DIPARTIMENTO DI
SCIENZE ECONOMICHE
UNIVERSITÀ DEGLI STUDI DI VERONA

Working Paper Series

Department of Economics

University of Verona

Moral Sentiments and Material Interests behind Altruistic Third-Party Punishment

Stefania Ottone, Ferruccio Ponzano, Luca Zarri

WP Number: 48 May 2008

ISSN: 2036-2919 (paper), 2036-4679 (online)

Moral Sentiments and Material Interests behind Altruistic Third-Party Punishment

Stefania Ottone (EconomEtica and University of Eastern Piedmont)
Ferruccio Ponzano (University of Eastern Piedmont)
Luca Zarri (University of Verona)

May 2008

Abstract

Social norms are ubiquitous in human life. Their role is essential in allowing cooperation to prevail, despite the presence of incentives to free ride. As far as norm enforcement devices are concerned, it would be impossible to have widespread social norms if second parties only enforced them. However, both the quantitative relevance and the motivations underlying altruistic punishment on the part of 'unaffected' third parties are still largely unexplored. This paper contributes to shed light on the issue, by means of an experimental design consisting of three treatments: a Dictator Game Treatment, a Third-Party Punishment Game Treatment (Fehr and Fischbacher, 2004) and a Metanorm Treatment, that is a variant of the Third-party Punishment Game where the Recipient can punish the third party. We find that third parties are willing to punish dictators (Fehr and Fischbacher, 2004; Ottone, 2008) and, in doing so, they are affected by 'reference-dependent fairness', rather than by the 'egalitarian distribution norm'. By eliciting players' normative expectations, it turns out that all of them expect a Dictator to transfer *something* – not half of the endowment. Consequently, the Observers' levels of punishment are sensitive to their subjective sense of fairness. A positive relation between the level of punishment and the degree of negative subjective unfairness emerges. Subjective unfairness also affects Dictators' behaviour: their actual transfers and their ideal transfer are not significantly different. Finally, we interestingly find that third parties are also sensitive to the receivers' (credible) threat to punish them: as the Dictator's transfer becomes lower and lower than the Observer's ideal transfer, the Observer's reaction is – other things being equal – significantly stronger in the Metanorm Treatment than in the Third-Party Punishment Game Treatment. Hence, despite their being to some extent genuinely nonstrategically motivated, also third parties – like second parties – are sensitive to the costs of punishing.

Keywords: Third-Party Punishment; Moral Sentiments; Material Interests; Subjective Unfairness; Social Norms.

JEL Classification: C72, C9, D63, Z13.

1. Introduction

Humans are the only species known to often incur costs in order to adopt various types of morally charged behaviors. In the last decades, experimental research has persuasively shown that the canonical model based on material self-interest only – the so called 'selfishness axiom' – is untenable in literally hundreds of experiments carried out in several countries by means of a large number of different game protocols. However, on more constructive grounds, a lot of disagreement still exists about the exact features of both fairness judgments and non-selfish behaviors. As we will make clear, our experimental data show that the *Homo Oeconomicus* model fails in a variety of new ways, but also that subtle interplays turn out to interestingly occur between selfishness and pro-sociality, in the domain of both individuals' normative beliefs and actual behavioral decisions.

In particular, we both confirm previous results and obtain new findings, with regard to the nature of unselfish behaviors acting as enforcement devices of a given 'social norm of fairness'. Social norms are ubiquitous among humans and play a relevant role in virtually any human society (see Henrich et al., 2001). Notwithstanding this, they are still poorly understood, to the point that they still constitute one of the big unsolved problems in the behavioral sciences (Fehr and Fischbacher, 2004). Experimental economics can contribute to greatly enhance our understanding of how social norms emerge and are endogenously enforced. Unlike field studies, properly designed experiments allow us to rigorously isolate the different forces shaping norm enforcement.

In the lab, subjects have often proved to be willing to display *altruistic* or *nonstrategic punishment*, that is forms of costly sanctioning which are not driven by (more or less sophisticated forms of) material payoff maximization[1]. Hence, this mechanism represents a potentially relevant endogenous norm enforcement device, especially within social environments where centralized institutions for the exogenous enforcement of legal rules are absent or ineffective. However, behavioral economics, so far, has mainly dealt with *second-party* altruistic punishment, by focusing on the 'vengeful' behavior of experimental subjects who had been directly hurt by other players (see, e.g. the famous work by Fehr and Gaechter, 2000). By contrast, as Fehr and Fischbacher (2004) correctly

[1] When nonstrategic sanctioning is both targeted at non-cooperators and associated with the willingness to cooperate with cooperators, such behavioral attitude is called *strong reciprocity*. For experimental evidence, see e.g. Fehr and Gaechter (2000) and the survey in Gintis et al. (2003). Strong reciprocity differs from the strategically motivated form of reciprocity which is typical of the so called 'Folk Theorem' literature (see e.g. Rubinstein, 1979 and Fudenberg and Maskin, 1986).

observe, “if only second parties imposed sanctions, a very limited number of social norms could be enforced because norm violations often do not directly hurt other people” (p. 64). In other words, it is often the case that the enforcement of a given norm of fairness depends on the (nonstrategic) action of both second and *third parties*. People often punish wrongdoers and an altruistic third-party punisher is an agent who is willing to incur material costs in order to inflict harm on violators of a social norm, even when such violations are not directed at him.

As Marlowe et al. (2008) observe: “Third-party punishment is an effective way to enforce the norms of strong reciprocity and promote cooperation”. According to Kurzban et al. (2007), third-party punishment is a crucial feature of human social life and forms the cornerstone of morality: “Humans everywhere seek and assess evidence of infractions, identify acts as morally right or wrong, and desire that wrongdoers be punished”[2]. (p. 75). Similarly, Carpenter and Matthews (2007) point out that “Although the logic of third-party punishment is not obvious, researchers have determined that it is crucial for the enforcement of social norms” (p. 2). They conclude their paper by highlighting that the emphasis on second-party punishment in the literature seems misplaced.

The main problem with second-party punishment has to do with Kandori’s (1992) distinction between *personal* and *community* enforcement mechanisms. The classic literature on the Folk Theorem in repeated games provides us with formal models of personal enforcement, in which cheating triggers retaliation by the victim. Such mechanisms, also described as relying on ‘direct’ – rather than ‘indirect’ reciprocity (Nowak and Sygmund, 1998) – can be effective only if *the same* agents frequently play the same stage game over time. By contrast, under community enforcement agents change their partners over time and punishment can be interpreted as a consequence of the fact that the punisher knows that the opponent adopts a ‘bad’ behavior in general, that is whenever she interacts with others in the population, not only in her direct interactions with the punisher. The key point is that nowadays, within increasingly interdependent, *large-scale* societies, the level of *anonymity* increases and a growing number of interactions take the form of *one-shot* games. Hence, it would be hard to believe that, in such contexts, endogenous norm enforcement occurs extensively due to second-party punishment only: insofar as we consider complex social scenarios where cheating becomes tempting and

[2] It is interesting to observe that some third-party punishment has been found also among nonhuman animals. For example, chimpanzees sometimes intervene on behalf of unrelated others (de Waal, 1996).

difficult to monitor, third-party punishment is likely to be critically at work in preventing a decay of cooperation[3].

However, despite the crucial role played by third parties in enforcing social norms in many real-life societies, the economics of third-party punishment – unlike second-party punishment studies – is still in its infancy, especially at experimental level[4]. Notable exceptions are the recent works by Fehr and Fischbacher (2004), Bernhard (2005), Shinada et al. (2005), Takahashi et al. (2005), Ottone (2005; 2008), Carpenter and Matthews (2006; 2007), Bernhard et al. (2006), Henrich et al. (2006), Marlowe et al. (2008) and Charness et al. (2007).

As to altruistic third-party punishment, we claim that one of the key questions to be addressed can be stated as follows: what triggers this form of punishment? What motives underlie 'uninvolved' third parties' willingness to incur costs in order to punish individuals who 'defected' on someone else? As Kurzban et al. (2007) observe: "Given that punishment is costly and can potentially draw retaliation, third-party punishment appears to be a tendency that would be selected against, raising the issue of how adaptations that give rise to moralistic punishment evolved" (p. 75). Takahashi et al. (2005) discover that also third-party punishment (like second-party punishment) is affected by the punisher's evaluation of the punishee's *intentions*, but in a way that significantly differs from what had been previously obtained in the domain of second-party sanctioning. The experimental works of Bernhard (2005), Bernhard et al. (2006), Henrich et al. (2006) and Marlowe et al. (2008) focus, inter alia, on third-party sanctioning across cultures. Bernhard et al. (2006) investigate the potentially parochial nature of altruistic norm enforcement by means of comparisons between third parties' behavior towards *ingroup* and *outgroup* subjects, in order to see whether group membership affects the punishment patterns of 'impartial' observers. Within their well-known cross-cultural project, Henrich et al. (2006) find that there is considerable variance in the levels of third-party punishment. Marlowe et al. (2008) report new findings that explain such cross-cultural variation. Charness et al. (2007) explore the effects of third-party intervention in different treatments of an Investment Game and find a strong and significant effect of this sanctioning mechanism. Ottone (2008) implements a design where the third party has the opportunity to both punish the Dictator and transfer money to the Receiver (Solomon's Game). What it turns out is that

[3] Marlowe et al. (2008) interestingly provide an experimental confirmation of this intuition, as they find that subjects belonging to larger, more complex societies engage in *significantly more* third-party punishment than people in small-scale societies.

[4] Sober and Wilson (1998) report field evidence on the relevance of third-party punishment.

the Observer's transfers to the Receiver appear to be complementary to the punishment of the Dictator at high levels of unfairness and to be substitutes of it at low levels. This may suggest that the attitude of human beings to help those who suffer from an unfair behavior is strong and multifaceted – the desire of revenge is not the only emotion stimulated by people's sense of justice. People care about the condition of the victims.

The main purpose of this paper is to contribute to shed light on the motivational dimension underlying sanctioning decisions taken by uninvolved third parties. We begin by wondering whether some key findings obtained within the domain of second-party punishment also carry over to third-party punishment. Hence, we address the following questions: do people engage in third-party punishment? Are third parties driven by 'moral sentiments', that is sensitive to the 'degree of unfairness' of the punishees? Is third-party punishment entirely nonstrategic or is it also affected by 'material interests' (as second-party punishment appears to be)?

The structure of the remainder of the paper is as follows. Section 2 illustrates the experimental design. Section 3 clarifies our experimental procedure. In Section 4 the theoretical predictions for our experimental game are derived. Section 5 contains our major results. Section 6 concludes.

2. Experimental design

The experimental design (Figure 1) consists of three treatments: the Dictator Game Treatment (DG), the Third-party Punishment Game Treatment (TPP) and the Metanorm Treatment (MN)[5]. In the DG the tool is the classic Dictator Game. At the beginning of the session each subject is randomly assigned a role (A or B) and groups of 2 participants are formed. In each group, participant A (the Dictator) and participant B (the Receiver) play a Dictator Game.

In the TPP, our vehicle is the 'third-party punishment in the dictator game' (TP-DG) originally proposed by Fehr and Fischbacher (2004; p. 66), that is a DG in which a third player with a punishment option is introduced. At the beginning of the first stage each subject is randomly assigned a role (A, B or C) and groups of 3 participants are formed. In each group, participant A (the Dictator) and participant B (the Receiver) play a Dictator Game. In the second stage, participant C (the Observer) enters the game and has to decide

[5] Instructions are available upon request.

whether to bear a cost in order to sanction A or to keep the whole initial endowment[6]. This design reflects the idea that violation of a given behavioral standard may be punished not only by second parties (participant B, in our design), but also by 'uninvolved' third parties (participant C).

In the MN, we study a variant of the TPP based on the notion of *metanorm* (Axelrod, 1986). After that players participate in TPP, a third stage begins. In this stage participant B has the possibility to become an active player, by punishing participant C. The MN may be seen as a combination of the TPP and the well-known Ultimatum Game (UG; see Guth et al., 1982): like in the TPP, the Observer can punish the Dictator at a cost and, like in the UG, the Receiver can punish the 'first party'. The key difference between the MN and the UG is that in the latter the Receiver can *directly* (though implicitly, that is by rejecting the offer) punish his coplayer[7], whereas in the former the Receiver is only allowed to *indirectly* punish the first party by punishing the Observer for not punishing (enough) the first party.

In each treatment, A's and C's initial endowment is the same (20 tokens), while B's initial endowment is 10 tokens. The cost for participant C to punish participant A for the amount of 2 tokens, is 1 token. In the MN, the cost for participant B to punish participant C for the amount of 2 tokens, is 1 token as well[8]. Each token's value is 0.50 Euro.

3. Experimental procedure

The experiment was run in the Experimental Economics Laboratory (EELAB), at the University of Milano-Bicocca in Milan, Italy. The experiment was programmed and conducted with the software z-Tree (Fischbacher, 2007). Overall, 2 sessions for each treatment were run, with a total of 157 participants (40 participants in the DG, 60 in the TPP, 57 in the MN), recruited through a web-based recruitment system. At the beginning of the experiment, participants were informed about the sequential nature of the game protocol. The instructions were read by participants on their computer screen while an experimenter read them loudly. After reading the instructions and before subjects were invited to take decisions, some control questions were asked in order to be sure that players understood the rules of the game. At the end of each session, subjects were asked to fill in

[6] It is important to make clear that, as it is customary in economic experiments on punishment, the instructions did not contain terms such as *punish* or *sanction*, but used instead the more neutral term *deduction*.

[7] The UG can then be seen as the most famous example of experimental analysis of second-party sanctions.

[8] Only transfers of entire tokens are allowed and no participant can earn a negative payoff.

a brief survey to check for socio-demographic data. Each subject participated in one session only and partners' identities were unknown even when the experiment was over. The strategy method at the Observer's stage was implemented[9]. Each session lasted for about 20 minutes for the DG, 40 minutes for the TPP, and about 50 minutes for the MN. Each subject earned on average 7.4 Euros.

4. Predictions

In this section we derive the theoretical predictions about subjects' behavioral decisions for our experimental game. First, we illustrate the economic prediction based on the assumption that players are rational, driven by classic material self-interest and that it is common knowledge that this is the case. Then, we present the predictions of recently developed social preference theories.

4.1. Prediction based on the 'selfishness axiom'

In the TPP, we expect that rational, selfish third parties never punish, since subjects never meet one another more than once and punishment is costly for them. For the same reason, under common knowledge, the same expectation holds for both Observers and Receivers in the MN. In other words, our design completely rules out the possibility of both self-interested third-party sanctions and self-interested punishment on the part of rational Observers and Receivers in the MN. Then, we also expect that, in all three treatments, if we assume that common knowledge holds and that, hence, A believes that C and (in the MN) B are selfish (so that their threat to punish is considered an incredible one), a selfish Dictator transfers nothing to the Receiver. As a consequence, under the above assumptions, the following subgame perfect equilibrium outcomes are predicted, in the three treatments: zero transfer by A in the DG; zero transfer by A and zero punishment by C in the TPP and zero transfer by A and zero punishment by both C and B in the MN.

[9] When the strategy method is used, subjects are asked to state their decision in correspondence of each possible case. In our experiment, this meant that C was asked to indicate the number of deduction points for each of A's possible transfer levels before knowing A's actual choice. The final payoff was then determined on the basis of A's actual choice. In order to help Cs to think carefully about their decisions, an overview of the resulting payoffs has been made available (see Appendix 3). Several tests in simple games have not found behavior induced by the strategy method to be significantly different from behavior induced by the standard direct-response method (Charness et al., 2007). Also in Henrich et al. (2006) the strategy method was used with regard to third-party punishers' decisions.

4.2. Predictions based on social preference theories

Let us now turn to the predictions to which recently developed social preference theories lead. According to the Fehr and Schmidt (1999) model, based on the assumption that individuals dislike inequality, only in the DG transfers from player A to player B (if $\beta \geq 0.5$) are possible, while both in the TPP and in the MN neither transfer nor punishment are predicted. We obtain the same results if we consider the models of Bolton and Ockenfels (2000) and Charness and Rabin (2002).

To sum up, according to Fehr and Schmidt (1999), Bolton and Ockenfels (2000) and Charness and Rabin (2002) we expect positive transfer levels in the DG only, while nothing is predicted in the other two treatments.

5. Results

In this section we present our major results. In particular, we focus on three relevant points. Of first note is the finding that the baseline data are qualitatively similar to results reported in previous works. Second, we wonder whether what we identify as players' subjective reference principle of fairness (i.e. the subjective perception of fairness emerging by eliciting players' normative beliefs) influences people's behavior. Finally, we investigate the effect of metanorms on third parties' choices. On the whole, we claim that, analogously to Gaechter and Riedl (2004), there are two dimensions to this empirical assessment. The first dimension is the *moral* or *normative* intuition that people have on what they think a fair solution is, with regard to transfer levels in a Dictator Game. To this goal, agents' normative judgments over the fair transfer level by the Dictator are elicited. We do this for all the subjects, by considering all player roles and all the three treatments composing our experiment. The second dimension is *actual behavior* that may or may not be influenced by fairness considerations.

The results we obtain provide empirical evidence on both the normative and the behavioral dimension. More specifically, as far as our analysis of players' behavioral decisions are concerned, we ask whether a relationship between normative judgments and actual punishing choices exists, on the part of third parties. To this end, we compare Observers' punishment pattern in the TPP and in the MN, in order to check whether significant differences emerge, once the principle of fairness that we elicit by focusing on normative beliefs is taken into account.

5.1 Robustness of the results: (objective) fairness and punishment decisions

Result 1. *In line with the existing experimental literature, in the TPP the Observer's level of punishment decreases as the Dictator's transfers increase. A similar negative correlation emerges in the MN*

As far as the TPP is concerned, this result (see Figure 2) is perfectly in line with the experimental evidence obtained in other works (see for instance Fehr and Fischbacher, 2004[10]; Bernhard, 2005; Ottone, 2005, 2008). Our data indicate that this negative correlation between the Observer's level of punishment and the Dictator's level of transfers holds also in the MN. A random-effects Tobit regression (see Appendix 2) of punishment on the Dictator's transfer confirms that a positive relation exists between the level of punishment and the degree of unfairness (DISTANCE, p = 0.000).

5.2 Subjective unfairness

After that subjects' decisions are taken, their first-order normative beliefs are elicited[11]. By doing this, our goal is to provide an answer to the following question: what are the normative views that players hold about the fair division?

In particular, each player is asked to identify her ideal transfer from the Dictator to the Receiver (see Table 3). This allows to investigate two relevant points. Firstly, we can check whether our agents, while playing different roles in the games, share a common norm of fairness and whether such norm significantly differs or not from the egalitarian one. Secondly, we can study whether the norm of fairness that we identify by eliciting subjects' normative beliefs affects their punishment patterns in the two punishment treatments we consider (TPP and MN).

Result 2. *People share a common norm of fairness such as 'selfishness aversion', rather than an 'egalitarian' one*

If we compare the average right transfers according to the Dictators, we find out that there is no significant difference along the treatments (Kruskall-Wallis test; p = 0.55). The same is true if we compare the average right transfers according to both the Recipients and the Observers along the treatments (Kruskall –Wallis test, p = 0.15; Mann-Whitney test, p

[10] In particular, Fehr and Fischbacher (2004) find that almost two-thirds of the third parties indeed punished the violation of the distribution norm and that their punishment increased the more the norm was violated.

[11] Bernhard et al. (2006), in their third-party punishment experiments, elicit players' empirical expectations about how the dictators would be punished at different transfer levels, but they do not elicit players' normative beliefs, as we do.

= 0.22). This means that players' ideal 'right' transfer is not influenced by the experimental game they play.

When we compare different participants' normative beliefs to check whether the role they are called upon to play is relevant in determining people's perception of fairness, it turns out that the Recipients' right transfer is significantly higher than the Dictators' and the Observers' ones (ttest, p = 0.0002 and p = 0.0038 respectively). On the other hand, Dictators' and Observers' beliefs are aligned (ttest, p = 0.69). However, we find that, regardless of their role, the players share the idea that the right transfer from the Dictator to the Receiver is significantly different from both 0 (the selfish choice) and 5 (the egalitarian choice). We call this subjective norm of fairness 'selfishness aversion'. For a more detailed description of subjects' normative beliefs, see Table 3[12].

Result 3. *The subjective perception of fairness plays a relevant role in people's decisional process: third parties display reference-dependent fairness*

During the experiment the Observers were asked to identify their ideal transfer from the Dictator to the Receiver (see Table 3). If we view such ideal transfer as a subjective reference point of fairness, each Dictator's transfer lower than the ideal transfer may be considered unfair on the part of the Observer. In this light, it seems natural to analyze the Observer's reaction when her subjective principle of fairness is violated, in order to see whether third parties display a form of 'reference-dependent fairness', in their punishment pattern.

What turns out is that in the TPP the Observers' levels of punishment are sensitive to their subjective sense of fairness. In Figure 3 the relation between the (subjectively perceived) unfairness of Dictators and the level of punishment from the Observer to the Dictator is depicted. It is clear that the level of punishment increases as the Dictator's transfer becomes lower and lower than the Observer's ideal transfer: the Observers' perceived sense of fairness is *reference-dependent*. A random-effects Tobit regression (see Appendix 2) of punishment on the variables *Negative Subjective Unfairness*[13] confirms the existence of a positive relation between the level of punishment and the degree of negative subjective unfairness (ΔNEG, p = 0.000). From Figure 3 it emerges also that some

[12] A series of Wilcoxon tests is performed to check whether, for all player roles, the ideal transfer from the Dictator to the Receiver is different from 0 and from 5. In all cases, p = 0.000.

[13] *Negative Subjective Unfairness* is defined as max {0, Ideal Transfer – Actual Transfer}

punishment still exists even when the Dictator transfers to the Receiver a sum which is higher than the Observer's ideal transfer. However, by regressing punishment on the variable *Positive Subjective Unfairness*[14], we see that such relationship turns out not to be significant (ΔPOS, $p = 0.18$).

Also the Dictators appear to be affected by reference-dependent fairness, in choosing their transfer levels: if we compare their actual transfers to their ideal transfer, there is no significant difference (ttest, $p = 0.33$).

5.3 Metanorms

Result 4. *When violations of subjective fairness occur, third-party punishment is harsher in the MN than in the TPP*

When we add the possibility for the Receiver to punish the Observer (e.g. if s/he thinks s/he did not sanction enough an unfair Dictator), the Observers' behavior seems not to change (see Figure 2). If we compare the Observer's level of punishment at each level of the Dictator's transfer (see Table 1), we find out that there is no significant difference (Mann-Whitney test; $p > 0.14$). A random-effects Tobit regression confirms this result (MN, $p = 0.35$).

However, when we focus on third parties' reference-dependent fairness, that is on the Observers' behavior when their principle of fairness is violated, it emerges that when the Dictator's transfer becomes lower and lower than the Observer's ideal transfer, the Observer's reaction is significantly stronger in the MN (ΔNEG*MN, $p = 0.000$; see Figure 3, on this).

6. Discussion

Third parties are willing to nonstrategically punish Dictators

Our analysis confirms Fehr and Fischbacher's (2004) major findings: most of third parties indeed punish 'unfair' Dictators and the amount of punishment is negatively related to the level of Dictators' transfers (Result 1). Hence, we confirm that the notion of strong negative reciprocity extends to the sanctioning behavior of 'unaffected' third parties. Further, we reach a similar conclusion by passing from the TPP to the MN treatment. Like

[14] *Positive Subjective Unfairness* is defined as max {0, Actual Transfer – Ideal Transfer}

Fehr and Fischbacher (2004), and unlike several experimental designs, we focus on one-shot interactions, so that punishment in the TPP cannot be strategically motivated by the desire to induce higher contributions from other subjects in later periods. In other words, our design allows us to isolate nonstrategic sanctioning. Further, we interestingly find that both third parties' propensity to costly punish and the existence of a negative correlation between the amount of punishment and the level of Dictators' transfers hold not only by considering 'objective' fairness (Result 1), but also when we focus on players' subjective fairness (Results 3 and 4).

Players share a common norm of fairness such as 'selfishness aversion', rather than an 'egalitarian' one

Social norms constitute standards of behavior based on widely shared beliefs on how individuals ought to behave in a given situation (Elster, 1989). Our results show that, when the experimental game takes the form of the well-known Dictator Game, experimental players do not subjectively perceive as 'fair' the so called 'egalitarian distribution norm' (see Result 2)[15]. While we confirm the previous finding that subjects are (altruistically) willing to enforce a norm of fairness even though the enforcement is costly for them, a subjectively perceived norm of fairness (i) exists, (ii) is to some extent shared by all the players across treatments but (iii) clearly differs from a purely egalitarian norm of fairness. Fehr and Fischbacher (2004) hypothesize that the salient distribution norm in the Dictator Game is the equal split, that is for A to transfer half of the 'pie' to B, arguing that since the players interact anonymously and are randomly assigned their roles, there is no reason why A should end up with more money than B. They also elicit fairness judgments which clearly indicated that the egalitarian solution is perceived to be the fair solution. Bernhard et al.'s (2006) experimental study was designed to capture the altruistic enforcement of egalitarian sharing norms documented by ethnographic studies (see e.g. Boehm, 1993). As they observe, such sharing norms appear to be beneficial to the group as they insure their members against the uncertainties in individual food acquisition success. Their data reveal that there is little punishment at and above the egalitarian level, while 'unfair' proposals are more heavily sanctioned the more the Dictator deviates from the 'equality norm'. They

[15] The fact that subjects' sense of fairness is context-dependent was clear also in Ottone (2008). When participants have to *earn* their endowment, the Observers both punish and transfer less than when the endowment is randomly assigned. This may imply that the fairness reference point changes as the situation changes.

conclude that “This finding suggests the existence of an egalitarian sharing norm in all four conditions (…). The third parties’ and the dictators’ beliefs further support this interpretation” (p. 913). By contrast, we reach a different conclusion as, by eliciting players’ normative beliefs, we find that all of them expect a Dictator to give *something*, not to give half of his endowment to the Receiver. In other words, it turns out from our experiment that *there is a salient distribution norm, but also that such norm differs from the equality norm*. We also find that players appear to be consistent in this normative evaluation across treatments. In our view, all this means that in all treatments Dictators, Observers and Receivers believe that, if you happen to play as a Dictator, playing entirely selfishly is unfair. However, we interestingly also see that fairness is subjectively perceived by all of them not egalitarianly but as *selfishness aversion*: while all the players agree that Dictators should avoid to keep the whole amount for themselves, they do not believe that splitting the pie equally is morally compulsory.

Are fairness norms socio-economically and culturally-specific?

A possible interpretation of the remarkable difference between our finding and Bernhard et al.’s result has to do with the fact that our subjects are university students, rather than indigenous group members living in Papua New Guinea. In other words, the point here, while speculative, is that when people’s life priority is to survive, then sharing rules are *essential* to this. As a consequence, the egalitarian distribution norm becomes salient within such groups, as widely documented by ethnographic studies. On the contrary, within wealthy societies in which primary needs have to some extent been satisfied, people are driven by a ‘less extreme’ norm of fairness – what we term ‘selfishness aversion’. Such interesting difference, due to a combination of socio-economic and cultural motives, appears to be in line with the more general finding obtained in the last years by cross-cultural experimental research (see e.g. Henrich et al., 2001), that is the even *larger behavioral variability* observed after expanding the diversity of cultural and economic circumstances of experimental subjects, compared to previous experiments mainly involving university students.

Receivers' right transfer is significantly higher than Dictators' and Observers' right transfer

As far as players' normative beliefs are concerned, let us finally observe that a remarkable difference between Receivers and the other two players exists: Receivers' right transfer is significantly higher than Dictators' and Observers' ones. This interestingly seems to suggest that when normative evaluations about what is fair within a given game protocol are asked to players involved in it, a Rawlsian 'veil of ignorance' would be necessary in order to elicit genuinely 'impartial' views about 'what is fair' from the players. By borrowing Loewenstein's (2000) well-known notion of 'hot-cold' empathy gap, we may speculatively argue that, in a TPP, both the Dictator and the Receiver are in a 'hot state', while the third party is, by definition, in a 'cold state'. Hence her normative judgments should reflect the viewpoint of an impartial spectator, due to her being a 'financially uninvolved' third party. We find that even though they are in a different state in making their normative judgments, Dictators' and Observers' normative beliefs are aligned. By contrast, Receivers' higher levels of right transfer are in line with their being in a 'hot state'. We may conclude that if the 'hot-cold' empathy gap strongly affects normative evaluations in a well-controlled laboratory environment, such distinction is likely to play an even more important role in outside-lab economic interactions.

When subjective fairness is taken into account, Observers' punishment is both nonstrategic and strategically motivated

Why do people costly punish others, when sanctioning yields neither present nor future material benefits? Fehr and Gaechter (2002) assert that negative emotions such as anger could be the proximate mechanism toward punishment. However, as Anderson and Putterman (2006) point out: "Although the presence of anger need not rule out systematic behavioral rules (…) if taken to its limit, the emotion approach might suggest behaviour which is simply not amenable to rational analysis" (p. 2). Anderson and Putterman's experiment tests whether the demand for punishment displays the usual downward slope with respect to price; hence, it also indirectly tests "whether a rational choice-with-social-tastes characterization, or an irrational anger description, is more accurate" (p. 3). Their analysis provides evidence that (i) nonstrategic punishment of free riders is common and that (ii) the demand for punishment does obey the Law of Demand (i.e. the quantity purchased is a decreasing function of the price). Therefore, a 'rational choice-with-social-tastes' characterization of punishers emerges from their study.

With regard to the nature of Observers' punishment in our experiment, it is interesting to ask a qualitatively similar question: is their attitude towards punishment *entirely nonstrategic*? In principle it is not clear whether punishment behavior underlies non-standard preferences or a relevant influence of emotional factors. While we cannot rule out that both factors are at work, our overall results tend to favor the first interpretation, as we find that Observers' punishment appears to be partially strategically motivated. More specifically, we can reject the hypothesis that punishing subjects are driven by non-rational factors (such as anger) – acting as 'moral sentiments' – only, since the agents playing as Observers in the MN are sensitive to the presence of potential punishers (i.e. the Receivers). In particular, insofar as the Observers see 'selfishness aversion' as the salient triggering norm, we find that such agents are not fully nonstrategically motivated, as the concern over the possibility of being punished by the Receivers seems to serve to discipline them and induces them to punish *significantly more* (see Result 4)[16].

This result appears to be in line with the more general finding obtained so far by experimental studies on sanctioning, that is the ordinary good nature of punishment. Anderson and Putterman (2006), through a series of experiments in which they randomly vary the cost of reducing the earnings of other group members following voluntary contribution decisions, show that the impulse to nonstrategically punish others is sensitive to the cost to the punisher. While we cannot reach exactly the same conclusion – as the costs of punishing do not vary in our experiment – we broadly confirm their general qualitative finding: even when it has a nonstrategic nature, punishment is sensitive to costs[17]. More specifically, while existing experiments characterize nonstrategic punishment as an ordinary good when the punisher is *not* a *third party*, we find that also third parties appear to be sensitive to the 'price' of altruistic punishment.

[16] However, while economics and game theory have traditionally favored 'consequentialist' interpretations of agents' behavior, it is important to make clear that, in principle, we cannot rule out that the Observers are also *sensitive to social presence per se* (see, on this, Kurzban et al., 2007) in the MN. In other words, in this context their increased punishment may be due to psychological and/moral reasons such as guilt aversion or the desire not to disappoint the Receivers, whose presence is made 'active' by the design of MN. The inclusion in the design of the MN of 'victims of unfairness' endowed with a punishing option may make such 'social presence' salient in the eyes of the Observers, in line with a thesis advanced in social psychology and often associated with the notion of so called 'moralistic punishment'.

[17] This also parallels Andreoni and Miller's (2002) finding, that is the consistency between altruistic behavior and rationality, as well as Carpenter's (2007) result that the demand for punishment is negatively related to its 'price'. Another experimental study documenting an inverse relationship between costs and frequency of punishing is Egas and Riedl (2005).

On the nature of the Observers' reasoning

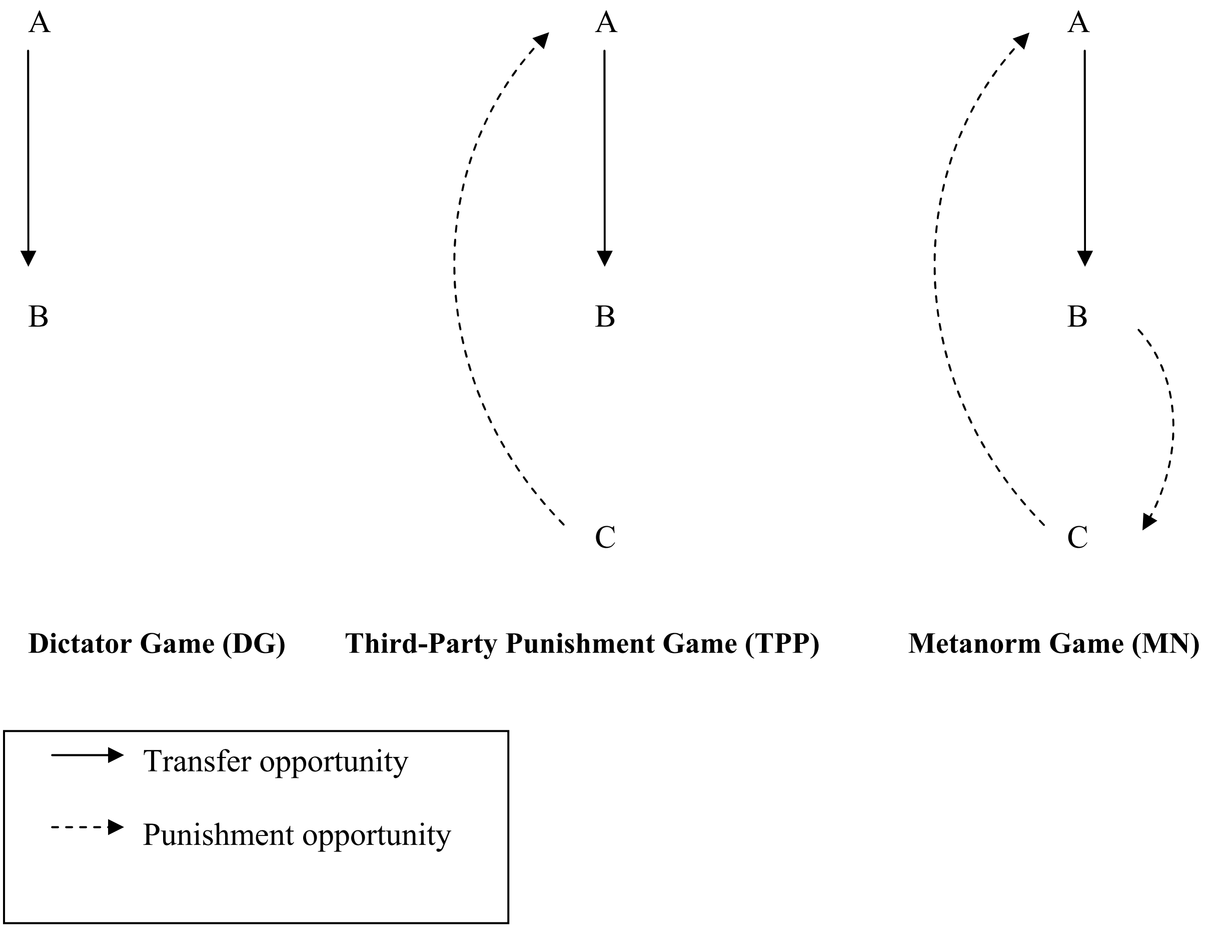
We can interestingly shed further light on player C's behavior across treatments by laying stress on the fact that, in the MN, B's potential punishment of C is entirely nonstrategic. In other words, such potential punishment turns out to be *credible* (as it affects C's punishment behavior) even though it is *nonstrategic*. Why does this occur? Why are players C affected by the possibility to be punished by Bs, if they know that Bs' potential punishment is nonstrategic? A possible explanation is that players C do not clearly understand that Bs' punishment is nonstrategic. A second interpretation is the following: players C know that Bs' potential sanctioning is nonstrategic; however, such sanctioning is credible, as they know that people can decide to nonstrategically punish others, as they are willing to do towards Dictators. This would mean that players C are willing to punish nonstrategically others, but they are also (consistently) afraid to be punished nonstrategically. In other words, it is interesting to find that players C are affected by both nonstrategic and strategic considerations in their punishment choices towards A and that the strategic component of their reasoning is crucially affected by (potential) nonstrategic punishment on the part of players B. On the whole, we then find that interesting and complex interplays between strategic and nonstrategic considerations take place, when 'rational' players displaying 'non-selfish' behavior are involved.

Comparison with proximate theories of social preferences

If we referred exclusively to the models of social preferences mentioned in Section 4, the data that come out from our experiment could not be satisfactorily explained. First of all, we expected transfers from A to B only in the Dictator Game. In our experiment, the positive level of transfer in the three treatments is not statistically different. Moreover, the level of transfer from A to B that subjects consider fair does not vary across treatments. Second, subjects C punish in TPP and in MN while the theoretical predictions in the models mentioned above do not consider intervention.

A possible explanation for these results could be that subjects' consideration of the others changes as their role changes. As argued by Fehr and Schmidt (2006), the relative relevance of each player in a game is an open issue. In order to understand the behavior of our players, we may hypothesize that some subjects A reveal a great interest in the condition of subjects B, while they are not interested in the pay-off of player C. At the same time, some players C could be not interested in their own payoff when they have to decide to punish or not subject A. Subjects C, in this case, could feel as a judge in the relationship between A and B. Moreover, we may think that C puts herself in B’s shoes and reacts to an unfair transfer of A as if she were acting as a Receiver in an Ultimatum Game. However, such hypotheses need further inquiry and are left for future research.

Appendix 1 – Figures and Tables



Dictator Game (DG) **Third-Party Punishment Game (TPP)** **Metanorm Game (MN)**

Fig. 1 The Experimental Games

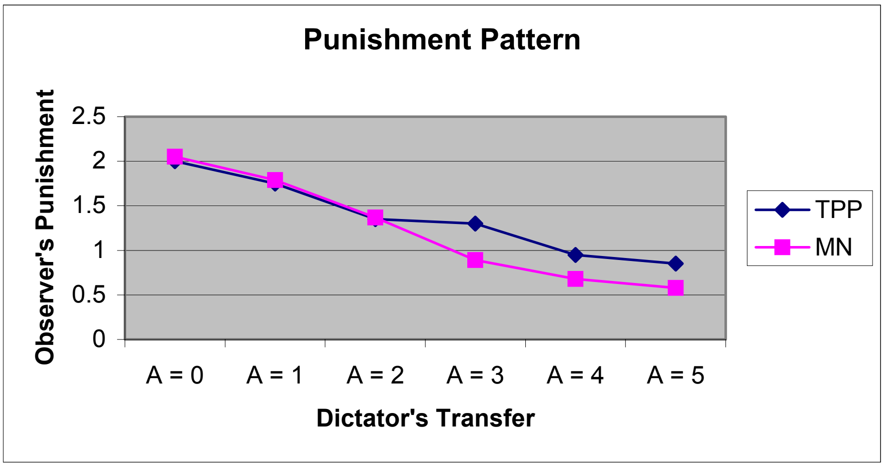


Fig. 2 Observer's behavior

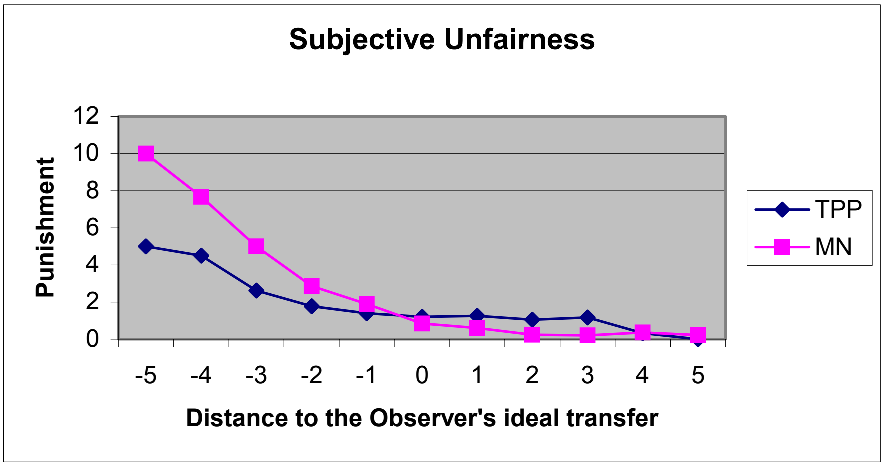


Fig. 3 Subjective Unfairness

Table 1.

	When the Dictator transfers …					
Average Punishment in …	**0**	**1**	**2**	**3**	**4**	**5**
TPP	2	1.75	1.35	1.3	0.95	0.85
MN	2.05	1.79	1.37	0.89	0.68	0.58
Mann-Whitney test	*p =0 .56*	*p =0 .48*	*p = 0.32*	*p =0 .14*	*p =0 .26*	*p =0 .72*

Table 2.

	A transfers to B (1)	**A thinks it is right to transfer to B (2)**	**B thinks it is right to transfer to B (3)**	**C thinks it is right to transfer to B (4)**	***Kruskall – Wallis test (2) = (3) = (4)***	***t- test (2) = (3) (3) = (4) (2) = (4)***
DG	1.4	1.5	3.1			*p =0 .0002* *p = 0.0038* *p =0 .69*
TPP	1.55	1.85	3	2	*p =0 .11*	
MN	1.53	1.42	2.1	1.42	*p =0 .35*	
Kruskall - Wallis test DG = TPP = MN	*p =0 .88*	*p =0 .55*	*p =0 .15*	*p =0 .22*		

Table 3.

	Percentage of cases of	0	1	2	3	4	5
DG	Dictators who transfer…	40%	15%	25%	10%	5%	5%
	Dictators who think the right transfer is …	30%	20%	30%	15%	0%	5%
	Recipients who think the right transfer is…	5%	10%	30%	15%	5%	35%
	Observers who think the right transfer is …	-	-	-	-	-	-
TPP	Dictators who transfer…	30%	20%	25%	20%	0%	5%
	Dictators who think the right transfer is …	30%	10%	25%	25%	0%	10%
	Recipients who think the right transfer is …	10%	0%	40%	15%	0%	35%
	Observers who think the right transfer is …	25%	5%	30%	30%	5%	5%
MN	Dictators who transfer…	21%	42%	21%	0%	11%	5%
	Dictators who think the right transfer is …	32%	26%	32%	0%	0%	11%
	Recipients who think the right transfer is …	26%	11%	32%	11%	0%	21%
	Observers who think the right transfer is …	47%	11%	16%	11%	11%	5%

Appendix 2 – The econometric analysis

$$P_i = \beta_0 + \beta_1 DISTANCE_i + \beta_2 MN_i + \beta_3 AGE + \beta_4 GENDER_i + \varepsilon_i$$

(S1)

$$P_i = \beta_0 + \beta_1 \Delta NEG_i + \beta_2 \Delta POS_i + \beta_3 \Delta NEG * MN_i + \beta_2 \Delta POS * MN_i + \beta_5 AGE + \beta_6 GENDER_i + \varepsilon_i$$

(S2)

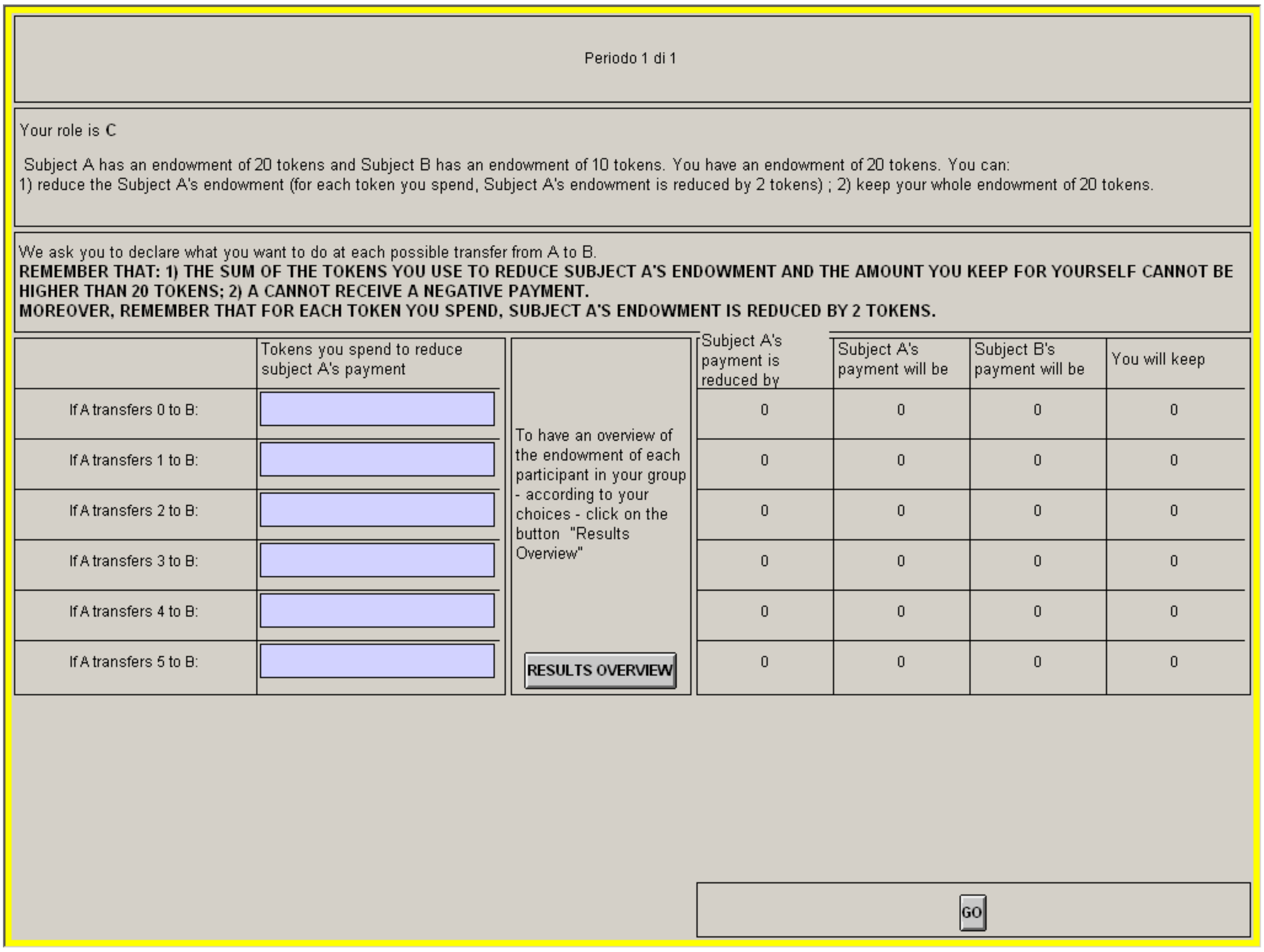
Dependent variable: punishment (P*i*)
Random-effects Tobit regression – censored at the low level (0)

Variables	*S1*	*S2*
DISTANCE	0.57*** (0.087)	-
ΔNEG	-	0.684*** (0.164)
ΔPOS	-	-0.19 (0.14)
MN	-0.625 (0.673)	
ΔNEG *MN		0.882*** (0.239)
ΔPOS *MN		-0.044 (0.20)
AGE	0.092 (0.13)	-0.02 (0.138)
GENDER	-0.30 (0.72)	0.063 (0.688)
Constant	-2.9 (2.77)	0.075 (2.787)
n	39	39
T	6	6
N	234	234
Log Likelihood	-280.68683	-261.51373
Sigma_u	3.2***	2.6***
Sigma_e	1.67***	1.42***

*****significance 1%**
Description of the variables used in the regression

Name	**Descripton**	**Mean**	**Standard Deviation**	**Min**	**Max**
DISTANCE	5 – transfer from A to B	2.5	1.711	0	5
ΔNEG	max {0, Ideal Transfer – Actual Transfer}	0.59	1.093	0	5
ΔPOS	max {0, Actual Transfer – Ideal Transfer}	1.37	1.589	0	5
MN	Dummy variable equal to 1 if the observation belongs to the MN	-	.501	0	1
ΔNEG*MN	Variable equal to max {0, Ideal Transfer – Actual Transfer} if the observation belongs to the MN; otherwise 0	0.25	0.791	0	5
ΔPOS*MN	Variable equal to max {0, Actual Transfer – Ideal Transfer} if the observation belongs to the MN; otherwise 0	0.77	1.419	0	5
AGE	Age	20.6	2.173	18	26
GENDER	Dummy variable equal to 1 if male	-	0.462	0	1

Appendix 3 – Observer's decision screen



Periodo 1 di 1

Your role is C

Subject A has an endowment of 20 tokens and Subject B has an endowment of 10 tokens. You have an endowment of 20 tokens. You can:
1) reduce the Subject A's endowment (for each token you spend, Subject A's endowment is reduced by 2 tokens) ; 2) keep your whole endowment of 20 tokens.

We ask you to declare what you want to do at each possible transfer from A to B.
REMEMBER THAT: 1) THE SUM OF THE TOKENS YOU USE TO REDUCE SUBJECT A'S ENDOWMENT AND THE AMOUNT YOU KEEP FOR YOURSELF CANNOT BE HIGHER THAN 20 TOKENS; 2) A CANNOT RECEIVE A NEGATIVE PAYMENT.
MOREOVER, REMEMBER THAT FOR EACH TOKEN YOU SPEND, SUBJECT A'S ENDOWMENT IS REDUCED BY 2 TOKENS.

	Tokens you spend to reduce subject A's payment		Subject A's payment is reduced by	Subject A's payment will be	Subject B's payment will be	You will keep
If A transfers 0 to B:		To have an overview of the endowment of each participant in your group - according to your choices - click on the button "Results Overview"	0	0	0	0
If A transfers 1 to B:			0	0	0	0
If A transfers 2 to B:			0	0	0	0
If A transfers 3 to B:			0	0	0	0
If A transfers 4 to B:			0	0	0	0
If A transfers 5 to B:		RESULTS OVERVIEW	0	0	0	0

References

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.

Andreoni, J., Miller, J., 2002. Giving according to GARP: An experimental test of the consistency of preferences for altruism. Econometrica, 70, 737-753.

Axelrod, R., 1986. An evolutionary approach to norms. American Political Science Review, 80 (4), 1095-1111.

Bernhard, H., 2005. Third party punishment within and across groups – An experimental study in Papua New Guinea, Institute for Empirical Research in Economics, University of Zürich, Preliminary Working Paper.

Bernhard, H., Fischbacher, U., Fehr, E., 2006. Parochial altruism in humans. Nature, 442, 912-915.

Bolton, G.E., Ockenfels, A., 2000. A theory of equity, reciprocity and competition. American Economic Review, 90, 166-93.

Boehm, C., 1993. Egalitarian behavior and reverse dominance hierarchy. Current Anthropology, 34, 227-254.

Carpenter, J., 2007. The demand for punishment. Journal of Economic Behavior and Organization, 62, 522-542.

Carpenter, J., Matthews P.H., 2006. Norm enforcement: anger, indignation or reciprocity?, unpublished manuscript.

Carpenter, J., Matthews, P.H., 2007. What norms trigger punishment?, unpublished manuscript.

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

Charness, G., Cobo-Reyew, R., Jimenez, N., 2007. An investment game with third-party intervention, Department of Economic Theory and Economics History, University of Granada, ThE Papers, 06/13.

Charness, G., Rabin, M., 2002. Social preferences: some simple tests and a new model. Quarterly Journal of Economics, 117, 817-69 .

Charness, G., Rabin, M., 2005. Expressed preferences and behavior in experimental games. Games and Economic Behavior, 53, 151-169.

Charness, G., Haruvy, E., Sonsino, D., 2007. Social distance and reciprocity: An Internet experiment. Journal of Economic Behavior and Organization 63, 88-103.

de Waal, F.B.M., 1996. Good natured: the origins of right and wrong in humans and other animals. Cambridge (Mass.), Harvard University Press.

Dufwenberg, M., Gneezy, U., 2000. Measuring beliefs in an experimental lost wallet game. Games and Economic Behavior, 30, 163-182.

Egas, M., Rield, A., 2005. The economics of altruistic punishment and the demise of cooperation, Tinbergen Institute Discussion Paper.

Elster, J., 1989. The cement of society. A study of social order. Cambridge, Cambridge University Press.

Falk, A., Fehr, E., Fischbacher, U., 2008. Testing theories of fairness – Intentions matter. Games and Economic Behavior, 62, 287-303.

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

Fehr, E., Gachter, S., 2000. Cooperation and punishment. American Economic Review, 90, 980-994.

Fehr, E., Schmidt, K., 1999. A theory of fairness, competition and cooperation. Quarterly Journal of Economics, 114, 817-51.

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

Forsythe, R., Horowitz, J.L., Savin, N.E., Sefton, M., 1994. Fairness in simple bargaining experiments. Games and Economic Behavior, 6, 347-369.

Fudenberg, D., Maskin, E., 1986. The Folk Theorem in repeated games with discounting or with incomplete information. Econometrica, 50, 533-554.

Gaechter, S., Riedl, A., 2004. Dividing justly in bargaining problems with claims. Tinbergen Institute Discussion Paper.

Gintis, H., Bowles, S., Boyd, R., Fehr, E., 2003. Explaining altruistic behavior in humans. Evolution and Human Behavior, 24, 153-172.

Gintis, H., Bowles, S., Boyd, R., Fehr, E. (eds), 2005. Moral sentiments and material interests. The foundations of cooperation in economic life, Cambridge (Mass.) and London, MIT Press.

Guth, W., Schmittberger, R., Schwartze, B., 1982. An experimental analysis of ultimatum games. Journal of Economic Behavior and Organization, 3, 367-388.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., 2001. Cooperation, reciprocity and punishment in fifteen small-scale societies. American Economic Review, 91, 73-78.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., 2006. Costly punishment across human societies. Science, 312, 1767-1770.

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

Kurzban, R., DeScioli, P., O'Brien, E., 2007. Audience effects on moralistic punishment. Evolution and Human Behavior, 28, 75-84.

List, J., 2007. On the interpretation of giving in dictator games, Journal of Political Economy, 115 (3), 482-493.

Marlowe, F.W., Berbesque, J.C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J.C., Ensminger, J., Gurven, M., Gwako, E., Henrich, J., Henrich, N., Lesorogol, C., McElreath, R., Tracer, D., 2008. More 'altruistic' punishment in larger societies. Proceedings of the Royal Society Biology, 275, 587-590.

Nowak, M.A., Sigmund, K., 1998. Evolution of indirect reciprocity by image scoring. Nature, 393, 573-577.

Ottone, S., 2005. Transfers and altruistic punishments in Solomon's game experiments, Paper n.57, AL.EX Series, POLIS, University of Eastern Piedmont.

Ottone, S., 2008. Are people Samaritans or avengers? Economics Bulletin, 3 (10), 1-8.

Rubinstein, A., 1979. Equilibrium in supergames with the overtaking criterion. Journal of Economic Theory, 21, 1-9.

Sally, D., 1995. Conversation and cooperation in social dilemmas: a meta-analysis of experiments. Rationality and Society, 7, 58-92.

Sober, E., Wilson, D.S., 1998. Unto others: the evolution and psychology of unselfish behavior, Cambridge (Mass.), Harvard University Press.

Takahashi, N., Miyahara, M., Mashima, R., 2005. Does intention matter in third-party punishment?, Paper presented at the 17th annual meeting of Human Behavior and Evolution Society, Austin, Texas, June 1-5.