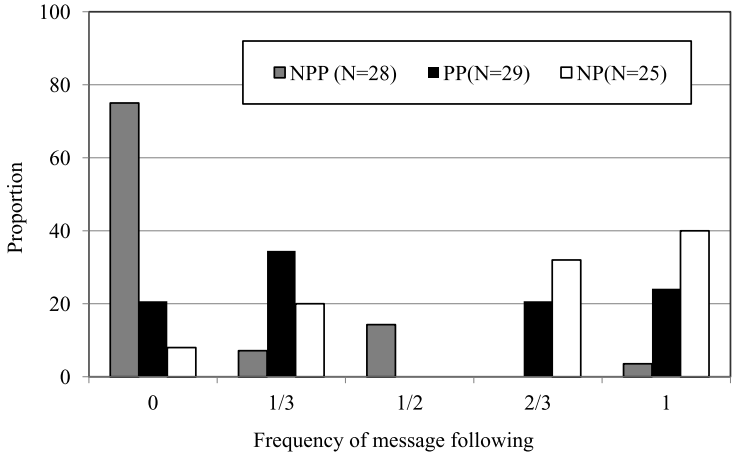
Table 2 reports how receivers, overall, follow senders' messages under different conditions in each treatment. As shown in Table 2, in the NPP treatment the enforcer did not punish the sender in 38 cases out of 87 observations. In these cases, receivers followed senders' messages 94.7% of the time (36 out of 38). In the PP treatment, we also found that receivers tend to follow senders' messages most of the time if the sender is not punished. Among the nine observations where punishment was not imposed, receivers followed seven times (77.8% of the messages).

To test Hypothesis 2, we next focus on treatment effects in cases where punishment is imposed. We compare receivers' tendencies to follow senders' messages among three cases: (1) when senders are punished in the NPP treatment; (2) when senders are punished in the PP treatment; and (3) when punishment is not available in the NP treatment or when punishment is invisible to the receiver (NPP′ or PP′).

The descriptive data in Table 2 shows that most receivers do not follow messages if the sender is punished in the NPP treatment (42 out of 49 decisions). Receivers in the PP treatment, however, seem to make either decision with equal probability (41 decisions follow the message, and 40 decisions do not). Consistent with previous findings (Gneezy, 2005; Sutter, 2009), we find that, in NP treatment receivers tend to follow senders' messages when the message is the only information they have (about 2/3 of decisions follow the sender's message). As expected, since the receivers receive the same instructions in the two treatments, the frequency of message following is about the same (about 56% of decisions follow the sender's message).

Individual level analyses. Since receivers get the same instructions in NPP′, PP′ and NP treatments, and their frequency of message following in NPP′ and PP′ is not significantly different from that in NP treatment (Mann–Whitney tests, two tail $p > 0.10$), we focus on the comparison among PP, NPP and NP in order to draw inferences about the norm communication function of punishment. Supporting Hypothesis 2, we find that the frequency of receivers following senders' messages is highest in the NP treatment and lowest in the NPP treatment (68% vs. 49% vs. 13%, Jonckheere test, $p < 0.01$; 49% vs. 13%, Mann–Whitney test, one-tail $p < 0.01$).[10] We classify receivers according to the fraction of times they follow the sender's decisions, and plot the subsequent distribution in Fig. 2. As shown, most receivers (75%) never follow the sender's message

[10] To test Hypothesis 2 and 3, we first run Jonckheere tests to establish the hypothesized trend and then use Mann–Whitney tests or Z-tests to provide evidence on specific differences between the NPP and PP treatments.



Note: The data excludes one receiver in each of the punishment treatment who did not see any punished message.
NPP: Non-profitable punishment and punishment is visible to receivers.
PP: Profitable punishment and punishment is visible to receivers.
NP: No-punishment treatment.

Fig. 2. Distributions of receivers' frequencies of following the message when the sender is punished in the two punishment treatments (NPP and PP), and in the no-punishment treatment (NP).

Table 3
Random effect probit regression analysis of receiver's decisions.

	Follow the message = 1, if yes; = 0, if no.	
	Coef.	Std. Err.
PP	0.01	0.18
NPP	−1.21	0.28
NP	0.54	0.20

Note: In the two punishment treatments, we only consider the cases where punishment is imposed.

when the sender is punished in the NPP treatment. In contrast, this percentage is only 21% in the PP treatment and 8% in NP treatment (Jonckheere test, $p < 0.01$; 75% vs. 21%, Z-test, one-tail $p < 0.01$). On the other hand, while only one out of 28 (4%) receivers in the NPP treatment always follows the sender's message when the sender is punished, about 24% of receivers do so in the PP treatment and 40% of receivers follow the sender's message in the NP treatment (Jonckheere test, $p < 0.01$; 4% vs. 24%, Z-test, one-tail $p = 0.01$).

We also ran a random individual effect probit regression analysis of receivers' decisions, using three treatment variables as explanatory variables (again, in the two punishment treatments, we only consider the cases where punishment is imposed). The regression result is reported in Table 3. As shown, and consistent with our findings above, the coefficient of NP is positive and highest while that of NPP is negative and lowest. All pairwise comparisons are significant (t-test, $p \leqslant 0.05$). These results support Hypothesis 2.

5.3. Sender's decisions

We report the descriptive data of senders' decisions and beliefs in Table 4. First, as shown in Table 4, the fraction of lies is higher in NPP′ than in NPP. To provide statistical evidence on this result, we calculate the percentage of lies (T_NF or F_F) and T_F for each sender. The deception rate in NPP′ is significantly higher than in NPP (67% vs. 34%, Mann–Whitney test, one tail $p < 0.01$) and the frequency of T_F in NPP′ is significantly lower than that in NPP (18% vs. 26%, Mann–Whitney test, one-tail $p = 0.01$). This result provides direct evidence supporting the norm communicating function of non-profitable punishment in reducing lying behavior.

Next, we compare senders' behavior between NPP′ and NP to investigate the extent to which the pure incentive effect of punishment decreases the deception rate in our experiment. We find that the frequency of lies in NPP′ treatment is not significantly different from that in NP treatment (67% vs. 60%, Mann–Whitney test, two tail $p > 0.40$) and neither of them is significantly different from that in PP′ treatment (53%) (Mann–Whitney tests, two tail $p > 0.10$). We find similar results for the rate of T_F. The frequency of T_F in NPP′ treatment (18%) is not significantly different from that in NP treatment (16%) (Mann–Whitney tests, two tail $p > 0.70$) and neither of them is significantly different from that in PP′ treatment

Table 4
Senders' decisions and beliefs by treatment.

Treatment (# of obs.)	Lie		Sender's decision_ belief		
	Freq.	%		Freq.	%
NPP (87)	30	34.48	T_NF	10	11.49
			F_F	20	22.99
			T_F	31	35.63
			F_NF	26	29.89
PP (90)	50	55.56	T_NF	36	40.00
			F_F	14	15.56
			T_F	18	20.00
			F_NF	22	24.44
NP (75)	45	60	T_NF	35	46.67
			F_F	10	13.33
			T_F	12	16.00
			F_NF	18	24.00
NPP′ (87)	58	66.67	T_NF	28	32.18
			F_F	30	34.48
			T_F	16	18.39
			F_NF	13	14.94
PP′ (93)	50	52.69	T_NF	26	27.96
			F_F	23	24.73
			T_F	19	20.43
			F_NF	25	26.88

T_NF: a sender sends a true message and expects the receiver not to follow.
F_F: a sender sends a false message and expects the receiver to follow.
T_F: a sender sends a true message and expects the receiver to follow.
F_NF: a sender sends a false message and expects the receiver not to follow.

(20%) (Mann–Whitney test, two tail $p > 0.70$). This suggests that, in our setting where punishment has no signaling value (not visible to the receiver), the incentive effect of non-profitable punishment does not affect lying decisions.[11]

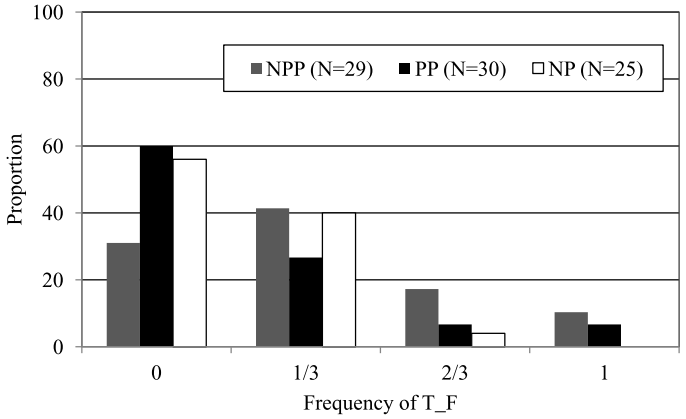
In contrast, Table 4 shows that the fraction of lies is higher in the NP and PP treatments than in the NPP treatment. The fraction of T_F is higher in NPP treatment than in the NP and PP treatments. Again, to provide statistical evidence, we calculate the percentage of lies (T_NF or F_F) and T_F for each sender. Supporting Hypothesis 3, we find that in the NPP treatment, on average, senders lie less frequently (34%) compared to those in NP treatment (60%) and PP treatment (56%) (Jonckheere test, $p < 0.01$; 34% vs. 56%, Mann–Whitney test, one-tail $p = 0.02$). On the other hand, the average percentage of T_F is higher in NPP treatment (36%) than that in NP treatment (16%) and PP treatment (20%) (Jonckheere test, $p < 0.01$; 36% vs. 20%, Mann–Whitney test, one-tail $p = 0.02$). We also found that the rate of lying in PP is not significantly different from that in PP′ (56% vs. 53%, Mann–Whitney test, two tail $p = 0.77$). This suggests that when punishment is profitable it is not able to perform the signaling function and, consequently, its ability to influence sender behavior is substantially diminished.

Fig. 3 plots the distribution of sender types according to the frequencies of T_F and lies. First, note from Fig. 3(A) that in the NPP treatment, about 69% of senders "send a true message and believe the receiver will follow the message" at least once. This percentage is only 40% in the PP treatment and 44% in the NP treatment (Jonckheere test, $p = 0.03$; 69% vs. 40%, Z-test, one-tail $p = 0.02$). Fig. 3(B) shows that about 77% of senders in the PP treatment and 88% of senders in the NP treatment lie at least once and only 55% do so in the NPP treatment (Jonckheere test, $p < 0.01$; 77% vs. 55%, Z-test, one-tail $p = 0.04$).

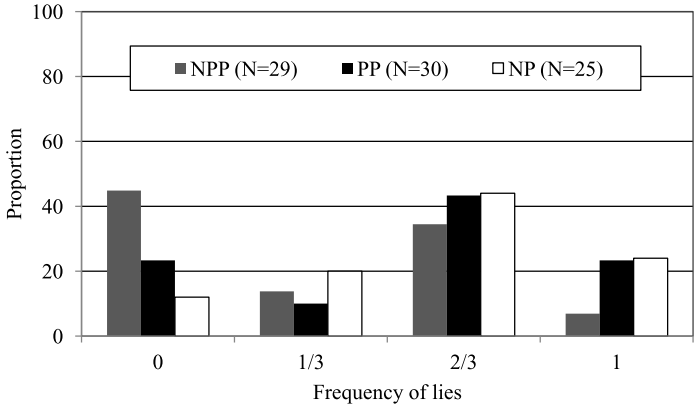
We also ran two random individual effect probit regression analyses of senders' decisions (see Table 5). In Regression (A), the dependent variable is whether the sender sent a true message and expected the receiver to follow (T_F). In Regression (B), the dependent variable is whether the sender lied (T_NF or F_F). Previous research suggests that the more one can profit from deception, the more one is likely to lie (Gneezy, 2005). In our experiment, the three games provide different

[11] It is interesting that in our experiment the incentive effect of non-profitable punishment does not decrease the deception rate compared with the no-punishment treatment. The insignificant incentive effect of punishment may be connected to the payoff structure of the games. To see this, we start with the cases when the incentive effect of non-profitable punishment is perfectly norm-enforcing (i.e., the enforcer punishes the sender if and only if she sends a false message). In NPP′ treatment, in Game 1, a sender will choose to lie as she expects to receive either \$5 or \$10 by lying which is better than \$0 by telling the truth. In Game 3, lying is no worse than being honest as a sender expects to receive \$8 or \$4 by lying and to receive \$4 or \$2 by not lying. In Game 2, if a sender believes the receiver will not follow her message, she prefers to lie (\$6 if she lies vs. \$3 if she tells the truth). On the other hand, if a sender believes the receiver will follow her message, she prefers to be honest (\$2 if she lies vs. \$4 if she tells the truth). Thus, the only case where we may see more truth-telling in NPP′ treatment is in Game 2 when a sender believes the receiver will follow her message. However the incentive to tell the truth in this case becomes small if the sender thinks that the enforcer may not punish a false message. In particular, it is easy to calculate that in Game 2 when the sender expects the receiver to follow her message, she will lie if she thinks the probability of being punished when she sends a false message is no more than 2/3.

(A). Distribution of the frequency of T_F by treatment



(B). Distribution of the frequency of lies (T_NF or F_F) by treatment



T_F: a sender sends a true message and expects the receiver to follow.
T_NF: a sender sends a true message and expects the receiver not to follow.
F_F: a sender sends a false message and expects the receiver to follow.

Fig. 3. Senders' behavior in NPP, PP and NP treatments.

Table 5
Random individual effects probit regression analysis of senders' behavior.

	(A) T_F = 1, if yes; = 0, if no.		(B) Lie = 1, if yes; = 0, if no.	
	Coef.	Std. Err.	Coef.	Std. Err.
PP	−0.44	0.31	−0.18	0.25
NPP	0.38	0.30	−0.90	0.27
NP	−0.64	0.33	−0.03	0.26
Game 1	−2.02	0.40	0.44	0.22
Game 3	−0.97	0.27	0.61	0.23

incentives for the senders to deceive. We therefore include dummies for each game in the regression.[12] The other three variables included are the dummies for the treatments. We use the same independent variables in the two regressions, the results of which are reported in Table 5.

Consistent with previous findings, in Regression (B), the coefficients of Game 1 and Game 3, although not significantly different, are significantly positive. This indicates that a higher potential profit from deception makes senders more likely to try to deceive. In Regression (A), the coefficient of Game 1 is significantly lower than in Game 3 (chi-square test, $p < 0.01$) and both are significantly negative. This suggests T_F is less likely when deception is more profitable.

[12] Note that although we have three different games, only senders' decisions should be affected by the nature of the game. Neither the enforcer nor the receiver should be affected by the specific payoff structures of the games, since they do not have any information regarding the real payoff structures of the games when making decisions.

In Regression (A), the coefficient of NPP is significantly higher than that of PP and NP (chi-square test, $p = 0.04$ and 0.02, respectively). In Regression (B), the coefficient of NPP is significantly lower than that of NP and PP (chi-square test, $p = 0.01$ and 0.02, respectively). The coefficients of NP and PP in both regressions are not significantly different (chi-square test, $p = 0.64$ in Regression (A) and $p = 0.62$ in Regression (B)).

All these results together provide evidence that the profitable punishment is less effective because it fails to communicate norm violations, not because it diminishes the incentive to obey the norm.

6. Discussion

We conduct experiments based on sender–receiver games in which norm-communicating and norm-irrelevant punishment yield different predictions about the decisions of senders and receivers. The results support our hypothesis that punishment can effectively express norm violations to receivers when enforcers do not benefit from the penalty. Its expression function is significantly diminished, however, when the punishment becomes a source of revenue for enforcers.[13] Importantly, when punishment is invisible to receivers, whether or not it is profitable does not affect senders' decisions. In contrast, when punishment is visible to the receivers punishment can discourage deception when it is non-profitable, but this effect is insignificant when enforcers can profit by punishing. In fact, deception occurs just as frequently in the case of profitable punishment as it does when punishment is not possible.

A substantial literature reveals that people care about equality and fairness and often share with others; even unaffected bystanders are willing to incur a cost to punish unfair behavior (Fehr and Fischbacher, 2004). We nevertheless find that when third-party enforcers can profit by punishing, a high percentage of them (about 80%) take advantage of this option and abuse their authority. Doing so benefits the enforcers but is costly both to senders and receivers. Given that enforcers often benefit from their duty to maintain social order, our paper calls attention to the role of third-party punishment in promoting cooperation when enforcement is profitable.

It is worthwhile to emphasize that in our experiment receivers do not know whether enforcers *are* "corrupt" because they do not receive any feedback. This suggests that the detrimental effect of profitable punishment can stem simply from the corrupting temptations embedded within enforcement institutions. In other words, an enforcement institution that enables corrupt behavior can have a detrimental impact on punishment's effectiveness, regardless of enforcers' actual decisions.

Although we design the three-person sender–receiver game so that it provides clean evidence on the norm communication function of punishment, the experiment setting also reflects the naturally occurring environments where punishment records can potentially signal the recipient's type and therefore affect the chance of being trusted by others in the future. For example, background checks, such as driving or criminal record checks, are often part of the process used to determine whether a job candidate is qualified for a position. Our results suggest that a candidate's record may have less weight in hiring decisions if employers believe legal systems are corrupt and may profit from creating violations. Future research might examine cross-country correlation between the extent of corruption in legal systems and the impact of backgrounds on hiring decisions.

Our study suggests new questions regarding how an enforcer's payment mechanism influences the expressive function as well as the outcome of punishment. For example, when enforcers are rewarded, it is reasonable to assume that we can preserve punishment's expressive function only under the conditions that (i) enforcers punish violators and (ii) enforcers do not punish non-violators. On the other hand, it is worth pointing out that even if these two conditions are satisfied, the opportunity to profit from punishment may motivate changing the rules in ways expected to increase the frequency of violations. One example of this in naturally occurring environments is shortening the duration of a yellow traffic light. Especially when the "right" way to design a rule is unclear (e.g., the duration of yellow traffic signals), profitable punishment may still have detrimental effects on the expressive function of punishment – even if enforcers only punish violators. The reason is that people may be suspicious of the intent of the new rules.

This study provides a new perspective on the potential causal relationship between corrupt legal systems and pervasive norm violations. By definition, corrupt societies include persistent and pervasive norm violations. This may be one important reason that corruption is widely perceived to be a major impediment to economic development. To curb corruption, scholars have highlighted the importance of punishment (Abbink, 2006). Unfortunately, in corrupt societies, authorities who enforce the punishments are typically also perceived as highly corrupt (Hunt, 2006; Transparency Internal, 2007). Our findings suggest that, rather than corrupt societies simply reflecting a culture of norm disobedience, legal institutions that embed corrupting temptations (e.g., profitable punishment) may be causally connected to systemic patterns of norm violations in society.

This finding speaks to the importance of establishing institutions that signal intolerance for corruption in the legal system. Such institutions have proven effective in natural environments. For example, during the economic growth period in Hong Kong, the ICAC (Independent Commission Against Corruption, http://www.icac.org.hk/en/home/index.html) was established to clean up its endemic corruption in law enforcement and many other government departments. Meeting challenges

[13] Our findings regarding the effect of profitability on the effectiveness of punishment are also consistent with Kuang et al. (2007)'s investigation of the effectiveness of advice. That paper reports that advice from a party with a monetary stake in whether the advice is followed is less effective than the same advice given by a neutral independent party.

initially, the ICAC ultimately proved vital in transforming Hong Kong from a graft-ridden city into a metropolis known for its cleanliness and lawfulness.

Our findings help to explain why the existence of organizations such as ICAC is important: their presence enhances the effectiveness of the legal system by restoring the norm communication function of law enforcement. Indeed, such organizations may be a necessary first step towards mitigating pervasive norm violations in severely corrupt societies, and doing so might help set the stage for rapid and peaceful economic expansion.

Acknowledgments

I thank Klaus Abbink, Jonathan Baron, Andreas Blume, Brit Grosskopf, Uri Gneezy, Daniel Houser, George Lowenstein, Roberto Weber, and participants at seminars at Texas University A&M; 2009 International ESA; 2009 North-American ESA; 2010 Behavioral Decision Research in Management Conference; 2010 American Law & Economics Association conference; CBDR seminar at Carnegie Mellon University for valuable comments. I gratefully acknowledge the National Science Foundation (SES-0961341), National Basic Research Program of China (973 Program) (Serial No. 2012CB955802) and Berkman Fund for funding that supported this research.

Appendix A

1. Sender's instructions

Instructions (Person 1)

(All the five treatments)

General information

Thank you for coming! You've earned $5 for showing up on time, and the instructions explain how you can make decisions and earn more money. So please read these instructions carefully! There should be no talking at any time during this experiment. If you have a question, please raise your hand, and an experimenter will assist you.

Each participant is in the role of either Person 1, or Person 2, or Person 3. You are in the role of Person 1.

This session consists of **three** rounds. At the beginning of each round, the computer will randomly group one Person 1 with one Person 2 and one Person 3. Thus, your counterpart in each round will change randomly throughout the experiment. No one will ever be informed of the identity of the two counterparts.

Below are the decision tasks in each round

In each round, two possible monetary payments are available to Person 1 and Person 2 in the experiment.

The two payment options are:

Option A : $\$W$ to Person 1 and $\$X$ to Person 2;

Option B : $\$Y$ to Person 1 and $\$Z$ to Person 2.

The payoff structure of Option A and Option B will be different in each round. Only Person 1 will know the exact values of W, X, Y, and Z in each round. Neither Person 2 nor Person 3 will know those values in any round. The computer will randomly assign W, X, Y, or Z, *with equal chance*, as Person 3's payoff.

- **Person 2's decision:** In each round, **Person 2** will decide to choose either Option A or Option B and thus decide the payoffs of Person 1 and Person 2.
- **Person 1's decision:** In each round, **Person 1** needs to decide which one of the following messages to send to Person 2 before Person 2 decides which option to choose.

Message A: "Option A will earn you more money than Option B."

Message B: "Option B will earn you more money than Option A."

(Two profitable-punishment treatments: PP and PP′)

- **Person 3's decision: Person 3** will NOT know the exact values of W, X, Y, and Z **but** will know whether the message sent by the matched Person 1 in each of the three rounds is true. (**Note:** Since the payoff structure of Option A and Option B changes from one round to another, which message is true in each round will also change accordingly.) After Person 1 decides which message to send to Person 2, Person 3 decides whether to assign a payoff-cut to Person 1. If Person 3 assigns the payoff-cut, Person 1's payoff (decided by the option Person 2 chooses) is cut by 50% and Person 3's payoff is increased by 50%. If Person 3 does not assign the payoff-cut, Person 1's payoff is not reduced by any amount and Person 3's payoff is not increased by any amount. Person 2's payoff will not change no matter what decision Person 3 makes. There is no cost for Person 3 to assign the payoff-cut.

Person 3 will make his/her decision prior to knowing his/her randomly assigned payoff.

(PP treatment)

After Person 1's and Person 3's decisions, Person 2 sees the decisions of both Person 1 and Person 3, and decides whether to choose Option A or Option B.

(PP′ treatment)

After Person 1's and Person 3's decisions, Person 2 sees Person 1's decision, and decides whether to choose Option A or Option B.

Note: Person 2 knows that Person 3 will observe the message sent by Person 1. But Person 2 does not know that Person 3 is able to assign a payoff-cut. That is, Person 2 will never see Person 3's payoff-cut decision and moreover, Person 2 does not even know that Person 3 has this option.

(PP and PP′ treatments)

For example:

Suppose, in one period, after Person 1 sent a message and Person 3 decided to impose the payoff-cut. Person 2 decided to choose Option B. The random payoff assigned to Person 3 is $\$X$.

Person 1's payoff in that period $= Y - 0.5 * Y$

Person 2's payoff in that period $= Z$

Person 3's payoff in that period $= X + 0.5 * X$

Suppose, in one period, after Person 1 sent a message and Person 3 decided NOT to impose the payoff-cut. Person 2 decided to choose Option B. The random payoff assigned to Person 3 is $\$X$.

Person 1's payoff in that period $= Y$

Person 2's payoff in that period $= Z$

Person 3's payoff in that period $= X$

(Two non-profitable punishment treatment: NPP and NPP′)

• Person 3's decision: Person 3 will NOT know the exact values of W, X, Y, and Z but will know whether the message sent by the matched Person 1 in each of the three rounds is true. (Note: Since the payoff structure of Option A and Option B changes from one round to another, which message is true in each round will also change accordingly.) After Person 1 decides which message to send to Person 2, Person 3 decides whether to assign a payoff-cut to Person 1. If Person 3 assigns the payoff-cut, Person 1's payoff (decided by the option Person 2 chooses) is cut by 50%. The payoffs of Person 2 and Person 3 will not change no matter what decision Person 3 makes. There is no cost for Person 3 to assign the payoff-cut.

Person 3 will make his/her decision prior to knowing his/her randomly assigned payoff.

(NPP treatment)

After Person 1's and Person 3's decisions, Person 2 sees the decisions of both Person 1 and Person 3, and decides whether to choose Option A or Option B.

(NPP′ treatment)

After Person 1's and Person 3's decisions, Person 2 sees Person 1's decision, and decides whether to choose Option A or Option B.

Note: Person 2 knows that Person 3 will observe the message sent by Person 1. But Person 2 does not know that Person 3 is able to assign a payoff-cut. That is, Person 2 will never see Person 3's payoff-cut decision and moreover, Person 2 does not even know that Person 3 has this option.

(NPP and NPP′ treatments)

For example:

Suppose, in one period, after Person 1 sent a message and Person 3 decided to impose the payoff-cut. Person 2 decided to choose Option B. The random payoff assigned to Person 3 is $\$X$.

Person 1's payoff in that period $= Y - 0.5 * Y$

Person 2's payoff in that period $= Z$

Person 3's payoff in that period $= X$

Suppose, in one period, after Person 1 sent a message and Person 3 decided NOT to impose the payoff-cut. Person 2 decided to choose Option B. The random payoff assigned to Person 3 is $\$X$.

Person 1's payoff in that period $= Y$

Person 2's payoff in that period $= Z$

Person 3's payoff in that period $= X$

(No-punishment treatment)

• Person 3 will NOT know the exact values of W, X, Y, and Z but will know whether the message sent by the matched Person 1 in each of the three rounds is true. (Note: Since the payoff structure of Option A and Option B changes from one round to another, which message is true in each round will also change accordingly.) After Person 1 decides which message to send to Person 2, Person 3 will see the message sent by Person 1.

Then, Person 2 sees the decisions of Person 1 and decides whether to choose Option A or Option B.

For example:

Suppose, in one period, after Person 1 sent a message, Person 2 decided to choose Option B. The random payoff assigned to Person 3 is $\$X$.

Person 1's payoff in that period $= Y$
Person 2's payoff in that period $= Z$
Person 3's payoff in that period $= X$

(All the five treatments)

After Person 2's decision, a new round starts. Each participant will be randomly paired with another *two* participants. Each round will proceed in the same way. You will not know the result of each round during the experiment. At the end of the experiment, one round will be randomly chosen to be your payoff round. Every participant will be informed of the result of that round and will be paid accordingly.

(PP treatment)

To repeat, Person 1 and Person 2 will earn the amounts specified in the option chosen by Person 2. However, Person 2 will never know what amounts were actually offered in the option not chosen (that is, he or she will never know whether Person 1's message was true or not). Moreover, Person 2 will never know the amounts to be paid to Person 1 according to the different options. Also, Person 2 and Person 1 will never know what payoff Person 3 randomly received. Person 1's earnings will also be affected by Person 3's payoff-cut decision. Person 3's earning will be increased if he/she assigns the payoff-cut to Person 1.

(PP′ treatment)

To repeat, Person 1 and Person 2 will earn the amounts specified in the option chosen by Person 2. However, Person 2 will never know what amounts were actually offered in the option not chosen (that is, he or she will never know whether Person 1's message was true or not). Moreover, Person 2 will never know the amounts to be paid to Person 1 according to the different options. Also, Person 2 and Person 1 will never know what payoff Person 3 randomly received. Person 1's earnings will also be affected by Person 3's payoff-cut decision. Person 3's earning will be increased if he/she assigns the payoff-cut to Person 1. Person 2 does not know that Person 3 has the option to assign a payoff-cut to Person 1.

(NPP treatment)

To repeat, Person 1 and Person 2 will earn the amounts specified in the option chosen by Person 2. However, Person 2 will never know what amounts were actually offered in the option not chosen (that is, he or she will never know whether Person 1's message was true or not). Moreover, Person 2 will never know the amounts to be paid to Person 1 according to the different options. Also, Person 2 and Person 1 will never know what payoff Person 3 randomly received. Person 1's earnings will also be affected by Person 3's payoff-cut decision.

(NPP′ treatment)

To repeat, Person 1 and Person 2 will earn the amounts specified in the option chosen by Person 2. However, Person 2 will never know what amounts were actually offered in the option not chosen (that is, he or she will never know whether Person 1's message was true or not). Moreover, Person 2 will never know the amounts to be paid to Person 1 according to the different options. Also, Person 2 and Person 1 will never know what payoff Person 3 randomly received. Person 1's earnings will also be affected by Person 3's payoff-cut decision. Person 2 does not know that Person 3 has the option to assign a payoff-cut to Person 1.

(No-punishment treatment)

To repeat, Person 1 and Person 2 will earn the amounts specified in the option chosen by Person 2. However, Person 2 will never know what amounts were actually offered in the option not chosen (that is, he or she will never know whether Person 1's message was true or not). Moreover, Person 2 will never know the amounts to be paid to Person 1 according to the different options. Also, Person 2 and Person 1 will never know what payoff Person 3 randomly received.

2. Receiver's instruction in NPP′ and PP′ treatment

(Receivers in the other three treatments were all given the full information as the senders received except the receivers were told they were Person 2. Thus, here we only provide instructions for receivers in NPP′ and PP′ treatment who were not given the information about the enforcer's punishment decision.)

Instructions (Person 2)

General information

Thank you for coming! You've earned $5 for showing up on time, and the instructions explain how you can make decisions and earn more money. So please read these instructions carefully! There should be no talking at any time during this experiment. If you have a question, please raise your hand, and an experimenter will assist you.

Each participant is in the role of either Person 1, or Person 2, or Person 3. You are in the role of Person 2.

This session consists of **three** rounds. At the beginning of each round, the computer will randomly group one Person 1 with one Person 2 and one Person 3. Thus, your counterpart in each round will change randomly throughout the experiment. No one will ever be informed of the identity of the two counterparts.

Below are the decision tasks in each round

In each round, two possible monetary payments are available to Person 1 and Person 2 in the experiment.

The two payment options are:

Option A: $\$W$ to Person 1 and $\$X$ to Person 2.

Option B : $\$Y$ to Person 1 and $\$Z$ to Person 2.

The payoff structure of Option A and Option B will be different in each round. Only Person 1 will know the exact values of W, X, Y, and Z in each round. Neither Person 2 nor Person 3 will know those values in any round. The computer will randomly assign W, X, Y, or Z, *with equal chance*, as Person 3's payoff.

- **Person 2's decision:** In each round, **Person 2** will decide to choose either Option A or Option B and thus decide the payoffs of Person 1 and Person 2.
- **Person 1's decision:** In each round, **Person 1** needs to decide which one of the following messages to send to Person 2 before Person 2 decides which option to choose.

  Message A: "Option A will earn you more money than option B."

  Message B: "Option B will earn you more money than option A."

- **Person 3** will NOT know the exact values of W, X, Y, and Z **but** will know whether the message sent by the matched Person 1 in each of the three rounds is true. (**Note:** Since the payoff structure of Option A and Option B changes from one round to another, which message is true in each round will also change accordingly.) After Person 1 decides which message to send to Person 2, Person 3 will see the message sent by Person 1.

Then, Person 2 sees Person 1's decision and decides whether to choose Option A or Option B.

Person 1 and Person 2 will receive the amount specified by the option chosen by Person 2.

For example, suppose, in one period, after Person 1 sent a message, Person 2 decided to choose Option B. In this case, Person 2 receives $\$Z$ and Person 1 receives $\$Y$ as specified by Option B.

After Person 2's decision, a new round starts. Each participant will be randomly paired with another two participants. Each round will proceed in the same way. You will not know the result of each round during the experiment. At the end of the experiment, one round will be randomly chosen to be your payoff round. Every participant will be informed of the result of that round and will be paid accordingly.

To repeat, Person 1 and Person 2 will earn the amounts specified in the option chosen by Person 2. However, Person 2 will never know what amounts were actually offered in the option not chosen (that is, he or she will never know whether Person 1's message was true or not). Moreover, Person 2 will never know the amounts to be paid to Person 1 according to the different options. Also, Person 2 and Person 1 will never know what payoff Person 3 randomly received.

Your ID _____

Appendix B

1. Survey questions for the sender

This session is finished. Please answer the following questions. You will be asked to guess your counterparts' decisions in each round. One of your answers will be randomly selected. If that answer is correct, you will earn an extra $2.

In Round 1: Option A ...
Option B ....
You sent Message ... to Person 2.

What payoff-cut decision do you think Person 3 made? (*only for the four punishment treatments*)

- ○ Person 3 chose to assign the payoff-cut.
- ○ Person 3 chose not to assign the payoff-cut.

Which option do you think Person 2 chose?

- Option A
- Option B

Same questions were asked for all the other two games.

2. Survey questions for the receiver

This session is finished. Please answer the following question.

How do you think Person 3 decides whether to assign the payoff-cut to Person 1? (*only for PP and NPP treatments*)

- Person 3 chose to assign the payoff-cut only if the message sent by Person 1 is not true.
- Person 3 chose to assign the payoff-cut only if the message sent by Person 1 is true.
- Person 3 chose to assign the payoff-cut in both cases.
- Person 3 chose to assign the payoff-cut in Neither case.

How do you decide which option to choose if Person 3 assigned the payoff-cut? (*only for PP and NPP treatments*)

- I would choose to follow Person 1's message (i.e., choose the one earns me more money according to Person 1's message).
- I would choose not to follow Person 1's message (i.e., choose the one earns me less money according to Person 1's message).
- I would randomly choose one of the two options.

How do you decide which option to choose if Person 3 did not assign the payoff-cut? (*only for PP and NPP treatments*)

- I would choose to follow Person 1's message (i.e., choose the one earns me more money according to Person 1's message).
- I would choose not to follow Person 1's message (i.e., choose the one earns me less money according to Person 1's message).
- I would randomly choose one of the two options.

How do you think Person 1 decides which message to send to you? (*only for NP, NPP′ and PP′ treatments*)

- Person 1 always sent a true message to me.
- Person 1 always sent a false message to me.
- Person 1 randomly chose one message to send.

How do you decide which option to choose? (*only for NP, NPP′ and PP′ treatments*)

- I always choose to follow Person 1's message (i.e., choose the one earns me more money according to Person 1's message).
- I always choose not to follow Person 1's message (i.e., choose the one earns me less money according to Person 1's message).
- I randomly choose one of the two options.

3. Survey questions for the enforcer

This session is finished. Please answer the following questions. You will be asked to guess your counterparts' decisions in each round. One of your answers will be randomly selected. If that answer is correct, you will earn an extra $2.

In Round 1, Message A was true and Message B was NOT true. Person 1 sent Message .... You decided to .... (In the NP, NPP′ and PP′ treatments, only remind the Message information)

Which option do you think Person 2 chose?

- Option A
- Option B

Same questions were asked for all the other two games

Appendix C

In the sender–receiver game, the receiver receives only: (1) a message from the sender claiming which option will give the receiver a higher payoff; and (2) in the NPP and PP treatments, the enforcer's punishment decision. The receiver does

not have any information regarding the payoff table. In particular, the receiver does not know how the sender's payoff is decided in each option. Nor does the receiver receive any feedback after each round. Thus, it is reasonable to assume that the receiver will make his/her decisions based on his/her own subjective assessment of the decision environment. In view of this feature, we apply Subjective Equilibrium Analysis (Kalai and Lehrer, 1995) to derive the predictions in each treatment. Our subjective equilibrium analysis (Kalai and Lehrer, 1995) of the NP treatment (a sender–receiver game without punishment) is similar to that in Rode (2010).

Each player assesses his/her own environment response function. This function specifies a probability distribution over all outcomes of a particular action. A set of strategies is subjectively rational if each player's strategy is optimal given the environment response function. This analysis assumes neither that the subjective beliefs are correct, nor that they coincide with the beliefs of other players.

1. NP treatment

Each game includes one sender and one receiver. Let Option A be (H, l) and Option B be (L, h), where H and L are the sender's payoff ($H > L \geqslant 0$), and h and l are the receiver's payoff ($h > l \geqslant 0$). Thus, Option B gives the receiver the higher payoff. The sender chooses a message $m \in M = \{A, B\}$ to tell the receiver which option has a higher payoff for the receiver. In our case, message B is the true message. The receiver sees the message and chooses an option $o \in \mathcal{O} = \{\text{A}, \text{B}\}$. We denote the true option that gives the receiver a higher payoff as $\theta \in \Theta = \{\text{A}, \text{B}\}$. Only the sender knows which option gives the receiver a higher payoff (h). Similar to Rode (2010), in the NP treatment, we apply the principle of insufficient reasoning (Laplace, 1824) and assume "Indifference", i.e., the receiver assigns the same probability to all $\theta \in \Theta$: $q(\theta = \text{A}) = q(\theta = \text{B}) = \frac{1}{2}$. Thus, we assume that ex ante, the receiver does not prefer a particular option to the other, and that this is common knowledge.

We denote $v_s = 1$ if $m \neq \theta$ (i.e., sender violates the norm of truth-telling); $v_s = 0$ if $m = \theta$ (i.e., sender follows the norm of truth-telling). We also denote $f_r = 1$ if $o = m$ (i.e., receiver follows the sender's message); $f_r = 0$ if $o \neq m$ (i.e., receiver does not follow the sender's message).

Receiver

The receiver's decision problem is to decide whether to follow the sender's message. Let $r \in \{0, 1\}$ indicate whether the receiver believes the sender's message is true. Let $r_A^{NP} \in \{0, 1\}$ indicate whether the receiver believes the sender's message is true when $m = A$ (i.e., the sender sends message A); and $r_B^{NP} \in \{0, 1\}$ indicate whether the receiver believes the sender's message is true when $m = B$. According to our assumption of "Indifference", $r_A^{NP} = r_B^{NP} = r$. We can specify the receiver's environment response function based on his/her subjective belief r.

Thus, the profit-maximizing strategy for a subjectively rational receiver is:

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

Sender

The sender's decision problem is to decide which message to send, given his/her belief as to whether the receiver will follow his/her message. We denote this belief as $f_s \in \{0, 1\}$.

Based on this subjective belief, we specify the sender's environment response function. Sender's expected payoff:

$$E\pi_s(m = A) = Hl_s + L(1 - f_s)$$

$$E\pi_s(m = B) = Ll_s + H(1 - f_s)$$

Thus, the profit-maximizing strategy for a subjective rational sender is:

$$m^*(s) = \begin{cases} A, & \text{if } f_s = 1 \\ B, & \text{if } f_s = 0 \end{cases}$$

That is, the sender's profit-maximizing strategy is always to lie.

2. Punishment treatments

Each game includes one sender, one receiver, and one enforcer. The interaction between the sender and the receiver in the four punishment treatments is the same as in the NP treatment. We assume that in this environment, the enforcer has a duty to monitor a society in which everyone is selfish and wants to maximize his/her own profit. The enforcer's duty is to punish norm violators. In this case, the only norm to enforce is the norm of truth-telling. A poor performance of his/her duty generates a negative utility for the enforcer. All this is common knowledge. We start our analysis from the enforcer's decision.

Enforcer

The enforcer's utility function consists of two parts: (1) the monetary payoff from the implementation of the duty; and (2) the disutility of how poorly he/she fulfills the duty:

$$U_e = \pi_e(d_e) - \sigma_e(d_e)$$

where $d_e = 1$ if the enforcer punishes the sender; $= 0$ if the enforcer does not punish the sender.

$$\sigma_e(d_e) = \begin{cases} \alpha, & \text{if } d_e \neq v_s \\ 0, & \text{if } d_e = v_s \end{cases}$$

where $\alpha > 0$ denotes the enforcer's dissatisfaction in the performance of his/her duty. Recall that $v_s = 1$ if $m \neq \theta$ (i.e., sender violates the norm of truth-telling); and $v_s = 0$ if $m = \theta$ (i.e., sender follows the norm of truth-telling). Thus, the enforcer incurs a disutility of α if he/she fails to punish a norm violator or punishes a norm follower. The enforcer decides whether to punish the sender in order to:

$$\max_{d_e} U_e = \pi_e(d_e) - \sigma_e(d_e)$$

• **When the punishment is non-profitable (NPP and NPP′ treatments):** $\pi_e(1) = \pi_e(0) = 0$.
The optimization problem becomes:

$$\max_{d_e} U_e = -\sigma_e(d_e)$$

It is easy to see that the solution to this problem is:

$$d_e^*(v_s) = \begin{cases} 0, & \text{if } v_s = 0 \\ 1, & \text{if} v_s = 1 \end{cases} \quad \text{or} \quad d_e^*(m) = \begin{cases} 0, & \text{if } m = B \\ 1, & \text{if } m = A \end{cases}$$

• **When the punishment is profitable (PP and PP′ treatments):** $\pi_e(1) > \pi_e(0) = 0$.

Case 1: If $v_s = 0$, $U_e(d_e = 1) = \pi_e(1) - \alpha$ and $U_e(d_e = 0) = 0$.
Thus, the optimal strategy for the enforcer is:

$$d_e^* = \begin{cases} 0, & \text{if } \pi_e(1) \leqslant \alpha \\ 1, & \text{if } \pi_e(1) > \alpha \end{cases}$$

Case 2: If $v_s = 1$, $U_e(d_e = 1) = \pi_e(1)$ and $U_e(d_e = 0) = -\alpha$
Thus, the optimal strategy for the enforcer is always to punish ($d_e^* = 1$).
In sum, when punishment is profitable, the enforcer's optimal strategy is as follows:

$$d_e^*(v_s, \pi_e, \alpha) = \begin{cases} 0, & \text{if } \pi_e(1) \leqslant \alpha \text{ and } v_s = 0 \\ 1, & \text{if} \pi_e(1) > \alpha \text{ and } v_s = 0 \\ & \text{or } v_s = 1 \end{cases}$$

or

$$d_e^*(m,\ \pi_e, \alpha) = \begin{cases} 0, & \text{if } \pi_e(1) \leqslant \alpha \text{ and } m = B \\ 1, & \text{if } \pi_e(1) > \alpha \text{ and } m = B \\ & \text{or } m = A \end{cases}$$

We assume the enforcer's optimal strategy above is common knowledge. Let $\delta \in \{0, 1\}$ be the subjective belief of the sender or the receiver about whether $\pi_e(1) > \alpha$ holds; $\delta = 1$ if they believe $\pi_e(1) > \alpha$ hold (i.e., the enforcer will punish regardless of whether the message is true or false); $\delta = 0$ if they believe $\pi_e(1) > \alpha$ does not hold (i.e., the enforcer punishes if and only if the message is false). Our predictions regarding the sender's behavior in the PP and PP′ treatments and the receiver's behavior in the PP treatment are derived based on this belief.

Receiver

• **When punishment is non-profitable (NPP and NPP′ treatments)**

NPP′ treatment

The receiver is not given any information about the enforcer's punishment opportunity in the NPP′ treatment. Thus, the receiver's decision is the same as that in the NP treatment:

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

NPP treatment

We show that when punishment is non-profitable, the sender's decision determines the enforcer's behavior with certainty. Given the knowledge of the enforcer's optimal strategy, the receiver can form his/her belief regarding whether the message is true based on the enforcer's punishment decision. Punishment is norm communicating.
Thus:

(1) when the enforcer did not punish the sender, the receiver believes that the sender sent a true message: $r^{NPP} = 1$; and
(2) when the enforcer punished the sender, the receiver believes the sender did not send a true message: $r^{NPP} = 0$.

Thus, the receiver's optimal strategy is not to follow the message when the enforcer punished the sender and to follow the message when the enforcer did not punish the sender.

$$f_r^*(d_e) = \begin{cases} 0, & \text{if } d_e = 1 \\ 1, & \text{if } d_e = 0 \end{cases}$$

When the punishment is profitable (PP and PP′ treatments)

PP′ treatment

As the receiver is not given any information about the enforcer's punishment opportunity, his/her decision is again the same as that in the NP treatment:

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

PP treatment

Let $r^{PP} \in \{0, 1\}$ be the receiver's subjective belief that the sender sends a true message in the PP treatment. We derive the receiver's profit-maximizing response in two cases depending on whether the enforcer punished the sender or not:

Case 1: $d_e = 1$

(1) If $\delta = 1$, the receiver believes the sender will be punished regardless of which message he/she sent. Punishment is norm irrelevant. Thus, $r^{PP} = r^{NP} = r$. In this case, the receiver's profit-maximizing response becomes the same as that of the NP treatment.

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

(2) When $\delta = 0$, the receiver's strategy will be the same as in the NPP treatment, thus $f_r^*(r) = 0$

Case 2: $d_e = 0$

According to the enforcer's best response function, $d_e = 0$ happens only if the true message is sent: $v_s = 0$. Thus, $r^{PP} = 1$. It is straightforward to see that in this case, the receiver will always follow the message when the enforcer does not punish the sender.

Taking all these together, and assuming: (1) there is a positive proportion of receivers who believe senders will send a true message in the NP treatment; and (2) this prior belief is constant across all treatments, our predictions about the relationship of the receiver's behavior among the five treatments are as follows:

(1) If all receivers believe $\pi_e(1) > \alpha$ (i.e., $\delta = 1$), we predict:

$$\begin{aligned} freq^{NPP}(\textit{follow message}|\textit{sender is punished}) &< freq^{PP}(\textit{follow message}|\textit{sender is punished}) \\ &= freq^{PP'}(\textit{follow message}) = freq^{NPP'}(\textit{follow message}) \\ &= freq^{NP}(\textit{follow message}) \end{aligned}$$

(2) If all receivers believe $\pi_e(1) > \alpha$ does not hold (i.e., $\delta = 0$), we predict:

$$\begin{aligned} freq^{NPP}(\textit{follow message}|\textit{sender is punished}) &= freq^{PP}(\textit{follow message}|\textit{sender is punished}) \\ &< freq^{PP'}(\textit{follow message}) = freq^{NPP'}(\textit{follow message}) \\ &= freq^{NP}(\textit{follow message}) \end{aligned}$$

(3) If a positive proportion of receivers believe $\pi_e(1) > \alpha$, we predict:

$$\begin{aligned} freq^{NPP}(\textit{follow message}|\textit{sender is punished}) &< freq^{PP}(\textit{follow message}|\textit{sender is punished}) \\ &\leqslant freq^{PP'}(\textit{follow message}) = freq^{NPP'}(\textit{follow message}) \\ &= freq^{NP}(\textit{follow message}) \end{aligned}$$

Thus, we can learn the norm communication function of non-profitable (profitable) punishment by comparing the frequency of message following when the sender is punished in the NPP(PP) treatment versus the NP treatment.

Sender

The sender receives a 50% payoff-cut if she is punished. We denote the sender's belief about whether she will be punished if she sends a true message as: $d_s(v_s = 0) = d_s(0) \in \{0, 1\}$. Similarly, $d_s(v_s = 1) = d_s(1) \in \{0, 1\}$ is the sender's belief about whether she will be punished if she sends a false message.

• When the receiver is informed of punishment decision (PP and NPP treatments)

In these two treatments, the sender holds subjective beliefs: (1) $f_s(f_r = 1|d_e = 1) = f_s(1) \in \{0, 1\}$ that the receiver will follow the message when the receiver sees that the sender was punished; and (2) $f_s(f_r = 1|d_e = 0) = f_s(0) \in \{0, 1\}$ that the receiver will follow the message when the receiver sees that the sender was not punished. As we demonstrated above, when the sender is not punished, the receiver will always follow the message. Therefore, $f_s(0) = 1$.

Sender forms her belief regarding whether the receiver will follow the message as follows:

(1) when sender sends message B (i.e. $v_s = 0$)

$$f_s^B = f_s(f_r = 1|m = B) = f_s(f_r = 1|v_s = 0) = f_s(1) * d_s(0) + f_s(0) * \big(1 - d_s(0)\big) = \big(1 - \big(1 - f_s(1)\big)\big)d_s(0) \quad (1)$$

(2) when sender sends message A (i.e. $v_s = 1$)

$$f_s^A = f_s(f_r = 1|m = A) = f_s(f_r = 1|v_s = 1) = f_s(1) * d_s(1) + f_s(0) * (1 - d_s(1)) = \big(1 - \big(1 - f_s(1)\big)\big)d_s(1) \quad (2)$$

Based on these beliefs, we next specify the sender's environment response function in each punishment condition.

NPP treatment

Given the sender's knowledge of the enforcer's optimal strategy in this case, the sender expects to be punished if she sends message A, and not to be punished if she sends message B.

$$d_s(0) = 0 \quad \text{and} \quad d_s(1) = 1$$

Thus, the sender's expected payoff:

$$E\pi_s(m = A) = \big(H * f_s^A + L * \big(1 - f_s^A\big)\big)(1 - 0.5)$$
$$E\pi_s(m = B) = L * f_s^B + H * \big(1 - f_s^B\big)$$

Given that the sender has full information regarding the payoffs, we can assume the sender's subjective beliefs f_s is consistent with the receiver's response function. Thus, $f_s^A = 0$ and $f_s^B = 1$.

We can obtain the sender's profit-maximizing strategy $m^* = B$ if $L > 0$; A or B if $L = 0$. Thus, in Game 2 and Game 3 where $L > 0$, the sender will always send a true message B, and believe the receiver will follow the message, i.e., the sender will always be honest. In Game 1, the sender is indifferent between sending A and believing the message not to be followed, and sending B and believing the message to be followed. Again, the sender will always tell the truth.

PP treatment

Given the knowledge of the enforcer's optimal strategy in this case,

$$d_s(0) = \delta \quad \text{and} \quad d_s(1) = 1$$

Thus, the sender's expected payoff:

$$E\pi_s(m = A) = E\pi_s(v_s = 1) = \big(H * f_s^A + L * \big(1 - f_s^A\big)\big)(1 - 0.5)$$
$$E\pi_s(m = B) = E\pi_s(v_s = 0) = \big(L * f_s^B + H * \big(1 - f_s^B\big)\big)(1 - \delta * 0.5)$$

(1) $\delta = 1$: $d_s(0) = d_s(1) = 1$.
In this case, punishment is perceived as completely norm irrelevant. From (1) and (2), we get: $f_s^A = f_s^B = f_s(1) = f_s$. Again, since senders have complete information regarding the receiver's payoff structure, we assume the sender's subjective belief $f_s(1)$ is consistent with the receiver's behavior when $\delta = 1$ in the PP treatment. In this case, it is straightforward to see that sender's profit-maximizing strategy becomes the same as that in the NP treatment. That is, senders' profit-maximizing strategy is to lie.

(2) $\delta = 0$ (the enforcer will never punish the true message). The sender's decision problem becomes the same as that in the NPP treatment. That is, senders' profit-maximizing strategy is to tell the truth.

• When the receiver is not informed of the punishment decision (PP′ and NPP′ treatments)

NPP′ treatment

Given the sender's knowledge of the enforcer's optimal strategy in this case, the sender expects to be punished if she sends message A, and not to be punished if she sends message B.

$$d_s(0) = 0 \quad \text{and} \quad d_s(1) = 1$$

Thus, the sender's expected payoff:

$$E\pi_s(m = A) = 0.5\big[Hf_s + L(1 - f_s)\big]$$
$$E\pi_s(m = B) = Lf_s + H(1 - f_s)$$

Again, f_s is the sender's belief as to whether the receiver will follow her message. Thus, the profit-maximizing strategy for a subjective rational sender:

$$m^*(s) = \begin{cases} A, & \text{if } f_s > \frac{2H-L}{3H-3L} \\ B, & \text{if } f_s < \frac{2H-L}{3H-3L} \\ A \text{ or } B, & \text{if } f_s = \frac{2H-L}{3H-3L} \end{cases}$$

It is straightforward to calculate that in Game 1: $\frac{2H-L}{3H-3L} = \frac{2}{3}$; Game 2: $\frac{2H-L}{3H-3L} = \frac{4}{3}$; Game 3: $\frac{2H-L}{3H-3L} = 1$.

Thus, we can obtain the sender's profit-maximizing strategy in each of the three games:

$$\text{Game 1:} \quad m^*(s) = \begin{cases} A, & \text{if } f_s = 1 \\ B, & \text{if } f_s = 0 \end{cases}$$

$$\text{Game 2:} \quad m^*(s) = B$$

$$\text{Game 3:} \quad m^*(s) = \begin{cases} A/B, & \text{if } f_s = 1 \\ B, & \text{if } f_s = 0 \end{cases}$$

Based on our definition of deception, in Game 1, the sender will always lie. In Game 2, the sender will always choose the true message B. The sender's decision is defined as a deception when $f_s = 0$, and is defined as truth-telling when $f_s = 1$. In Game 3, the sender will send B if she believes the receiver will not follow her message (i.e. lie); if the sender believes the receiver will not follow her message, she is indifferent between the two messages.

PP′ treatment

Given the knowledge of the enforcer's optimal strategy in this case,

$$d_s(0) = \delta \quad \text{and} \quad d_s(1) = 1$$

Thus, the sender's expected payoff:

$$E\pi_s(m = A) = E\pi_s(v_s = 1) = 0.5\big[Hf_s + L(1 - f_s)\big]$$
$$E\pi_s(m = B) = E\pi_s(v_s = 0) = (1 - \delta * 0.5)\big[Hf_s + L(1 - f_s)\big]$$

(1) $\delta = 1$: $d_s(0) = d_s(1) = 1$.
In this case, it is straightforward to see that the sender's profit-maximizing strategy becomes the same as that in the NP treatment.
(2) $\delta = 0$ (the enforcer will never punish the true message). The sender's decision problem becomes the same as the case when punishment is non-profitable (NPP′ treatment).

Taking all these together, our predictions about the relationship of the receiver's behavior among the five treatments are as follows:

(1) If all senders believe $\pi_e(1) > \alpha$ (i.e., $\delta = 1$), we predict:
freq(*deception*):
Game 1: $NPP < NPP'$; $NPP < PP = NP$; $NPP' = PP' = NP$
Game 2/3: $NPP \leqslant NPP'$; $NPP < PP = NP$; $NPP' \leqslant PP' = NP$
freq(*send a true message and believe the message to be followed*):
Game 1: $NPP > NPP'$; $NPP > PP = NP$; $NPP' = PP' = NP$
Game 2/3: $NPP \geqslant NPP'$; $NPP > PP = NP$; $NPP' \geqslant PP' = NP$
(2) If all senders believe $\pi_e(1) > \alpha$ does not hold (i.e., $\delta = 0$), we predict:
freq(*deception*):
Game 1: $NPP < NPP'$; $NPP = PP < NP$; $NPP' = PP' = NP$
Game 2/3: $NPP \leqslant NPP'$; $NPP = PP < NP$; $NPP' = PP' \leqslant NP$
freq(*send a true message and believe the message to be followed*):
Game 1: $NPP > NPP'$; $NPP = PP > NP$; $NPP' = PP' = NP$
Game 2/3: $NPP \geqslant NPP'$; $NPP = PP > NP$; $NPP' = PP' \geqslant NP$
(3) If a positive proportion of senders believe $\pi_e(1) > \alpha$:
freq(*deception*):
Game 1: $NPP < NPP'$; $NPP < PP \leqslant NP$; $NPP' = PP' = NP$
Game 2/3: $NPP \leqslant NPP'$; $NPP < PP \leqslant NP$; $NPP' \leqslant PP' \leqslant NP$
freq(*send a true message and believe the message to be followed*):
Game 1: $NPP > NPP'$; $NPP > PP \geqslant NP$; $NPP' = PP' = NP$
Game 2/3: $NPP \geqslant NPP'$; $NPP > PP \geqslant NP$; $NPP' \geqslant PP' \geqslant NP$

As the punishment is not visible to the receivers in the NPP′ treatment, the differences in the senders' behavior between the NPP and NPP′ treatments provide information on the norm-signaling effect of non-profitable punishment on cooperation. The difference in senders' behavior between the NPP′ and NP treatments provides information on the incentive effect of non-profitable punishment.

Under the profitable punishment mechanism, when senders hold the belief $\delta = 1$, the profit opportunity of punishment can increase the deception rate not only when the receivers are aware of the punishment (NPP vs. PP), but also when the receivers are not aware of the punishment (NPP′ vs. PP′). The difference in senders' behavior between the NPP and PP treatment can be attributed to: (1) incentive effect, i.e., senders expect that their payoff will be deducted regardless of what message they send in the PP treatment, but that they will be punished only when they send a false message in the NPP treatment; and (2) signaling effect, i.e., senders expect the receivers to view the punishment as signaling norm violation in NPP, but not in the PP treatment. On the other hand, given that receivers are not aware of the enforcer's punishment option, any deduction in the deception rate in NPP′ compared with PP′ treatment can only be attributed to the diminished incentive effect due to the profitability of punishment in the PP′ treatment.

If the diminished signaling effect of profitable punishment leads to a higher deception rate, we should expect the difference in deception rate between the NPP and PP treatments to be higher than that between NPP′ and PP′ treatments.

References

Abbink, K., 2006. Laboratory experiments on corruption. In: Rose-Ackerman, S. (Ed.), International Handbook on the Economics of Corruption. Elgar, Cheltenham, pp. 418–437.

Andreoni, J., Harbaugh, W., Vesterlund, L., 2003. The carrot or stick: reward, punishment and cooperation. Amer. Econ. Rev. 93 (3), 893–902.

Ariely, D., Bracha, A., Meier, S., 2009. Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. Amer. Econ. Rev. 99 (1), 544–555.

Becker, G.S., 1968. Crime and punishment: an economic approach. J. Polit. Economy 1968 (76), 169.

Benabou, R., Tirole, J., 2003. Intrinsic and extrinsic motivation. Rev. Econ. Stud. 70, 489–520.

Bicchieri, C., 2006. The Grammar of Society: The Nature and Dynamics of Social Norms. Cambridge University Press.

Blume, A., DeJong, D.V., Kim, Y.G., Sprinkle, G.B., 2001. Evolution of communication with partial common interest. Games Econ. Behav. 37 (1), 79–120.

Boles, T.L., Croson, R.T.A., Murnighan, J.K., 2000. Deception and retribution in repeated ultimatum bargaining. Org. Behav. Hum. Decis. Process. 83, 235–259.

Bolton, G.E., Ockenfels, A., 2000. ERC: A theory of equity, reciprocity, and competition. Amer. Econ. Rev. 90, 166–193.

Brandts, J., Charness, G., 2003. Truth or consequence: an experiment. Manage. Sci., 116–130.

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

Cooter, R.D., 1998. Expressive law and economics. J. Legal Stud. 27 (2), 585–608.

Crawford, V.P., 1998. A survey of experiments on communication via cheap talk. J. Econ. Theory 78 (2), 286–298.

Crawford, V.P., 2003. Lying for strategic advantage: rational and boundedly rational misrepresentation of intentions. Amer. Econ. Rev. 93 (1), 133–149.

Crawford, V.P., Sobel, J., 1982. Strategic information transmission. Econometrica 50, 1431–1451.

Croson, R., Boles, T., Murnighan, J.K., 2003. Cheap talk in bargaining experiments: lying and threats in ultimatum games. J. Econ. Behav. Organ. 51, 143–159.

Dawes, C.T., Fowler, J.H., Johnson, T., McElreath, R., Smirnov, O., 2007. Egalitarian motives in humans. Nature 446 (12), 794–796.

Dickinson, D., 2001. The carrot vs. the stick in work team motivation. Exper. Econ. 4, 107–124.

Dreber, A., Johannesson, M., 2008. Gender differences in deception. Econ. Letters 99, 197–199.

Ellingsen, T., Johannesson, M., 2004. Promises, threats and fairness. J. Econ. 114, 397–420.

Ellingsen, T., Johannesson, M., 2008. Pride and prejudice: the human side of incentive theory. Amer. Econ. Rev. 98 (3), 990–1008.

Ellingsen, T., Johannesson, M., Lilja, J., Zetterqvist, H., 2009. Trust and truth. J. Econ. 119, 252–276.

Elster, J., 1989. Social norms and economic theory. J. Econ. Perspect. 3 (4), 99–117.

Erat, S., Gneezy, U., 2012. White lies. Manage. Sci. 58 (4), 723–733.

Falk, A., Kosfeld, M., 2006. The hidden cost of control. Amer. Econ. Rev. 96 (5), 1611–1630.

Fehr, E., Falk, A., 2002. Psychological foundation of incentives. Europ. Econ. Rev. 46, 687–724.

Fehr, E., Fischbacher, U., 2004. Third party punishment and social norms. Evol. Hum. Behav. 25, 63–87.

Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. Amer. Econ. Rev. 90 (4), 980–994.

Fehr, E., List, J., 2004. The hidden costs and rewards of incentives. J. Eur. Econ. Assoc. 2 (5), 743–771.

Fehr, E., Rockenbach, B., 2003. Detrimental effects of sanctions on human altruism. Nature 422, 137–140.

Fehr, E., Schmidt, K., 1999. A theory of fairness, competition, and cooperation. Quart. J. Econ. 114, 817–868.

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

Frey, B., Oberholzer-Gee, F., 1997. The cost of price incentives: an empirical analysis of motivation crowding-out. Amer. Econ. Rev. 87, 746–755.

Funk, P., 2007. Is there an expressive function of law? An empirical analysis of voting laws with symbolic fines. Amer. Law Econ. Rev. 9 (1), 135–159.

Fuster, A., Meier, S., 2010. Another hidden cost of incentives: the detrimental effect on norm enforcement. Manage. Sci. 56 (1), 57–70.

Galbiati, R., Vertova, P., 2008. Obligations and cooperative behaviour in public good games. Games Econ. Behav. 64, 146–170.

Galbiati, R., Schlag, K., van der Weele, J., 2009. Can sanctions induce pessimism? An experiment. Working paper 24/2009, University of Siena.

Gneezy, U., 2005. Deception: the role of consequences. Amer. Econ. Rev. 95, 384–394.

Gneezy, U., Rustichini, A., 2000. A fine is a price. J. Legal Stud. 29 (1), 1–17.

Herrmann, B.C., Thöni, C., Gächter, S., 2008. Antisocial punishment across societies. Science 319 (5868), 1362–1367.

Houser, D., Xiao, E., McCabe, K., Smith, V., 2008. When punishment fails: research on sanctions, intentions and non-cooperation. Games Econ. Behav. 62 (2), 509–532.

Hunt, J., 2006. Why are some public officials more corrupt than others? In: Rose-Ackerman, S. (Ed.), International Handbook on the Economics of Corruption. Edward Elgar, Northampton, MA.

Hurkens, S., Kartik, N., 2009. Would I lie to you? On social preferences and lying aversion. Exper. Econ. 12, 182–192.

Kahan, D.M., 1998. Social meaning and the economic analysis of crime. J. Legal Stud. 27 (2), 661–672.

Kalai, E., Lehrer, E., 1995. Subjective games and equilibria. Games Econ. Behav. 8 (1), 123–163.

Kuang, X.J., Weber, R.A., Dana, J., 2007. How effective is advice from interested parties? An experimental test using a pure coordination game. J. Econ. Behav. Organ. 62 (4), 591–604.

Kube, S., Traxler, C., 2011. The interactions of legal and social norms enforcement. J. Public Econ. Theory 13 (5), 639–660.

Laplace, P., 1824. Essai Philosophique sur les Probabilités. Dover, New York. English translation.

Lundquist, T., Ellingsen, T., Gribbe, E., Johannesson, M., 2009. The aversion to lying. J. Econ. Behav. Organ. 70, 81–92.
Masclet, D., Noussair, C., Tucker, S., Villeval, M., 2003. Monetary and non-monetary punishment in the voluntary contributions mechanism. Amer. Econ. Rev. 93, 366–380.
Mazar, N., Amir, O., Ariely, D., 2008. The dishonesty of honest people: a theory of self-concept maintenance. J. Marketing Res. 45, 633–644.
Rode, J., 2010. Truth and trust in communication. An experimental study of behaviour under asymmetric information. Games Econ. Behav. 68 (1), 325–338.
Ostrom, E., Walker, J., Gardner, R., 1992. Covenants with and without a sword: self-governance is possible. Amer. Polit. Sci. Rev. 86, 404–417.
Sánchez-Pagés, S., Vorsatz, M., 2007. An experimental study of truth-telling in a sender–receiver game. Games Econ. Behav. 61, 86–112.
Sánchez-Pagés, S., Vorsatz, M., 2009. Enjoy the silence: an experiment on truth-telling. Exper. Econ. 12, 220–241.
Sefton, M., Shupp, R., Walker, J., 2007. The effect of rewards and sanctions in provision of public goods. Econ. Inquiry 45, 671–690.
Shavell, S.M., 2004. Foundations of Economic Analysis of Law. Belknap Press of Harvard University Press.
Sliwka, D., 2007. Trust as a signal of a social norm and the hidden costs of incentive schemes. Amer. Econ. Rev. 97 (3), 999–1012.
Sunstein, C.R., 1996. On the expressive function of law. Univ. PA. Law Rev. 144, 2021–2031.
Sutter, M., 2009. Deception through telling the truth? Experimental evidence from individuals and teams. J. Econ. 119, 47–60.
Transparency Internal, 2007. Report on the transparency international global corruption barometer 2007. http://www.transparency.org/content/download/27256/410704/file/GCB_2007_report_en_02-12-2007.pdf.
Tyran, J., Feld, L., 2006. Achieving compliance when legal sanctions are non-deterrent. Scand. J. Econ. 108 (1), 135–156.
Van Der Weele, J., 2012. The signaling power of sanctions in social dilemmas. J. Law, Econ., Organ. 28 (1), 103–126.
Weibull, J., Villa, E., 2005. Crime, punishment and social norms. Working Paper Series in Economics and Finance, no. 610. Stockholm School of Economics.
Xiao, E., Houser, D., 2011. Punish in public. J. Public Econ. 95, 1006–1017.
Yamagishi, T., 1986. The provision of a sanctioning system as a public good. J. Pers. Soc. Psychol. 51 (1), 110–116.