University of Heidelberg

Department of Economics

Discussion Paper Series | No. 545

Non-Strategic Punishment when Monitoring is Costly: Experimental Evidence on Differences between Second and Third Party Behavior

Timo Goeschl and Johannes Jarke

June 2013

Non-Strategic Punishment when Monitoring is Costly: Experimental Evidence on Differences between Second and Third Party Behavior*

TIMO GOESCHL† Heidelberg University

JOHANNES JARKE‡ Hamburg University

June 11, 2013

Abstract

This paper studies monitoring and punishment behavior by second and third parties in a cooperation experiment with endogenous information structures: Players are uninformed whether the target player cooperated or defected at the cooperation stage, but can decide to resolve the information imperfection at non-negative cost at the punishment stage. We examine how monitoring and punishment respond to changes in monitoring costs, and exploit the evidence to gain new insights about commonalities and differences between second and third party behavior. We establish three effects of positive monitoring costs relative to the zero-cost baseline and find that each one affects third parties differently than second parties: A «direct punishment cost effect» (the supply of non-strategic punishment decreases), a «blind punishment effect» (players punish without resolving the information imperfection) and a «diffusion effect» (defectors make up a smaller share of the punished and receive weaker punishment). The first effect affects third parties less, the other two more. As a result, third party punishment leads to increasingly weaker incentives for cooperation relative to second party punishment as monitoring costs rise. In addition, the differences between second and third parties suggest the presence of a «pure role effect»: Taking into account elicited beliefs and risk preferences, third parties punish *differently* from second parties, not just more weakly. (*JEL* C92, C72, D03, D80, Z13)

Keywords: monitoring; punishment; sanctions; information; cooperation

*We are grateful to the German Federal Ministry of Education and Research and the University of Heidelberg for financial support. Valuable inputs by (in alphabetical order) Catherine Eckel, Marco Faravelli, Dirk Engelmann, Ernst Fehr, Urs Fischbacher, Alia Gizatulina, Freddy Huet, William Neilson, Cornelia Neuert, Jean-Robert Tyran and Rick Wilson, as well as various participants at the 2nd Thurgau Experimental Economics Meeting, the 15th Annual Conference of The International Society for New Institutional Economics at Stanford University, the 2011 Annual International Meeting of the Economic Science Association at the University of Chicago, the International Association for Research in Economic Psychology 2011 Conference at the University of Exeter, the 65th European Meeting of the Econometric Society at the University of Oslo and seminars at the Universities of Mannheim, Heidelberg, East Anglia, Leuven, Regensburg, and Tel Aviv (Bar-Ilan University) are appreciated. All remaining errors are ours.

†Alfred-Weber-Institute of Economics, Bergheimer Strasse 20, D-69115 Heidelberg, Germany. Phone: +49 6221 54 8010. E-mail: goeschl@eco.uni-heidelberg.de.

‡Corresponding author. School of Business, Economics and Social Sciences, Department of Socioeconomics, Welckerstrasse 8, D-20354 Hamburg, Germany. Phone: +49 40 42838 8768. E-mail: johannes.jarke@wiso.uni-hamburg.de.

1 Introduction

Social norms, normative standards of behavior that are enforced by informal social sanctions, exist in every human society, and economists increasingly recognize their importance in accounting for cooperative behavior in a broad range of economic contexts such as team production, trade, credit or resource management. A recent stream of experimental research has produced valuable insights into how they are enforced by non-materially motivated social sanctions (see Balliet et al., 2011, for an overview). In this, *observability* of behavior has emerged as a critical condition for social sanctions. As Elster (2009) argues, the application of social rewards and punishments is difficult if there is uncertainty about the target behavior.[1] However, whether and to what extent a potential sanctioner is in the position to observe, or otherwise acquire information about the target individual's behavior, is rarely fixed exogenously. In many situations, information on the target individual's actions is not immutably imperfect, but players are in a position to augment available information through costly effort, either *ex ante* (e.g. putting oneself in a better «vantage point») or *ex post* (e.g. collecting evidence). In other words, a player's choice whether to impose social sanctions on another player is frequently preceded by a *choice* on whether to invest resources into monitoring his or her actions.[2]

In the present paper we report on an experiment that exploits this insight to inform our understanding of how non-strategic punishment by second and third parties is affected by information imperfections and the availability of costly remedies. Since the choice whether to observe the behavior of the target player is costly, the setting also lends itself as an eliciting device with which to learn more about the commonalities and differences between non-strategic punishment by second parties and third parties. These commonalities and differences, and the motivational constructs that underpin them, are poorly understood despite the argument that social sanctions by materially unaffected third parties are of particular importance to social norms in larger communities (e.g. Bendor and Mookherjee, 1990; Kandori, 1992; Bendor and Swistak, 2001). Since the vast majority of the literature focuses on sanctions imposed by directly affected second parties, we know little more than that third parties usually punish more weakly than second parties.[3] To advance our understanding of whether and how players' behavior as second or third parties differs, we draw on a seminal experimental paradigm of Fehr and Fischbacher (2004) that is designed to study both non-strategic second and third party punishment in the laboratory. The design assures that, since punishment is costly and there are no conceivable pecuniary returns for sanctioning parties from punishment, any punitive behavior must be driven by subjective benefits alone. We extend this paradigm in a new direction by allowing important parts of the information structure of the game to be *endogenously* determined by the players. Namely, while in Fehr and Fischbacher (2004) sanctioning parties were automatically and freely informed about the target player's behavior before they had the opportunity to punish, in our experiment they decide whether to acquire this information at non-negative cost. This gives rise to three different

[1] In fact, recent experimental research suggests that the efficacy of social sanctions is hampered by exogenous restrictions on observability (Carpenter, 2007b; Bornstein and Weisel, 2010; Grechenig et al., 2010; Ambrus and Greiner, 2012).

[2] Such monitoring effort to overcome imperfect information on coplayers' actions is well known in a variety of economically relevant contexts, such as shared resource management (Ostrom, 1990; Ostrom and Gardner, 1993; Seabright, 1993; Rustagi et al., 2010), production teams (Alchian and Demsetz, 1972; Kandel and Lazear, 1992; Dong and Dow, 1993; Craig and Pencavel, 1995) and alliances (Acheson, 1975, 1987, 1988; Palmer, 1991), labor relations (Shapiro and Stiglitz, 1984; Kanemoto and MacLeod, 1991; Lazear, 1993), micro-finance (Armendáriz and Morduch, 2005) or neighborhood watch (Sampson et al., 1997), to name just a few.

[3] A noteworthy recent attempt to understand more about motivations underlying second and third party punishment has been made by Leibbrandt and López-Pérez (2012).

types of behavior. If players decide not to acquire the information, their decision in the sanctioning stage remains uninformed about the target player's action at the cooperation stage of the game. In this case, if they decide not to punish, we term their behavior «non-punishment». If they punish despite being uninformed about the target player's action, we term this «blind punishment». By choosing to acquire information they are able to condition their sanctioning on complete information regarding the coplayer's action, a behavior we term «targeted punishment».

On the basis of this design, we compare second and third party sanctioning behavior in a baseline condition with zero information costs and two conditions with different levels of strictly positive information costs. We focus on three main questions: First, what is the information acquisition behavior when observing coplayers' actions is costly, and how do second and third parties differ? Second, how does information acquisition behavior respond to changes in the price of information, and how do second and third parties differ? Third, how does the endogeneity of information impact on the occurrence and incidence of mutual sanctions, and how do these patterns differ between a second and third party sanctioning setting?

We establish three effects of positive monitoring costs relative to the zero-cost baseline and find that each one affects third parties differently than second parties. First, information costs have the impact of decreasing the gross supply of punishment. We call this the «direct punishment cost effect». Second, while virtually all sanctioning is discriminate at zero information costs, in the presence of positive information costs a distinct share of subjects punish without resolving the information imperfection rather than refraining from sanctioning. We term this the «blind punishment effect». The emergence of costly punishment that is *deliberately* blind when both not punishing and targeted punishment are available alternatives is a new finding of this paper. Finally, positive information costs lead to defectors making up a smaller share of the punished and receiving weaker punishment compared to zero information costs, with the result that incentives get too weak to render cheating unprofitable. We call this loss of clear focus on defectors at the punishment stage the «diffusion effect».

We also provide a comparative assessment to test whether and how players' behavior is influenced by the three effects depending on their role as second or third parties. If the drivers of third party punishment are simply weaker than those of second party punishment, then compared to second parties an increase in monitoring costs could be expected to lead to a greater reduction in punishment supplied and less blind punishment. In our experiment, however, we observe that the «direct punishment cost effect» affects third parties less than second parties while the «blind punishment effect» and the «diffusion effect» are stronger for third parties. As a result, third party punishment leads to increasingly weaker incentives for cooperation relative to second party punishment as monitoring costs rise. In addition, the differences between second and third parties suggest the presence of a pure role effect: Taking into account elicited beliefs and risk preferences, third parties punish *differently* from second parties, not just more weakly.

In the remainder of the paper we proceed as follows. In section 2 we describe the aim, design, and procedures of the experiment. We proceed to present the results in section 3. We conclude in section 4.

2 The Experiment

The aim of the design is to generate a setting in which costly punishment behavior under endogenous monitoring by second and third parties can be compared. The previous literature has been careful to identify and isolate non-strategic sanctioning. In keeping with this focus,

we design the experiment such that strategic and altruistic motives for monitoring coplayers' actions are ruled out as far as possible and the observed monitoring and sanctioning should be driven only by subjective rewards to the individual carrying out these activities.

We draw on the seminal study by Fehr and Fischbacher (2004), who use a two-stage design consisting of a cooperation stage and a sanctioning stage. The innovative step in our experiment consists of adding a monitoring decision into the sanctioning stage, in which sanctioning parties decide whether to give up material rewards in order to receive full information about the actions of their target player at the cooperation stage. If they choose not to buy the information, they forgo the option of being able to discriminate between a cooperating and defecting target player at the sanctioning stage. If they do, sanctioning can be conditioned on observed target player behavior. We exogenously vary the magnitude of this fee as our treatment variable to observe how behavior is adjusted.

2.1 Experimental game form

Each subject played two two-stage experimental games, a second party monitoring game (SPMG) and a third party monitoring game (TPMG) with randomly matched and unknown coplayers in random sequence. Both games had two stages, wheres the first stage was identical in both games. In the first stage, both players in a given group of two were endowed with 10 tokens and interacted with one another in a strategic game of the Prisoner's Dilemma type with monetary payoffs. Subjects' first-stage decision was binary: Players could either keep their tokens or transfer all of them to the other group member, in which case the experimenter tripled them.

The second stage differed between the SPMG and the TPMG. At the beginning of the second stage, each player received an additional endowment of 40 tokens in both games. Subsequently, in the SPMG both players in each group could choose to monitor the coplayer's action in the first stage and possibly punish the other. In the TPMG, the first (second) player in a group had the opportunity to monitor and punish a player from another (a third) group. The matching protocol was identical to the one used by Fehr and Fischbacher (2004) and rules are any reciprocity between third party monitors. In order to avoid potential biases due to a variable cost-impact ratio (see Casari, 2005), we chose a linear sanctioning technology commonly employed in the literature (Fehr and Gächter, 2002; Fehr and Fischbacher, 2004; Gächter et al., 2008): any token spent by player i for punishing player j subtracted three tokens from the latter's payoff. Expenditures were restricted to integers and damage was limited to sixty tokens.

In both games, subjects reported their choices in the monitoring and punishment stages jointly using a modified version of the strategy method. Specifically, in the SPMG players announced a contingent punishment plan, indicating the number of deduction points for each of the target player's possible transfer decisions, without being informed about the actual decision (see Brandts and Charness, 2010, for a detailed exposition of the strategy method). The presence of a monitoring decision was implemented in the strategy method by essentially allowing subjects a choice between two punishment plans. If a subject conditioned punishment on the target player's action, a payment of an exogenously set monitoring fee *in addition* to the applicable punishment costs fell due. The alternative was a punishment plan in which the subject did not condition the punishment on the target player's action. This plan did not involve a fee over and above the cost of punishment. The advantage of this design is that all subjects in all conditions were confronted with *exactly* the same task and instructions that differed *only* in one number: the fee level.

Likewise, in the TPMG each third party, while being informed about the first stage trans-

fer decision of the other member in their own group, had to indicate a punishment plan over four possible contingencies (one for each possible transfer combination in the other group). Again, any punishment conditioned on the target player's first stage behavior required payment of the monitoring fee in addition to the applicable punishment costs. After the punishment stage, both games ended.

2.2 Additional tasks

We supplemented the two games by incentivized elicitation of first-order beliefs and risk preferences. Both are potentially relevant in accounting for observed behavior (Keser and van Winden, 2000; Fischbacher et al., 2001; Fischbacher and Gächter, 2010). Since subjects receive no actual information in the course of play, they respond to their beliefs on their coplayer's behavior, and this response involves a decision under strategic risk. Thus, in addition to the games described above, every subject performed the following tasks, which were identical in each treatment and for every subject. Following their decision in the first stage, we asked subjects to state their belief about the transfer decision of the target player and, following their decision in the second stage, their belief about the punishment they expected to receive. Those statements were incentivized with a opportunity to earn extra tokens in case of accurate predictions (see appendix).

At the end of each session we employed the Holt-Laury lottery choice method (Holt and Laury, 2002) to obtain an indication of each subject's risk preferences (see appendix). Subjects were only informed about this task after the two experimental games. It should be noted that *a priori* it is possible that the method may not capture the specific kind of risk involved in the situation, but since it is the most common method to elicit risk preferences we thought it is worth investigating.

2.3 Design

Our primary focus is on the relationship between information costs and monitoring and punishment behavior. The objective of the treatments is to introduce an exogenous variation in the cost of information acquisition with the aim of forcing a materially consequential choice by subjects regarding the desired information status prior to the punishment decision. Specifically, we employed between-subjects variation in the monitoring fee and examine the changes in monitoring and punishment behavior associated with this variation, both in the SPMG and the TPMG. Evidence on monitoring and sanctioning activity when monitoring is costless, i.e. when the fee is zero, is an obvious starting point: It provides a direct replication of the PD experiment by Fehr and Fischbacher (2004) and therefore a meaningful baseline treatment. In what follows, this treatment is labeled $M0$. Given the payoff structure of the game, we expected a fee of ten tokens close to prohibitive such that we have chosen this level as a reasonable boundary of the considered fee interval. In order to achieve efficient identification, and to account for possible non-linearities, we implemented three conditions with fee levels set at the extremes (zero and ten tokens, respectively) and the midpoint (five tokens) of the interval 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; List et al., 2011, for efficient sample arrangement methods). The two treatment conditions are labeled $M5$ and $M10$, respectively. In order to counterbalance for possible order effects, one part of the subjects in each treatment cell played the SPMG before the TPMG, the other part the other way around.[4]

[4]It turned out that there were no systematic differences between sequences anyway. Using Mann-Whitney tests we find no significant differences among sequences in monitoring of second parties ($p = .160$) and third parties

2.4 Subjects and procedures

All experiments were conducted at the experimental laboratory of the Alfred-Weber-Institute (AWI-Lab) at Heidelberg University 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, of which 49.3 percent were female. A share of 64.2 percent had never participated in a laboratory experiment before. The mean age was 21.8 years. No subject participated in more than one session and treatment. Based on the sample arrangement described above, 22 subjects were assigned to $M0$, 66 to $M5$, and 46 to $M10$.

Upon entering the laboratory subjects were randomly assigned to the computer terminals. Direct communication among them was not allowed for the duration of the entire session. Booths separated the participants visually, ensuring that they made their decisions anonymously and independently. Furthermore, subjects did not receive any information on the personal identity 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 the procedural rules. The experiment was framed in a sterile way using neutral language and avoiding value laden terms in the instructions (see supplementary material). 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.

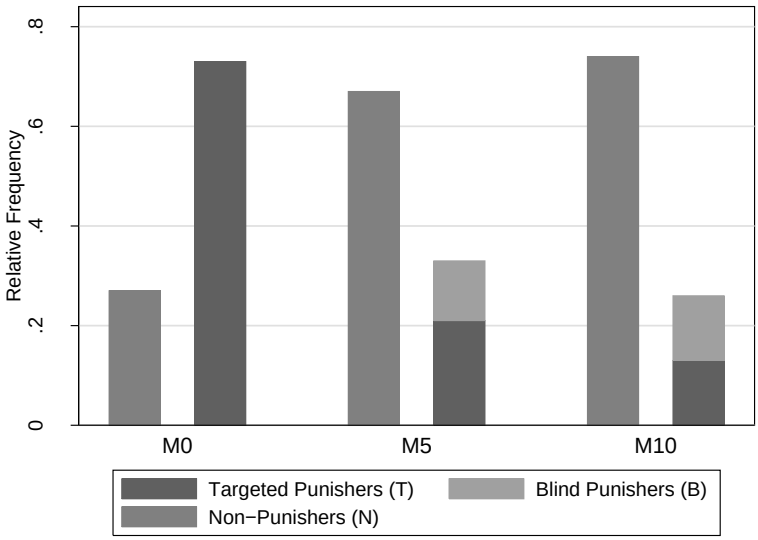
The experiment was programmed and conducted with z-Tree (Fischbacher, 2007). The exact timing of events was as follows. First, the subjects were randomly matched into groups of two. Then each subject made her or his decisions and reported expectations in the SPMG or TPMG, depending on the sequence.[5] 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 game. After being randomly re-matched into new groups, each subject made her or his decisions and reported expectations in the TPMG or SPMG, depending on the sequence. After being informed about the payoffs, the experimenter announced that another (and definitely final) experiment will now be conducted and distributed additional instructions that explained the supplementary lottery choice task. Thus, subjects did not know until this announcement that the experiment involved another task in order to rule out confounding effects of the lottery-choice task on the main experimental games. We cannot rule out reverse confounds which is, however, of minor significance. 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 (according to a random number matched to their decisions; no personal identities were used throughout the whole experiment) and dismissed.

In every session subjects received a fixed show-up fee of €2, which was not part of their endowment. The whole experiment lasted approximately 90 minutes and subjects earned an average of €18.46 (€0.10 per token earned), including the fixed show-up fee. Earnings exceed the local average hourly wage of a typical student job and can hence be considered meaningful to the participants.

(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). Behavior was significantly different between sequences with respect to third party punishment of cooperators when the target's coplayer defected (p = .006). Fisher exact tests yield similar results which are available on request.

[5] Until the end of the first game, subjects did not know that another game will follow. This ambiguity was a deliberate design choice aimed at minimizing confounding effects of the second on the first game.

Figure 1: Distribution of behavioral types at different monitoring cost levels among second parties.



3 Results

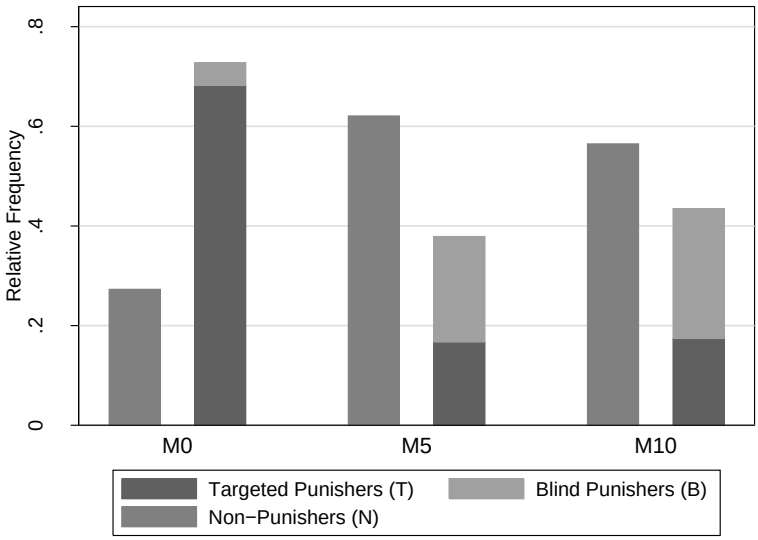
The analysis of the experimental data will focus on three main behavioral types. One type are «non-punishing» subjects that decide not to monitor and do not punish. These subjects are referred to as behavioral type N. Among the punishing subjects, there are two types. «Blind punishers» devote no resources to monitoring, but still punish indiscriminately. This type is denoted as B. The other type T are «targeted punishers» that devote resources both to monitoring and to punishment. The B-types and T-types together form the class of «punishers», denoted P. The forth possible type (monitoring without punishment) is excluded by design.[6]

We organize the presentation of the experimental results in the following way. We begin by discussing our results from the baseline treatment that replicate previous experimental evidence on sanctioning behavior. Against this benchmark, we then describe the core results of the experiment, namely the «punishment cost effect», the «blind punishment effect», and the «diffusion effect» resulting from costly monitoring, and the respective differences between second and third parties. Finally, we provide additional results that are informative with respect to the underlying drivers of those differences.

Figures 1 and 2 depict the relative frequencies of behavioral types N, B, and T for the three monitoring fee levels in the SPMG and the TPMG, respectively. Normalization is required on account of the between-subjects design with a variable number of subjects in each treatment. A readily apparent first diagnosis of the evidence is that variations in the cost of information are associated with different patterns of monitoring and punishment behavior. The detailed nature of these differences is the subject of the following sections.

[6]We did this in order to have control over the possible motivations underlying monitoring: in our setting monitoring was only possible for «instrumental» reasons in the sense of using this information in order to impose punishments discriminately.

Figure 2: Distribution of behavioral types at different monitoring cost levels among third parties.



3.1 Punishment with costless monitoring: Replication of previous findings

The baseline treatment *M*0 replicates the experiment run by Fehr and Fischbacher (2004) which is among the seminal ones in a recent stream of research on costly sanctioning.[7] These experiments find that, under perfect observability, there is a marked propensity among subjects to reduce others' incomes («punish») even at a personal net cost, both as a second and as a third party. Even after eliminating strategic motives for punishment, the share of subjects that apply costly sanctions consistently exceeds one half. Fehr and Fischbacher (2004) find that 67 percent of second parties and 59 percent of third parties applied costly sanctions on cheaters (unilateral defectors). Our results, displayed in the left-hand column in figure of figures 1 and 2, respectively, are consistent with theirs.

Result 1. *When information costs are zero, most second and third parties make use of the non-strategic punishment option. The fraction of punishers is about equal among second and third parties.*

In particular, we find a large majority of 73 percent (16 out of 22) willing to impose costly sanctions on their coplayers for non-strategic reasons, both among second and third parties.[8] The class of punishers *P* therefore dominates when there are no monitoring costs.

Previous experiments have also generated evidence with respect to the direction of punishment under perfect observability: it (i) is predominantly supplied by cooperators and (ii) is predominantly targeted upon cheaters. With respect to the origin of punishment, Fehr and Fischbacher (2004) do not report the composition of the group of punishers, but the following numbers are indicative: 69 percent of cooperating second parties and 68 percent (36 percent) of cooperating third parties imposed sanctions on unilateral (mutual) defectors as

[7] See Gächter and Herrmann (2009), Balliet et al. (2011), and Bowles and Gintis (2011) for recent overviews.

[8] It should be noted the 16 subjects are not identical in the SPMG and TPMG, respectively. Two subjects who did not impose sanctions as a third party did so as a second party. Likewise, two subjects who did not impose sanctions as a second party did so as a third party. We come back to this below.

opposed to 50 percent of defecting second parties and 40 percent (27 percent) of defecting third parties imposing sanctions on other unilateral (mutual) defectors. The two groups of punishers diverge even more in terms of the volume of sanctions. There, cooperating second parties punishing unilateral (mutual) defectors spent on average 9.2 tokens and cooperating third parties on average 3.7 tokens (1.7 tokens) on sanctions while defecting second parties punishing unilateral (mutual) defectors spent 2.7 tokens and defecting third parties spent 1.9 tokens (0.9 tokens).

Result 2. *When information costs are zero, most second and third party punishment is supplied by cooperators.*

In the baseline condition of the SPMG, 19 punishment actions were carried out by 14 cooperators and 2 defectors.[9] Of these punishment actions, 15 (79 percent) were taken by cooperators and 4 (21 percent) by defectors. In the baseline condition of the TPMG, 36 punishment actions were carried out of whom 28 (78 percent) were taken by cooperating third parties and 8 (22 percent) by defecting third parties. Taking intensity into account, cooperating second parties accounted for 84 percent and cooperating third parties for 70 percent of total investment into punishment while defecting second parties accounted for 16 percent, defecting third parties for 30 percent. Among the class of punishers, therefore, cooperators clearly supplied most of the punishment.

With respect to targeting, previous experiments have generated clear evidence that when information about coplayer behavior at the cooperation stage is perfect, then punishment is, on average, targeted on cheaters. For example, Fehr and Gächter (2000) find evidence of a clear relationship between deviation from average contribution in a public goods game. Fehr and Fischbacher (2004) also find that behavior at the cooperation stage is a highly significant predictor of receiving punishment. But previous experiments have also uncovered considerable heterogeneity among subjects with respect to sanctioning behavior.[10] For example, alongside the targeted punishment of defectors by cooperators, punishment of cooperators is a less frequent but also robust empirical regularity (Cinyabuguma et al., 2006; Gächter et al., 2006; Herrmann et al., 2008; Ertan et al., 2009; Gächter and Herrmann, 2010). In Fehr and Fischbacher (2004), 8.3 percent of cooperators were targeted at the punishment stage of the second-party punishment game, and 15 percent (independently from the behavior of the target player's coplayer) in the third-party punishment game. Our experiment replicates these findings closely.

Result 3. *When information costs are zero, second and third party punishment is generally targeted on cheating, but third party punishment is less targeted and weaker than second party punishment. On average, cheating is unprofitable in both the SPMG and the TPMG.*

Note that observability (monitoring) is a precondition for targeting. The result therefore means that the information about the target player's first stage behavior is, if it is costless to acquire, used to apply punishment discriminately. In fact, in $M0$ *every* second party and 15 out of 16 third parties that imposed a punishment used the information about the target player's first stage behavior. In terms of our behavioral types, there are 73 percent T-types, 27 percent N-types, and no B-types among the second parties, and 68 percent T-types, 27

[9] Recall that the strategy method allows each subject to specify a punishment up to two times in the SPMG and up to four times in the TPMG, once for each possible first stage action profile given the subject's own action.

[10] In an experiment designed to differentiate between a variety of different motivations to impose costly sanctions, Leibbrandt and López-Pérez (2009, 2011) find substantial heterogeneity, with a combination of inequity aversion, self-interest, spite and direct reciprocal motivations accounting for observed patterns best.

percent N-types, and 5 percent B-types among the third parties.[11] Thus, second and third party behavior exhibits almost identical distributions of our behavioral types when monitoring is costless.

However, while we will see below that monitoring costs create significant differences in those distributions, this type of analysis hides some important differences between second and third party behavior present in $M0$ already. First, it is illuminating to compare second and third party behavior at the individual level. There, 17 out in 22 (77 percent) revealed the same behavioral type as both a second and a third party. The majority of those (13 subjects, or 76 percent) are universal T-types, that is, T-types as both a second party and as a third party. The rest (4 subjects, or 24 percent) are universal N-types. There exist no universal B-types when monitoring is costless. Of the remaining five subjects that revealed a different behavioral type as a second and a third party, respectively, two where type T as a second party and type N as a third party, two the reverse, and one subject was type T as a second party and type B as a third party.

A second important difference is the stiffness of punishment and the degree of targeting. In 16 out of 19 second party punishment decisions (84 percent) and in 26 out of 36 third party punishment decisions (72 percent) the target of a punishment action was a defector (in the latter case, 16 actions against unilateral defectors and 10 actions against mutual defectors). Cooperators were the target of 14 percent of second party punishment actions and of 28 percent of third party punishment actions. Taking intensity into account, in the SPMG defectors received 93 percent of all the punishment imposed while cooperators received 7 percent. In the TPMG defectors received 77 percent of all the punishment imposed, 60 percent unilateral defectors and 17 percent mutual defectors, while cooperators received 23 percent. In terms of average punishments received, cheaters received damages of 22.8 tokens on average in the SPMG and 18.0 tokens in the TPMG, while cooperators received on average 1.8 tokens in the SPMG and 4.8 tokens in the TPMG.[12]. Thus, third party punishment is somewhat weaker and less targeted, although the average net punishment of cheating exceeded the gain (10 tokens) by a sufficient margin in both the SPMG ($22.8 - 1.8 = 21$ tokens) and the TPMG ($18.0 - 4.8 = 13.2$ tokens).[13]

3.2 Monitoring and punishment with positive information costs

Results 1 through 3 underline the benchmark quality of the baseline treatment by providing additional evidence to support similar findings in the literature. Against this benchmark, we now report results on the behavioral implication of making monitoring, the acquisition on the target player's first stage action, costly. We establish three effects of positive monitoring costs relative to the zero-cost baseline and show that each one affects third parties differently than second parties: A «direct punishment cost effect» that is stronger in second parties (result 4), as well as a «blind punishment effect» (result 5) and a «diffusion effect» (result 6) which are

[11] Third parties also make a difference on whether defection of the target player is mutual or unilateral, that is, they also monitored the target player's coplayer. Among the 15 T-types, 13 (or 87 percent) also conditioned punishment on the first stage behavior of the target player's coplayer.

[12] Defectors whose coplayer also defected received average punishments of 6.9 tokens in the TPMG.

[13] At the individual level, of the 18 subjects that imposed punishments on cheaters, eight (44 percent) invested strictly more in punishment in the role of a second party than in the role of a third party, with an average expenditure difference of 6.6 tokens. Six (33 percent) made no difference. Thus, a majority of 14 subjects (78 percent) indeed punished at least as strong as a second party than as a third party. The remaining four subjects (22 percent), however, punished strictly stiffer in the role of a third party than in the role of a second party, with an average expenditure difference of 4.5 tokens. Although this individual level comparison reveals some additional heterogeneity, the conclusion that most people are willing to spend more on punishment when it is themselves who is cheated compared to when it is someone else is reinforced.

stronger in third parties, respectively. As a result, third party punishment leads to increasingly weaker incentives for cooperation relative to second party punishment as monitoring costs rise (result 7).

When monitoring is costly, the relative shares of behavioral types diverge from the baseline in two important ways. First, a higher overall cost of sanctioning is associated with a higher share of second and third parties that neither monitor nor punish, which is in line with previous evidence showing that increasing cost of punishment reduce the supply of sanctions (Suleiman, 1996; Oosterbeek et al., 2004; Anderson and Putterman, 2006; Carpenter, 2007a; Egas and Riedl, 2008).

Result 4. *In the presence of positive information costs, the propensity to punish decreases relative to the baseline stronger in second than in third parties.*

We term this the «punishment cost effect». Figures 1 and 2 illustrate this first finding: Among second parties, the share of subjects that neither monitor nor punish (type N) increases from 27 percent in $M0$ to 67 percent in $M5$ and 74 percent in $M10$. The hypothesis that the frequency of N-types and monitoring costs are independent can be clearly rejected (Kruskal-Wallis test, $p < .001$). Pairwise, only the former change is statistically significant (Mann-Whitney test, $p = .001$), the latter is not ($p = .414$).

Likewise, among third parties the share of type-N subjects increases from 27 percent in $M0$ to 62 percent in $M5$ and 57 percent in $M10$. The hypothesis that the frequency of N-types and monitoring costs are independent can again be clearly rejected (Kruskal-Wallis test, $p = .017$), whereas pairwise, only the change associated with increasing the fee from zero to five is statistically significant (Mann-Whitney test, $p = .005$), the change associated with increasing the fee from five to ten is not ($p = .554$). Comparing second and third parties, it is evident that the share of N-types is slightly less increasing in third than in second parties, both absolutely (30 vs. 47 percentage points) and relatively (111 vs. 174 percent). The frequency of N-types is apparently not different between second and third parties in the baseline condition, but significantly different when monitoring is costly (Wilcoxon signed-rank test, $p = .016$; $p = .366$ in $M5$ and $p = .011$ in $M10$). As an alternative demonstration, a dummy indicating treatment (both costly monitoring conditions) is significantly positively correlated with a dummy indicating type N (Phi/Cohen's $w = 0.282$, $p = .000$, or Kendall's $\tau_b = 0.282$, $p = .000$), whereas this correlation is somewhat stronger for second parties (Phi/Cohen's $w = 0.325$, $p = .000$, or Kendall's $\tau_b = 0.325$, $p = .000$) than for third parties (Phi/Cohen's $w = 0.242$, $p = .005$, or Kendall's $\tau_b = 0.242$, $p = .005$).

Second, in addition to the result on the propensity to punish at all, our design explicitly allows to study punitive behavior of those individuals that opt to remain ignorant about their target player's first stage behavior.

Result 5. *In the presence of positive information costs, the propensity to punish indiscriminately increases relative to the baseline stronger in third than in second parties.*

We term this the «blind punishment effect». Figures 1 and 2 report the share of blind punishers (type B) that do not acquire information and yet impose sanctions. Among second parties, this share increases from zero in the baseline to 12 and 13 percent as the monitoring costs increase to 5 and 10 tokens, respectively. The former change is statistically marginally significant (Mann-Whitney test, $p = .089$), the latter not ($p = .885$). However, both frequencies under positive monitoring costs are statistically significantly different from zero (Wilcoxon signed-rank tests, $p = .005$ and $p = .014$, respectively). The frequency of indiscriminate punishers as a fraction of all punishers increases successively from zero in $M0$ to 36 percent in $M5$ and to 50 percent in $M10$.

Among third parties, the share of blind punishers increases from 5 percent (one out of 22) in the baseline to 21 and 26 percent as the monitoring costs increase to 5 and 10 tokens, respectively. Here, the change associated with increasing the fee from zero to five is statistically marginally significant (Mann-Whitney test, $p = .073$), the change associated with increasing the fee from zero to ten is statistically significant ($p = .036$), while the change associated with increasing the fee from five to ten is not significant ($p = .550$). However, both frequencies under positive monitoring costs are again statistically significantly different from zero (Wilcoxon signed-rank tests, $p < .001$, respectively). The frequency of indiscriminate punishers as a fraction of all punishers increases successively from 6 percent in $M0$ to 56 percent in $M5$ and to 60 percent in $M10$. Again comparing second and third parties, it is evident that third parties have clearly a stronger tendency to punish indiscriminately than second parties when monitoring is costly (absolute increase of B-types 21 vs. 13 percentage points). The frequency of B-types is not significantly different between second and third parties in the baseline condition (Wilcoxon signed-rank test, $p = .317$), but significantly different when monitoring is costly (Wilcoxon signed-rank test, $p = .014$; $p = .083$ in $M5$ and $M10$). Again, as an alternative demonstration, a dummy indicating treatment (both costly monitoring conditions) is significantly positively correlated with a dummy indicating type B (Phi/Cohen's $w = 0.160$, $p = .009$, or Kendall's $\tau_b = 0.160$, $p = .009$), whereas this correlation is somewhat weaker for second parties (Phi/Cohen's $w = 0.151$, $p = .080$, or Kendall's $\tau_b = 0.325$, $p = .082$) than for third parties (Phi/Cohen's $w = 0.172$, $p = .050$, or Kendall's $\tau_b = 0.172$, $p = .043$).

A corollary of the previous two results is that the share of targeted second party punishers (type T) decreases from 73 percent for costless monitoring to 21 percent in $M5$ and 13 percent in $M10$, while the third party T-types decreases from 68 percent in $M0$ to 17 percent in $M5$ and $M10$, respectively. Again, in both cases the hypothesis that the frequency of T-types and monitoring costs are independent can be clearly rejected (Kruskal-Wallis tests, $p < .001$). Pairwise the changes associated with increasing the fee from zero to five and zero to ten, respectively, are statistically significant (Mann-Whitney tests, $p < .001$ and $p < .001$, respectively, in second parties; $p = .000$ and $p = .036$, respectively, in third parties), while the change associated with increasing the fee from five to ten is not ($p = .269$ in second parties, $p = .920$ in third parties). A dummy indicating treatment (both costly monitoring conditions) is significantly negatively correlated with a dummy indicating type T (Phi/Cohen's $w = 0.447$, $p = .000$, or Kendall's $\tau_b = -0.447$, $p = .000$), whereas this correlation is a little stronger for second parties (Phi/Cohen's $w = 0.459$, $p = .000$, or Kendall's $\tau_b = -0.459$, $p = .000$) than for third parties (Phi/Cohen's $w = 0.436$, $p = .000$, or Kendall's $\tau_b = -0.436$, $p = .000$).

We now proceed to the second step of the analysis by investigating the implications for intensity and direction of second and third party punishment, respectively. Recall (result 3) that in the baseline condition both second and third party punishment was clearly targeted on cheating, but that the latter was somewhat weaker and less targeted than the former. It turns out that the first aspect is weakened and that the second is strengthened when monitoring is costly.

Result 6. *At positive information costs, cooperators remain the dominant origin of the direction of second and third party punishment relative to the baseline, but cheaters dominate less as its target.*

We term the latter the «diffusion effect». Concerning the origin of punishment action, the fraction of P-types among second parties was between 29 ($M10$) and 31 percent ($M5$) among the cooperators and between 18 ($M10$) and 39 percent ($M5$) among the defectors.[14] Stated differently, the group of second party punishers (P-types) consisted out of 68 percent

[14] Among second parties, 31 percent (15 out of 48) and 29 percent (10 out of 35) of cooperators imposed sanctions

cooperators and 32 percent defectors in $M5$, and out of 83 percent cooperators and 17 percent defectors in $M10$. Likewise, among third parties the fraction of P-types was between 35 ($M5$) and 44 percent ($M10$) among the cooperators and 43 percent (in both $M5$ and $M10$) among the defectors.[15] In other words, the group of third party punishers (P-types) consisted out of 60 percent cooperators and 40 percent defectors in $M5$, and out of 70 percent cooperators and 30 percent defectors in $M10$. Thus, cooperators remain the dominant origin of punishment action, and there are evidently no noteworthy differences between second and third parties along those lines.

As already suggested by results 4 and especially 5, this is different with respect to the target locus of sanctions. A qualitative feature that is common among second and third parties is that, despite the rise in monitoring costs, defectors rather than cooperators remain the prevailing target of sanctions on average, although less pronounced. In the SPMG (see also table 1), defection is still punished significantly more often than cooperation in both $M5$ (32 vs. 17 percent, $p = .004$, Wilcoxon signed-rank test), and $M10$ (26 vs. 13 percent, $p = .014$, Wilcoxon signed-rank test). Also, defectors continue to receive stronger second party punishment on average than cooperators in both $M5$ (9.8 vs. 3.3 tokens, $p = .006$, Wilcoxon signed-rank test), and $M10$ (5.6 vs. 1.1 tokens, $p = .014$, Wilcoxon signed-rank test).

Table 1: Frequency of punishment and mean damages (punishments received, or punishment expenditures times three) in the SPMG.

	Relative frequency		Mean damages		
κ_m	Cooperation	Defection	Cooperation	Defection	Net Damages
0	.136	.727	1.8	22.8	21.0
5	.167	.318	3.3	9.8	6.5
10	.130	.261	1.1	5.6	4.6

In the TPMG (see also tables 2 and 3), in $M5$ unilateral defection is still punished significantly more often than mutual defection (36 vs. 26 percent, $p = .020$, Wilcoxon signed-rank test) or cooperation (24 percent, $p = .011$). The same is true for $M10$, where cheating is punished 41 percent of the time, and mutual defection (30 percent) or cooperation (33 percent) less often, although the differences are not statistically significant in this condition ($p = .059$

on defectors in $M5$ and $M10$, respectively, with respective average expenditures of 11.6 tokens and 8.2 tokens. Five and four of the cooperators in $M5$ and $M10$, respectively, also imposed punishments on other cooperators, with respective average expenditures of 6.8 tokens and 3 tokens. 11 out of 18 (61 percent) and 9 out of 11 (82 percent) of the defectors did not punish in $M5$ and $M10$, respectively. Of the remaining seven defectors in $M5$, four punished indiscriminately (B-type) with an average expenditure of 5.8 tokens, and three targeted (T-type) with an average expenditure of 6 tokens on defectors and 5 tokens on cooperators, plus the five tokens monitoring fee. Of the remaining two defectors in $M10$, both punished indiscriminately (B-type) with an average expenditure of 2 tokens.

[15] Among third parties, 35 percent (15 out of 43) and 44 percent (14 out of 32) of cooperating third parties imposed sanctions on defectors in $M5$ and $M10$, respectively, with respective average expenditures for punishing unilateral (mutual) defectors of 2.9 tokens (1.3 tokens) and 3.2 tokens (2.0 tokens). Eight (19 percent) and ten (31 percent) of the cooperators in $M5$ and $M10$, respectively, also imposed punishments on other cooperators, with respective average expenditures for punishing unilateral (mutual) cooperators of 0.7 tokens (1.2 tokens) and 0.6 tokens (1.6 tokens). 13 out of 23 (57 percent) and 8 out of 14 (57 percent) of the defecting third parties did not punish in $M5$ and $M10$, respectively. Of the remaining ten defectors in $M5$, seven punished indiscriminately (B-type) with an average expenditure of 4.6 tokens in case the target player's coplayer defected, and 4.3 tokens in case the target player's coplayer cooperated. Three punished discriminately (T-type) with an average expenditure of 6.7 tokens on unilateral defectors, 5 tokens on mutual defectors, and 1.7 tokens on cooperators, plus the five tokens monitoring fee. Likewise, of the remaining six defectors in $M10$, four punished indiscriminately (B-type) with an average expenditure of 4.3 tokens in case the target player's coplayer cooperated, and 1.8 tokens in case the target player's coplayer defected. Two targeted (T-type) with an average expenditure of 6 tokens on unilateral defectors, 2.5 tokens on mutual defectors, and between 0.5 and 1 tokens on cooperators, plus the ten tokens monitoring fee.

and $p = .103$, respectively). Also, unilateral defectors continue to receive stronger punishment on average than mutual defectors or cooperators in both $M5$ (2.6 vs. 1.6 vs. 1.4 tokens, $p = .038$ for the first and $p = .018$ for the second difference), and $M10$ (2.8 vs. 1.7 vs. 1.5 tokens, $p < .001$ for the first and $p = .055$ for the second difference).

Table 2: Frequency of punishment in the TPMG.

	Cooperation		Defection	
κ_m	Unilateral	Mutual	Unilateral	Mutual
0	.227	.227	.727	.455
5	.200	.227	.364	.258
10	.174	.326	.413	.304

Table 3: Mean damages (punishments received, or punishment expenditures times three) in the TPMG.

	Cooperation		Defection	
κ_m	Unilateral	Mutual	Unilateral	Mutual
0	3.4	4.1	18.0	6.8
5	3.0	4.0	7.9	4.8
10	1.7	4.6	8.5	5.0

Importantly, both in the SPMG and the TPMG cheating was no longer unprofitable as there was a strictly positive monitoring fee.

Result 7. *In both the SPMG and the TPMG, cheating is only unprofitable when monitoring is costless. The incentives to cooperate are weaker in the TPMG than in the SPMG at zero information costs already, and the presence of positive information costs increases this difference.*

For the SPMG this result is summarized in table 1. We first observe that there is no significant variation in punishment of cooperation (Kruskal-Wallis test, $p = .756$). Punishment of defection, however, is significantly different between treatments ($p < .001$), exhibiting a diminishing trend as monitoring costs rise (Kendall's $\tau_b = -.266$, $p < .001$). The same is true for the mean net punishment of defection, which is the difference between damages imposed on defection and cooperation, respectively (Kendall's $\tau_b = -.299$, $p < .001$). Hence, *on average* defecting gets less costly as monitoring costs rise. As a result, defection *did* pay when monitoring costs were positive.

The analogous result for the TPMG is summarized in table 3. Again, there is no significant variation in punishment of cooperation (Kruskal-Wallis tests, $p = .796$ for unilateral and $p = .600$ for mutual cooperation) and mutual defection ($p = .291$), but punishment of cheating is significantly different between treatments ($p = .005$), its intensity being markedly lower when monitoring is costly compared to the baseline condition. In consequence cheating also *did* pay when monitoring costs were positive.

The difference between average punishment of cheaters and average punishment of cooperators, however, decreases less strongly in second parties (from 20.1 tokens in $M0$ over 6.5 tokens in $M5$, a 68 percent decrease, to 4.5 tokens in $M10$, a further 31 percent decrease, for a total decrease of 78 percent) than in third parties (from 14.3 tokens in $M0$ over 1.2 tokens in $M5$, a 92 percent decrease, to 1.3 tokens in $M10$, for a total decrease of 91 percent) in relative terms. This reinforces the conclusion that while both second and third party punishment tends to become less targeted on cheaters through the introduction of monitoring costs,

the effect is stronger in third parties. In other words, while third party punishment provides weaker incentives to cooperate at costless monitoring already, the introduction of monitoring costs renders the difference even larger.

It should be noted, however, that subjects did not anticipate the effects of monitoring costs on punishment and the incentives to cooperate entirely.[16] In the baseline condition of the present experiment, subjects' average expectations match actual punishment patterns remarkably accurately.[17] In the presence of positive monitoring costs, on the other hand, the punishment anticipation hypothesis fails. As shown in tables 5 and 6 in the appendix, subjects anticipate the actual impact of monitoring costs on punishment patterns *qualitatively* (the direction), but *quantitatively* they significantly underestimate the effects of increasing monitoring costs on their coplayers' monitoring and punishment behavior. The difference is that in the subjects' expectations, positive monitoring costs have a much weaker impact on the punishment of defectors (and hence the net punishment of defecting) than is actually the case. For example, the *actual* net cost of defecting in the SPMG amounted to 6.5 tokens on average in $M5$ and 4.6 on average in $M10$, whereas the *expected* net costs were 12.3 and 10.0, respectively, on average. As a result, cheating was still unprofitable in subjects' expectation.[18]

To sum up, monitoring costs lead to three important qualitative differences compared to the costless monitoring condition which are common in second and third parties: (1) The supply of social sanctions decreases as if punishment costs increased. (2) A small, but significant share of subjects sacrifices resources to actively monitor their coplayer and thus target their punishment. (3) A small, but significant share of subjects chooses to punish without information on whether the coplayer cooperated. The first-order impact of increasing monitoring costs is therefore to reduce the supply of sanctions.[19] But this variation also brings to light important heterogeneities across subjects with respect to the type of sanctioning they will supply, targeted or blind. Targeted sanctioning is conducive to maintaining cooperation because it rewards cooperative behavior in relative terms. The deliberately blind sanctioning observed in the experiment on the other hand is unproductive for maintaining cooperation. Here is where we find the key difference between second and third party behavior: While second parties respond to the introduction of monitoring costs more often than third parties with refraining from sanctioning altogether contingent on remaining «blind», third parties opt more often than second parties to indiscriminate punishment in this case. We turn to this

[16] Fehr and Gächter (2002, p. 139, emphasis added) noted that defection «may cause strong negative emotions among the cooperators [which] may trigger their willingness to punish ... and *that most people expect these emotions*.» The anticipation of non-strategic punishment is the key link to the cooperation decision in the first stage.

[17] In the SPMG, subjects expected on average to receive punishment of 21.2 (1.0) tokens in case of defection (cooperation). This is a remarkably accurate prediction of the actual punishments of 22.8 (1.8) tokens. The rank correlation between expectations and realizations is significantly positive (Kendall's $\tau_b = .547$, $p = .001$, in case of defection, $\tau_b = .622$, $p = .004$, in case of cooperation). In the TPMG subjects expected on average to receive punishment of 17.9 tokens in case of unilateral defection, 8.7 in case of mutual defection, and at most 4.2 in case of cooperation. Those predictions match the the actual punishments of 18.0 tokens in case of unilateral defection, 6.8 in case of mutual defection, and at most 4.1 in case of cooperation closely as well. The rank correlation between expectations and realizations is significantly positive (Kendall's $\tau_b = .440$ in case of unilateral defection, $\tau_b = .552$ in case of mutual defection, $\tau_b = .426$ in case of unilateral cooperation, $\tau_b = .453$ in case of mutual cooperation, all $p = .000$). López-Pérez and Kiss (2012), whose study explicitly focuses on whether subjects anticipate positive and negative sanctions, obtain similar results as ours. Notably, they find that punishments are better anticipated than rewards.

[18] This might explain the fact that we did find no significant differences in the frequency of cooperation across conditions.

[19] This qualifies results reported in a related paper by Grechenig et al. (2010) in the sense that if it is the subject's *own choice* of whether to acquire information on their coplayer's behavior, the vast majority refrain from punishment altogether if they opted to remain ignorant.

difference in more detail in the following section.

3.3 Why is there more «blind punishment» among third parties?

The aggregate differences between the SPMG and the TPMG are driven by individuals that behave differently when they find themselves in the role of a second or third party, respectively. Across all treatments, 37 out of 134 subjects change their behavioral type between the roles. This fraction is disproportionally large in the costly monitoring treatments, where 32 out of 112 subjects (29 percent) reveal a different behavioral type as a second and a third party, respectively, the rest exhibit the same type. Those subjects are represented by the off-main-diagonal cells in table 4, where each cell illustrates a different transition path. The key insight from this data is that there are disproportionally more subjects (18) that punish blindly as a third but not as a second party than the other way around (six subjects).

Table 4: «Transition matrix» of behavioral types across the SPMG and the TPMG for the costly monitoring treatments.

	TPMG Type			
SPMG Type	*N*	*T*	*B*	Margin
N	(62) .554	(6) .054	(10) .089	(78) .696
T	(2) .018	(10) .089	(8) .071	(20) .179
B	(3) .027	(3) .027	(8) .071	(14) .125
Margin	(67) .598	(19) .170	(26) .232	(112) 1.000

Relative frequencies. Absolute frequencies in parantheses.

There are at least two possible and not mutually exclusive explanations why subjects behave differently in the two roles.[20] First, an obvious and conventional explanation would be that the subjects' prior about the target player's first stage behavior differs between the SPMG and the TPMG. A second explanation for the differences in behavior is that the SPMG and the TPMG bring out differences in the social or moral preferences attached to the role of being a second party or third party. This would point to the presence of a «pure role effect» that makes the same individual punish for different reasons depending on their position in a social dilemma. The experimental design elicited subjects' beliefs and risk preferences. The first candidate explanation can therefore be addressed directly and the second hypothesis indirectly, provided that our elicitation procedure measures the subject's beliefs and risk preferences adequately. As a first step, we therefore conduct a plausibility test on our measures of beliefs and risk preferences. Having passed this test, we examine in a second step the evidence for the two candidate explanations.

First stage behavior is consistent with «conditional cooperation»

To establish plausibility that our measures of beliefs and risk preferences are valid, we investigate their correlations with first-stage cooperation behavior and evaluate the consistency with known motivational traits. The results suggest that the elicitation produced valid measures of subjects' beliefs and risk preferences. To see this, recall the current agreement in the literature that a majority of people have «conditionally cooperative» or «strongly reciprocal» preferences (see e.g. Fehr et al., 2002; Gintis et al., 2005; Gächter, 2007, for overviews).

[20] One might add that the shape of the risk preferences differs between the two games, but this appears not very plausible.

For a player with such preferences, defect is *not* a dominant strategy in a cooperation game of the type implemented in our experiment. Instead, the player prefers to cooperate if (and only if) the coplayer cooperates as well. This implies that a simultaneous-move cooperation game involves a decision under strategic risk, such that the probability of a cooperative move should be (i) increasing in a player's belief in the coplayer's cooperation, and (ii) decreasing in a player's degree of risk aversion.[21] Previous experiments have confirmed these predictions (e.g. Eckel and Wilson, 2004; Houser et al., 2010; Fischbacher and Gächter, 2010), and our results are affirmative as well: A dummy indicating an optimistic belief about the coplayer's cooperation (70 or more percent sure that the coplayer cooperates) is significantly positively correlated (Phi/Cohen's $w = 0.459$, $p = .000$, or Kendall's $\tau_b = 0.459$, $p = .000$), and a dummy indicating a pessimistic belief (70 or more percent sure that the coplayer defects) is significantly negatively correlated with own cooperation (Phi/Cohen's $w = 0.515$, $p = .000$, or Kendall's $\tau_b = -0.515$, $p = .000$). Furthermore, a dummy indicating risk aversion is negatively correlated with own cooperation, although not statistically significantly so (Phi/Cohen's $w = 0.130$, $p = .133$, Kendall's $\tau_b = -0.130$, $p = .135$). However, if we restrict the analysis to those whose belief is indeterminate (reference category to optimistic and pessimistic belief, respectively) we get statistical significance and a stronger effect (Phi/Cohen's $w = 0.276$, $p = .041$, Kendall's $\tau_b = -0.276$, $p = .044$), while for those with a either optimistic or pessimistic beliefs correlation is not significantly different from zero (Phi/Cohen's $w = 0.008$, $p = .947$, Kendall's $\tau_b = -0.008$, $p = .953$). These results are retained in multiple regression analysis with all indicators included simultaneously.[22] This consistency with known motivational patterns is indirect evidence in favor of the validity of our measures of beliefs and risk preferences.

Is it because of different beliefs?

With plausible measures of beliefs and risk preferences in hand, we now proceed to using elicited beliefs about target player behavior to address the first hypothesis directly. Overall, the subjects' beliefs about the target player's behavior did not differ significantly between the SPMG and the TPMG (Wilcoxon signed-rank test, $p = .340$): The average second party expected the target player to cooperate with probability .667, the average third party with probability .643.[23] For investigating the differential effects of monitoring costs on second and third party information acquisition and punishment, however, the relevant population is not the whole sample but only those 37 subjects that changed their behavioral type. Those subjects' beliefs about the target player's behavior did not differ significantly between the SPMG and the TPMG either (Wilcoxon signed-rank test, $p = .766$): The average second party expected the target player to cooperate with probability .703, the average third party with probability .676.[24]

The invariance of beliefs with respect to the setting (SPMG or TPMG) can also be shown by calculating the correlation between the belief indicators used above and the behavioral type indicators. The dummy indicating an optimistic belief about the target player's cooperation is not correlated with either one of the type indicators.[25] The same is true for the dummy

[21] For a player with «standard» payoff-monotone preferences defect is a dominant strategy, such that beliefs and risk preferences are irrelevant.

[22] We estimated various specifications using a number of different fitting methods, obtaining essentially the same results. They are available from the authors on request.

[23] Separated by treatment, only when monitoring is costless there is a significant difference ($p = .046$), the average second party expecting the target player to cooperate with probability .618, the average third party with probability .718.

[24] Separated by treatment, there are also no significant differences.

[25] The Phi coefficient or Cohen's w is 0.059 for the type N indicator, 0.028 for the type B indicator, and 0.089 for

indicating a pessimistic belief about the target player's cooperation.[26] In sum, differences in beliefs do not provide a satisfactory explanation for the differential effects of monitoring costs on second and third party information acquisition and punishment in this experiment.

Is it because of different preferences?

The second candidate explanation focuses on the possibility of a «pure role effect», possibly tied to different social or moral preferences activated by occupying the position of a second or third party. To examine this possibility, we exploit the idea that preferences that underpin conditionally punitive behavior are conceivably related to the sensitivity towards committing punishment errors: Punishing someone who cooperated would be perceived differently from leaving a cheater escape unpunished. If those preferences differ between the SPMG and the TPMG, then this sensitivity should differ between the games as well, as long as risk preferences do not differ too much between the SPMG and TPMG. In other words, given a player's beliefs and risk preferences (and punishment costs), behavior can only differ in the two roles if the subjective benefits accruing through punishment differ.

The elicited risk preferences provide an opportunity to examine the possible presence of a «pure role effect» to some degree. In a population of subjects whose preferences embody a norm of conditional cooperation, these players should have a preference towards targeting punishments on cheaters. Behaving as a N-type or a B-type then both implies a risk of error: The N-type risks letting a defector escape unpunished, the B-type risks erroneously harming a cooperator. The only risk-free option is behaving as a T-type. It follows that risk averse subjects should have a stronger tendency to sort into T-type, while more risk tolerant subjects should be more willing to sort into N- or B-type at the margin. With one puzzling exception, the data is consistent with this reasoning: Restricting the analysis to the treatment conditions,[27] the dummy indicating risk aversion is negatively correlated with a dummy indicating type N (Phi/Cohen's $w = 0.244$, $p = .000$, or Kendall's $\tau_b = -0.244$, $p = .000$), and positively correlated with dummies indicating type B (Phi/Cohen's $w = 0.181$, $p = .007$, or Kendall's $\tau_b = 0.181$, $p = .007$) and type T (Phi/Cohen's $w = 0.124$, $p = .064$, or Kendall's $\tau_b = 0.124$, $p = .065$), respectively. Thus, the predictions with respect to N-types and T-types are confirmed. The coefficient for B-types, however, goes exactly in the opposite direction. B-types are *more* risk averse than the average rather than less. A possible explanation is that for these subjects, the error of letting a defector escape unpunished is evaluated as much worse than the error of erroneously harming a cooperator. A test of this hypothesis is an interesting avenue for future research.

The presence of a pure role effect requires that the correlation between risk preferences and observed behavior differ between second and third parties. A difference would be indirect evidence that the average subject is motivated differently in the role of second and a third party, respectively. Comparing the correlations between the risk aversion indicator and the type indicators, every single one is stronger for second than for third parties: For the dummy indicating type N, the Phi coefficient or Cohen's w is 0.267 ($p = .005$) in second parties

the type T indicator (Kendall's rank correlation coefficients are identical, except for the negative sign in the first two cases), but all fail the test for being statistically significantly different from zero ($p > .144$). This remains to be true when the analysis is carried out for the baseline and treatment conditions separately, respectively.

[26] The Phi coefficient or Cohen's w is 0.039 for the type N indicator, 0.062 for the type B indicator, and 0.094 for the type T indicator (again, Kendall's rank correlation coefficient is identical, except for the negative sign in the latter case), but again all fail the test for being statistically significantly different from zero ($p > .126$). This remains to be true when the analysis is carried out for the baseline and treatment conditions separately, respectively.

[27] We do so since the risk free alternative (type T) is available at no cost in the baseline condition and hence there is not much systematic variance in type assortment.

and in 0.224 (p = .018) third parties. Likewise, for the type-B indicator correlation is 0.267 (p = .005) in second parties and 0.224 (p = .018) in third parties. Finally, for the type T indicator it is 0.156 (p = .099) in second parties and 0.092 (p = .333) in third parties. This means that second parties appear to be somewhat more sensitive to the risk of punishment error.

4 Conclusion

Social sanctions have been credited with providing an important instrument for resolving social dilemmas. These sanctions are provided by materially affected (second) parties and materially unaffected (third) parties. The supply of such sanctions for non-strategic reasons and at a cost have attracted the attention of many researchers. This paper reports on an experiment that examines how second and third parties in a social dilemma situation respond to the presence of monitoring costs that need to be incurred if the sanctioning party wants to be able to differentiate between cooperators and defectors in the social dilemma. The presence of such costs is both an empirically salient feature of many social dilemmas and an opportunity to examine the phenomenon of endogenous information structures in games of cooperation and punishment. The choice of whether to incur the monitoring cost is, at the same time, an opportunity to exploit revelation of preferences through choice and learn something about the preferences underpinning second and third party punishment. The commonalities and differences of these preferences depending on the role in which the player finds himself are a key interest of this paper.

Building on a well-established paradigm, the experimental variation of monitoring costs provides both a close replication of previous evidence on the presence of non-strategic sanctioning and the observation of new types of differences between second party and third party behavior. Relative to a baseline with zero monitoring costs, three effects are present. First, a «direct punishment cost effect» of decreasing the gross supply of punishment, which is the first-order effect. Second, the propensity to punish indiscriminately increases, an impact we termed «blind punishment effect». Finally, positive information costs lead to defectors making up a smaller share of the punished and receiving weaker punishment compared to zero information costs, with the result that incentives get too weak to render cheating unprofitable. We called this the «diffusion effect». Comparing second and third party behavior, the «punishment cost effect» is more pronounced in the former, whereas the «blind punishment effect» and the «diffusion effect», respectively, is stronger in the latter. As a result, third party punishment leads to increasingly weaker incentives for cooperation relative to second party punishment as monitoring costs rise. In addition, the differences between second and third parties suggest the presence of a «pure role effect»: Taking into account elicited beliefs and risk preferences, third parties do not simply provide weaker punishment than second parties. Third parties provide *different* punishment because they appear to punish for both the same, but also for different reasons.

Interest within management and policy-making in exploring and implementing informal team and community governance is increasing. Against the background of this increasing interest, we believe that evidence on the impact of monitoring cost on punishment behavior and the results on commonalities and differences between second and third party sanctioning behavior are a valuable input to a more fine-grained collection of organization principles. Several avenues for future research along these lines suggest themselves, for example extending the setting to interactions with a longer horizon, increasing the group size, and varying the conditions and nature of information acquisition.

References

Acheson, J.M., 1975. The lobster fiefs: Economic and ecological effects of territoriality in the maine lobster industry. Human Ecology 3, 183–207.

Acheson, J.M., 1987. The lobster fiefs revisited: Economic and ecological effects of territoriality in the Maine lobster industry, in: McClay, B., Acheson, J.M. (Eds.), The Problem of the Commons: The Culture and Ecology of Communal Resources. University of Arizona Press, Tucson, Arizona, USA, pp. 37–65.

Acheson, J.M., 1988. The Lobster Gangs of Maine. University Press of New England, Hanover, NH, USA.

Alchian, A.A., Demsetz, H., 1972. Production, information costs, and economic organization. American Economic Review 62, 777–795.

Ambrus, A., Greiner, B., 2012. Imperfect public monitoring with costly punishment – an experimental study. American Economic Review .

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, 1–24.

Armendáriz, B., Morduch, J., 2005. The Economics of Microfinance. MIT Press, Cambridge, MA, USA.

Balliet, D., Mulder, L., van Lange, P.A.M., 2011. Reward, punishment, and cooperation: A meta-analysis. Psychological Bulletin 137, 594–615.

Bendor, J., Mookherjee, D., 1990. Norms, third-party sanctions, and cooperation. Journal of Law, Economics, and Organization 6, 33–63.

Bendor, J., Swistak, P., 2001. The evolution of norms. American Journal of Sociology 106, 1493–1545.

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

Bowles, S., Gintis, H., 2011. A Cooperative Species: Human Reciprocity and its Evolution. Princeton University Press, Princeton, NJ, USA.

Brandts, J., Charness, G., 2010. The strategy versus the direct-response method: A survey of experimental comparisons. Mimeograph. Universidad Autònoma de Barcelona. Barcelona, Spain.

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

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

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

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

Craig, B., Pencavel, J., 1995. Participation and productivity: A comparison of worker cooperatives and conventional firms in the Plywood industry. Brookings Papers on Economic Activity: Microeconomics 1995, 121–174.

Dong, X.Y., Dow, G.K., 1993. Monitoring costs in Chinese agricultural teams. Journal of Political Economy 101, 539–553.

Eckel, C.C., Wilson, R.K., 2004. Is trust a risky decision? Journal of Economic Behavior & Organization 55, 447–465.

Egas, M., Riedl, A., 2008. The economics of altruistic punishment and the maintenance of cooperation. Philosophical Transactions of the Royal Society: Biological Sciences 275, 871–878.

Elster, J., 2009. Social norms and the explanation of behavior, in: Hedström, P., Bearman, P. (Eds.), The Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford, UK, pp. 195–217.

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, 495–511.

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

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

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

Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415, 137–140.

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

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

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

Gächter, S., 2007. Conditional cooperation: Behavioral regularities from the lab and the field and their policy

implications, in: Frey, B.S., Stutzer, A. (Eds.), Economics and Psychology. A Promising New Cross-Disciplinary Field. MIT Press, Cambridge, Massachusetts, USA, pp. 19–50.

Gächter, S., Herrmann, B., 2009. Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. Philosophical Transactions of the Royal Society: Biological Sciences 364, 791–806.

Gächter, S., Herrmann, B., 2010. The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia. European Economic Review 55, 193–210.

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

Gächter, S., Renner, E., Sefton, M., 2008. The long-run benefits of punishment. Science 322, 1510.

Gintis, H., Bowles, S., Boyd, R., Fehr, E. (Eds.), 2005. Moral Sentiments and Material Interests. The Foundations of Cooperation in Economic Life. MIT Press, Cambridge, Massachusetts, USA.

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, 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.

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

Holt, C.A., Laury, S.K., 2002. Risk aversion and incentive effects. American Economic Review 92, 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, 72–81.

Kandel, E., Lazear, E.P., 1992. Peer pressure and partnerships. Journal of Political Economy 100, 801–817.

Kandori, M., 1992. Social norms and community enforcement. Review of Economic Studies 59, 63–80.

Kanemoto, Y., MacLeod, W.B., 1991. The theory of contracts and labor practices in Japan and the United States. Managerial and Decision Economics 12, 159–170.

Keser, C., van Winden, F., 2000. Conditional cooperation and voluntary contributions to public goods. Scandinavian Journal of Economics 102, 23–39.

Lazear, E.P., 1993. Labor economics and the psychology of organizations. Journal of Economic Perspectives 5, 89–110.

Leibbrandt, A., López-Pérez, R., 2009. An exploration of third and second party punishment in ten simple games. Mimeo. Universidad Autonoma de Madrid. Madrid, Spain.

Leibbrandt, A., López-Pérez, R., 2011. Individual heterogeneity in punishment and reward. Economic Analysis Working Paper Series 1/2011. Universidad Autonoma de Madrid. Madrid, Spain.

Leibbrandt, A., López-Pérez, R., 2012. An exploration of third and second party punishment in ten simple games. Journal of Economic Behavior & Organization 84, 753–766.

List, J.A., Sadoff, S., Wagner, M., 2011. So you want to run an experiment, now what? some simple rules of thumb for optimal experimental design. Experimental Economics 14, 439–457.

López-Pérez, R., Kiss, H.J., 2012. Do people accurately anticipate sanctions? Southern Economic Journal forthcoming.

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

Oosterbeek, H., Sloof, R., van de Kuilen, G., 2004. Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. Experimental Economics 7, 171–188.

Ostrom, E., 1990. Governing the commons: The evolution of institutions for collective action. Cambridge University Press, Cambridge, UK.

Ostrom, E., Gardner, R., 1993. Managing local commons: Theoretical issues in incentive design. Journal of Economic Perspectives 7, 113–134.

Palmer, C.T., 1991. Kin-selection, reciprocal altruism, and information sharing among Maine lobstermen. Ethology and Sociobiology 12, 221–235.

Rustagi, D., Engel, S., Kosfeld, M., 2010. Conditional cooperation and costly monitoring explain success in forest commons management. Science 330, 961–965.

Sampson, R.J., Raudenbush, S.W., Earls, F., 1997. Neighborhoods and violent crime: A multilevel study of collective efficacy. Science 277, 918–924.

Seabright, P., 1993. Coping with asymmetries in the commons: Self-governing irrigation systems can work. Journal of Economic Perspectives 7, 93–112.

Shapiro, C., Stiglitz, J.E., 1984. Equilibrium unemployment as a worker discipline device. American Economic Review 74, 433–444.

Suleiman, R., 1996. Expectations and fairness in a modified ultimatum game. Journal of Economic Psychology 17, 531–554.

Table 5: Frequency of subjects expecting to be punished and mean expected damages in the SPMG.

	Relative frequency		Mean damages		
κ_m	Cooperation	Defection	Cooperation	Defection	Net Damages
0	.136	.864	1.0	21.1	20.2
5	.212	.636	3.2	15.5	12.3
10	.196	.652	2.4	12.4	10.0

Table 6: Mean damages (punishment points times three) expected by the target players in the TPMG.

	Cooperation		Defection	
κ_m	Unilateral	Mutual	Unilateral	Mutual
0	2.6	4.2	17.9	8.7
5	4.4	2.9	11.2	7.5
10	3.1	2.7	11.4	6.0

Table 7: «Transition matrix» of behavioral types across the SPMG and the TPMG including the baseline condition.

	TPMG Type			
SPMG Type	*N*	*T*	*B*	Margin
N	(66) .493	(8) .060	(10) .075	(84) .627
T	(4) .030	(23) .172	(9) .067	(36) .269
B	(3) .022	(3) .022	(8) .060	(14) .104
Margin	(73) .545	(34) .254	(27) .201	(134) 1.000

Relative frequencies. Absolute frequencies in parantheses.