The Demand for Mutual Monitoring: Theory and Evidence*

Timo Goeschl
Department of Economics
University of Heidelberg

Johannes Jarke†
Department of Economics
University of Heidelberg

June 6, 2011

Abstract

In this paper we study monitoring behavior, punishment behavior, and their interaction in a simple exchange experiment. We assess whether and to what extent subjects will monitor, that is, engage in costly information acquisition prior to potential punishment in order to be able to condition damage inflicted on the target player's behavior. We do so by manipulating the monitoring costs both in a second party and a third party monitoring condition. This allows a bifocal investigation of (1) responses to changes in those costs and (2) potential differences in behavior between second and third parties. We find that some individuals withhold punishment altogether when monitoring gets costly while others switch to untargeted punishment with the result that the demand for monitoring information obeys the law of demand. Controlling for risk preferences, third parties are more likely to go for the risk of erroneous punishment. Overall, average net costs inflicted on defectors diminish considerably as monitoring gets costly.

JEL-Classification: D02, C92, C72, D03, D80
Keywords: experiment, exchange, Prisoners' Dilemma, sanction, punishment, monitoring, strategy method, social norms

*We are grateful to the German Federal Ministry of Education and Research and the University of Heidelberg for financial support. Valuable inputs by Catherine Eckel, Marco Faravelli, Ernst Fehr, Urs Fischbacher, Alia Gizatulina, Freddy Huet, Cornelia Neuert and Rick Wilson, as well as various participants at the 2nd Thurgau Experimental Economics Meeting, the 2011 Annual International Meeting of the Economic Science Association, the International Association for Research in Economic Psychology 2011 Conference, the 65th European Meeting of the Econometric Society and seminars at the Universities of Mannheim and Heidelberg are appreciated. All remaining errors are ours.

†Corresponding author. Contact: Bergheimer Str. 20, D-69115 Heidelberg, Germany. Phone: +49 6221 54 8015. Fax: +49 6221 54 8020 . E-mail: jarke@eco.uni-heidelberg.de.

1 Introduction

In this paper we study monitoring behavior, punishment behavior, and their interaction in a simple exchange experiment. The purpose is the assessment whether and to what extent subjects will monitor, that is, engage in costly information acquisition prior to potential punishment, in order to be able to condition damage inflicted on the target player's behavior. In addition, we are interested in identifying possible interactions between monitoring and punishment behavior. We do so by manipulating the monitoring costs both in a second party and a third party monitoring condition. This allows a bifocal investigation of (1) responses to changes in those costs and (2) potential differences in behavior between second and third parties. This investigation contributes to the understanding of spontaneous cooperation in informal exchanges and the enforcement mechanisms underlying social norms.

Specifically, we conducted a laboratory experiment inspired by a design used in Fehr & Fischbacher (2004b). Subjects play a two stage one-shot game. In the first stage they are matched into dyads to interact in a simple exchange exhibiting a Prisoners' Dilemma structure in pecuniary payoffs. Then both players in each group are given the opportunity to monitor and punish each others' first stage behavior in a second party monitoring condition, while in a third party monitoring condition the first (second) player in a group could only monitor and punish a player from another (a third) group. We implemented the monitoring and punishment stage with a modified version of the strategy method (Selten 1967). If the subject wanted to condition punishment on the target player's first stage behavior, (s)he had to incur a monitoring fee in addition to eventual punishment costs, while there was no fee if damage was inflicted unconditionally (or not at all). Meting out conditional punishment is equivalent to purchasing the information on whether the target player cooperated or defected beforehand. This design isolates the effect of changing monitoring costs in a particularly clean way by ensuring that any subject in any treatment did exactly the same moves on exactly the same screens instructed by exactly the same instructions. Furthermore, we supplemented the core experiment by performing incentivized elicitations of beliefs and risk attitudes in order to increase its explanatory power.

The principle questions we want to investigate with this experiment are threefold. Do subjects incur personal costs to acquire monitoring information in singular interactions? How do subjects respond to changes in those costs? How does monitoring and punishment behavior interact? Our results demonstrate that monitoring information appears to obey the *law of demand*, that is, its quantity demanded increases as its price decreases and vice versa. This effect can be decomposed into two distinct effects that mediate the interaction between monitoring and punishment behavior: some individuals withhold punishment altogether when monitoring costs increase while others switch to untargeted punishment. Interestingly, the second effect is stronger for third than for second parties, indi-

cating that third parties are more likely to go for the risk of erroneous punishment. Overall, average net costs inflicted on defectors diminish considerably as monitoring gets costly. Hence, our results suggest that the effect of (exogenous) increases or decreases of mutual monitoring costs can be predicted by standard microeconomic arguments provided that social preferences are taken into account.

We contribute to several streams in the literature. First of all, we contribute to the literature studying informal sanctions with an institutional focus (see Yamagishi; Ostrom et al.; Fehr & Gächter for seminal contributions. This literature is vast and no attempt is made to review it here. Instead the reader is referred to a recent survey by Jarke 2011). A well established stylized fact from this literature is that the extension of the players' strategy set in social dilemmata by opportunities for direct social sanctions drastically changes both play in one-shot games as well as dynamics in iterated interactions. Great progress has been made isolating proximate psychological mechanisms shaping social sanctions. From an institutional perspective, however, many questions remain unanswered. An obvious but important question to ask is, for example, to what extent the results are robust to more realistic information conditions? What the majority of previous research has shown is that the existence of *targeted* punishment options induced the observed changes in behavior and the increase of cooperation over time, while cooperation unraveled in the absence of those options. Perfect targeting, however, was only possible in those experiments by essentially equipping subjects with a perfect monitoring technology by design (that is, they were informed by the experimenter about their coplayers' behavior).

In fact, recent research indicates that punishment may fail to sustain cooperation if the players' capacities of monitoring their coplayers' characteristics (Bornstein & Weisel 2010), actions (Carpenter 2007b; Grechenig et al. 2010; Ambrus & Greiner 2010), or both (Patel et al. 2010) are imperfect by design. In those cases penalties inflicted are much less targeted on defectors and consequently cooperation breaks down. In social interactions outside the laboratory, however, individuals are rarely stuck with information of exogenously given quality. Rather, it is a *choice* how much costly effort to put into monitoring others' behavior. Anecdotal evidence suggests that people can and often do actively invest in the acquisition of more accurate information and thereby endogenously mitigate informational asymmetries due to absent or noisy signals on their coplayers' behavior. The relevant question, therefore, is not whether individuals *lack* information on characteristics and/or behavior of their coplayers to one degree or another but what is the cost of *acquiring* such information. It can be expected that monitoring, punishment, and cooperative behavior interact in interesting ways. This paper is a first step towards an understanding of this interaction. In this respect, our research is closest to the studies on monitoring imperfections in peer punishment experiments, to which has been referred to above, as well as the literature on mutual monitoring in teams and exchange networks (e.g. Pitariu 2007; Grosse et al. 2008; Carpenter et al. 2009, 2010). In the former, monitoring

imperfections are exogenous while they are endogenous in ours. The latter usually considers mutual monitoring without separating the act of monitoring from the act of punishment (see also Carpenter et al. 2004 for this in a social norm enforcement context), while our paper provides evidence that this generally not appropriate.[1]

Secondly, we contribute to the experimental literature studying punishment with a behavioral focus by adding to the understanding of proximate psychological mechanisms underlying costly punishment behavior (e.g. Falk et al. 2005; Nikiforakis & Normann 2008; Carpenter & Matthews 2009, see again Jarke 2011 for a review). It is recognized that a strong reciprocity preference (Gintis 2000; Fehr et al. 2002) is the dominant motive underlying costly punishment behavior (Fehr & Fischbacher 2005). In particular, a strong reciprocator has a preference for *instrumental* punishment, that is, for a reduction of a defector's payoff relative to that of a cooperator (in the strict sense, a strong reciprocator should never decrease the payoff of a cooperator). On the other hand, there is experimental evidence for significant punishment of cooperative behavior (Cinyabuguma et al. 2006; Gächter et al. 2006; Ertan et al. 2009). Spite may be motive to explain this. However, consistent with neuroscientific evidence indicating that subjects enjoy to punish *per se* due to brain-internal rewards (de Quervain et al. 2004), Casari & Luini (2009, p. 274) speculated that "the punisher derives her utility from the act of punishment in itself and not from achieving (...) a total amount of punishment" when they discovered in their experiment that subjects seem to condition their sanction not onto the punishment already levied to the same person by other subjects. We call such behavior *hedonic* punishment. We suggest that both motives are present in individuals and individual heterogeneity exist in their relative strength. Specifically, the strong reciprocity motive implies a preference for punishing conditionally, that is, to target penalties on defectors or, conversely, wishing to avoid erroneous punishment of cooperation. Hence, this motive implies some willingness to invest additional resources in monitoring. Conversely, the willingness to invest in monitoring should be the lower the stronger the hedonic (or spite) motive, according to this motive it should matter less whether the target player is a defector or a cooperator.[2] More precisely, the demand for monitoring should decrease the faster in its price, the stronger hedonic motives are, or analogously, should decrease slower the stronger the reciprocity motives.

Finally, by studying monitoring and sanctioning behavior by third parties we contribute to the literature on social norms. We refer to Hechter & Opp (2001) for an overview and to Fehr & Fischbacher (2004b) for an experimental design we draw on. To our knowledge, monitoring has not been studied as an activity distinct from the imposition of sanctions generally, and in the enforcement of

[1] Our paper is also related to the studies by Anderson & Putterman (2006) and Carpenter (2007a) who examine the demand for punishment by varying the fine-to-fee ratio (under a perfect monitoring technology).

[2] We define these notions and arguments precisely in section 2.2.

social norms in particular.

The setup studied in this paper also has implications for (and is motivated by) organizational design as well. Indexing information on social behaviors and displaying it together with the identity of the individual emitting it in a (more or less) publicly accessible location or register have been used for centuries in human history. The aim of such positive (*white*) and negative (*black*) lists is to disclose information publicly in order facilitate social reinforcement or punishment. For example, positive lists are frequently used by charity organizations and internet communities, pursuing the goal to induce positive social sanctions, that is, social approval in order to reinforce donors and casually creating new imitators (Holländer 1990; Glazer & Konrad 1996; Harbaugh 1998a,b; Soetevent 2005).[3] Perhaps even older and widespread are negative lists and public exposure with the goal of inducing negative social sanctions, either direct or indirect.[4] However, even when it is physically feasible to access those locations or registers, it is typically costly (in terms of time, effort or material costs) to do so. From an economic viewpoint, then, making such information easier accessible means nothing else than making its acquisition less costly.

We proceed as follows. The experimental design and procedures are described in section 2. In section 2.2 we develop a simple theoretical model and use it to derive the main hypotheses to be tested. The results are presented in section 3. We conclude in section 4.

[3]Indeed, revealing contributions and identities have been shown to increase contribution levels in experimental public good games (Sell & Wilson 1991; Croson & Marks 1998; Andreoni & Petrie 2004; Rege & Telle 2004; Anderson & Stafford 2009). This is still true if the acquisition of this information is costly, even though the identities of contributors are viewed seldomly (in less than ten percent of the time) in this case (Savikhin & Sheremeta 2010).

[4]Stocks, pillories and prangers have been used for public humiliation as punishment from medieval to early modern times. They were set up to hold petty criminals in marketplaces, crossroads, and other public places and were typically placed on platforms to increase public visibility of the offender (Pettifer 1992; Kellaway 2003). Examples for modern negative lists are for example the Brazilian *lista negra*, an inventory listing corporations and estate owners tolerating unpaid labor within their businesses, which is published annually by all brazilian newspapers, or the EU air carrier blacklist (based on Regulation EC 2111/2005 of the European Parliament and Council) which lists airlines found to violate safety standards and is published quarterly by the European Commission online. In the United States, the *National Sex Offender Registry* is a website coordinated by the US Department of Justice that enables every citizen to access information on the identity and location of known sex offenders, Furthermore, so-called *creative sentencing*, which includes public shaming as a form of punishment, is increasingly used by US courts as a response to excessive prison costs (Brook 1999; Kahan & Posner 1999; Owens 2000). Inventories that may be considered as combinations of positive and negative lists are pollution registers, such as the *Toxic Release Inventory* (TRI) in the United States or the *Pollutant Release and Transfer Registers* (PRTRs) established in many countries around the world.

2 Method

2.1 Design

The experiment involves six treatments in a two-by-three design combining crossover and between-subjects variation. In the first dimension, any subject played both a second and a third party monitoring game. Both games had two stages. In the first stage both players in a given group were endowed with 10 tokens and interacted with one another in an exchange exhibiting a Prisoners' Dilemma structure in monetary payoffs: each player could either keep her or his tokens or transfer all of them to the other group member, in which case the experimenter tripled them. The second stage differs between the second and the third party monitoring conditions.

At the beginning of the second stage, each player received an additional endowment of 40 tokens. Then both players in each group had the opportunity to monitor and punish each other in the second party monitoring condition, while in the third party monitoring condition the first (second) player in a group could only monitor and punish a player from another (a third) group. This matching protocol, invented by Fehr & Fischbacher (2004b), rules out any reciprocity between punishers. The punishment technology used was neutral in the sense of a constant penalty-to-fee ratio: the cost for one punishment point was one token for the punishing individual and three tokens for the target player.[5] Punishment was limited to twenty points.

We implemented the monitoring and punishment stage with a modified version of the strategy method (Selten 1967). Following the strategy method, subjects first state contingent choices for every decision node they may face before the appropriate choices are carried out for the nodes that are reached while the other contingent choices are ignored.[6] Specifically, in the second party monitoring condition each player had to indicate the number of deduction points for each of the other group member's possible transfer decisions before knowing the actually realized one. In the third party monitoring condition, each player, while being informed about the transfer decision of the other member in *their* group, had to indicate the number of deduction points to be assigned to a member of a different group for each possible transfer combinations in that group without knowing the actually realized combination. We added the following feature: If the subject wanted to condition punishment on the transfer decision of the target player, (s)he had to incur a monitoring fee κ_m in addition to eventual punishment costs, while there was no fee for unconditional punishment. Carrying out conditional punishment is equivalent to buying the information whether the target player

[5] Casari (2005) has shown that non-neutral punishment technologies, as adopted for example in Fehr & Gächter (2000), can significantly bias the results.

[6] This contrasts with the direct response method in which subjects make a single choice only for those nodes that are actually reached in the course of play.

transferred or not at a fee κ_m beforehand.

In summary, the payoff function of each subject i in the second party monitoring condition is

$$w_i = 10\,(1 - a_i + 3a_j) + 40 - p_{ij} - 3p_{ji} - m_i\kappa_m$$

where $a_i, a_j \in \{0, 1\}$ are the transfer decisions, p_{ij} is the amount of punishment points i inflicts on j, p_{ji} is the amount of punishment points i receives from j, and $m_i \in \{0, 1\}$ is the monitoring decision. In the third party monitoring condition the payoff function of each subject i is

$$w_i = 10\,(1 - a_i + 3a_j) + 40 - p_{ik} - 3p_{li} - m_i\kappa_m$$

where p_{ik} is the amount of punishment points i inflicts on the target player in group B, k, and p_{ij} is the amount of punishment points i receives from l, the player in group C that is eligible to punish i. We implemented variation between the two conditions using a crossover procedure (see the following subsection). In the second dimension, between-subjects variation was introduced by manipulating the monitoring fee, $\kappa_m = \{0, 5, 10\}$. We used three fee levels since we had strong priors for a non-linear (specifically, a convex) effect, which turned out to be the case. In order to achieve efficient identification, we set the values at the extremes and the midpoint of the range and allocated approximately half of the subjects to the midpoint cell, a third to maximum cell, and a sixth to the minimum cell, respectively (see for example McClelland 1997 for efficient sample arrangement methods). In what follows we label the three cells $M0$, $M5$ and $M10$, respectively.

In summary, this design allows us to address the two focal questions stated in the introduction in that we can study responses to changes in monitoring costs (between subjects) and possible differences in second and third party behavior (within subjects).

Elicitation of beliefs

After subjects made their decisions in the first and second stage, respectively, we asked them to state their beliefs on the transfer decision of the target player as well as the punishment they expect from the other player that is allowed to punish them.

Those statements were incentivized with a opportunity to earn extra tokens in case of good predictions. To elicit the subjects' beliefs at the transfer stage, we use an affine transformation of the one-dimensional Brier score (Brier 1950). Let $\tilde{a}_{ij} \in [0, 1]$ subject i's probability prediction that the target player j transferred her or his endowment, then the Brier score is given by $S_{\text{Brier}} = (\tilde{a}_{ij} - a_j)^2 \in [0, 1]$. We implemented the affine transformation

$$S = 8 - 8S_{\text{Brier}} = 8 - 8\,(\tilde{a}_{ij} - a_j)^2 \in [0, 8]$$

such that a perfect prediction was rewarded by eight tokens and the opposite by zero tokens.[7] This scoring rule apparently belongs to the class of quadratic score functions which are known to be proper (see for example Selten 1998 and Gneiting & Raftery 2007). Physically, we implemented the belief elicitation at the first stage by a screen with a slider with which subjects could specify their probabilistic belief that the target player transferred her or his endowment in ten-percent steps.

At the punishment stage we asked the subjects for point predictions of the deductions assigned to them by their coplayers and incentivized statements with a distance score similar to those used, for example, by Gächter & Renner (2010) or Fischbacher & Gächter (2010). Here, each perfect prediction was rewarded by four tokens, a deviation of one point by two tokens, a deviation by two points by one token, and deviations of three of more points received no reward. This elicitation was physically implemented by a screen with input boxes shown at the end of the second stage.

Elicitation of risk preferences

At the end of each session we used the Holt-Laury lottery choice method (Holt & Laury 2002) to obtain an indication of each subject's risk attitude. Subjects faced a set of choices between two lotteries for each of ten decisions. The payoffs were identical for each lottery A (20 tokens or 16 tokens) and each lottery B (38.5 tokens and 1 token).[8] Probabilities for high and low payoffs are the same for both alternatives for each decision. Thus lottery B always has higher variance. As the subject moves down the ten decisions, the probability gradually shifts from the lower to the higher payoff. The expected return is higher for lottery A for the first four decisions, and for lottery B after that. We were interested in the point where subjects switch from the more risky option (lottery B) to the less risky option (lottery A) and how often they switch if applicable. A risk neutral agent is expected to switch in line three or four; the further down the switching point the higher the agent's degree of risk aversion. Physically, subjects made their choices on a screen with two check boxes for option A and B, respectively, for any of the ten decisions that were arranged row-wise on the screen. The lotteries had been resolved by throwing a ten-sided die, simulated by a random number generator program. One of the ten choices was actually paid out. Which one was determined by a second virtual ten-sided die.[9]

[7] We restricted $\tilde{a}_{ij}$ to be in the set of one-digit decimals on the unit interval.

[8] These payoffs correspond to the payoffs in the low-stakes condition in Holt & Laury 2002.

[9] Laury (2005) provides evidence that supports the validity of using this random-choice payment method.

2.2 Model and predictions

We develop an simple model in order to derive our main hypotheses formally. Since or focus is on monitoring and punishment behavior, our model begins with the second stage of our experiment, that is, the first stage actions are exogenous to the model.[10] Consider a player i in the position to monitor a target player (j). Player j's first stage action is denoted by $a_j \in \{0,1\}$. First, player i decides whether to monitor or not, $m_{ij} \in \{0,1\}$, where $m_{ij} = 1$ denotes monitoring. Monitoring costs $\kappa_m \in \mathbb{R}_+$. Player i does not receive any signal on the realization of a_j before his move such that it is a Bernoulli random variable from his perspective. Assume that player i is a subjective expected utility maximizer and holds (first order) prior belief $\dot{a}_j^i \in [0,1] \subset \mathbb{R}$ that $a_j = 1$. In line with the literature on subjective games (Kalai & Lehrer 1993, 1995; Oechssler & Schipper 2003) we do not impose any restrictions on the players' prior beliefs. After choosing $m_{ij} \in \{0,1\}$, i obtains a signal $\sigma_{ij}(m_{ij}) : \{0,1\} \to \{\varnothing, a_j\}$, where $\sigma_{ij}(0) = \varnothing$ and $\sigma_{ij}(1) = a_j$, and updates his belief Bayesian, such that his posterior belief (after receiving the monitoring signal) is given by

$$\dot{a}_j^{i'} = \begin{cases} a_j & m_{ij} = 1 \\ \dot{a}_j^i & m_{ij} = 0 \end{cases}$$

After receiving $\sigma_{ij}(m_{ij})$, i decides whether to inflict damage $p_{ij} \in \mathbb{R}_+$ on player j at constant marginal costs $\kappa_p \in \mathbb{R}_{++}$.

A terminal history $z \in \{0,1\}^4 \times \mathbb{R}_+^2$ consists of a sequence of transfer indicators (a_i, a_j), a sequence of monitoring indicators (m_{ij}, m_{ji}), and a sequence of punishment levels (p_{ij}, p_{ji}). Assume that the preferences on the set of terminal histories held by each individual in the population can be described by the following (quasi-linear) utility function,

$$u_i(z) = \underbrace{w_i^1 - p_{ji} - \kappa_p p_{ij} - \kappa_m m_{ij}}_{\text{pecuniary payoff}} + \underbrace{(a_i(1-a_j)\alpha_i - \theta_i)\chi^{-1} p_{ij}^{\chi}}_{\text{psychological reward}}$$

where $\alpha_i \in \mathbb{R}_+$ is an individual parameter capturing i's anger towards defectors, $\theta_i \in \mathbb{R}$ is an individual parameter capturing i's residual benevolence $(\theta_i > 0)$ or malevolence $(\theta_i < 0)$, and $\chi \in]0,1[\subset \mathbb{R}$ is a convexity parameter which is assumed to be equal for all individuals; for simplicity we assume $\chi = \frac{1}{2}$.[11] As indicated, the

[10] In doing so, we maintain the implicit assumption commonly made in the (experimental) literature on direct punishment that the expected punishment inflicted on the sanctioning player under consideration does not matter for his punishment decision. Note, however, that this assumption may be violated if players are motivated by inequity aversion, for example.

[11] The convexity parameter implies a decreasing marginal willingness to pay for decreasing the target's payoff. This assumption is made to allow for interior optima, but is not crucial for any of the results. The scaling factor χ^{-1} is included for arithmetical convenience and can be readily dropped.

former component is the pecuniary payoff accruing to i along terminal history z, whereas the latter component is i's psychological reward from decreasing j's payoff by p_{ij} units, given any previous history.

This utility function is compatible with the ideas behind a variety of recently proposed preference models, such as inequity aversion (Fehr & Schmidt 1999, Bolton & Ockenfels 2000) as well as proximate (Dufwenberg & Kirchsteiger 2004, Falk & Fischbacher 2006, Cox et al. 2007) and evolutionary (Gintis 2000) theories of strong reciprocity.[12] Though, we emphasize that it is not the purpose of this paper to explicitly test any of those theories.

Punishment

Denote by $p_{ij}^{\star}\left(a_i, \dot{a}_j^{i'}\right)$ the preferred punishment level by i given his own action, a_i, and his posterior belief about the target player's action, $\dot{a}_j^{i'}$. Any interior solution is defined by the first order condition

$$-\kappa_p + \left(a_i\left(1-\dot{a}_j^{i'}\right)\alpha_i - \theta_i\right) p_{ij}^{-\frac{1}{2}} = 0$$

such that

$$p_{ij}^{\star}\left(a_i, \dot{a}_j^{i'}\right) = \begin{cases} 0 & \theta_i \geq a_i\alpha_i\left(1-\dot{a}_j^{i'}\right) \\ \left(\frac{\theta_i + a_i\alpha_i\left(\dot{a}_j^{i'}-1\right)}{\kappa_p}\right)^2 & \theta_i < a_i\alpha_i\left(1-\dot{a}_j^{i'}\right) \end{cases}$$

It is straightforward to show that the level of punishment is decreasing in θ_i , $\dot{a}_j^{i'}$ and κ_p, and increasing in α_i. Thus, the model is consistent with and predicts a variety of stilized regularities from experimental research.

First, behavioral data from a variety of experimental games fits a model in which individuals have both an unconditional (residual) social motivation towards individuals about whom they know nothing and a reciprocal or normative motivation which is conditional on behavior. For example, individuals with a motive to behave generously towards strangers, but willing to reduce the payoffs of an individual who reveal a bad type even at a cost to himself when his reciprocal preferences outweigh his unconditional benevolence are frequent, and have been dubbed *strong reciprocators* (Gintis 2000; Fehr et al. 2002; Carpenter et al. 2009). Using structural modelling techniques, Cox et al. (2007) showed how behavioral data from a variety of experimental games can be neatly decomposed into such a residual social motive and an affective reciprocity component.

Second, it has been shown that roughly one third of a given population never metes out any punishment (Fischbacher et al. 2001; Fischbacher & Gächter 2010). This is predicted by our model either if the player is a purely individualistic pecuniary payoff maximizer ($\theta_i = \alpha_i = 0$) or if anger towards defectors is too weak to overcome the player's residual benevolence ($\theta_i > \alpha_i > 0$, see below). The

[12] See Carpenter et al. (2009) for a similar specification.

latter is consistent with the cross-cultural evidence provided by Henrich et al. (2006, Science) showing that costly punishment positively covaries with altruistic behavior across populations.

Third, the model predicts that both defection and cooperation is punished by some individuals (Cinyabuguma et al. 2006; Gächter et al. 2006; Ertan et al. 2009, see also Fehr & Fischbacher 2004b). However, it also predicts that defection is typically punished more strongly, and this is usually driven by negatively valenced affect towards defectors (Fehr & Fischbacher 2004b,a).[13]

Fourth, punishment inflicted by defectors is usually significantly weaker than that inflicted by cooperators (Fehr & Fischbacher 2004b). This is predicted by our model as well (observe that the first term in the denominator disappears if i defected).

Fifth, punishment is predicted to be negatively price elastic, which has shown to be the case for non-linear (Suleiman 1996, see also Oosterbeek et al. 2004) and linear punishment technologies (Anderson & Putterman 2006; Carpenter 2007a).

Finally, in experiments with linear punishment technology, punishment levels are (except for those who do not punish at all) rarely at the boundaries of the feasible interval but accumulate more in the interior. The introduction of the convexity parameter χ accounts for this regularity.

To see all this more clearly, it is illuminating to distinguish a couple of cases and parameter constellations explicitly. First, if $a_i = 0$, that is, i defects, then he does not condition punishment on the target player's behavior and metes out punishment only if he is malevolent ($\theta_i < 0$),

$$p_{ij}^{\star}\left(0, \dot{a}_j^{i'}\right) = \begin{cases} 0 & \theta_i \geq 0 \\ \left(\frac{\theta_i}{\kappa_p}\right)^2 & \theta_i < 0 \end{cases}$$

Further, the punishment level is decreasing in θ_i and κ_p, and increasing in χ.

Second, if $a_i = 1$, that is, i cooperates, then

$$p_{ij}^{\star}\left(1, \dot{a}_j^{i'}\right) = \begin{cases} 0 & \theta_i \geq \left(1 - \dot{a}_j^{i'}\right)\alpha_i \\ \left(\frac{\theta_i + \alpha_i\left(\dot{a}_j^{i'} - 1\right)}{\kappa_p}\right)^2 & \theta_i < \left(1 - \dot{a}_j^{i'}\right)\alpha_i \end{cases}$$

such that punishment is meted out if and only if $\left(1 - \dot{a}_j^{i'}\right)\alpha_i > \theta_i \Leftrightarrow \dot{a}_j^{i'} < 1 - \frac{\theta_i}{\alpha_i}$, that is, i's anger about defectors and his belief that the target in fact defected are suffiently large to overcome his inhibitory residual benevolence motive. Specifically, if $\theta_i > 0$, then the individual generally dislikes doing harm to others, but if his affective preference to express disapproval to cheaters is strong enough (and he is sufficiently confident that the target in fact defected), then he inflicts punishment anyway. Needless to mention, that the above condition is always satisfied

[13] Burnham (2007) showed that punishment is positively related to high testosterone levels in men.

if $\theta_i < 0$, that is, malevolent individuals always mete out punishment (if the cost is not too high).

Apparently, since $\dot{a}_j^{i'}$ is a posterior, the actual punishment level depends on the previous monitoring decision. If $m_{ij} = 0$, then $\sigma_{ij}(0) = \emptyset$, such that $\dot{a}_j^{i'} = \dot{a}_j^i$ and everything above applies with just the posterior replaced by i's prior. On the other hand, if $m_{ij} = 1$, then $\sigma_{ij}(1) = a_j$, such that $\dot{a}_j^{i'} = a_j$. If $a_j = 1$, that is, the target player cooperated, then i only inflicts punishment only if he is malevolent, exactly like in the case where i defected (see above),

$$p_{ij}^{\star}(1,1) = \begin{cases} 0 & \theta_i \geq 0 \\ \left(\frac{\theta_i}{\kappa_p}\right)^2 & \theta_i < 0 \end{cases}$$

If $a_j = 0$, that is, the target player defected, then i only inflicts punishment only if his anger about defectors is suffiently large to overcome his inhibitory residual benevolence motive (which is of course always true if i is malevolent),

$$p_{ij}^{\star}(1,0) = \begin{cases} 0 & \theta_i \geq \alpha_i \\ \left(\frac{\theta_i - \alpha_i}{\kappa_p}\right)^2 & \theta_i < \alpha_i \end{cases}$$

Notably, i conditions punishment on the target player's behavior if and only if $\alpha_i > 0$.

Example 1. For illustration, consider the case with $\kappa_p = \frac{1}{3}$ as in our experiment. Then punishment meted out by player i on target player j is given by

$$p_{ij}^{\star}\left(0, \dot{a}_j^{i'}\right) = \begin{cases} 0 & \theta_i \geq 0 \\ 9\theta_i^2 & \theta_i < 0 \end{cases}$$

if i defected,

$$p_{ij}^{\star}\left(1, \dot{a}_j^{i'}\right) = \begin{cases} 0 & \theta_i \geq \left(1 - \dot{a}_j^{i'}\right)\alpha_i \\ 9\left(\theta_i + \alpha_i\left(\dot{a}_j^{i'} - 1\right)\right)^2 & \theta_i < \left(1 - \dot{a}_j^{i'}\right)\alpha_i \end{cases}$$

if he cooperated. Specifically, for the fully informed cases we get

$$p_{ij}^{\star}(1,1) = \begin{cases} 0 & \theta_i \geq 0 \\ 9\theta_i^2 & \theta_i < 0 \end{cases}$$

and

$$p_{ij}^{\star}(1,0) = \begin{cases} 0 & \theta_i \geq \alpha_i \\ 9\left(\theta_i - \alpha_i\right)^2 & \theta_i < \alpha_i \end{cases}$$

respectively.

Note that depending on the parameter vector (α_i, θ_i), player i can be classified as one of four possible types.

- The *cool benevolent type*, $\theta_i \geq \alpha_i \geq 0$, never inflict any punishment, independent from their belief, $p^\star_{ij}\left(1, \dot{a}^{i'}_j\right) = 0 \ \forall \dot{a}^{i'}_j \in [0,1]$
- The *hot benevolent type*, $\alpha_i > \theta_i \geq 0$, inflicts punishment on defectors, and only those, if sufficiently angry and confident that the target player in fact defected:
  - if $\left(1 - \dot{a}^{i'}_j\right)\alpha_i > \theta_i$, then $p^\star_{ij}(1,0) \geq p^\star_{ij}\left(1, \dot{a}^{i'}_j\right) > 0 = p^\star_{ij}(1,1)$
  - if $\left(1 - \dot{a}^{i'}_j\right)\alpha_i \leq \theta_i$, then $p^\star_{ij}(1,0) > 0 = p^\star_{ij}\left(1, \dot{a}^{i'}_j\right) = p^\star_{ij}(1,1)$
- The *cool malevolent type*, $\alpha_i = 0 > \theta_i$, inflict the same amount of punishment, independent from their belief, $p^\star_{ij}\left(1, \dot{a}^{i'}_j\right) = \left(-\frac{\kappa_p}{\theta_i}\right)^{\frac{1}{\chi-1}} > 0 \ \forall \dot{a}^{i'}_j \in [0,1]$
- The *hot malevolent type*, $\alpha_i > 0 > \theta_i$

Monitoring

We now show that the dicision to monitor critically depends on which of the above types player i is.

First, it is intuitive that cool players ($\alpha_i = 0$) and players who defected themselves ($a_i = 0$) never have a positive willingness to pay for monitoring such that they are completely price insensitive. To see this, consider first an arbitrary player who defected in the first stage. Since he always inflicts the same amount of damage (which may be zero), we have

$$E\left[u_i | p^\star_{ij}, m_{ij} = 0\right] = \begin{cases} w^1_i & \theta_i \geq 0 \\ w^1_i + \frac{\theta^2_i}{\kappa_p} & \theta_i < 0 \end{cases}$$

$$E\left[u_i | p^\star_{ij}, m_{ij} = 1\right] = \begin{cases} w^1_i - \kappa_m & \theta_i \geq 0 \\ w^1_i - \kappa_m + \frac{\theta^2_i}{\kappa_p} & \theta_i < 0 \end{cases}$$

such that $m^\star_{ij} = 0$ for all priors and monitoring prices.[14] Exactly the same applies for cooperating cool types.

Second, only hot players ($\alpha_i > 0$) who cooperated in the first stage ($a_i = 1$) may (but need not) have a willingsness to pay for monitoring, depending on their prior. To see this, observe that

$$E\left[u_i | p^\star_{ij}, m_{ij} = 1\right] = \begin{cases} w^1_i - \kappa_m & \theta_i \geq \alpha_i \\ w^1_i - \kappa_m + \frac{\left(1 - \dot{a}^i_j\right)(\theta_i - \alpha_i)^2}{\kappa_p} & \theta_i < \alpha_i \end{cases}$$

[14] For $\kappa_m = 0$, $m^\star_{ij} = 0$ is only a weak best response. We assume that a player always opts for non-monitoring when indifferent.

$$E\left[u_i|p^{\star}_{ij}, m_{ij}=0\right] = \begin{cases} w_i^1 & \theta_i \geq \left(1-\dot{a}_j^i\right)\alpha_i \\ w_i^1 + \frac{\alpha_i\left(\dot{a}_j^i\alpha_i + 2\dot{a}_j^i\theta_i - 2\dot{a}_j^i\alpha_i - 2\dot{a}_j^i\theta_i + \alpha_i\right) + \theta_i^2}{\kappa_p} & \theta_i < \left(1-\dot{a}_j^i\right)\alpha_i \end{cases}$$

Thus, three cases need to be distinguished. First, if $\theta_i \geq \alpha_i$, then

$$E\left[u_i|p^{\star}_{ij}, m_{ij}=1\right] = w_i^1 - \kappa_m < w_i^1 = E\left[u_i|p^{\star}_{ij}, m_{ij}=0\right]$$

$\forall \kappa_m > 0$, such that $m^{\star}_{ij} = 0$ for all priors. Thus, those types who are hot but also sufficiently benevolent never monitor (because they never punish) and are hence insensitive of κ_m.

Second, if $\theta_i < \alpha_i$ but also $\theta_i \geq \left(1-\dot{a}_j^i\right)\alpha_i$, then $m^{\star}_{ij} = 1$ if and only if

$$w_i^1 - \kappa_m + \frac{\left(1-\dot{a}_j^i\right)\left(\theta_i-\alpha_i\right)^2}{\kappa_p} > w_i^1 \Leftrightarrow \frac{\left(1-\dot{a}_j^i\right)\left(\theta_i-\alpha_i\right)^2}{\kappa_p} > \kappa_m \Leftrightarrow \dot{a}_j^i < 1 - \frac{\kappa_m\kappa_p}{\left(\theta_i-\alpha_i\right)^2}$$

If this condition is satisfied, then the player monitors and inflicts punishment if the target player defected and no damage if the target player cooperated; if this condition is violated, then the player does neither monitor nor inflict any damage. Individuals of this type generate the *scale effect* of increasing monitoring costs on punishment: they have a threshold price

$$\check{\kappa}_m = \frac{\left(1-\dot{a}_j^i\right)\left(\theta_i-\alpha_i\right)^2}{\kappa_p}$$

such that they monitor and punish conditionally as long as $\kappa_m < \check{\kappa}_m$, while switching to non-monitoring together with non-punishment if the price increases above $\check{\kappa}_m$.

Third, if $\theta_i < \left(1-\dot{a}_j^i\right)\alpha_i$, then $m^{\star}_{ij} = 1$ if and only if

$$\frac{\alpha_i\dot{a}_j^i\left(1-\dot{a}_j^i\right)}{\kappa_p} > \kappa_m \Leftrightarrow \dot{a}_j^i \in \left(\frac{1}{4} - \frac{\kappa_m\kappa_p}{\alpha_i^2}, \frac{3}{4} - \frac{\kappa_m\kappa_p}{\alpha_i^2}\right)$$

If this condition is satisfied, then the player monitors and inflicts punishment if the target player defected and no damage if the target player cooperated; if this condition is violated, then the player does not monitor but inflict some damage unconditionally. Individuals of this type generate the *substitution effect* of increasing monitoring costs on punishment: they have a threshold price

$$\hat{\kappa}_m = \frac{\alpha_i\dot{a}_j^i\left(1-\dot{a}_j^i\right)}{\kappa_p}$$

such that they monitor and punish conditionally as long as $\kappa_m < \hat{\kappa}_m$, while switching to non-monitoring together with untargeted punishment if the price increases above $\hat{\kappa}_m$.

In sum, there is a class of types (consisting of the cool types and the hot types which are sufficiently confident that the target player cooperated) which are completely price insensitive, and a class of types (consisting of the hot types which are sufficiently confident that the target player defected) which are sensitive to changes in the monitoring costs.

Hypotheses

We now derive our hypotheses from the model by making two mild assumptions on the distribution of parameters in the population. First, we assume that there is some prohibitive price $\bar{\kappa}_m \in \mathbb{R}_{++}$ at which no individual in the population monitors. Second, for all prices $\kappa_m \in [0, \bar{\kappa}_m] \subset \mathbb{R}$ we assume that a marginal individual who is just indifferent between monitoring and non-monitoring exists.

First, we expect information on the target player's behavior to obey the *law of demand*, that is, the demand for monitoring to be decreasing in its cost.

Hypothesis 1. The frequency of monitoring second and third parties is decreasing in κ_m.

More specifically, we expect the relative frequency of subjects that punish without monitoring not to be significantly different from zero if monitoring is costless, but significantly positive if monitoring is costly.

The next three hypotheses concern the impact of monitoring costs on punishment behavior, specifically the crowding and substitution effects identified in the model.

Hypothesis 2. The frequency of second and third parties punishing in at least one contingency is decreasing in κ_m.

This means that we expect the relative frequency of second and third parties punishing at all is decreasing in the monitoring costs. This is the crowding effect. On the other hand, we expect the relative frequency of second and third parties punishing unconditionally, that is, defection and cooperation alike, is increasing in the monitoring costs. This is the substitution effect and is stated in the following

Hypothesis 3. The frequency of second and third parties punishing unconditionally is increasing in κ_m.

Finally, by the crowding und substution effect and the fact that $p^\star_{ij}(0) \geq p^\star_{ij}\left(\dot{a}^i_j\right) \geq p^\star_{ij}(1) \geq 0$ we get the following

Hypothesis 4. Average punishment of defection (absolute and relative) decreases as κ_m increases.

We do not predict any differences between second and third parties, since *ad hoc* assumptions on the distribution of preference parameters would be necessary. We leave this to be resolved by the data.

Table 1: Sample arrangement

	Number of subjects		
	Sequence		
Cell	2P-3P	3P-2P	Marginal frequency
M0	14	8	22
M5	36	30	66
M10	26	20	46
Marginal frequency	76	58	134

2.3 Procedures

All experiments were conducted at the experimental laboratory of the Alfred-Weber-Institute (AWI-Lab) at the University of Heidelberg in late 2010. Participants were recruited from the general undergraduate student population using the online recruitment system ORSEE (Greiner 2004). In total 134 subjects participated (49.3% have been female, 93.3% German citizen, and 64.2% had never participated in a laboratory experiment before). The mean age was 21.8 years. No subject participated in more than one session. The sample arrangement is shown in table 1.

At the beginning of each session, subjects were randomly assigned to the computer terminals upon entering the laboratory. Any direct communication among subjects has been strictly forbidden during the whole session and booths separated the participants visually, ensuring that they made their decisions anonymously and independently. Further, subjects did not receive any information on personal identities of any other participant, neither before nor while nor after the experiment.

At the beginning of the experiment, that is before any decisions were made, subjects received detailed written instructions that explained the exact structure of the game and procedural rules.[15] Participants had to answer a set of control questions individually at their respective seats in order to ensure comprehension of the rules. We did not start the experiment before all subjects had answered all questions correctly.

The experiment was programmed and conducted with z-Tree (Fischbacher 2007). The exact timing of events was as follows. First, the subjects were randomly matched to groups of two. Then each subject made her or his decisions and reported expectations under the second party or third party monitoring condition, respectively, depending on the sequence. After being informed about the payoffs, the experimenter announced that a second experiment will be conducted and distributed additional instructions that explained the differences to the first

[15]The experiment was framed in a sterile way using neutral language and avoiding value laden terms in the instructions. The instructions (in German) are available upon request.

condition. After being randomly re-matched in new groups, each subject made her or his decisions and reported expectations under the third party or second party monitoring condition, respectively, depending on the sequence. After being informed about the payoffs, the experimenter announced that a third and definitely final experiment will now be conducted and distributed additional instructions that explained the supplementary lottery choice experiment. After being informed about the payoffs, subjects were asked to answer a short questionnaire while the experimenter prepared the payoffs. Subjects were then individually called to the experimenter booth, payed out and discarded.

In any session subjects received a fixed show-up fee of €2, which was not part of their endowment but is included in the mean earnings reported below. A session lasted around 90 minutes and subjects earned €18.46 (€0.10 per token earned) on average. Earnings exceed the local average hourly wage of a typical student job.

3 Results

We begin by showing that the results under the baseline treatment $M0$ replicate the results of Fehr & Fischbacher (2004b). We then proceed to test or main hypotheses presented in the previous section.

Since we have, in contrast to Fehr & Fischbacher (2004b), no significant order effects in any of the key decision variables, we are able to pool the two sequences for the statistical analysis.[16] In appendix A summary data given in tabular form, including separations by sequence.

3.1 Baseline

Our baseline condition $M0$ replicates the experiment reported in Fehr & Fischbacher (2004b, pp. 81-84), so we present the results from that condition first and compare them to the results obtained by those authors. When monitoring was costless, the two main results obtained by Fehr & Fischbacher (2004b) are confirmed. Firstly, subject target punishment on defection, that is, both second

[16] From an investigation of the tables in appendix A it is already apparent that differences between sequences are not systematic. Using Mann-Whitney tests we find no significant differences among sequences in monitoring of second parties ($p = .160$) and third parties ($p = .775$), second party punishment of defectors ($p = .132$), second party punishment of cooperators ($p = .066$), third party punishment of defectors ($p = .122$ when the target's coplayer defected, $p = .116$ when the target's coplayer cooperated) and of cooperators when the target's coplayer cooperated ($p = .187$), as well as transfer decisions in the second party monitoring ($p = .084$). Behavior was significantly different between sequences with respect to third party punishment of cooperators when the target's coplayer defected ($p = .006$) and with respect to transfers in the third party monitoring condition ($p = .037$). Fisher exact tests yield similar results which are available on request.

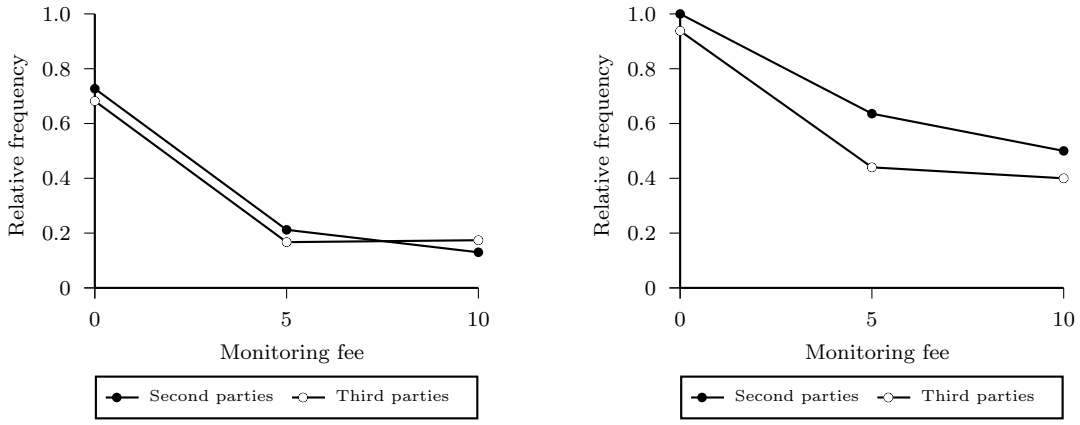
and third parties strongly punished defectors whereas punishment of cooperators was weak but existent. Secondly, second party punishment was somewhat stronger than third party punishment (see figures 4(a), 5(a) and 5(b)).

Concerning the first result, defection was always punished stronger than cooperation, both by second and by third parties. The differences are statistically significant for second party punishment (Wilcoxon signed-rank test, $p = .000$) and third party punishment in case the target player's coplayer cooperated (Wilcoxon signed-rank test, $p = .000$) and marginally significant for third party punishment in case the target player's coplayer defected (Wilcoxon signed-rank test, $p = .099$). Tables 5 and 7 further supports this by showing that the share of subjects who punished defection was considerably greater (between 45.5% and 72.7%) than the share of subjects who punished cooperation (between 13.6% and 22.7%). Nevertheless, the fraction of subjects punishing cooperation is non-negligible even when targeting is costless.

Concerning the second result, figures 4(a), 5(a) and 5(b) reveal that second parties punished defection with 7.59 points on average, while third parties assigned 6.00 points on average to defectors whose coplayers cooperated and 2.27 points to defectors whose coplayers defected. The difference is significant in case target's coplayer defected (Wilcoxon signed-rank test, $p = .000$) but insignificant in case the coplayer cooperated (Wilcoxon signed-rank test, $p = .217$). This divergence from the result of Fehr & Fischbacher (2004b) is due to the stronger punishment by third parties in our experiment (5.25 punishment points on average in the 3P-2P sequence in ours, 3.09 points in theirs). This implies that, while defection was always profitable in the 3P condition in the experiment by Fehr & Fischbacher, defection was only profitable in the 3P condition if the subject's coplayer also defected. In the 2P condition, a defector's average (pooled over sequences) income was reduced by 22.77 tokens whereas the gain from defection was only 10 tokens. Likewise, in the 3P condition, a defector whose coplayer cooperated (defected) incurred a cost of 18.00 tokens (6.81 tokens), and defection led to an overall income of 22.00 tokens (33.19 tokens) whereas a cooperative choice led to a small residual punishment cost of 1.71 tokens (4.56 tokens) and an overall income of 28.29 tokens (25.56 tokens). Thus, given the punishment pattern, defection in the 3P condition was profitable if the subject's coplayer also defected but unprofitable if the subject's coplayer cooperated.

Since Fehr & Fischbacher (2004b) did not elicit beliefs in their experiment, we are able to contribute to a deeper understanding of third party punishment on top of replicating their results. Under costless monitoring, subjects' expectations correspond quite accurately to actual play. Subjects expected to receive on average 7.05 (0.32) punishment points from second parties in case of defection (cooperation), while 7.59 (0.59) points have been actually levied on average, such that the net cost of defecting (21 tokens) was predicted very accurately with 20.19 tokens. Similarly, subjects expected to receive on average 5.95 (1.41) punishment points from third parties in case of defection (cooperation) when their partner co-

Figure 1: Relative frequency of monitoring subjects in the complete sample (a) and the subsample of those subjects who punished.



operated and 2.91 (0.86) when their partner defected, while 6.00 (1.36) and 2.27 (1.14) points have been actually levied, respectively, on average. These results provide quite convincing evidence that the game has been well understood by the subjects even before playing it for the first time, contributing to invalidating concerns about true one-shot experiments (see the discussion in section **??**).

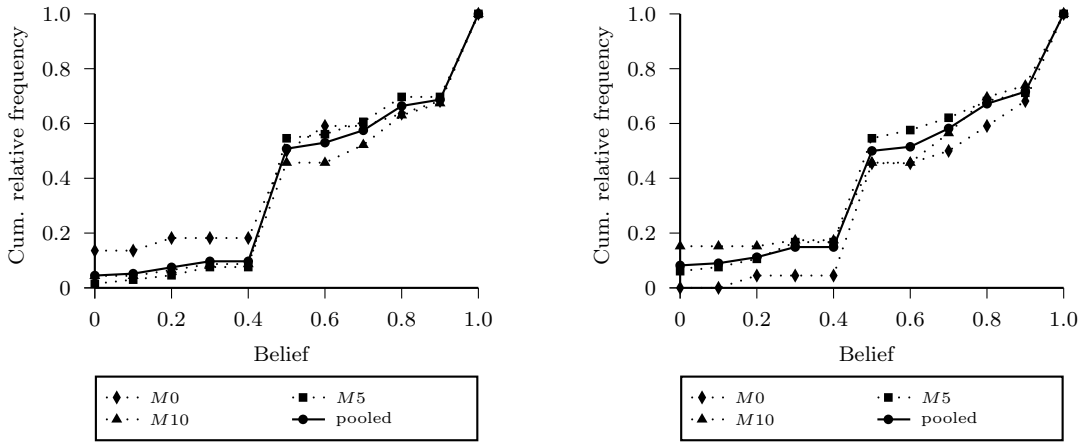
3.2 Monitoring

Two first, conservative statistical null hypothesis implied by hypothesis 1 is the frequency of monitoring subjects to be no different under the different monitoring fee levels. This null can be rejected for both second and third parties (Kruskal-Wallis tests, adjusted for ties, $p = .000$, respectively). Thus, the level of monitoring costs does have an impact on the propensity to monitor. Secondly, we predicted in hypothesis 1 monitoring to obey the law of demand. This is indeed shown by the data. As depicted in 2(a), the frequency of monitoring subjects diminishes as monitoring costs get positive. Negative correlation is significant for both second (Kendall's $\tau_b = -.363$, $p = .000$, continuity corrected) and third parties (Kendall's $\tau_b = -.287$, $p = .001$, continuity corrected).[17] We consider this a full support for hypothesis 1, although we note that the impact of increasing monitoring costs seems to be strongest at low levels and substantially weaker at higher levels.

Of course, those individuals which do not punish at all (and therefore do not

[17] We prefer Kendall's (Kendall 1938) over Spearman's (Spearman 1904) rank correlation coefficient since the former is can be considered both more robust and more intuitive to interpret. Estimates of Spearman's rho are available on request.

Figure 2: Distributions of beliefs on the transfer decision of the target subject under the second and third monitoring conditions

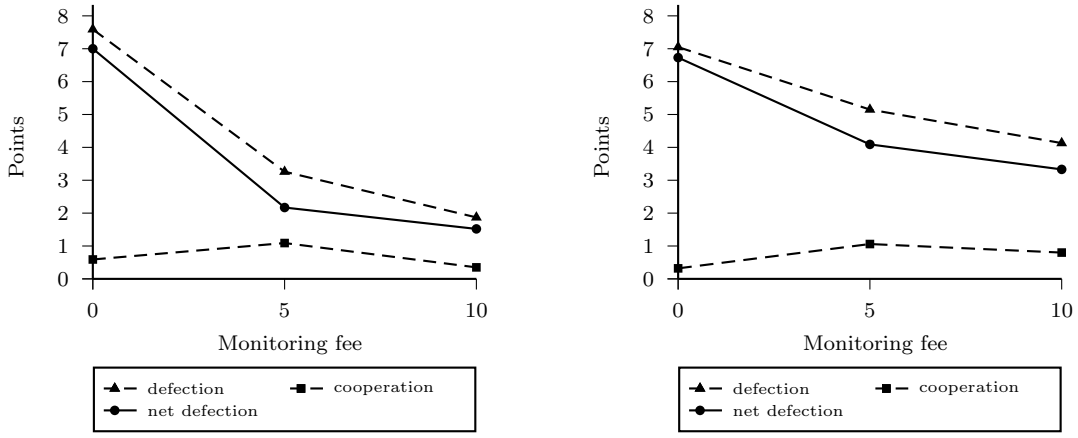


monitor) even at zero monitoring costs are irrelevant for changes in demand for monitoring as the price changes. About 27% of both second and third parties are of this kind in our experiment, which is in line with previous experimental evidence that classifies around one quarter to one third of subjects as purely individualistic (Fischbacher et al. 2001, Fischbacher & Gächter 2010). Thus, the effects should be much clearer as only those subjects that punish in at least one contingency are considered. Excluding non-punishers, every second party monitored (16 of 16), and only one of 16 third parties (still, the relative frequency of monitoring third parties is not significantly different from one, Wilcoxon signed-rank test, $p = .317$) punished unconditionally (i.e. without monitoring) when monitoring was costless. Further, as depicted in 2(b), as considerable share of subjects begin to punish without prior monitoring as monitoring costs get positive. We come back to this below. At this point, we identify an even stronger negative relationship between monitoring costs and the demand for monitoring if non-punishers are excluded (Kendall's $\tau_b = -.407$, $p = .003$, continuity corrected, for second parties, $\tau_b = -.370$, $p = .003$ for third parties).

What is the role of beliefs on the target players' behavior. Figures 3(a) and 3(b) show the distributions of beliefs under the second and third party monitoring conditions for all monitoring cost treatments. The modal belief is always at 0.5 with another cluster at 1 (note that the beliefs are coded as the probability that the target player cooperates).[18] The mean belief of those second (third)

[18]Increasing monitoring costs seem to had no effect on beliefs: In the the second party monitoring condition, 50.0% have a tendency to believe that their coplayer cooperated (a belief strictly greater than one half) under $M0$, 45.5% under $M5$, and 54.4% under $M10$.

Figure 3: Mean magnitude of second party punishment and mean magnitude of expected second party punishment.



parties that monitored was .693 (.755) under $M5$ and .700 (.700) under $M10$, while the mean belief of those that did not monitor was .658 (.604) and .693 (.613), respectively, with non of those differences being statistically significant (Mann-Whitney tests, $p > .130$). The null hypothesis that the decision to monitor and beliefs are independent cannot be rejected (Kendall's $\tau_b = .002$, $p = .987$, continuity corrected, for the whole sample, $\tau_b = .030$, $p = .738$ for $c_m > 0$).

3.3 Punishment

In hypothesis 4 we predicted mean damages imposed on defectors to be decreasing if monitoring costs increase.

Figure 4(a) shows how the mean magnitudes of second party punishment develop as the monitoring costs get positive. While there is no significant variation in punishment of cooperation (Kruskal-Wallis test, adjusted for ties, $p = .756$), punishment of defection is significantly different between treatments (Kruskal-Wallis test, adjusted for ties, $p = .000$), exhibiting a diminishing trend as monitoring costs rise (Kendall's $\tau_b = -.266$, $p = .000$). The same is true for the net punishment of defecting, which is the difference of punishment of defection and punishment of cooperation (Kendall's $\tau_b = -.299$, $p = .000$). Hence, *on average* defecting gets less costly as monitoring costs rise.

This pattern is qualitatively anticipated by the subjects, as shown in figure 4(b). Again, there is no significant variation in expected punishment of cooperation (Kruskal-Wallis test, adjusted for ties, $p = .626$), punishment of defection is significantly different between treatments (Kruskal-Wallis test, adjusted for ties, $p = .045$). However, the impact on punishment of defection (Kendall's $\tau_b = -.156$,

p = .033) and hence the net punishment of defecting (Kendall's τ_b = −.166, p = .024) is somewhat underestimated by the subjects. Specifically, while the actual net cost of defecting amounted to 6.50 tokens on average under $M5$ and 4.57 on average under $M10$, the expected net costs were 12.27 and 9.98, respectively, on average. Thus, while the *actual* sanctions by second parties were not sufficient to render defection unprofitable under costly monitoring, cooperation was still consistent with the subjects' *beliefs*.

Similar patterns resulted in the third party monitoring condition. Figures 5(a) and 5(b) show the mean magnitudes of third party punishment for the cases in which the target player's coplayer cooperated and defected, respectively. As under the second party monitoring condition, while there is no significant variation in punishment of cooperation (Kruskal-Wallis tests, adjusted for ties, p = .600 in case the target player's coplayer cooperated, p = .796 in case the target player's coplayer defected), punishment of defection is significantly different between treatments in case the target player's coplayer cooperated (Kruskal-Wallis test, adjusted for ties, p = .005), but not in case the target player's coplayer defected (Kruskal-Wallis test, adjusted for ties, p = .291). Thus, punishments targeted at unilateral defection are the only ones affected by costly monitoring, its strength being diminishing as monitoring costs rise (Kendall's τ_b = −.148, p = .049), although to a weaker extent as second party punishments targeted at defection. The same is true for the net punishment of defecting unilaterally, (Kendall's τ_b = −.308, p = .000). Hence, as for second party punishment, *on average* defecting, in particular unilaterally, gets less costly as monitoring costs rise.

Again, subjects anticipate this pattern quite well qualitatively but underestimate it quantitatively, as shown in figures 5(c) and 5(d). No significant variation in expected punishment of cooperation exists (Kruskal-Wallis tests, adjusted for ties, p = .288 in case the target player's coplayer cooperated, p = .701 in case the target player's coplayer defected). Expected punishment of defection is marginally significantly different between treatments in case the target player's coplayer cooperated (Kruskal-Wallis test, adjusted for ties, p = .070), but not in case the target player's coplayer defected (Kruskal-Wallis test, adjusted for ties, p = .856). However, the impact on punishment of unilateral defection (Kendall's τ_b = −.113, p = .121) and hence the net punishment of defecting unilaterally (Kendall's τ_b = −.090, p = .217) is underestimated by the subjects. Specifically, while the actual net cost of defecting amounted to 3.95 (1.73) tokens on average under $M5$ and 3.91 (3.33) on average under $M10$ when the other group member cooperated (defected), the expected net costs were 8.32 (3.09) and 8.74 (2.93), respectively, on average.

In summary, these patterens are consistent with hypothesis 4.

Figure 4: Mean magnitude of third party punishment and mean magnitude of expected third party punishment.

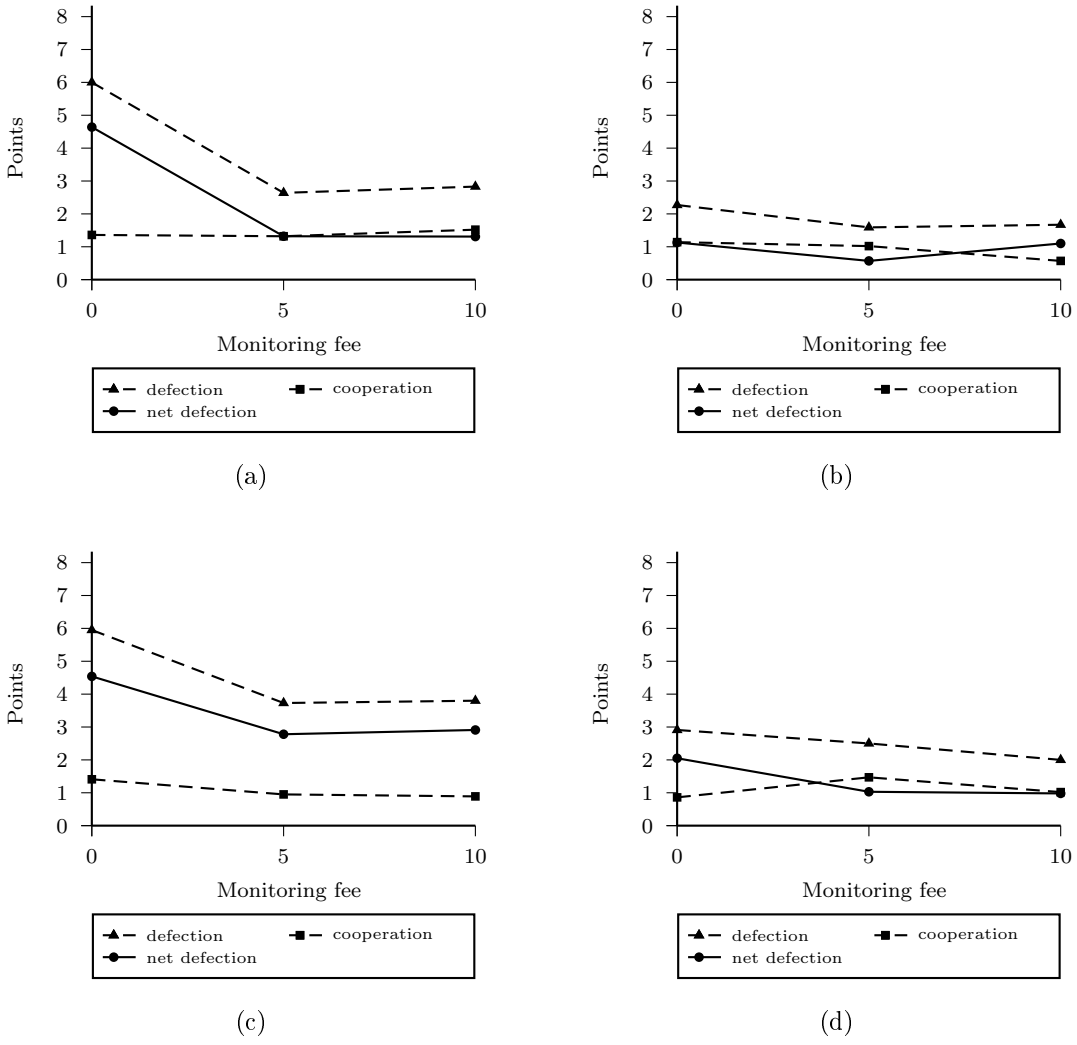
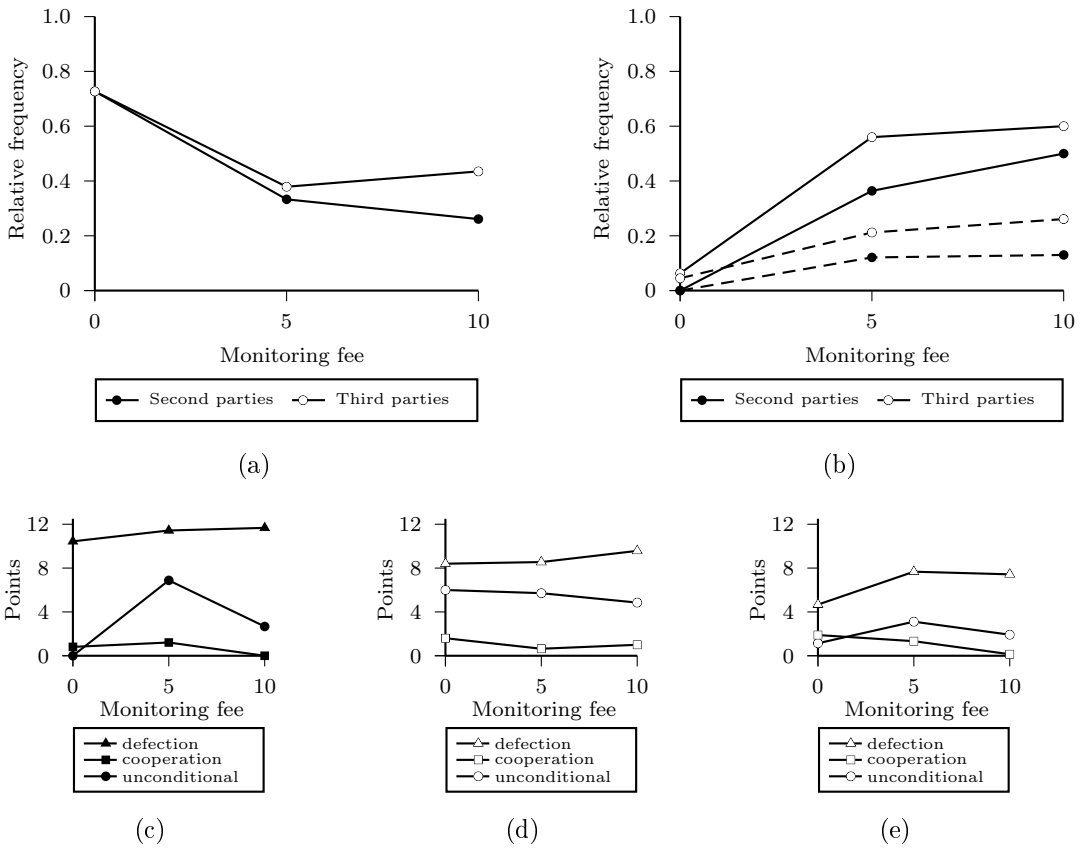


Figure 5: Relative frequency of subjects punishing in at least one contingency (a), relative frequency of subjects punishing unconditionally (b), average punishment levied by second parties punishing in at least one contingency (c), and average punishment levied by third parties punishing in at least one contingency, in case the target player's coplayer cooperated (d) and defected (c), respectively.



Crowding effect

How can we explain those patterns? Here our hypotheses on the crowding and substitution effects are of help. Consider second party monitoring and punishment first. In hypothesis 2 we predicted the frequency of subjects punishing in at least one contingency to be decreasing in monitoring costs. Figure 6(a) shows the fraction those subjects. Apparently, less subjects are willing to punish at all as monitoring becomes costly (Kruskal-Wallis test, adjusted for ties, $p = .001$ for second parties, $p = .017$ for second parties), where the relationship is monotonically decreasing for second parties (Kendall's $\tau_b = -.260$, $p = .002$), but sightly more third parties punish as the monitoring fee rises from five to ten tokens such that there is no significant rank correlation (Kendall's $\tau_b = -.128$, $p = .123$). While 72.7% of both second and third parties punish if monitoring is costless, which is in line with previous experimental evidence that classifies around one quarter to one third of subjects as purely individualistic (see references above), somewhat less than half of them are deterred from punishing as monitoring becomes costly. Thus, the evidence for a crowding effect is, at least for low monitoring cost levels, is strong such that we evaluate hypothesis 2 as supported.

Substitution effect

In hypothesis 3 we predicted the frequency of second parties punishing unconditionally (that is, without monitoring) to be increasing in monitoring costs. Figure 6(b) depicts the fraction of those subjects, that becomes quite substantial as monitoring costs increase, in particular among third parties. 36.4% and 50% of the sanctioning second parties (12.1% and 13% of all second parties) decreased their coplayers' income without conditioning on the latters' first stage behavior when monitoring costs are five and ten tokens, respectively, and even 56% and 60% of the sanctioning third parties (21.2% and 26.1% of all third parties) did so. Differences are significant in all instances (Kruskal-Wallis tests, adjusted for ties, $p < .008$), and correlation between unconditional punishment and level of the monitoring fee is significantly positive (Kendall's $\tau_b = .407$, $p = .002$, continuity corrected for second parties, $\tau_b = .370$, $p = .003$) such that 3 may be considered as supported, too.

Punishment of defection under costly monitoring

The consequences of these patterns in terms of average punishment strength is depicted in figures 6(c) through 6(e). Figure 6(c) shows the average sanction points levied by those second parties that sanctioned in at least one case. While conditional punishment of defection is about constant (slightly increasing on average, but differences are not significant according to a Kruskal-Wallis test, adjusted for ties, $p = .928$) at about 11 points (implying a deduction of 33 tokens), and conditional punishment of cooperation is at low levels decreasing to zero (differences

are again not significant according to a Kruskal-Wallis test, adjusted for ties, $p = .488$), average unconditional punishment amounts to 6.9 points (implying a deduction of 20.70 tokens) if the monitoring fee is five, decreasing to 2.7 points (implying a deduction of 8.1 tokens) if the monitoring fee is ten (the difference is significant, $p = .033$, according to a Mann-Whitney test).

Figures 6(d) and 6(e) show the average sanction points levied by those third parties that sanctioned in at least one case. Figure 6(d) is the case in which the target player's coplayer cooperated. Average conditional punishment of defection is slightly increasing and unconditional punishment slightly decreasing, but differences are not significant (Kruskal-Wallis test, adjusted for ties, $p = .845$ and $p = .726$, respectively). Conditional punishment of cooperation is about constant at low levels (Kruskal-Wallis test, adjusted for ties, $p = .726$) while not being significantly different from zero under $M0$ (Wilcoxon signed-rank test, $p = .317$). Figure 6(d) is the case in which the target player's coplayer defected. Here, average conditional punishment of defection is increasing more pronounced and conditional punishment decreasing down to zero under $M10$ (Wilcoxon signed-rank test, $p = .002$), but still differences are not significant (Kruskal-Wallis test, adjusted for ties, $p = .358$ and $p = .532$, respectively). Unconditional punishment is slightly higher under costly monitoring, but not significantly so (Kruskal-Wallis test, adjusted for ties, $p = .377$), however, unconditional punishment is significantly weaker if the target player's coplayer defected than if the latter cooperated (Wilcoxon signed-rank test, $p = .003$).[19]

3.4 Multivariate regression analysis

Additional variables

A trust-control survey instrument In the post-experimental questionnaire we implemented questions on trust and monitoring. First, we asked the three items used in the German Socio-Economic Panel (GSOEP) to construct a survey trust measure, (1) "In general, one can trust people", (2) "Nowadays, you can't rely on anybody", and (3) When dealing with strangers, it is better to be cautious before trusting them".[20] The answer categories were on a five-point Likert scale from strong disagreement (coded as −2) to strong agreement (coded as 2). The mean scores on these three items were 0.507, −0.925, and 0.388, respectively.

We augmented this battery by two further items, (4) "I can trust fellows I don't (yet) know", and (5) "Trust is good, control is better", coded in the same way as the previous three items. The fourth item is more personal than the general questions in the first three items, and the fifth item is specifically tailored

[19]Note that each third party could costlessly condition punishment on the behavior of the target player's coplayer.

[20]The one-dimensional items in the GSOEP are considered to be superior to the items used in the American General Social Survey (GSS) or the Wold Values Survey (WVS) (Miller & Mitamura 2003)

Table 2: Results of the principal-component factor analysis on the trust-monitoring battery

		Factor analysis/correlation				
		Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
unrotated	Eigenvalue	2.4355	0.7662	0.6990	0.6312	0.4681
	Difference	1.6692	0.0672	0.0678	0.1630	–
	Proportion	0.4871	0.1523	0.1389	0.1262	0.0936
	Cumulative	0.4871	0.6403	0.7801	0.9064	1.0000
rotated	Variance	2.4355				
	Difference	–				
	Proportion	0.4871				
	Cumulative	0.4871				
		Factor loadings (pattern matrix) and unique variances				
	Variable	Factor 1		Uniqueness		Scoring coefficient
	Item 1	−0.7489		0.4378		−0.3079
	Item 2	0.7522		0.4342		0.3089
	Item 3	0.6884		0.5261		0.2827
	Item 4	−0.6355		0.5962		−0.2609
	Item 5	0.6555		0.5703		0.2692

Method: principal-component factors. $N = 134$, 5 parameters, 1 factor retained. Rotation: orthogonal verimax (Kaiser off). LR test: independent vs. saturated, $\chi^2(10) = 125.37$, $p = .000$.

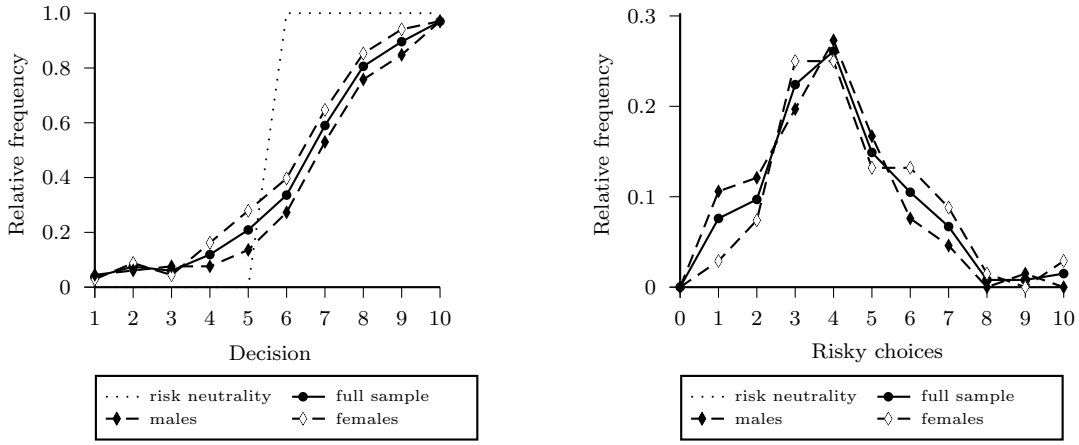
at the trust-vs.-control attitude. The mean scores on these two items were 0.366 and 0.388, respectively.

Using principal-component factor analysis, we obtained a single factor from these five items, that we call *trust-control instrument* (see table **??**). Using the scoring coefficients based on verimax rotated factors, we estimated the individual scores (mean −1.50, std. dev. 1, min. −2.17, max. −2.23, note that the variable is coded such that higher scores indicate greater mistrust), and used the instrument as a predictor of the monitoring probability in the regression analysis reported below. Its coefficient is positive and significant, suggesting that the measure captures some preference based (mis)trust because this effect is present even when controlling for beliefs.

Risk preferences Table 3 contains summary data for the choices in the Holt-Laury task. The data is coded by the number of risky choices (option B).[21] Zero (ten) risky choices indicate extreme risk aversion (preference), six risky choices indicate risk neutrality. The subjects in our sample are somewhat risk averse, with an average of 4.1 risky choices. Consistent with most findings in the risk literature (see Eckel & Grossman 2008), women are more risk averse than men,

[21] Note that Holt & Laury (2002) report their data by the number of safe choices. We made the appropriate adjustments in comparing our data with theirs.

Figure 6: Summary statistics of the risk preference instrument.



(a) (b)

Table 3: Summary statistics of the Holt-Laury instrument

	Share by gender		Total share		
Risky choices	Male	Female	All	Consistent	Holt & Laury
0 – 1	.029	.106	.067	.076	.01
2	.074	.121	.097	.102	.03
3	.250	.197	.224	.237	.13
4	.250	.273	.261	.271	.23
5	.132	.167	.149	.136	.26
6	.132	.076	.105	.110	.26
7	.088	.046	.067	.059	.06
8	.015	.000	.008	.000	.01
9 – 10	.029	.015	.022	.009	.01
Mean	4.41	3.77	4.10	3.92	4.8

In the fifth column subjects whose choices were inconsistent with the typical one-crossover pattern (option A at the top, switching to B at some point) are removed.

and this difference is marginally statistically significant in our sample (Mann-Whitney test, $p = .078$). The distribution of risk preference is similar to Holt & Laury (2002), however, the the distribution in our sample is stronger skewed to the right indicating our subjects to be somewhat more risk averse. About 12 percent of the subjects had more than one crossover between the two options. The subsample of those inconsistent subjects is significantly different from the consistent subjects (Mann-Whitney test, $p = .009$), the former choosing on average more risky options (5.44) than the latter (3.92). In subsequent analysis, we include all subjects' choices.

Finally, since we conducted the lottery choice task at the end of each session, it may be possible that the prior tasks had some impact on the subjects' choices. In particular, choices in the lottery choice task may be different for subject's whose coplayer defected than for those whose coplayer cooperated in the first part of the session. However, Mann-Whitney tests indicate that this is not the case: in terms of risky choices in the lottery choice task the subsample whose coplayer defected is not significantly different from the subsample whose coplayer cooperated, neither in the second party monitoring ($p = .721$) nor the third party monitoring condition ($p = .162$).

We used the instrument as a predictor of the monitoring probability in the regression analysis reported in appendix 3.4 and found no significant relationship between a subject's risk attitude (as measured by the Holt-Laury instrument) and the probability that this subject monitors This fact is, although speculative, perhaps related to the converging evidence that also trust decisions are relatively independent from an individual's risk attitudes in trust games (Houser et al. 2010).

Estimation results

We estimated various non-linear models with nested random effects that take into account clustering by session using GLLAMM (Skrondal & Rabe-Hesketh 2003, Rabe-Hesketh et al. 2005).[22] The results are shown in table 4.

4 Conclusion

In this paper we studied monitoring behavior, punishment behavior, and their interaction in a simple exchange experiment. The purpose was the assessment whether and to what extent subjects will monitor, that is, engage in costly information acquisition prior to potential punishment, in order to be able to condition punitive responses on the target player's behavior. In addition, we were interested

[22]Generally, ignoring clustering by session and treating subjects as completely independent observations may result in underestimation of standard errors and henceforth an overstatement of statistical significance.

Table 4: Estimates of binary response model regressions: Monitoring

(a) Second parties

	Nested RE Logit				Nested RE Probit	
Variable	Model 1		Model 2		Model 1	
Monitoring fee	−0.333	(0.103)	−0.367	(0.117)	−0.211	(0.065)
	.001		*.002*		*.001*	
Cooperated	1.150	(0.648)	1.045	(0.597)	0.628	(0.342)
	.076		*.080*		*.067*	
Trust	0.470	(0.218)	0.382	(0.205)	0.225	(0.122)
	.031		*.063*		*.065*	
Belief	0.046	(0.915)				
	.960					
Risk preference	0.013	(0.112)				
	.905					
Gender	−0.447	(0.457)				
	.328					
Age	−0.159	(0.107)				
	.139					
Constant	3.267	(2.423)	−0.073	(0.702)	−0.066	(0.412)
	.117		*.918*		*.872*	
RE level 1 Variance	0.140	(0.018)	0.145	(0.018)	0.145	(0.018)
RE level 2 Variance	0.244	(0.347)	0.201	(0.333)	0.088	(0.124)
Level 1 units	134		134		134	
Level 2 units	9		9		9	
Log likelihood	−60.213		−62.458		−62.667	

(b) Third parties

	Nested RE Logit		Nested RE Probit
Variable	Model 1	Model 2	Model 1

The dependent variable is the probability that the a subject monitors the target player. *Cooperated* is an indicator variable with value 1 if a subject cooperated and 0 otherwise. *Trust* is the trust-monitoring instrument constructed from questionaire data in section **??**. *Belief* is the subject's probability estimate that the target player cooperated. *Risk preference* is the the number of risky choices in the Holt-Laury task. *Gender* is an indicator variable with value 1 if a subject is female and 0 otherwise. Tabulated are the parameter estimates. The standard errors in parantheses and the *p*-values are in italics. These models are estimated using GLLAMM for STATA. Standard errors take into account clustering by session.

in identifying possible interactions between monitoring and punishment behavior. We do so by manipulating the monitoring costs both in a second party and a third party monitoring condition. This allows a bifocal investigation of (1) responses to changes in those costs and (2) potential differences in behavior between second and third parties.

The main results are the following. First, the outcomes under the baseline treatment (zero monitoring costs) replicate known patterns, specifically the results obtained by Fehr & Fischbacher (2004b): the large majority of second and third parties use punishment, strongly targeted on defectious behavior. *All* subjects (with one exception) that used punishment conditioned it on the target player's first stage behavior.

Second, as monitoring gets costly, subjects do invest in monitoring over and above the punishment costs incurred and monitoring information appears to obey the *law of demand* for both second and third parties, that is, less individuals demand information the higher its price. This relationship is strongly convex, with high elasticity at lower and smaller elasticity at higher monitoring cost levels.

Third, changes in demand can be decomposed into two distinct effects that both point in the same direction with respect to monitoring but exhibit different implications for punishment behavior. On the one hand, as monitoring costs rise some individuals withdraw from punishment altogether (that is, they punish neither defection nor coopration). We termed this *scale effect* of increasing monitoring costs. On the other hand, others begin to switch from targeted to unconditional punishment (that is, punishment of cooperation and defection alike) as monitoring gets costlier. We termed this *composition effect* of increasing monitoring costs. Notably, among those individuals that monitor, punishment of cooperative behavior is gradually repressed up to complete elimination as monitoring costs rise. Thus, it appears that subjects making relatively little quantitative difference between punishing defection and cooperation, respectively, seem to be crowded out (towards non-punishment and the unconditional punishment groups) while the sharply targeted sanctions survive. Nevertheless, overall, average net costs inflicted on defectors diminish considerably as monitoring gets costly.

Fourth, while both effects are present in both second and third parties, the composition effect is stronger for third than for second parties with the fraction of unconditionally third parties being significantly greater than the fraction of unconditionally punishing second parties at all positive monitoring cost levels, indicating that third parties are much less reluctant to punish untargeted and go for the risk of erroneous punishment. Since this comparison is within-subjects, risk preferences cannot account for the difference. Rather, it is consistent with the application of within-subject differences in direct strong reciprocity and social reciprocity motives (Carpenter et al. 2004; Carpenter & Matthews 2004)

In summary, from an institutional perspective our results suggest that the effect of (exogenous) increases or decreases of mutual monitoring costs can be

predicted by standard microeconomic arguments provided that social preferences are taken into account while, from a behavioral perspective, a large share of subjects appear to exhibit a relatively low valuation of monitoring information, implying a rather weak reciprocity motive.

References

Ambrus, A. & Greiner, B. (2010). *Imperfect public monitoring with costly punishment - An experimental study.* Mimeograph, Harvard University, Cambridge, Massachusetts.

Anderson, C. M. & Putterman, L. (2006). Do non-strategic sanctions obey the law of demand? the demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior*, 54(1), 1–24.

Anderson, L. R. & Stafford, S. L. (2009). *An experimental study of the effects of announcements on public goods contributions.* Department of Economics Working Paper 82, College of William and Mary, Williamsburg, Virginia.

Andreoni, J. & Petrie, R. (2004). Public goods experiments without confidentiality: A glimpse into fund-raising. *Journal of Public Economics*, 88(7-8), 1605–1623.

Bolton, G. E. & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1), 166–193.

Bornstein, G. & Weisel, O. (2010). Punishment, cooperation, and cheater detection in "noisy" social exchange. *Games*, 1(1), 18–33.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.

Brook, A. S. (1999). Shame on you: An analysis of modern shame punishment as an alternative to incarceration. *William and Mary Law Review*, 40(2), 653–686.

Carpenter, J. (2007a). The demand for punishment. *Journal of Economic Behavior & Organization*, 62(4), 522–542.

Carpenter, J., Bowles, S., Gintis, H., & Hwang, S.-h. (2009). Strong reciprocity and team production: Theory and evidence. *Journal of Economic Behavior & Organization*, 71(2), 221–232.

Carpenter, J., Kariv, S., & Schotter, A. (2010). *Network architecture and mutual monitoring in public goods experiments.* IZA Discussion Paper Series 5307, Forschungsinstitut zur Erforschung der Arbeit, Bonn.

Carpenter, J. P. (2007b). Punishing free-riders: how group size affects mutual monitoring and the provision of public goods. *Games and Economic Behavior*, 60(1), 31–51.

Carpenter, J. P., Matthews, H. P., & Org'ong'a, O. (2004). Why punish? social reciprocity and the enforcement of prosocial norms. *Journal of Evolutionary Economics*, 14(4), 407–429.

Carpenter, J. P. & Matthews, P. H. (2004). *Social reciprocity*. IZA Discussion Paper Series 1347, Forschungsinstitut zur Erforschung der Arbeit, Bonn.

Carpenter, J. P. & Matthews, P. H. (2009). What norms trigger punishment? *Experimental Economics*, 12(3), 272–288.

Casari, M. (2005). On the design of peer punishment experiments. *Experimental Economics*, 8(2), 107–115.

Casari, M. & Luini, L. (2009). Cooperation under alternative punishment institutions: An experiment. *Journal of Economic Behavior & Organization*, 71(2), 273–282.

Cinyabuguma, M., Page, T., & Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics*, 9(3), 265–279.

Cox, J. C., Friedman, D., & Gjerstad, S. (2007). A tractable model of reciprocity and fairness. *Games and Economic Behavior*, 59(1), 17–45.

Croson, R. & Marks, M. (1998). Identifiability of individual contributions in a threshold public goods experiment. *Journal of Mathematical Psychology*, 42(23), 167–190.

de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, 305(5688), 1254–1258.

Dufwenberg, M. & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2), 268–298.

Eckel, C. C. & Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. In C. R. Plott & V. L. Smith (Eds.), *Handbook of experimental economics results* (pp. 1061–1073). Amsterdam: North-Holland.

Ertan, A., Page, T., & Putterman, L. (2009). Who to punish? individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, 53(5), 495–511.

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, 73(6), 2017–2030.

Falk, A. & Fischbacher, U. (2006). A theory reciprocity. *Games and Economic Behavior*, 54(2), 293–315.

Fehr, E. & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Science*, 8(4), 185–190.

Fehr, E. & Fischbacher, U. (2004b). Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2), 63–87.

Fehr, E. & Fischbacher, U. (2005). The economics of strong reciprocity. In H. Gintis, S. Bowles, R. Boyd, & E. Fehr (Eds.), *Moral sentiments and material interests. The foundations of cooperation in economic life* (pp. 151–192). Cambridge, Massachusetts: MIT Press.

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, 13(1), 1–25.

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

Fehr, E. & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3), 817–868.

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

Fischbacher, U. & Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, 100(1), 541–556.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, 71(3), 397–404.

Gächter, S., Herrmann, B., & Thöni, C. (2006). Cross-cultural differences in norm enforcement. *Behavioral and Brain Sciences*, 28(6), 822–823.

Gächter, S. & Renner, E. (2010). The effects of (incentivized) belief elicitation in public goods experiments. *Experimental Economics*, 13(3), 364–377.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, 206(2), 169–179.

Glazer, A. & Konrad, K. A. (1996). A signaling explanation for charity. *American Economic Review*, 86(4), 1019–1028.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

Grechenig, K., Nicklisch, A., & Thöni, C. (2010). Punishment despite reasonable doubt - a public goods experiment with sanctions under uncertainty. *Journal of Empirical Legal Studies*, 7(4), 847–867.

Greiner, B. (2004). *The online recruitment system ORSEE 2.0 - A guide for the organization of experiments in economics.* Working Paper Series in Economics 10, University of Cologne, Cologne.

Grosse, S., Putterman, L., & Rockenbach, B. (2008). *Monitoring in teams: Using laboratory experiments to study a theory of the firm.* Mimeograph, University of Erfurt, Erfurt, Germany.

Harbaugh, W. T. (1998a). The prestige motive for making charitable transfers. *American Economic Review*, 88(2), 277–282.

Harbaugh, W. T. (1998b). What do donations buy? a model of philanthropy based on prestige and warm glow. *Journal of Public Economics*, 67(2), 269–284.

Hechter, M. & Opp, K.-D., Eds. (2001). *Social norms.* New York: Russell Sage Foundations.

Holländer, H. (1990). A social exchange approach to voluntary cooperation. *American Economic Review*, 80(5), 1157–1167.

Holt, C. A. & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5), 1644–1655.

Houser, D., Schunk, D., & Winter, J. (2010). Distinguishing trust from risk: An anatomy of the investment game. *Journal of Economic Behavior & Organization*, 74(1-2), 72–81.

Jarke, J. (2011). *Enforced cooperation. Part II: Punishment.* Working paper, University of Heidelberg, Heidelberg.

Kahan, D. M. & Posner, E. A. (1999). Shaming white-collar criminals: A proposal for reform of the federal sentencing guidelines. *Journal of Law and Economics*, 42(1), 365–391.

Kalai, E. & Lehrer, E. (1993). Subjective equilibrium in repeated games. *Econometrica*, 61(5), 1231–1240.

Kalai, E. & Lehrer, E. (1995). Subjective games and equilibria. *Games and Economic Behavior*, 8(1), 123–163.

Kellaway, J. (2003). *The history of torture and execution: From early civilization through medieval times to the present.* London: Chaucer Press.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1-2), 81–93.

Laury, S. K. (2005). *Pay one or pay all: Random selection of one choice for payment.* Andrew Young School Policy Studies Research Paper 06-13, Department of Economics, Georgia State University, Atlanta.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods*, 2(1), 3–19.

Miller, A. S. & Mitamura, T. (2003). Are surveys on trust trustworthy? *Social Psychology Quarterly*, 66(1), 62–70.

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

Oechssler, J. & Schipper, B. (2003). Can you guess the game you are playing? *Games and Economic Behavior*, 43(1), 137–152.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: self-governance is possible. *American Political Science Review*, 86(2), 404–417.

Owens, J. B. (2000). Have we no shame? thoughts on shaming, 'white collar' criminals, and the federal sentencing guidelines. *American University Law Review*, 49, 1047–1058.

Patel, A., Cartwright, E., & van Vugt, M. (2010). *Punishment cannot sustain cooperation in a public good game with free-rider anonymity.* Discussion Paper in Economics 451, University of Gothenburg, Gothenburg.

Pettifer, E. W. (1992). *Punishments of former days.* Winchester, UK: Waterside Press, new edition.

Pitariu, A. H. (2007). *Monitoring in teams.* PhD thesis, University of South Carolina, Columbia, South Carolina.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, 128(2), 301–323.

Rege, M. & Telle, K. (2004). The impact of social approval and framing on cooperation in public good situations. *Journal of Public Economics*, 88(7-8), 1625–1644.

Savikhin, A. & Sheremeta, R. M. (2010). *Visibility of contributions and cost of information: An experiment on public goods.* Mimeograph, University of Chicago, Chicago, Illinois.

Sell, J. & Wilson, R. K. (1991). Levels of information and contributions to public goods. *Social Forces*, 70(1), 107–124.

Selten, R. (1967). Die strategiemethode zur erforschung des eingeschränkt rationalen verhaltens im rahmen eines oligopolexperiments. In H. Sauermann (Ed.), *Beiträge zur experimentellen Wirtschaftsforschung*, volume I (pp. 136–168). Tübingen: Mohr Siebeck.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1(1), 43–62.

Skrondal, A. & Rabe-Hesketh, S. (2003). Multilevel logistic regression for polytomous data and rankings. *Psychometrika*, 68(2), 267–287.

Soetevent, A. (2005). Anonymity in giving in a natural context - a field experiment in 30 churches. *Journal of Public Economics*, 89(11-12), 2301–2323.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15(1), 72–101.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, 51(1), 110–116.

A Tables

Table 5: Relative frequency of punishing second parties and of players expecting to be punished by a second party

		Second party punishment		
Target player is a	Treatment	2P-3P sequence	3P-2P sequence	pooled
Defector	$M0$	.643 (.133)	.875 (.125)	.727 (.097)
	$M5$	.472 (.084)	.133 (.063)	.318 (.058)
	$M10$	.231 (.084)	.300 (.105)	.261 (.065)
Cooperator	$M0$	.214 (.114)	.000 (.000)	.136 (.075)
	$M5$	.250 (.073)	.067 (.046)	.167 (.046)
	$M10$	.115 (.064)	.150 (.082)	.130 (.050)
		Expected second party punishment		
Player is a	Treatment	2P-3P sequence	3P-2P sequence	pooled
Defector	$M0$	.857 (.097)	.875 (.125)	.864 (.075)
	$M5$	.694 (.078)	.567 (.092)	.636 (.060)
	$M10$	.731 (.089)	.550 (.114)	.652 (.071)
Cooperator	$M0$	.214 (.114)	.000 (.000)	.136 (.075)
	$M5$	.306 (.078)	.100 (.056)	.212 (.051)
	$M10$	.231 (.084)	.150 (.082)	.196 (.059)

In the upper panel the conditional sample means of an indicator variable that has value 1 if the subject punished in the respective case and value 0 otherwise is tabulated. In the lower panel the conditional sample means of an indicator variable that has value 1 if the subject expected to be punished in the respective case and value 0 otherwise is tabulated. Standard errors are in parantheses.

Table 6: Mean magnitude of second party punishment and of expected second party punishment

		Second party punishment		
Target player is a	Treatment	2P-3P sequence	3P-2P sequence	pooled
Defector	$M0$	6.86 (1.73)	8.88 (2.70)	7.59 (1.45)
	$M5$	4.67 (1.10)	1.57 (0.87)	3.26 (0.73)
	$M10$	1.54 (0.85)	2.30 (1.21)	1.87 (0.71)
Cooperator	$M0$	0.93 (0.54)	0.00 (0.00)	0.59 (0.35)
	$M5$	1.53 (0.51)	0.57 (0.50)	1.09 (0.36)
	$M10$	0.31 (0.21)	0.40 (0.27)	0.35 (0.16)
		Expected second party punishment		
Player is a	Treatment	2P-3P sequence	3P-2P sequence	pooled
Defector	$M0$	6.21 (0.98)	8.50 (2.20)	7.05 (1.01)
	$M5$	5.67 (0.96)	4.53 (1.10)	5.15 (0.72)
	$M10$	4.73 (0.87)	3.35 (1.18)	4.13 (0.71)
Cooperator	$M0$	0.50 (0.36)	0.00 (0.00)	0.32 (0.23)
	$M5$	1.56 (0.45)	0.47 (0.26)	1.06 (0.28)
	$M10$	1.00 (0.46)	0.55 (0.34)	0.80 (0.30)

In the upper panel the conditional sample means of punishment points imposed on the target player are tabulated. In the lower panel the conditional sample means of punishment points expected by the target player are tabulated. Standard errors are in parantheses.

Table 7: Relative frequency of punishing third parties and of players expecting to be punished by a third party

		Third party punishment					
		if target player's coplayer cooperates			if target player's coplayer defects		
Target player is a	Treatment	2P-3P sequence	3P-2P sequence	pooled	2P-3P sequence	3P-2P sequence	pooled
Defector	$M0$	.786 (.114)	.625 (.183)	.727 (.097)	.429 (.137)	.500 (.189)	.455 (.109)
	$M5$	.417 (.083)	.300 (.085)	.364 (.060)	.361 (.081)	.133 (.063)	.258 (.054)
	$M10$	.423 (.099)	.400 (.112)	.413 (.073)	.308 (.092)	.300 (.105)	.304 (.069)
Cooperator	$M0$	.286 (.125)	.125 (.125)	.227 (.091)	.286 (.125)	.125 (.125)	.227 (.091)
	$M5$	.306 (.078)	.133 (.063)	.227 (.052)	.306 (.078)	.067 (.046)	.197 (.049)
	$M10$	.308 (.092)	.350 (.109)	.326 (.070)	.231 (.084)	.100 (.069)	.174 (.057)
		Expected third party punishment					
		if player's coplayer cooperates			if player's coplayer defects		
Player is a	Treatment	2P-3P sequence	3P-2P sequence	pooled	2P-3P sequence	3P-2P sequence	pooled
Defector	$M0$	.714 (.125)	1 (.000)	.818 (.084)	.357 (.133)	.625 (.183)	.455 (.109)
	$M5$	.528 (.084)	.700 (.085)	.606 (.061)	.444 (.084)	.367 (.089)	.409 (.061)
	$M10$	.654 (.095)	.700 (.105)	.674 (.070)	.462 (.100)	.500 (.115)	.478 (.074)
Cooperator	$M0$	.429 (.137)	.250 (.164)	.364 (.105)	.286 (.125)	.125 (.125)	.227 (.091)
	$M5$	.194 (.067)	.200 (.074)	.197 (.049)	.250 (.073)	.267 (.082)	.258 (.061)
	$M10$	.231 (.084)	.200 (.092)	.217 (.061)	.269 (.089)	.100 (.069)	.196 (.059)

In the upper panel the conditional sample means of an indicator variable that has value 1 if the subject punished in the respective case and value 0 otherwise is tabulated. In the lower panel the conditional sample means of an indicator variable that has value 1 if the subject expected to be punished in the respective case and value 0 otherwise is tabulated. Standard errors are in parantheses.

Table 8: Mean magnitude of third party punishment and of expected third party punishment

		Third party punishment					
		if target player's coplayer cooperates			if target player's coplayer defects		
Target player is a	Treatment	2P-3P sequence	3P-2P sequence	pooled	2P-3P sequence	3P-2P sequence	pooled
Defector	$M0$	6.43 (1.30)	5.25 (2.41)	6.00 (1.18)	2.14 (0.87)	2.50 (1.30)	2.27 (0.71)
	$M5$	3.64 (0.95)	1.43 (0.49)	2.64 (0.57)	2.22 (0.60)	0.83 (0.53)	1.59 (0.41)
	$M10$	2.65 (0.82)	3.05 (1.18)	2.83 (0.68)	1.69 (0.66)	1.65 (0.80)	1.67 (0.51)
Cooperator	$M0$	1.71 (0.83)	0.75 (0.75)	1.36 (0.59)	1.36 (0.75)	0.75 (0.75)	1.14 (0.54)
	$M5$	2.00 (0.62)	0.50 (0.29)	1.32 (0.37)	1.67 (0.52)	0.23 (0.18)	1.02 (0.31)
	$M10$	1.31 (0.48)	1.80 (0.68)	1.52 (0.40)	0.77 (0.33)	0.30 (0.22)	0.57 (0.21)
		Expected third party punishment					
		if player's coplayer cooperates			if player's coplayer defects		
Player is a	Treatment	2P-3P sequence	3P-2P sequence	pooled	2P-3P sequence	3P-2P sequence	pooled
Defector	$M0$	5.71 (1.26)	6.38 (1.69)	5.95 (0.99)	1.86 (0.74)	4.75 (1.66)	2.91 (0.80)
	$M5$	3.64 (0.77)	3.83 (0.79)	3.73 (0.55)	3.17 (0.71)	1.70 (0.53)	2.50 (0.46)
	$M10$	4.35 (0.95)	3.10 (0.89)	3.80 (0.66)	2.62 (0.77)	1.20 (0.36)	2.00 (0.47)
Cooperator	$M0$	1.86 (0.64)	0.63 (0.42)	1.41 (0.45)	1.00 (0.53)	0.63 (0.63)	0.86 (0.40)
	$M5$	0.78 (0.32)	1.17 (0.52)	0.95 (0.29)	1.50 (0.50)	1.43 (0.63)	1.47 (0.39)
	$M10$	1.08 (0.48)	0.65 (0.32)	0.89 (0.30)	1.65 (0.71)	0.20 (0.16)	1.02 (0.42)

In the upper panel the conditional sample means of punishment points imposed on the target player are tabulated. In the lower panel the conditional sample means of punishment points expected by the target player are tabulated. Standard errors are in parantheses.

Table 9: Relative frequency of monitors that actually monitor and of players expecting to be monitored

		Monitoring		
Role	Treatment	2P-3P sequence	3P-2P sequence	pooled
Second party	$M0$	.643 (.133)	.875 (.125)	.727 (.097)
	$M5$	.333 (.080)	.067 (.046)	.212 (.051)
	$M10$	.115 (.064)	.150 (.082)	.130 (.050)
Third party	$M0$	.786 (.114)	.500 (.189)	.682 (.102)
	$M5$	.139 (.058)	.200 (.074)	.167 (.046)
	$M10$	.154 (.072)	.200 (.092)	.174 (.057)
		Expected monitoring		
Monitor	Treatment	2P-3P sequence	3P-2P sequence	pooled
Second party	$M0$	.857 (.097)	.875 (.125)	.864 (.075)
	$M5$	.611 (.082)	.500 (.093)	.561 (.062)
	$M10$	.538 (.100)	.500 (.115)	.522 (.074)
Third party	$M0$	.786 (.114)	1 (0)	.864 (.075)
	$M5$	.472 (.084)	.667 (.088)	.561 (.062)
	$M10$	.654 (.095)	.700 (.105)	.674 (.070)

In the upper panel the conditional sample means of an indicator variable that has value 1 if the subject monitored in the respective case and value 0 otherwise are tabulated. In the lower panel the conditional sample means of an indicator variable that has value 1 if the subject expected to be monitored in the respective case and value 0 otherwise are tabulated. Standard errors are in parantheses.