

# The Importance of Emotions for the Effectiveness of Social Punishment\*

Astrid Hopfensitz

Ernesto Reuben

## Abstract

This paper experimentally explores how the enforcement of cooperative behaviour in a social dilemma is facilitated through institutional as well as emotional mechanisms. Recent studies emphasize the importance of anger and its role in motivating individuals to punish free riders. However, we find that anger also triggers retaliatory behaviour by the punished individuals. This makes the enforcement of a cooperative norm more costly. We show that in addition to anger, ‘social’ emotions like guilt need to be present for punishment to be an effective deterrent of uncooperative actions. They play a key role by subduing the desire of punished individuals to retaliate and by motivating them to behave more cooperatively in the future.

---

\*We would like to thank Sam Bowles, Dirk Engelmann, Nikos Nikiforakis, Arno Riedl, Arthur Schram, Martine Visser, and Frans van Winden for useful comments and suggestions. Financial support from CREED of the University of Amsterdam is gratefully acknowledged.

An important mechanism for the promotion of cooperation is the enforcement of social norms (Ostrom, 1998; Fehr and Gächter, 2000a; Boyd and Richerson, 2005). As shown by Fehr and Gächter (2000b), cooperative behaviour can persist when there is an opportunity to punish defectors. However, although punishment can have desirable consequences, it can also have negative effects. For example, punishment can crowd out positive reciprocity (Fehr and Rockenbach, 2003), lead to welfare losses (Egas and Riedl, 2005), and sometimes produce only a negligible increase in cooperation levels (Gächter and Herrmann, 2006; Nikiforakis and Normann, forthcoming). Studying the choices of individuals who punish as well as the reaction of those who are punished can help us predict in which situations punishment produces negative results. Considering the role of emotions seems to be necessary to understand this kind of behaviour (Loewenstein, 1996; Elster, 1999; Thaler, 2000).

The goal of this paper is to understand the type of motivations that must be present, among both the punishers and the punished, for punishment to be an effective institution for the promotion of cooperation. We concentrate on the role of social emotions, such as shame and guilt, as an essential component for the successful enforcement of cooperative norms. In particular, we are interested in their role as inhibitors of retaliatory behaviour by the punished individuals.

Although it has attracted little attention, antisocial behaviour such as retaliation or the punishment of cooperative individuals has been often observed in laboratory experiments, including, for example, public good games (Fehr and Gächter, 2000b; Nikiforakis, 2008), prisoner dilemma games (Falk

*et al.*, 2005), and moonlighting games (Abbink *et al.*, 2000). This type of behaviour is widespread and is observed in around one quarter of all subjects (e.g., Cinyabuguma *et al.*, 2006; Falk *et al.*, 2008).

We study, by means of an experiment, antisocial behaviour in a social dilemma game. We introduce a punishment institution where individuals who are punished always have the opportunity to retaliate. After all, if a punishment technology exists, it is likely that both the punisher and the punished have access to it. Our results show that many individuals do retaliate after being punished. In various cases, this escalates as individuals punish each other in turns. In order to observe the effect of retaliation on future behaviour, subjects played the game twice. We find that although retaliation considerably increases the cost of punishing selfish behaviour, it does not deter future cooperation or punishment. Hence, its effect seems to be restricted to welfare losses caused by the destruction of resources.

Recent research has revealed that emotions motivate individuals to punish opportunistic behaviour. In particular, anger has been shown to be of influence when subjects have to decide whether to punish or not. Unkind behaviour induces anger, and the angrier people are the more likely they are to incur costs in order to penalize such behaviour (Bosman and van Winden, 2002; Quervain *et al.*, 2004). We replicate these findings and extend this line of research by studying the emotional reaction of punished individuals.

In order to explain the behaviour of both punishers and punished, we measure their emotional response. Our results show that individuals who act in an unkindly manner do nevertheless feel angry when punished. Furthermore, we find that high intensities of anger are related to positive retaliation.

Consequently, anger alone induces multiple rounds of punishment and a significant destruction of resources. Therefore, anger cannot explain whether punishment will effectively promote prosocial behaviour. The effectiveness of punishment depends on the reaction of the individuals who are punished.

What is missing to make punishment effective is a ‘moral’ reaction of the punished. Namely, after receiving punishment the punished individual should act more cooperatively and abstain from retaliation. We show that the social emotions of shame and guilt motivate individuals to react in precisely this way. In other words, individuals do not retaliate when feelings of guilt restrain their anger-induced desire to fight back.

The paper is organized as follows. In Section 1 we describe the experiment’s design. Section 2 describes the subjects’ behaviour. In Section 3 we analyze the relationship between the emotions and the behaviour of the punishers and the punished. Section 4 discusses the results and concludes.

## 1 Experimental Design

Lately, punishment mechanisms have been mostly analyzed in the context of public good games (using the design of Fehr and Gächter, 2000b). However, in this study we use a simpler setting where the causes and effects of emotions can be easily observed and analyzed. To study the impact of social emotions, we used a two-person social dilemma game with and without punishment opportunities. Our game is similar to many of the social dilemma games in the literature, such as, the sequential prisoners’ dilemma, the trust game (Berg *et al.*, 1995), the gift exchange game (Akerlof, 1982; Fehr *et al.*, 1993)

and others.

## 1.1 The Game

We first describe the baseline game without punishment opportunities and then we explain how punishment is introduced. The game consists of two players, a ‘first mover’ and a ‘second mover’, taking part in a one-shot game (see Figure 1). At the start of the game, the first mover receives 150 points and the second mover 100 points. In the first stage, the first mover decides to either defect or cooperate. If the first mover defects, he keeps his 150 points, the second mover keeps her 100 points, and the game ends. If the first mover cooperates, 50 of his 150 points are multiplied by six and transferred to the second mover (i.e. the second mover receives 300 points). In the second stage, the second mover returns either 150 points (an equal split of the gains), 50 points (returning exactly the points lost by the first mover), or 0 points. After the decision of the second mover the game ends. Hence, if the first mover cooperates his payoff is  $\pi_1 = 100 + r$  and the payoff of the second mover is  $\pi_2 = 100 + 6 \times 50 - r$  where  $r$  is the amount returned by the second mover.

— Figure 1 around here —

In the game with punishment both players can assign punishment points. We denote  $p_{it}$  as the amount of points assigned by player  $i$  in punishment round  $t$ . After the second mover decides how much to return, the first round of punishment starts. First, the first mover has the opportunity to assign punishment points to the second mover ( $p_{11} \geq 0$ ). Every point assigned costs

1 point to the first mover and reduces the payoff of the second mover by 4 points. In order to avoid losses during the experiment, the first mover can assign punishment points only as long as the second mover has a positive number of points (i.e.  $\frac{1}{4}(100 + 6 \times 50 - r) \geq p_{11}$ ). If the first mover chooses  $p_{11} = 0$  the game ends. However, if the first mover chooses  $p_{11} > 0$  the second mover is given the opportunity to assign punishment points to the first mover ( $p_{21}$ ). In order to avoid confusion, we will refer to punishment by the second mover as *retaliation*. Punishment by first movers and retaliation by second movers has the same cost and does the same amount of harm. Thus for each retaliation point assigned, the first mover loses four points. Once more, the second mover can assign retaliation points only as long the first mover has a positive number of points (i.e.  $\frac{1}{4}(100 + r - p_{11}) \geq p_{21}$ ).<sup>1</sup> If  $p_{21} = 0$  the game ends, but if  $p_{21} > 0$  the game continues to a second round of punishment. That is, the first mover has the opportunity to assign *additional* punishment points to the second mover ( $p_{12}$ ). As before, if  $p_{12} = 0$  the game ends but if  $p_{12} > 0$ , the second mover has the opportunity to assign *additional* retaliation points ( $p_{22}$ ), and so on. The process repeats itself until either one of the players has zero points and cannot be punished further, or one of the players assigns zero punishment points. Therefore, if the first mover cooperates his payoff is  $\pi_1 = 100 + r - \sum_t p_{1t} - 4 \times \sum_t p_{2t}$  and that of the second mover is  $\pi_2 = 100 + 6 \times 50 - r - \sum_t p_{2t} - 4 \times \sum_t p_{1t}$ .

If we use the standard assumption of rational individuals with self-regarding preferences, the unique subgame-perfect Nash equilibrium of the game with and without punishment is for second movers to return zero points and thus for first movers not to cooperate (since punishment is always costly, it is

not credible at any stage). The predictions can change if individuals possess other-regarding preferences such as a concern for unequal payoffs, efficient outcomes, and/or reciprocating kind and unkind actions (e.g., Rabin, 1993; Fehr and Schmidt; 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Dufwenberg and Kirchsteiger, 2005; Falk and Fischbacher, 2006). In the game without punishment, if the frequency of selfish individuals is sufficiently low then there can be equilibria where some second movers return positive amounts and some first movers cooperate. In the game with punishment, in addition to individuals who are willing to act kindly, there might be individuals who are willing to punish selfish behaviour. If the expectation of punishment leads to higher returns from the second movers, then first movers have an additional incentive to cooperate.<sup>2</sup> Certainly, the first movers' willingness to punish depends on the amount of retaliation by second movers, which in turn depends on the willingness of first movers to punish once again, and so on. This, in our opinion is a more realistic way of modeling social punishment. If both the punisher and the punished have access to the punishment technology, the punished will always have the opportunity to retaliate. Moreover, both players have the option to avoid further interaction by deciding not to punish and thus ending the game. To our knowledge, no other study examines punishment behaviour in such a setting.<sup>3</sup>

## **1.2 Measurement of Emotions**

During the experiment participants were presented with a number of questionnaires concerning emotions and evaluations of expectations and fairness.

Emotions were measured on a seven-point scale<sup>4</sup> after subjects observed the choice of the other player but before they made their own choice. The list of measured emotions included: anger, gratitude, guilt, happiness, irritation, shame, and surprise. Expectations about the behaviour of the other player were measured by asking for a point estimate of the most likely action (this was done after the subjects made their decision but before they observed the other player's choice). Fairness perceptions were measured on a seven-point scale at the end of the experiment in the debriefing questionnaire (see Appendix A).

Despite their widespread use in social psychology and some areas of economics, such as research on happiness (e.g., Frey and Stutzer, 2002; Kahneman and Krueger, 2006) and trust (e.g., Glaeser *et al.*, 2000), the use of self reports is sometimes regarded with suspicion by experimental economists (Smith and Walker, 1993). This suspicion is warranted when individual incentives are not aligned with truthful reporting. An example is voluntary contributions to public goods, where self reports will often overstate the true amount that individuals are willing to pay (Murphy *et al.*, 2005). Self reports are further not applicable in situations where participants are not aware of their own motivations or goals but are asked to describe them.<sup>5</sup> To incentivize truthful reports, many ingenious techniques have been developed, for example by paying subjects depending on the accuracy of their expectations (Murphy and Winkler, 1970). However these methods require a measure of accuracy, which is not always available.

Since emotions are internal, difficult to observe states, self reports of emotions are an often used technique in social psychology (Robinson and Clore,



2002). Psychological research has shown that self reports of emotional experiences are consistently correlated with a number of different physiological measures (e.g. heart rate and facial muscle activity see, Bradley and Lang, 2000). In experimental economics self reports of anger have been related to skin conductance levels for emotional reactions in the power to take game (Ben-Shakhar *et al.*, 2007). Physiological measures are however difficult to apply to social emotions (Adolphs, 2002). To quote Tangney and Dearing (2002): “Shame and guilt are internal affective states that are difficult, if not impossible, to assess directly. For example, unlike most of the ‘primary’ emotions (e.g., anger, sadness, joy), shame and guilt do not involve clearly definable, codable facial expressions [...]. There is no clear-cut ‘guilt’ expression, nor a readily recognizable ‘shame’ expression”. It is therefore considered that the best way to measure guilt and shame are self reports. That these are indeed reliable can be deduced from a recent neuroimaging study (Takahashi *et al.*, 2004) that showed correlations of self reported guilt ratings and activation of theory of mind related brain areas (i.e. medial prefrontal cortex and posterior superior temporal sulcus).<sup>6</sup>

For all these reasons, we believe that self-reports of emotions are reliable with respect to our research questions. Concerning social emotions they clearly are the preferred measurement method. In addition, since we concentrate on relative differences in emotional reactions, a general bias in a given direction will not affect our results. Lastly, to avoid focusing subjects on a specific emotion and to limit post-rationalization, we confront subjects with a number of different emotions, which they rate before making their decision.

## 2 Observed Behaviour

In total 162 students participated in the experiment (26 in the baseline treatment and 136 in the punishment treatment). Approximately 54% were students of economics and the rest came from a variety of fields. The average age was 22 years and 58% of the participants were male. Each subject played *twice* the social dilemma game (we refer to each play as part 1 or part 2). We used a perfect strangers matching protocol to avoid any reputation effects. The average earnings were 10.55 euros. The experimental procedures as well as the instructions are found in Appendix A.

In this section, we give an overview and a brief discussion of the behaviour of first and second movers. A summary of the behavioural data can be found in Appendix B. We start by investigating how often first movers cooperate and how much second movers return. Comparing the baseline and the punishment treatments allows us to observe the effect of the punishment institution on the subjects' behaviour. Then, in order to explain any differences induced by punishment, we analyze the punishment behaviour of first movers as well as the retaliatory behaviour of second movers. Finally, we examine whether punishment and retaliation in part 1 have an effect on their behaviour in part 2.

### 2.1 Cooperation and Returns

Figure 2 summarizes the main differences between the baseline and the punishment treatment. Namely, first movers cooperate more often and second movers return more in the presence of punishment.

— Figure 2 around here —

As can be seen in Figure 2A, in both treatments, almost all first movers cooperate in the first part (more than 84.6%). However, in the absence of punishment, cooperation decreases substantially in the second part. In contrast, if the opportunity to punish others exists, first movers cooperate equally often in both parts. Testing for differences between treatments confirms this observation. There is no significant difference in the frequency of cooperation in the first part ( $p = 0.90$ ) but a highly significant difference in the second part ( $p = 0.02$ ).<sup>7</sup> There is an even starker difference between treatments when we consider the behaviour of second movers. In part 1 and 2, second movers return noticeably less in the absence of punishment ( $p = 0.01$  and  $p = 0.07$ ).

Given the behaviour of second movers, it is easy to understand the decrease in cooperation in the baseline treatment. Remember that first movers who cooperate send 50 points. In the baseline treatment, they receive on average a smaller amount in return. In contrast, first movers who cooperate in the punishment treatment receive back roughly twice the amount sent. Thus, even when it is possible to retaliate, punishment limits the opportunistic behaviour of second movers and helps sustain high cooperation levels. It is still a puzzle, however, why first movers do not foresee the low returns of second movers in the first part of the baseline treatment. This could be the result of subjects forming overoptimistic beliefs concerning the cooperativeness of others due to the similarity between the experiment and common everyday interactions in which cooperation is an equilibrium (e.g. due to punishment

or repeated play).<sup>8</sup>

In spite of more cooperation, punishment did not lead to overall higher earnings. In part 1, the average earnings of all participants are actually higher in the baseline treatment (230.8 vs. 189.0 points), whereas in part 2, average earnings are higher in the punishment treatment (187.3 vs. 182.7 points). In neither case is the difference significant ( $p > 0.23$ ). In the following paragraphs, we examine how subjects punish and retaliate.

## 2.2 Punishment and Retaliation

As Figure 3A illustrates (see also Table B1), a large number of subjects are willing to spend some or all of their earnings in order to punish second movers or to retaliate against first movers. In the cases where the second mover returned less than 150 points, about two thirds end up in punishment (68.1%). Retaliation against first movers is common but somewhat less frequent (40.0%). We even observe that a large fraction (55.6%) of first movers who had the opportunity to punish once again decide to do so (we refer to this as ‘additional punishment’).<sup>9</sup>

Figure 3B shows that the amount spent on punishment by first movers who got back less than 150 points was clearly higher than the amount spent on retaliation by second movers who got punished ( $p = 0.01$ ). Surely, this is partly explained by the fact that the earnings of first movers when they faced retaliation were lower than the earnings of second movers when they faced punishment. Therefore, since the amount of punishment or retaliation is limited by the earnings of the other player, first movers were able to spend

more on reducing the other's payoff. Still, if we normalize both punishment and retaliation using the maximum amount of points that an individual could assign to the other, we see that first movers are more aggressive punishers than second movers ( $p = 0.09$ ).

— Figure 3 around here —

Although it is not predicted by traditional economic theory (assuming own-payoff maximization), the punishment behaviour of first movers has been observed in similar experiments (see, Camerer, 2003). Also in parallel with other experimental findings, we observe that the amount and frequency of punishment increases as the amount returned decreases. First movers who received 150 points punish less than first movers who received 50 or 0 points (in each part  $p < 0.01$ ). Comparing first movers who received 50 points with those who received 0 points shows that the latter punish significantly more only in the second part ( $p = 0.02$ , otherwise  $p > 0.28$ ).

We find more intriguing the willingness of second movers to retaliate. After all, these subjects had behaved in an unkindly manner by returning less than 150 points. Furthermore, when they had to decide whether they wanted to retaliate, 65.0% of the second movers had equal or higher earnings than the first mover. It is remarkable that 7 (i.e. 53.8%) of these 13 second movers chose a positive amount of retaliation.<sup>10</sup> Unlike for first movers, the retaliatory behaviour of second movers does not seem to depend on the actions of the other player.<sup>11</sup> For instance, there is no significant difference in the amount or the frequency of retaliation between second movers who received a large amount of punishment and second movers who received a

small amount (above or below median punishment,  $p > 0.50$ ).

It is instructive to calculate how retaliation affects the first movers' 'real' cost of punishment. Whenever first movers punish, they not only incur the cost of reducing the second mover's earnings, but they also risk further losses if the second mover decides to retaliate.<sup>12</sup> If there is no retaliation, the cost of punishment is 0.250 points per point reduced. Including the actual losses due to retaliation increases the average costs of punishment to 0.399 points per point reduced. Nonetheless, even though this is a substantial increase of 59.4%, punishment can still be considered an inexpensive tool for the reduction of the second mover's earnings. This might explain why cooperation is sustained in spite of frequent retaliation. However, more generally the impact of retaliation on the costs of punishment will depend on the game played and its parameters. It is possible that in some cases retaliation will drive the costs of punishment to the point where punishment fails to sustain cooperation.<sup>13</sup> A similar analysis for the real cost of retaliation (given losses due to additional punishment) gives that second movers incur a cost of 1.013 points per point reduced. This remarkable 305.6% cost increase might explain why second movers punish less aggressively than first movers do. We now turn to how first and second movers adjust their behaviour from part 1 to part 2.

## 2.3 Dynamics

As already noted, the starkest difference between treatments concerning the behaviour of first movers is the large decrease in cooperation from part 1 to

part 2 in the baseline treatment compared to the punishment treatment. On closer inspection, this difference is due to two reasons. First, as shown in Figure 4A, in the baseline treatment 66.7% of the first movers who got back less than 150 points in part 1 defected in part 2. In contrast, in the punishment treatment it was only 19.0% (the difference is significant,  $p = 0.04$ ). Second, in the baseline treatment more second movers chose to return less than 150 points (81.8% in the baseline treatment vs. 35.6% in the punishment treatment,  $p = 0.01$ ). Hence, it appears that punishment has two desirable effects. On one hand, second movers anticipate punishment and as a result increase the amount returned. On the other hand, after experiencing selfish behaviour, first movers are more willing to keep on cooperating if they have the opportunity to punish. In fact, if we examine how first movers in the punishment treatment adjust their behaviour, we find that, among first movers who receive less than 150 points, those who punish seem to be more likely to cooperate once again than those who did not punish (see Figure 4B).

— Figure 4 about here —

We now turn to the effects of punishment on the future behaviour of second movers. If we concentrate on second movers who had a good chance of being punished (i.e. those who returned less than 150), we find that, on average, second movers who were not punished decrease the amount returned by 25.0 points whereas those who were punished increase it by 10.0 points ( $p = 0.22$ ). Hence, although actual punishment does promote prosocial behaviour, the effect is not particularly strong. In other words, punishment has

a bigger impact by deterring second movers from returning low amounts in the first place than by increasing the returns of those who behave selfishly in spite of the threat of punishment. For example, if none of the second movers who returned a low amount had been punished in period 1, the average return in period 2 would have been 87.7 points (instead of 96.6 points). In contrast, if the threat of punishment had not been there at all then the average return would have been as in the baseline treatment (i.e. 41.7 points).

Lastly, we analyze the impact of retaliation on both future cooperation and punishment by first movers. In general, retaliation in part 1 does not deter first movers from cooperating in part 2. For instance, among first movers who punished a low return in part 1, those who received retaliation were as likely to cooperate in part 2 as those who received no retaliation ( $p = 0.64$ ). It is also the case that retaliation does not deter first movers from punishing. Among the first movers who punished in part 1 and then received a low return in part 2, those who had received positive retaliation punished in part 2 as often as those who had received no retaliation ( $p = 0.80$ ). In fact, they punished as often as those who received a low return in part 2 after they had received a high return in part 1 ( $p = 0.36$ ). The main findings from the behavioural data are summarized in the following result:

**Result 1:** *In the presence of punishment opportunities, cooperation is sustained at high levels. This is because, second movers return more, and first movers who punish do not stop cooperating after experiencing opportunistic behaviour. Punishment of opportunistic behaviour is common and persistent despite the fact that in numerous cases punishment leads to retaliation by second movers.*



### 3 Emotions and Behaviour

In the following section we investigate the relationship between the subjects' behaviour and their emotional response. To begin, we concentrate on the emotions of first movers, and particularly on the relationship between anger and punishment. Subsequently, we analyze what triggers first movers to feel high intensities of anger. We then turn to study the emotional reaction of second movers. First, we investigate the relationship between guilt and the second movers' decision to retaliate. Second, we try to determine what causes second movers to feel guilt.

— Table 1 about here —

Throughout the section, we report the results of tests done with the emotion of anger and the emotion of guilt. However, we should note that we find very similar results and significance levels if we substitute anger with irritation or (lack of) happiness, or if we substitute guilt with shame. This hints at the possibility that some of these emotions are in fact measuring the same underlying effect. To confirm whether this is the case we applied principal factor analysis to the emotions data.<sup>14</sup> As can be seen in Table 1, we find three factors that summarize the subjects' emotional response. For both first and second movers, the first factor can be interpreted as anger-like emotions and the second factor as guilt-like emotions.<sup>15</sup> It is interesting to note that the subjects' negative emotional response is described by these two factors. This indicates that, in the experiment, the experience of guilt and shame is distinct from other negative emotions and is not a reflection of a more general negative mood. Furthermore, as argued by Zeelenberg and

Pieters (2006), it gives support to a more complex view of emotions than a one-dimensional positive-negative scale (as is implied in many studies e.g., Kahneman, 1999).

### 3.1 Anger and Punishment

Throughout the experiment, anger is clearly related to the punishment decision. As is illustrated in Figure 5, first movers who were angry after observing the amount returned by the second mover punish more than first movers who were not angry ( $p < 0.01$  in both parts).<sup>16</sup> Furthermore, although there are few observations, a similar pattern is observed in the second punishment round. On average, after observing the amount of retaliation assigned to them, first movers who felt angry punish more than first movers who did not feel as angry ( $p = 0.11$  for the amount of additional punishment and  $p = 0.41$  for its frequency). As one might expect, neither surprise nor the social emotions of guilt nor shame are related to the first movers punishment decision.

— Figure 5 about here —

Having found that punishment is related to experienced anger, the question arises what explains the different intensities of anger (the emotional reaction of first movers to the amount returned can be found in Appendix B). In both treatments, the most important trigger of high intensities of anger is simply receiving back less than 150 points. First movers who received 150 points felt lower intensities of anger than first movers who received either 50 or 0 points ( $p = 0.01$ , see Table B3). It is also the case that first movers who

received 0 points back were angrier than those who received 50 points back ( $p = 0.03$ ).

In addition to the amount returned, the first movers' expectations have an effect on the intensity of anger. In particular, first movers who overestimated the amount returned by the second mover tended to be angrier than first movers who underestimated it. For example, if we control for the amount that was actually returned by concentrating on first movers who got back 50 points, we find that first movers who were expecting back 150 points were angrier than first movers who were expecting back 50 or 0 points ( $p = 0.01$ ).

Lastly, we also observe that the amount of anger experienced by first movers is related to their fairness perceptions. First movers who thought it is unfair to return low amounts were angrier than those who thought that it is fair to return low amounts (below or above median fairness). For instance, if we look again only at first movers who got back 50 points, we find that first movers who thought returning 50 was unfair were angrier than first movers who thought returning 50 was fair ( $p = 0.01$ ).

We get similar results in a regression. Specifically, we estimate anger using the amount returned, the expected amount returned, the perceived fairness of returning 50 points, and some demographic variables. We find first movers feel angrier the less is returned. Especially if they were expecting a return of 150 points or considered low returns to be very unfair (see Table C1).

Focusing on the emotional reaction of first movers to the amount of retaliation received from the second mover gives a comparable finding. Namely, first movers who faced no retaliation experienced lower intensities of anger than first movers who faced positive retaliation ( $p = 0.04$ , see Table B4).

Unfortunately, in this case we do not have enough observations to test for the effects of expectations and fairness perceptions. These findings are summarized in the following result.

**Result 2:** *First movers who punish do so because they are angry. High intensities of anger are triggered by selfish behaviour by the second mover, especially if it is unexpected and considered unfair. Retaliation by second movers also makes first movers angry and leads to additional punishment.*

### 3.2 Guilt and Retaliation

We now turn to the relationship between the emotions and behaviour of second movers (the emotional reaction of second movers can be found in Table B5). We find that feelings of guilt are clearly related to retaliation. In particular, second movers who felt no guilt are more likely to retaliate than other second movers. Furthermore, we also find that, for second movers who were punished, experiencing guilt induces them to correct their behaviour. For anger-like negative emotions as well as the positive emotions, there no direct relationship with retaliation. However as will be explained below, this might be due to the effect of guilt itself.

As can be seen in Figure 6, second movers who felt no guilt after being punished are more likely to retaliate than second movers who felt guilty ( $p = 0.04$ ). We also get a similar result if we test for differences in the amount of points spent on retaliation ( $p = 0.08$ ).

— Figure 6 about here —

The effect of guilt can be further described if we analyze the interaction

of guilt and anger. Given that anger motivates first movers to punish, one could think that, if second movers get angry when they are punished, anger could motivate second movers to retaliate. Indeed, a simple look at the relationship between anger and retaliation suggests that second movers who are angry retaliate more often than second movers who are not angry (see Figure 6). However, these differences are not significant ( $p = 0.77$  for the differences in the amount of retaliation and  $p = 0.82$  for the differences in frequency).

Examining the interaction of anger and guilt clarifies why some of the angry second movers do not retaliate. Second movers who were angry and felt no guilt retaliate more frequently than second movers who were angry and felt guilt ( $p = 0.02$  and  $p = 0.02$ ). For second movers who were not angry, there are no significant differences between those who felt no guilt and those who did ( $p > 0.79$ ). Hence, guilt appears to influence the behaviour of second movers by suppressing their anger-induced desire to retaliate.

In addition to retaliation, guilt seems to be related to how second movers adjust their behaviour from part 1 to part 2. In Section 2 it was shown that second movers who were punished tend to return more in the subsequent part than second movers who were not punished. However, this effect is not very strong. The emotional reaction of second movers hints that the propensity of second movers to adjust their behaviour after being punished depends on whether they felt guilty or not. On average, second movers who felt guilt after being punished increase the amount returned by 50.0 points whereas those who felt no guilt decrease the amount returned by 9.1 points ( $p = 0.11$ ). Next, we explain the differences in the intensities of anger and

guilt experienced by second movers.

The most important reason why second movers get angry is simply receiving a positive amount of punishment (see Table B5). For example, second movers who were punished at least once reported significantly more anger than those who were never punished ( $p = 0.01$ ).<sup>17</sup> We further investigate the effect of punishment on anger through a regression. We estimate anger among second movers who received a positive amount of punishment using demographic variables and three variables capturing the interaction between the amount of punishment and the amount returned. The regression is available in Table C2. We find that higher amounts of punishment trigger higher intensities of anger. Furthermore, the increase in anger is bigger when the second mover returns a high amount.<sup>18</sup> This is understandable given that the more a second mover returns, the more undeserved is the punishment.

The clearest trigger of high intensities of guilt is acting selfishly. Second movers who returned 150 points reported lower intensities of guilt than those who transferred less ( $p = 0.02$ ).<sup>19</sup> If anticipated, this type of emotional reaction supports the idea that some individuals will not act selfishly in order to avoid feelings of guilt. We do not find, however, a difference between the intensity of guilt reported by second movers who returned 50 points and those who returned 0 points ( $p = 0.53$ ).

Interestingly, punishment does not seem to influence the intensity of guilt experienced by second movers. For example, among second movers who returned less than 150, there is no significant difference between the amount of guilt reported by those who were punished and by those who were not ( $p = 0.58$ ).<sup>20</sup>

We do not find that, for a given transferred amount, fairness perceptions influence the intensity of guilt. However, we do find that second movers who thought it is unfair to return low amounts transferred significantly more than those who thought that it is fair to return low amounts (117.5 vs. 59.4,  $p < 0.01$ ). Hence, the apparent disconnection between guilt and fairness perceptions might be due to the correlation between fairness perceptions and the amount returned. A possible explanation for this is that fair-minded second movers feel more guilt when transferring a low amount. Hence, they return a high amount in order to avoid high intensities of guilt. The following result summarizes the findings concerning guilt.

**Result 3:** *Second movers who retaliate do so because they are angry and do not feel guilt. In addition, following the feeling of guilt, second movers are more likely to rectify selfish behaviour. High intensities of anger are triggered by punishment, especially if the second mover had returned a high amount. High intensities of guilt are triggered by selfish behaviour and are not affected by punishment.*

## 4 Discussion and Conclusions

In this paper, we have shown that a realistic punishment institution, in which multiple rounds of punishment and retaliation are possible, is an effective tool for the support of cooperative behaviour. However, retaliation is commonly observed and can result in an extreme reduction of the payoffs of the involved individuals. We further confirmed that anger-like emotions are an important motivation for punishment. Selfish behaviour induces anger in others and

thus increases the likelihood of being punished. Finally, we showed that the experience of prosocial emotions, namely shame and guilt, restrains angry individuals from retaliating. Therefore, prosocial emotions can be regarded as a mechanism adjusting the behavioural reactions of anger.

Many economists acknowledge that emotions are an essential component of decision making (e.g., Loewenstein, 2000). Knowing which emotions are at play in a given situation can help us understand why people take certain actions and help us build more accurate models. These insights can further point to effects or variables that would have otherwise been ignored. In this section we discuss some of our findings that illustrate these two points.

As expected, we find that anger motivates individuals to punish opportunistic behaviour (Bosman and van Winden, 2002). Anger, unlike other emotions such as envy, is known to be strongly affected by expectations and the intentionality of an action (Haidt, 2003). This helps explain why punishment increases the more unexpected the opportunistic act is (Ben-Shakhar *et al.*, 2007), and decreases if the action was unintentional (Falk *et al.*, 2008; Charness and Levine, 2007). In fact, our results show that unfulfilled expectations can be as important for punishment as fairness concerns. In economics, expectations are usually assumed to affect behaviour only if the uncertainty remains unresolved. However, our results indicate that for punishment, this is not the case (see also, Reuben and van Winden, 2008). Note that, if anger had not been associated to punishment by previous research, this role of expectations might have been overlooked.

Another characteristic of anger is that its action tendency is to attack even if triggered by unfair behaviour (Lazarus, 1991). An individual who



experiences unfairness (e.g. a deviation from an equality or maximin norm) can have various reasons to punish the initiator of the unfair action. One reason could be to correct the resulting unfair material outcome (e.g. by equalizing payoffs). A second one could be to try to alter the future behaviour of the norm transgressor by teaching him that acting unfairly does not pay. Lastly, a third reason to punish could simply be to harm the other party (anger-induced punishment indeed seems to give pleasure to the punisher, Quervain *et al.*, 2004). Although all three reasons might be present, only the third is easy to reconcile with an excessive use of punishment. For example, if first movers who got back 50 points used punishment to equalize payoffs or to make the selfish action unprofitable, then they should not punish more than 266.67 or 100 points respectively.<sup>21</sup> However, of the first movers who punished at least once in this situation, 31.3% did so by more than 266.67 points and 81.3% by more than 100 points. Thus, restricting anger-induced punishment in settings similar to this experiment might be desirable as its shown in this paper, its excessive use can lead to retaliation from punished individuals and to further destruction of resources.<sup>22</sup> In this sense, outcome based models of social preferences such as Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) miss an important characteristic of punishment behaviour.

An important and yet overlooked aspect of punishment is the emotional reaction of the punished. As was shown in this paper, prosocial emotions such as guilt play a crucial role for the enforcement of social norms. In Section 3 we have shown that feeling guilty helps explain why some individuals who acted selfishly adjust their behaviour whereas others do not. It has been observed

that in public good games, the use of non-monetary punishment has a positive effect on contribution levels.<sup>23</sup> Non-monetary punishment has the desirable property that it can affect behaviour without destroying resources. However, as shown by Noussair and Tucker (2005), the lack of real consequences for free-riders makes this effect deteriorate over time. This suggests that the effect of guilt is not very strong when punishment is only symbolic. Our results indicate that it is the combination of feeling guilty and receiving monetary punishment that has a significant effect on behaviour. In this sense, the best performing punishment institution might be one in which both symbolic and monetary punishments are available (Noussair and Tucker, 2005).

Another essential role for guilt is the prevention of retaliation by punished individuals. As we stated in Result 3, even if individuals acted unkindly, they still feel angry when they are punished. However, it is only those individuals who are angry and do not feel guilty that decide to retaliate. Therefore, if it were not for some individuals experiencing guilt, retaliation would be much more common and punishment of selfish behaviour much more costly. For example, if second movers who felt guilty had behaved as second movers who did not feel guilty (controlling for anger) then retaliation would have been 69.8% more frequent and 72.4% higher. Furthermore, the decrease in the amount returned from part 1 to part 2 would have been 42.2% more severe. Social emotions like guilt are thus essential for the effectiveness of a punishment institution. This fits the assumption that social emotions coevolved with institutions and anger-like emotions to limit antisocial actions (Bowles and Gintis, 2005).<sup>24</sup> An interesting question for further exploration is the

specific evolutionary mechanisms that lead to this situation.

Shame and guilt are both ‘self-reproach’ emotions elicited by the individuals’ own blameworthy actions (Ortony *et al.*, 1988). This study was not designed to differentiate between the effects of these two emotions. However, we should note that the emotions’ action tendencies are different (Tangney and Dearing, 2002). Guilt is related to the blameworthiness of an act and is thus more likely to result in reparation and action. Shame is related to a devaluation of the self and is more likely to result in avoidance of further contact.<sup>25</sup> Knowing which emotion dominates can be important. For example increasing feelings of shame (e.g. through framing) does not necessarily lead to an increase in prosocial behaviour. If individuals are given the possibility to avoid contact altogether, they might prefer to do so instead of participating in an activity where feelings of shame ‘force’ them to act prosocially (Lazear *et al.*, 2006). Finally, already the anticipation of shame and guilt might induce norm-abiding behaviour.

In economics, most of the emphasis on emotions has been given to their causes and valence (Zeelenberg and Pieters, 2006). That is, the triggers of emotional reactions and whether emotions increase or decrease utility. However, emotions are a more complex phenomenon. As we showed in this paper, it is also important to take the emotions’ action tendencies into account. Negative emotions such as guilt and anger can both be considered to decrease utility, but at the same time they produce very different behaviours. The study of how specific emotions and their action tendencies affect individual behaviour can greatly increase our understanding of economic decision making.

## Footnotes

1. Players can have negative earnings if by assigning punishment points to another player they reduce their own earnings below zero. This way subjects cannot avoid punishment or retaliation by reducing the earnings of others to zero. A show-up fee was given to cover any losses incurred during the experiment.

2. For example, using the same assumptions they use about the distribution of types, the model of Fehr and Schmidt (1999) predicts that in the case of no punishment 40% of second movers return 150 points. In this situation only 30% of first movers cooperate (the other 70% prefers to avoid the chance of ending up with extremely disadvantageous inequality). In the case of punishment, there are enough first movers that would punish so that all second movers return 150 points and hence all first movers cooperate.

3. Nikiforakis (2008) studies punishment in a public good game in which retaliation was possible, but in which the punishment phase ends automatically after retaliation. Furthermore, various bargaining games have a similar punishment structure (e.g. rejections in alternating bargaining games see, Binmore *et al.*, 1989, 1991). However, in these cases punishment is less intuitive as subjects generally make new offers after a rejection decision (this makes the cost of punishment unclear and in some cases not even positive).

4. From: 1 = “not at all” to 7 = “very intensely”; see Appendix A.

5. An extreme case is patients whose brain hemispheres are separated and therefore one side might not be aware of the motivations of the other (LeDoux, 1996).

6. Participants had to read and contemplate guilt-inducing phrases and

were later asked to rate on a six-point scale how guilty they felt.

7. Throughout the paper, unless it is otherwise noted, we use two-sided Wilcoxon-Mann-Whitney tests. We use each subject as an independent observation for tests concerning either part 1 or part 2. If we combine the data of both parts, we first calculate for each subject the mean for the variable in question and then compute the test using these means as the independent observations. There are subjects from whom we have, for various variables, data from only one of the parts (e.g. a second mover who faces a first mover who cooperates in part 1 and a first mover who defects in part 2). In these cases, we take the data from the part for which we have information as that subject's mean.

8. Subjects in the baseline treatment are indeed very optimistic. They expected a return of 31.8 points more than the one they actually got. In the punishment treatment, subjects expected a return of 3.4 points less than the one they got.

9. We only observe one case in which the second mover retaliated once again ( $p_{22} > 0$ ). However, this is because in all the other pairs where the first mover punished a second time ( $p_{12} > 0$ ), the first mover ended up with zero points or less and hence the punishment stage ended automatically. We also observe only one case in which a second mover was punished for returning 150 points (for similar behaviour see, Cinyabuguma *et al.*, 2006; Gächter and Herrmann, 2006). In this case further rounds of retaliation and additional punishment lead to both subjects receiving 0 points.

10. This tendency to retaliate against punishers could be the reason why we see 'perverse' punishment in public good games (Cinyabuguma *et al.*,

2006). In other words some low contributors might punish high contributors because they *expect* to be punished by them.

11. We also find that retaliation does not depend on the amount transferred. There is no significant difference between second movers who returned 50 points and those who returned 0 points ( $p > 0.55$ ).

12. The only case in which second movers cannot retaliate after being punished occurs when first movers who get back 0 points spend all of their earnings punishing the second mover. In this case, both subjects end up with 0 points and no further retaliation is possible. Overall, this occurred in 24.3% of the cases in which there was positive punishment.

13. In public good settings, punishment stops sustaining cooperation when the cost of punishing increases over 0.500 per point reduced (Nikiforakis and Normann, forthcoming).

14. The table reports the results for the emotional response of first movers to the amount returned and for the emotional response of second movers to the amount of punishment. In both cases, we use orthogonal varimax rotation. Results do not vary significantly if we combine the emotional response of first and second movers or if we use other rotation methods.

15. The interpretation of the third factor is less clear. For first movers it represents mostly surprise, and for second movers it is a combination of gratitude and happiness.

16. Throughout the paper, we refer to a person feeling ‘angry’ if the reported value for anger was above the median, and as feeling ‘not angry’ if the value was below or equal to the median. The same is true in the case of guilt.

17. This is also true if we test only among second movers who returned less than 150 points ( $p = 0.01$ ).

18. We use three variables  $I^r$  with  $r \in \{0, 50, 150\}$ .  $I^r = 0$  if the amount returned was different from  $r$ , and  $I^r =$  the amount of punishment received if the amount returned was  $r$ . We obtain positive and significant coefficients for  $I^0$ ,  $I^{50}$ , and  $I^{150}$  ( $p < 0.02$ ) with the coefficient for  $I^0$  being the smallest and the one for  $I^{150}$  being the largest. The coefficient of  $I^{150}$  is significantly different from those of  $I^0$  and  $I^{50}$  (Wald tests,  $p < 0.01$ ). The coefficient of  $I^{50}$  is higher but not significantly different from the coefficient of  $I^0$  (Wald test,  $p = 0.21$ ). See Table C2 for details.

19. This result is not driven by the different punishment rates faced by subjects who returned 150 points and by those who returned less. For example, second movers who returned 150 points and were not punished felt lower intensities of guilt than second movers who returned less than 150 points and were not punished ( $p = 0.01$ ).

20. The lack of a significant relationship between guilt and punishment could be due guilt being affected only in certain circumstances. Punishment might have little effect on guilt among individuals that felt little guilt before being punished, and a larger effect on individuals that were already feeling guilty. However, exploring this would require the measurement of guilt before and after punishment, which we did not do.

21. If the first mover knows there are people who do not punish, then he must punish more in order to make the second mover's action unprofitable on average. Given the frequency of punishment in the experiment this amounts to punishing by 168.6 points.

22. In this respect, punishment by an unaffected third party might be better. If observing a norm violation (as opposed to experiencing it) triggers less anger but retains the desire to correct unfair behaviour, then punishment is less likely to be disproportionate to the norm violation. This is in line with research that finds third party punishment to be less severe (Fehr and Fischbacher, 2004; Carpenter and Matthews, 2005).

23. For instance, Masclet *et al.* (2003) use symbolic punishment points and find that, in the short run, they work almost as well as real punishment points. Barr (2001) reports that the public blaming of the free-rider can increase cooperation in future rounds.

24. Although some social emotions such as empathy have been documented among animals (Preston and de Waal, 2002), guilt and shame are still considered to be characteristic human emotions (see, Kagan, 2004).

25. Economists usually distinguish shame and guilt by the visibility of behaviour. Shame is said to be triggered in social situations in which actions are seen by others, whereas guilt is more related to internalized values and hence is not influenced by the presence of others (e.g., Kandel and Lazear, 1992). However, research by psychologists has shown that people feel shame even when their actions are unobserved (Tangney *et al.*, 1996), and that the experience of guilt varies considerably depending on the interpersonal context (Baumeister *et al.*, 1994).



# **A Experimental Procedures and Instructions**

## **A.1 Experimental Procedures**

The computerized experiment was conducted in March 2005 in the CREED laboratory at the University of Amsterdam. Subjects were recruited through the CREED recruitment website and the experiment was programmed with z-Tree (Fischbacher, 2007). The experiment lasted approximately 30 minutes and the average earnings, including a 1 euro showup fee, were 10.55 euros (30 experimental points equalled 1 euro).

Each subject played twice the social dilemma game described in Section 1. We used a perfect strangers matching protocol to avoid reputation effects. In total, 26 subjects participated in the baseline treatment, without punishment opportunities. The remaining 136 subjects participated in the punishment treatment.

After arrival in the reception room, subjects were randomly assigned to a desk in the laboratory. Once everyone was seated, subjects were given the instructions for the experiment (see below). Subjects were told that the experiment consisted of two independent parts. We emphasized the fact that they will interact with different individuals in each part, and that, their choices in the first part will not affect their earnings in the second part. After this, the one-shot social dilemma game was described as the first part of the experiment. Once all subjects finished reading the instructions, they answered a few questions to ensure their understanding of the game. Subsequently, the subjects played the social dilemma game via the computer (part 1). At the end of the first part, instructions were distributed concerning the

second part of the experiment. Subjects were told they would be in the same position as in the first part (i.e. first or second mover), and with certainty, their partner would be different partner from the one they had played with in the first part. After they finished the second part (part 2), subjects filled in a debriefing questionnaire. Thereafter, they were paid their earnings in private and dismissed.

## **A.2 Instructions**

These are the instructions for first movers in the punishment treatment. The instructions for second movers and for the baseline treatment are available upon request.

### **A.2.1 Instructions for Part 1**

There are two types of participants in this part, participants A and participants B. Half of the persons participating in the experiment will be in the role of participant A, and the other half in that of participant B. *You are a participant A.*

In part 1 of the experiment, you will be randomly assigned a participant B. During this part, you will interact only with this participant B. Moreover, you will *not* interact again with this participant in part 2 of the experiment. Part 1 consists of three steps. In step one, you must decide whether you will transfer points to participant B or if you will retain the points for yourself. In step two, participant B will decide if he will transfer points to you or if he will keep them himself. In step three, both of you must again make a

decision. There are various options in step three, which will be explained below. We will also describe the exact experimental procedure on the next pages.

### **A.2.2 Procedure for the Three Steps**

At the beginning of part 1 you and participant B will each receive 100 points as earnings.

*Step one:* At the beginning of the first step you will receive 50 decision points. Participant B will receive no decision points. In step one, you must decide whether you want to transfer your 50 decision points to participant B or transfer no points to participant B. If you transfer the 50 points, they will be multiplied by six, meaning that participant B will receive  $6 \times 50 = 300$  points. Then, step two begins. If you decide to transfer nothing part 1 will end here.

*Step two:* In step two, participant B has to decide whether he will transfer 150, 50 or 0 points to you. You will then receive exactly the number of points B transferred.

Therefore, four possibilities exist after the first two steps:

1. You retain your decision points: your additional earnings = 50 points, B's additional earnings = 0 points.
2. You transfer your decision points and B transfers 150 points: your additional earnings = 150 points, B's additional earnings = 150 points.
3. You transfer your decision points and B transfers 50 points: your additional earnings = 50 points, B's additional earnings = 250 points.
4. You transfer your decision points and B transfers nothing: your addi-

tional earnings = 0 points, B's additional earnings = 300 points.

Hence, after step two your total earnings will be:  $100 +$  the additional earnings indicated above.

*Step three:* In step three, you will be informed how many points participant B transferred to you. Now, you can assign penalty points to participant B. The assignment of penalty points has financial consequences for both participants, A and B. Each penalty point which you assign costs you one point, while four points are deducted from your participant B. If you assign three penalty points to participant B, this will cost you three points and participant B will have twelve points deducted.

You cannot deduct more points from participant B than his total earnings in that part (i.e.  $100 +$  B's additional earnings). If participant B has 250 points after step 2, then with your assignment of penalty points you can reduce his earnings by at most 250 points. Hence, as long as your participant B has positive earnings, you can assign him as many penalty points as you want. You can also assign him no penalty points.

Participant B will then be informed how many penalty points you assigned him and how many points were deducted from his earnings. If you decided not to assign penalty points, part 1 will end here. If you assigned penalty points to participant B, he can decide to assign penalty points to you. The assignment of penalty points has the same financial consequences as described above. Each penalty point that participant B assigns to you costs him one point, while four points are deducted from your earnings. You can not be deducted more points than the total earnings you own at that moment. If participant B decides to assign no penalty points to you, part 1 will end here.

Note: Participant B can assign penalty points even if his earnings at that point are zero. If he does so, he will lose points in part 1 of the experiment.

If participant B assigned you penalty points, you and participant B will have the option to assign penalty points to each other in turns. Part 1 will end when either you or participant B decides to assign no penalty points, or if either you or participant B can not be assigned penalty points because your or his earnings are zero or less. In other words, as long as one of you assigns a positive amount of penalty points, the other will have the opportunity to assign penalty points back. Note that, you will be able to assign penalty points *even if your earnings at that point are zero*. Furthermore, you *cannot* be assigned penalty points if your own earnings are zero.

### **A.2.3 Finally**

Remember that, you participate in part 1 only once. Therefore consider your decisions carefully. At the end of part 1 you will receive instructions for part 2 of the experiment.

### **A.2.4 Instructions for Part 2**

We will now give you the instructions for part 2 of the experiment.

Also in this part there will be two types of participants, participants A and participants B. Every person participating in the experiment will be in the role they had in part 1. Therefore, you are a participant A. As in part 1 you will be randomly assigned a participant B. During this part, you will interact only with this participant B. You can be certain that this participant B is not the same person as in part 1.

This part will consist of the same three steps as part 1. Therefore exactly the same instructions apply for part 2 as for part 1. Remember that you will participate in this part only once. Therefore consider your decisions carefully.

### **A.3 Examples of questions in the self-reports**

*To measure emotions:* “Indicate how intensely you feel each of the following emotions right now, after knowing the amount that B transferred to you?” The subject then filled in a series of seven-point scales that ranged from ‘not at all’ (1) to ‘very intensely’ (7).

*To measure expectations:* “Player A can now assign you penalty points. How many penalty points do you think A will assign to you?” The subject then entered a point estimate.

*To measure fairness perceptions:* “Suppose that participant A transfers the 50 decision points to participant B. Participant B has to choose to transfer back either 150 points, 50 points or 0 points. In your opinion, how fair do you believe is each of these choices: If participant B transfers back 150 (50, 0) points this choice is ... ?” The subject then filled in three seven-point scales (one for each choice) that ranged from ‘very unfair’ (1) to ‘very fair’ (7).

## **B Descriptive Statistics**

Table B1 and B2 summarize the behavioural data for the punishment and baseline treatments respectively. Statistics are presented for each part separately and averaged. To be precise, the last column displays the average of

the mean behaviour of each subject across both parts (first we take the mean behaviour across parts for each subject and then we take the mean across all subjects). In the cases where a subject had only one action, we take the data from that part as that subject's mean.

– Table B1 about here –

– Table B2 about here –

Tables B3 and B4 show the emotional reaction of first movers in the punishment treatment. In the baseline treatment, the emotional reaction of first movers was statistically indistinguishable from the one in the punishment treatment. The emotional reaction of second movers is summarized in Table B5.

– Table B3 about here –

– Table B4 about here –

– Table B5 about here –

## **C Additional Data Analysis**

Table C1 presents a model estimating the intensity of anger experienced by first movers after they observed the amount of points returned by the second mover in the punishment treatment. Ordered probit estimates using robust standard errors and clustering on each subject. Note that in the regression we take into account the effect of perceived fairness norms, by

estimating the models using the variable ‘Fairness of returning 50 points’. The reason for this is that this variable exhibited the most variance among the three variables measuring fairness perceptions. For the variable ‘Fairness of returning 150 points’, 85.3% of subjects agreed that it was very fair. For the variable ‘Fairness of returning 0 points’, 83.1% of subjects agreed that it was very unfair. The coefficients of ‘Return = 50’ and ‘Return = 0’ are significantly different from each other (Wald test,  $p = 0.05$ ). That is, anger is higher when the return is 0 points.

– Table C1 about here –

Table C2 presents a model estimating the intensity of anger experienced by second movers who received a positive amount of punishment. Ordered probit estimates using robust standard errors and clustering on each subject.

– Table C2 about here –

The coefficient of the variable ‘Punishment if Return = 150’ is significantly different from those of ‘Punishment if Return = 50’ and ‘Punishment if Return = 0’ (Wald tests,  $p < 0.01$ ). This indicates that second movers get angrier if they are punished for transferring a high amount. The coefficient of ‘Punishment if Return = 50’ is higher but not significantly different from the coefficient of ‘Punishment if Return = 0’ (Wald test,  $p = 0.21$ ).

*TSE (GREMAQ), University of Toulouse*

*Kellogg School of Management, Northwestern University*



## References

Abbink, K., Irlenbusch, B., and Renner, E. (2000). 'The moonlighting game: an experimental study on reciprocity and retribution', *Journal of Economic Behavior & Organization*, vol. 42(), pp. 265-77.

Adolphs, R. (2002). 'Recognizing emotion from facial expressions: psychological and neurological mechanisms', *Behavioral and Cognitive Neuroscience Reviews*, vol. 1(), pp. 21-61.

Akerlof, G.A. (1982). 'Labor contracts as partial gift-exchange', *The Quarterly Journal of Economics*, vol. 97(), pp. 543-69.

Barr, A. (2001). 'Social dilemmas and shame-based sanctions: experimental results from rural Zimbabwe', working paper 149, Centre for the Study of African Economies.

Baumeister, R.F., Stillwell, A.M., and Heatherton, T.F. (1994). 'Guilt: an interpersonal approach', *Psychological Bulletin*, vol. 115(), pp. 243-67.

Ben-Shakhar, G., Bornstein, G., Hopfensitz, A., and van Winden, F. (2007). 'Reciprocity and emotions in bargaining using physiological and self-report measures', *Journal of Economic Psychology*, vol. 28(), pp. 314-23.

Berg, J., Dickhaut, J., and McCabe, K. (1995). 'Trust, reciprocity, and social history', *Games and Economic Behavior*, vol. 10(), pp. 122-42.

Binmore, K., Morgan, P., Shaked, A., and Sutton, J. (1991). 'Do people exploit their bargaining power? an experimental study', *Games and Economic Behavior*, 3(), pp. 295-322.

Binmore, K., Shaked, A., and Sutton, J. (1989). 'An outside option experiment', *The Quarterly Journal of Economics*, vol. 104(), pp. 753-70.

Bolton, G.E. and Ockenfels, A. (2000). 'A theory of equity, reciprocity, and competition', *American Economic Review*, vol. 90(), pp. 166-93.

Bosman, R. and van Winden, F. (2002). 'Emotional hazard in a power to take experiment', *ECONOMIC JOURNAL*, vol. 112(), pp. 147-69.

Bowles, S. and Gintis, H. (2005). 'Pro-social emotions', in (L.E. Blume and S.N. Durlauf, eds.) *The Economy as a Complex Evolving System III: Essays in Honor of Kenneth Arrow*, pp. 337-67, Oxford: Oxford University Press.

Boyd, R. and Richerson, P.J. (2005). 'Solving the puzzle of human cooperation', in (S.C. Levinson and P. Jaisson, eds.) *Evolution and Culture*, pp. , Cambridge, MA: MIT Press.

Bradley, M.M. and Lang, P.J. (2000). 'Measuring emotion: behavior, feeling and physiology', in (R.D. Lang and L. Nadel, eds.) *Cognitive Neuroscience of Emotions*, Oxford: Oxford University Press.

Camerer, C.F. (2003). *Behavioral Game Theory*, Princeton, NJ: Princeton University Press.

Carpenter, J.P. and Matthews, P.H. (2005). 'Norm enforcement: anger, indignation or reciprocity', working paper, Middlebury College.

Charness, G. and Levine, D.K. (2007). 'Intention and stochastic outcomes: an experimental study', *ECONOMIC JOURNAL*, vol. 117(), pp. 1051-72.

Charness, G. and Rabin, M. (2002). 'Understanding social preferences with simple tests', *The Quarterly Journal of Economics*, vol. 117(), pp. 817-69.

Cinyabuguma, M., Page, T., and Putterman, L. (2006). 'Can second-order punishment deter perverse punishment?' *Experimental Economics*, vol. 9(), pp. 265-79.

de Quervain, D.J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., and Fehr, E. (2004). 'The neural basis of altruistic punishment', *Science*, vol. 305(), pp. 1254-8.

Dufwenberg, M. and Kirchsteiger, G. (2005). 'A theory of sequential reciprocity', *Games and Economic Behavior*, vol. 47(), pp. 268-98.

Egas, M. and Riedl, A. (2008). 'The economics of altruistic punishment and the maintenance of cooperation', *Proceedings of the Royal Society B – Biological Sciences*, vol. 275(), pp. 871-8.

Elster, J. (1999). *Alchemies of the Mind*, Cambridge: Cambridge University Press.

Falk, A., Fehr, E., and Fischbacher, U. (2005). 'Driving forces behind informal sanctions', *Econometrica*, vol. 73(), pp. 2017-30.

Falk, A., Fehr, E., and Fischbacher, U. (2008). 'Testing theories of fairness: intentions matter', *Games and Economic Behavior*, vol. 62(), pp. 287-303.

Falk, A. and Fischbacher, U. (2006). 'A theory of reciprocity', *Games and Economic Behavior*, vol. 54(), pp. 293-315.

Fehr, E. and Fischbacher, U. (2004). 'Third party punishment and social norms', *Evolution and Human Behavior*, vol. 25(), pp. 63-87.

Fehr, E. and Gächter, S. (2000). 'Cooperation and punishment in public goods experiments', *American Economic Review*, vol. 90(), pp. 980-94.

Fehr, E. and Gächter, S. (2000). 'Fairness and retaliation: the economics of reciprocity', *Journal of Economic Perspectives*, vol. 14(), pp. 159-81.

Fehr, E., Kirchsteiger, G., and Riedl, A. (1993). 'Does fairness prevent market clearing? an experimental investigation', *The Quarterly Journal of Economics*, vol. 108(), pp. 437-59.

Fehr, E. and Rockenbach, B. (2003). 'Detrimental effects of sanctions on human altruism', *Nature*, vol. 422(), pp. 137-40.

Fehr, E. and Schmidt, K.M. (1999). 'A theory of fairness, competition, and cooperation', *The Quarterly Journal of Economics*, vol. 114(), pp. 817-68.

Fischbacher, U. (2007). 'z-tree: Zurich toolbox for ready-made economic experiments', *Experimental Economics*, vol. 10(2), pp. 171-8.

Frey, B. and Stutzer, A. (2002). 'What can economists learn from happiness research?', *Journal of Economic Literature*, vol. 40(), pp. 402-35.

Gächter, S. and Herrmann, B. (2007). 'The limits of self-governance in the presence of spite: experimental evidence from urban and rural Russia', discussion paper no. 2006-13, CeDEx.

Glaeser, E.L., Laibson, D.I., Scheinkman, J.A., and Soutter, C.L. (2000). 'Measuring trust', *The Quarterly Journal of Economics*, vol. 115(), pp. 811-846.

Haidt, J. (2003). 'The moral emotions', in (R.J. Davidson, K.R. Scherer, and H.H. Goldsmith, eds.), *Handbook of Affective Sciences*, Oxford: Oxford University Press.

Kagan, J. (2004). 'The uniquely human in human nature', *Daedalus*, vol. 133(), pp. 77-88.

Kahneman, D. (1999). 'Objective happiness', in (D. Kahneman, E. Diener, and H. Schwartz, eds.) *Well-being: The Foundations of Hedonic Psychology*. New York, NY: Russell Sage Foundation.

Kahneman, D. and Krueger, A.B. (2006). 'Developments in the measurement of subjective well-being', *Journal of Economic Perspectives*, vol. 20(), pp. 3-24.

Kandel, E. and Lazear, E.P. (1992). 'Peer pressure and partnerships', *Journal of Political Economy*, vol. 100(), 801-17.

Lazarus, R.S. (1991). *Emotion and Adaptation*. Oxford: Oxford University Press.

Lazear, E.P., Malmendier, U., and Weber, R.A. (2006). 'Sorting in experiments', working paper 12041, NBER.

LeDoux, J. (1996). *The Emotional Brain – The Mysterious Underpinnings of Emotional Life*. Touchstone.

Loewenstein, G. (1996). 'Out of control: visceral influence on behavior', *Organizational Behavior and Human Decision Processes*, vol. 65(), pp. 272-92.

Loewenstein, G. (2000). 'Emotions in economic theory and economic behavior', *American Economic Review*, vol. 90(), pp. 426-32.

Masclot, D., Noussair, C., Tucker, S., and Villeval, M.-C. (2003). 'Monetary and non-monetary punishment in the voluntary contribution mechanism', *American Economic Review*, vol. 93(), pp. 366-80.

Murphy, A.H. and Winkler, R.L. (1970). 'Scoring rules in probability assessment and evaluation', *Acta Psychologica*, vol. 34(), 273-86.

Murphy, J.J., Allen, P.G., Stevens, T.H., and Weatherhead, D. (2005). 'A meta-analysis of hypothetical bias in stated preference valuation', *Environmental and Resource Economics*, vol. 30(), pp. 313-25.

Nikiforakis, N. (2008). 'Punishment and counter-punishment in public good games: can we really govern ourselves', *Journal of Public Economics*, vol. 92(), pp. 91-112.

Nikiforakis, N. and Normann, H.T. (forthcoming). 'A comparative statistics analysis of punishment in public-good experiments', *Experimental Economics*.

Noussair, C. and Tucker, S. (2005). 'Combining monetary and social sanctions to promote cooperation', *Economic Inquiry*, vol. 43(), pp. 649-60.

Ortony, A., Collins, A., and Clore, G.L. (1988). *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.

Ostrom, E. (1998). 'A behavioral approach to the rational choice theory of collective action: presidential address American Political Science Association, 1997', *American Political Science Review*, vol. 92(), pp. 1-22.

Preston, S.D. and de Waal, F.B.M. (2002). 'Empathy: its ultimate and proximate bases', *Behavioral and Brain Sciences*, vol. 25(), 1-72.

Rabin, M. (1993). 'Incorporating fairness into game theory and economics', *American Economic Review*, vol. 83(), pp. 1281-1302.

Reuben, E. and van Winden, F. (2008). 'Social ties and coordination on negative reciprocity: the role of affect', *Journal of Public Economics*, vol. 92(), 34-53.

Robinson, M.D. and Clore, G.L. (2002). 'Belief and feeling: evidence for an accessibility model of emotional self-report', *Psychological Bulletin*, vol.

128(), pp. 934-60.

Smith, V. and Walker, J. (1993). 'Monetary rewards and decision cost in experimental economics', *Economic Inquiry*, vol. 31(), pp. 245-61.

Takahashi, H., Yahata, N., Koeda, M., Matsuda, T., Asai, K., and Okubo, Y. (2004). 'Brain activation associated with evaluative processes of guilt and embarrassment: an fMRI study', *NeuroImage*, vol. 23(), 967-74.

Tangney, J.P. and Dearing, R.L. (2002). *Shame and Guilt*. New York, NY: The Guilford Press.

Tangney, J.P., Miller, R., Flicker, L., and Barlow, D.H. (1996). 'Are shame, guilt and embarrassment distinct emotions?', *Journal of Personality and Social Psychology*, vol. 70, pp. 1256-69.

Thaler, R.H. (2000). 'From homo economicus to homo sapiens', *Journal of Economic Perspectives*, vol. 14, pp. 133-41.

Zeelenberg, M. and Pieters, R. (2006). 'Feeling is for doing: a pragmatic approach to the study of emotions in economic behavior', in (D. de Cremer, M. Zeelenberg, and K. Murnighan, eds.), *Social Psychology and Economics*. Erlbaum, Mahwah.

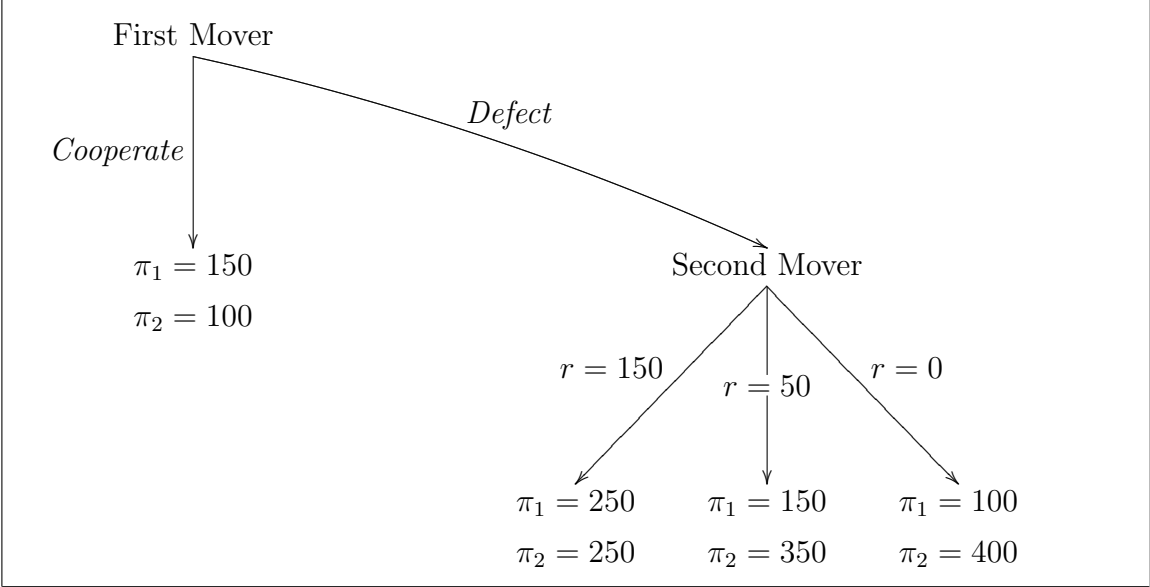


Fig. 1. Game tree without punishment opportunities



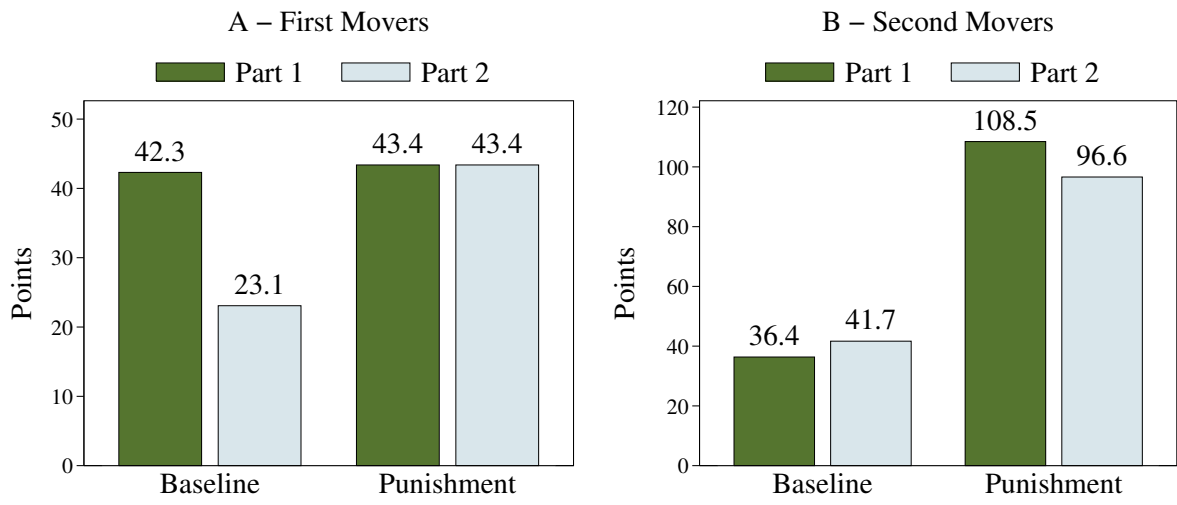


Fig. 2. *Points sent by first movers and returned by second movers*

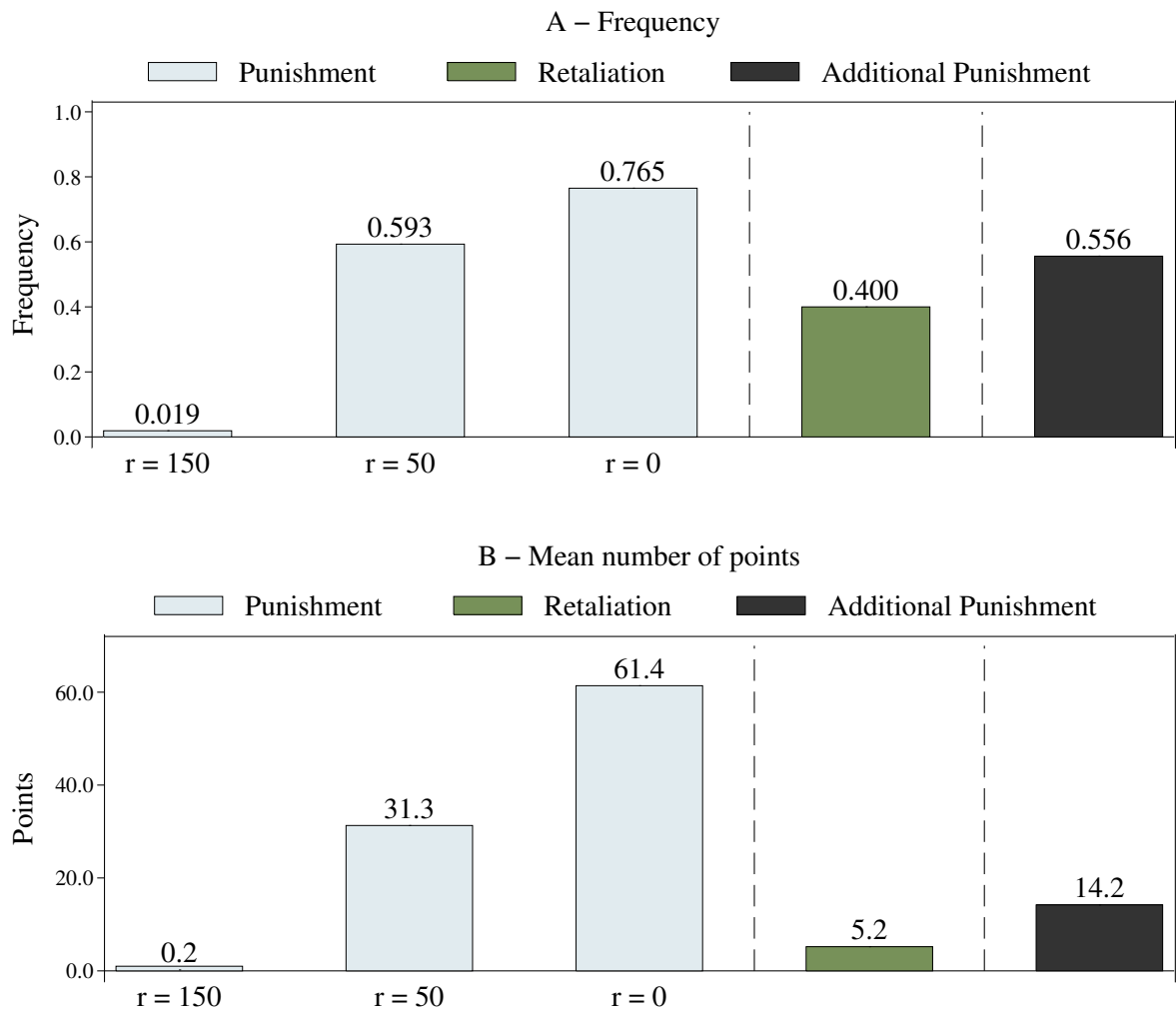


Fig. 3. *Punishment and retaliation*

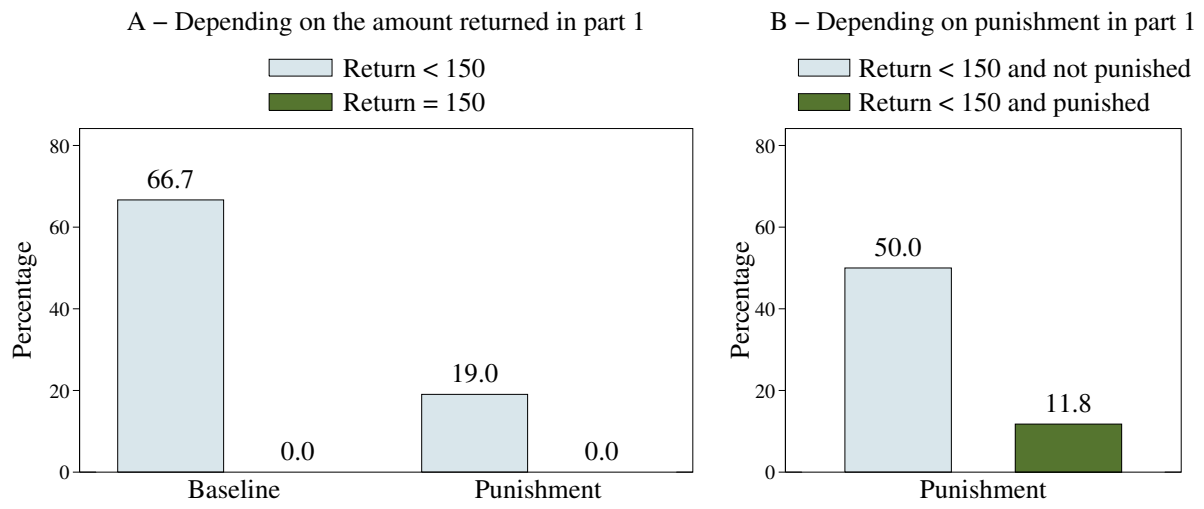


Fig. 4. Percentage of first movers who defect in part 2 depending on outcomes in part 1

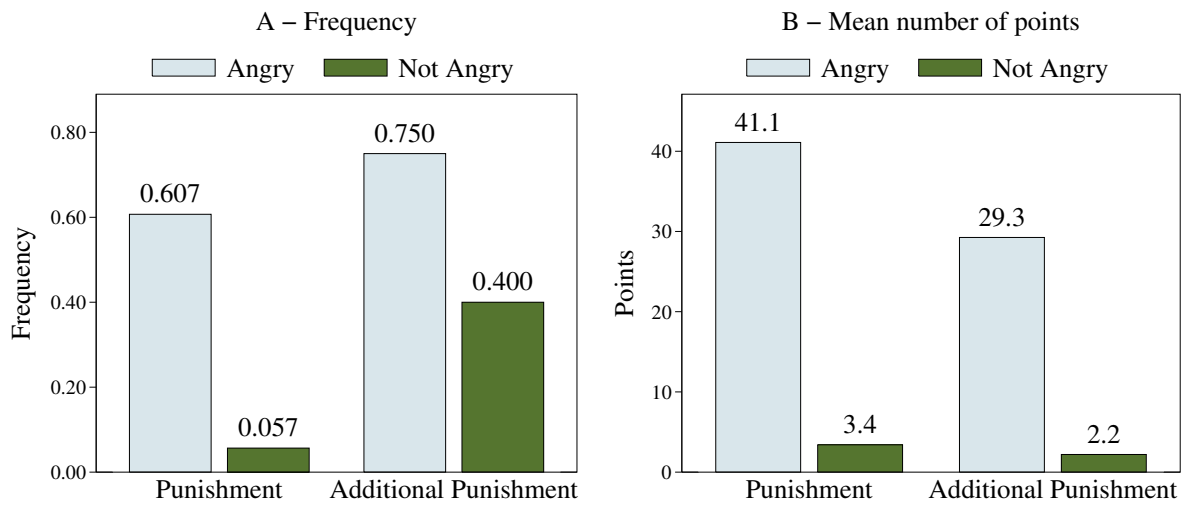


Fig. 5. *Punishment depending on anger*

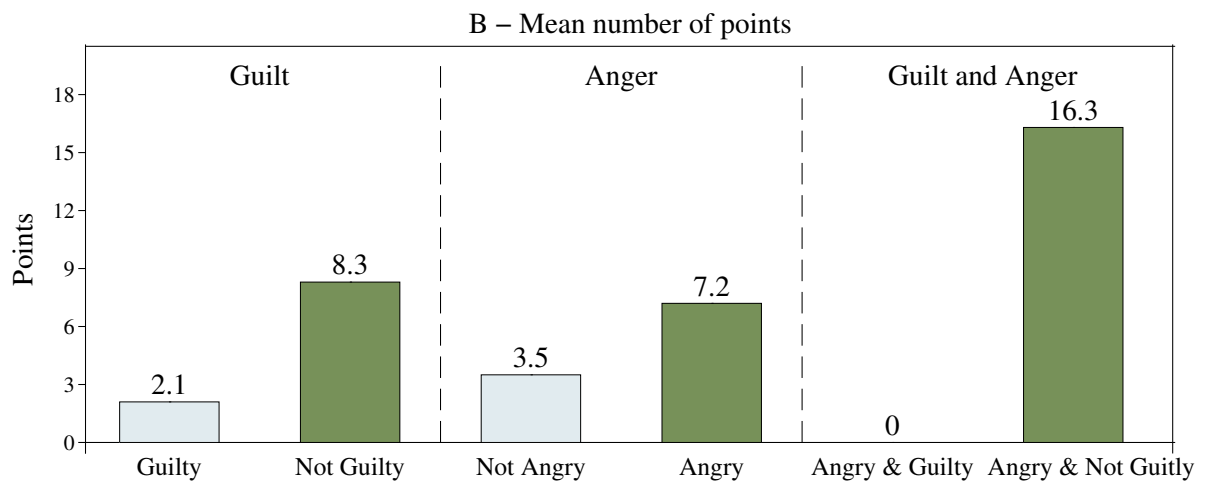
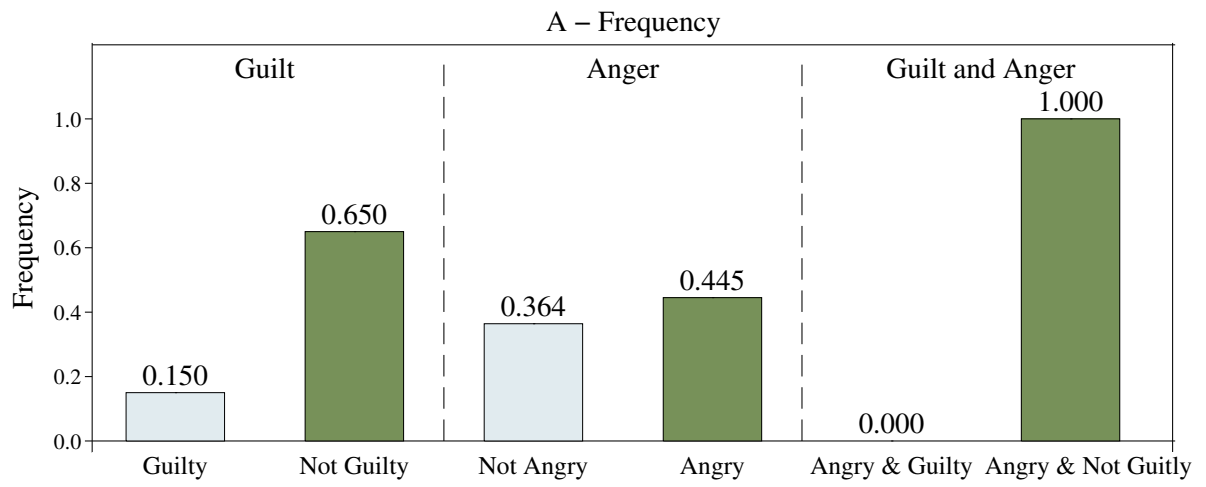


Fig. 6. *Retaliation depending on guilt and anger*

Table 1

*Factors underlying the subjects' emotional response*

Factors	First Movers				Second Movers			
	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	1 - h <sup>2</sup>	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	1 - h <sup>2</sup>
Explained variance	0.83	0.18	0.13	-	0.65	0.34	0.15	-
	Factor loadings							
Anger	0.85	0.18	0.07	0.23	0.89	0.23	-0.09	0.15
Irritation	0.90	0.12	0.06	0.17	0.86	0.18	-0.13	0.21
Happiness	-0.84	-0.10	0.28	0.21	-0.57	0.05	0.49	0.43
Gratitude	-0.67	-0.04	0.33	0.44	-0.22	0.25	0.51	0.63
Shame	0.33	0.54	0.04	0.59	0.27	0.72	0.08	0.40
Guilt	0.18	0.53	-0.04	0.69	0.23	0.71	0.07	0.43
Surprise	-0.07	0.03	0.50	0.74	0.57	0.24	0.09	0.61
	KMO test = 0.78				KMO test = 0.74			

Table B1

*Summary of the behavioural data in the punishment treatment*

	Part 1	Part 2	Both
Cooperation by first movers			
Number of observations	68	68	68
Frequency of cooperation	86.4	86.4	86.4
Mean points sent (cooperation)	43.4	43.4	43.4
(standard deviation)	(17.1)	(17.1)	(14.7)
Returns by second movers			
Number of observations	59	59	66
Frequency of returning 150	0.644	0.559	0.614
Frequency of returning 50	0.237	0.254	0.227
Frequency of returning 0	0.119	0.186	0.159
Mean points returned	108.5	96.6	103.4
(standard deviation)	(58.1)	(62.9)	(57.5)
Punishment by first movers			
Number of observations	59	59	63
Frequency of punishment	0.305	0.254	0.278
Mean punishment points	17.3	18.7	18.1
(standard deviation)	(31.4)	(35.5)	(26.2)
Retaliation by second movers			
Number of observations	16	9	20
Frequency of retaliation	0.375	0.444	0.400
Mean retaliation points	5.5	5.9	5.2
(standard deviation)	(8.7)	(10.0)	(8.2)
Additional punishment by first movers			
Number of observations	5	4	9
Frequency of additional punishment	0.600	0.500	0.556
Mean additional punishment points	6.2	24.3	14.2
(standard deviation)	(8.8)	(28.0)	(20.6)

Table B2

*Summary of the behavioural data in the baseline treatment*

	Part 1	Part 2	Both
Cooperation by first movers			
Number of observations	13	13	13
Frequency of cooperation	84.6	46.2	65.4
Mean points sent (cooperation)	42.3	23.1	32.7
(standard deviation)	(18.8)	(25.9)	(15.8)
Returns by second movers			
Number of observations	11	6	12
Frequency of returning 150	0.182	0.167	0.167
Frequency of returning 50	0.182	0.333	0.208
Frequency of returning 0	0.636	0.500	0.625
Mean points returned	36.4	41.7	35.4
(standard deviation)	(59.5)	(58.5)	(56.9)



Table B3

*Emotional reaction of first movers to returns (punishment treatment)*

Mean Intensity	Got back 150	Got back 50	Got back 0
Anger	1.1	4.5	5.8
(standard deviation)	(0.5)	(1.9)	(1.5)
Irritation	1.2	5.0	6.1
(standard deviation)	(0.7)	(1.5)	(1.5)
Happiness	6.1	2.3	1.8
(standard deviation)	(1.0)	(1.4)	(1.1)
Gratitude	4.9	2.4	1.6
(standard deviation)	(1.8)	(1.7)	(1.1)
Shame	1.2	1.9	2.9
(standard deviation)	(0.5)	(1.6)	(2.3)
Guilt	1.1	1.3	1.8
(standard deviation)	(0.5)	(0.9)	(1.7)
Surprise	4.2	3.9	4.5
(standard deviation)	(1.6)	(1.7)	(2.5)
Number of observations	53	27	17

Table B4

*Emotional reaction of first movers to received retaliation*

Mean Intensity	No Retaliation	Positive Retaliation
Anger	1.9	3.6
(standard deviation)	(1.5)	(2.2)
Irritation	2.2	4.7
(standard deviation)	(1.7)	(2.2)
Happiness	3.4	2.6
(standard deviation)	(1.8)	(1.3)
Gratitude	2.4	2.7
(standard deviation)	(2.0)	(1.9)
Shame	2.1	1.5
(standard deviation)	(1.3)	(0.9)
Guilt	2.1	1.5
(standard deviation)	(1.9)	(1.1)
Surprise	4.8	2.3
(standard deviation)	(1.9)	(1.6)
Number of observations	14	10

Table B5

*Emotional reaction of second movers to received punishment*

Mean Intensity	No Punishment	Positive Punishment
Anger	1.1	3.7
(standard deviation)	(0.8)	(1.9)
Irritation	1.3	4.1
(standard deviation)	(1.2)	(2.3)
Happiness	5.0	2.0
(standard deviation)	(1.6)	(1.2)
Gratitude	4.0	2.5
(standard deviation)	(2.0)	(1.5)
Shame	1.2	1.5
(standard deviation)	(0.9)	(0.9)
Guilt	1.4	1.9
(standard deviation)	(1.1)	(1.3)
Surprise	2.5	4.6
(standard deviation)	(1.9)	(2.1)
Number of observations	55	25

Table C1

*Ordered probit model estimating first movers' anger*

---

Variable	Coefficient	Std. Error	<i>p</i> -value
Return = 50	2.648	0.337	0.000
Return = 0	3.352	0.438	0.000
Expected return = 50	-0.368	0.338	0.276
Expected return = 0	-0.891	0.473	0.059
Fairness of returning 50	-0.226	0.115	0.049
Economist	-0.043	0.302	0.888
Female	-0.322	0.290	0.267
Number of obs. = 118		$\chi^2(7) = 111.03$	
Log likelihood = -96.765		Prob > $\chi^2 = 0.000$	

---

*Note:* Variables: 'Return =  $x$ ' and 'Expected return =  $x$ ' = 1 if the (expected) return was  $x$ , 0 otherwise; Economist = 1 if economics mayor, 0 otherwise; Female = 1 if female, 0 if male; 'Fairness of returning 50' ranges from 1 = 'very unfair' to 7 = 'very fair'.

Table C2

*Ordered probit model estimating second movers' anger*

Variable	Coefficient	Std. Error	<i>p</i> -value
Punishment if return = 150	0.208	0.074	0.005
Punishment if return = 50	0.028	0.010	0.004
Punishment if return = 0	0.021	0.008	0.014
Economist	-0.107	0.411	0.794
Female	0.735	0.622	0.237
Number of obs. = 33		$\chi^2(5) = 14.18$	
Log likelihood = -58.228		Prob > $\chi^2 = 0.015$	

*Note:* Variables: 'Punishment if return =  $x$ ' equals amount of punishment if the return was  $x$ , 0 otherwise. For other variables see Table C1.