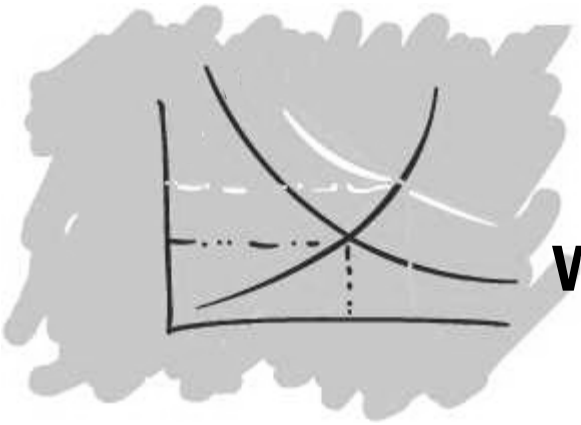
I.S.S.N: 1885-6888



ECONOMIC ANALYSIS WORKING PAPER SERIES

Individual Heterogeneity in Punishment and Reward

♦

Andreas Leibbrandt and Raúl López-Pérez

Working Paper 1/2011

UAM
UNIVERSIDAD AUTONOMA
DE MADRID

DEPARTAMENTO DE ANÁLISIS ECONÓMICO:
TEORÍA ECONÓMICA E HISTORIA ECONÓMICA

Individual Heterogeneity in Punishment and Reward[*]

Andreas Leibbrandt and Raúl López-Pérez[†]

February 1, 2011

Abstract: We design experiments to study the extent to which individuals differ in their motivations behind costly punishment and rewarding. Our findings qualify existing evidence and suggest that the largest fraction of players is motivated by *a mixture of both* inequity-aversion and reciprocity, while smaller fractions are primarily motivated by pure inequity-aversion and pure reciprocity. These findings provide new insights into the literature on other-regarding preferences and may help to reconcile important phenomena reported in the experimental literature on punishment and reward.

Keywords: Heterogeneity, inequity aversion; monetary punishment/reward; reciprocity; social norms.

JEL Classification: C70, C91, D63, D74, Z13.

[*] We are grateful for useful comments and suggestions from Hubert Kiss, Antonio Martín-Arroyo, and participants at the IMEBE 2010 and THEEM conferences. In addition, María García-Sola provided helpful research assistance. We also gratefully acknowledge financial support from the Spanish Ministry of Science and Innovation through the research project ECO2008-00510.

[†] Leibbrandt: University of Chicago, Chicago, IL 60637, USA and Workshop in Political Theory and Policy Analysis, Indiana University, USA. E-mail: leibbrandt@uchicago.edu. López-Pérez: Department of Economic Analysis, Universidad Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain. E-mail address: raul.lopez@uam.es.

1. Introduction

A large body of experimental data suggests that people have heterogeneous non-selfish preferences. For instance, laboratory studies show that subjects differ in their cooperativeness in social dilemma games (Ledyard, 1995; Fehr and Gächter, 2000), rejection behavior in ultimatum games (Güth, 1995; Roth, 1995), generosity in dictator games (Andreoni and Miller, 2002; Charness and Rabin, 2002), and effort provisions and wage offers in labor market games (Fehr et al, 1993). As these examples suggest, an accurate picture of heterogeneity in motivations may significantly improve our understanding of important phenomena. In addition, heterogeneity may help us to understand costly punishment and rewarding, behaviors which appear to be strongly coined by different non-selfish motivations (Blount, 1995; Offerman, 2002; Falk et al, 2003; Charness, 2004; Falk et al, 2005; Dawes et al., 2007; Falk et al, 2008; Nikiforakis, 2008; Johnson et al, 2009) and are crucial to explain cooperation and compliance with social norms (Fehr and Gächter, 2002).[1]

In this paper we present four simple games, use a within-subject experimental design, and apply a maximum-likelihood classification procedure to study whether and how individuals differ in their motivations behind costly punishment and rewarding. In these four games, a first mover chooses between two allocations of payoffs for her and a second mover. The second mover then observes this choice and decides whether and how strongly she wants to punish or reward the first mover. The within-subjects design and the different payoff constellations of the games make it possible to disentangle which motivation each second mover follows most likely. In this respect, we investigate five basic motivations to punish or reward which are prominent in the literature on non-selfish preferences: Inequity-aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), reciprocity (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Cox et al., 2007), altruism (Andreoni and Miller, 2002), competitiveness (as suggested in Levine, 1998), and spitefulness (see also Levine, 1998). In addition, our design allows us to analyze more complex motivations such as combinations of some of these previous forces (for an example, see Falk and Fischbacher, 2006).[2]

[1] Analog to other experimental papers, we define punishment in two-player games as a costly action that reduces the payoff of the co-player. Rewards are defined analogously.

[2] Although other papers have studied heterogeneity of non-selfish preferences before (e.g., Charness and Rabin, 2002), this has been mainly done in the context of dictator games, which do not take into account reciprocity; or in public goods games, which do not permit to disentangle motivations like reciprocity and inequity aversion. Moreover, few papers have simultaneously studied punishment and reward (see Abbink et al., 2000, and Falk et al., 2008, for exceptions), which seems crucial to test the different models of social preferences.

Our classification analysis suggests that the largest fraction of subjects is motivated by a mixture of both reciprocity and equity concerns. More precisely, a second mover following this motivation punishes the first mover only if this player harmed her *and* has a larger payoff than her – the first pattern indicates reciprocal concerns, the second egalitarian concerns. Analogously, she rewards the first mover only if this player helped her *and* has a smaller payoff. In addition, there are considerable fractions of subjects who are *purely* inequity-averse, *purely* reciprocal, or selfish, and smaller fractions of players who are predominantly altruistic, competitive, or spiteful. However, the analysis indicates that models with more than three types do not provide significant improvements for the interpretation of the observed behavioral patterns. Hence, our results do not only clarify whether heterogeneity is relevant, but also provide an idea about the extent to which heterogeneity should be accounted for.

This paper contributes to the experimental and theoretical literature on non-selfish motivations, punishment, and rewards in at least four ways. *First*, we design experimental games which render it possible to discriminate between many prominent motivations for punishing and rewarding, *including* combinations of some of the basic models. This is important among other reasons because there seems to be little consensus in the literature with regard to the main force(s) behind costly punishment and reward. With respect to punishment, for instance, Falk et al. (2005, p. 2017) conclude that "retaliation […] seems to be the most important motive behind fairness-driven informal sanctions", whereas Johnson et al. (2009, p. 192; see also Dawes et al., 2007) suggest that "individuals who care about equality are those who are most willing to punish free-riders in public goods games". *Second*, we show that a large fraction of individuals is motivated by a mixture of reciprocity and equity considerations, whereas relatively smaller fractions tend to behave as purely inequity-averse or reciprocal. As we explain in section four, these findings may be useful to reconcile earlier seemingly contradicting results in the experimental literature on punishment (Blount, 1995; Offerman, 2002; Falk et al, 2003; Charness, 2004; Falk et al, 2005; Dawes et al., 2007; Falk et al, 2008; Johnson et al, 2009), and possibly also on reward. *Third*, we provide evidence for the statistical significance of heterogeneity with the help of a rigorous classification analysis, encouraging the development of models which incorporate different motivations. *Fourth*, we use a punishment/rewarding technology which prevents potential confounds due to non-linear preferences when studying the determinants of punishment/reward.

The rest of the paper proceeds as follows. The next section describes the experimental design and procedures. Section three shows the predictions of the main theories in our games, reports our main results, and applies the classification procedure. Section four reviews related

literature and suggests how our results may help to organize earlier experimental findings. The fifth section concludes.

2. Experimental Design and Procedures

Participants in our experiment play four games, and all of these games are two-player games of perfect information and an identical two-stage structure. In the first stage, one player (called A) chooses between a left-hand and right-hand allocation of money between herself and another player (called B). Table 1 shows the two (A, B) payoff allocations available in each game (the different payoff constellations render it possible to discriminate between several models of non-selfish preferences; see section 3.1). Payoffs are presented in points; the exchange rate was 10 points = 1 Euro.

TABLE 1—THE ALLOCATIONS IN THE 4 GAMES

		Game			
		1	2	3	4
Allocation	Left	(250, 100)	(250, 100)	(100, 200)	(100, 200)
	Right	(200, 150)	(250, 250)	(150, 150)	(100, 300)

In the second stage, B observes A's choice and can then punish or reward A at a *fixed* cost of five points for B. More precisely, B can either decrease or increase A's allocation by up to 100 points, but only if B decided to pay five points from her allocation share. If B does not want to affect A's balance, no points are deducted from her allocation share. For example, suppose that A chooses allocation (x_A, x_B) in a game. If B decides not to pay the five points, allocation (x_A, x_B) is implemented. If she pays the five points, however, she can choose a score $s \in [-100, 100]$ so that A's payoff in the game is $x_A + s$, while B gets a payoff of $x_B - 5$. For simplicity, s had to be a multiple of 10.

We chose a punishment/reward technology with a fixed cost to facilitate the test of the theories and prevent possible confounds due to non-linear preferences. In effect, since additional units of punishment/reward have no additional costs in our design, any theory predicts that the players who punish/reward will do it with the same intensity independently of the curvature of their preferences. As an example, punishers/rewarders who are inequity-averse should punish/reward with equal strength even if they are heterogeneous with respect to their marginal disutility from inequity. If additional units of punishment were costly, in contrast, players need not punish/reward equally strongly, and that would greatly complicate

the theoretical analysis. In addition, this technology also prevents the existence of multiple equilibria in reciprocity models in our games.

We conducted four paper-and-pencil sessions at University of (location), with a total of 92 participants. Participants were students from different disciplines (14 percent were economics students), and they were not students of the experimenters. Before each session started, we distributed instruction and decision sheets (dependent on role; see appendix I for the instructions of the B-players) in a class room, leaving enough space between seats to ensure anonymity. When the participants entered the room, they were informed that the decision sheets had been randomly distributed. The sheets were initially covered and the subjects could freely choose their seat; in that manner, we assigned them to be either an A- or a B-player. Subjects could read the instructions at their own pace and we answered their questions in private. We used neutral language and avoided terms such as "punishment". Before proceeding with their decisions, participants had to fill out control questions to make sure that they understood the rules.

Each subject played the games in the same role and with the same anonymous co-player. Yet reputation effects were impossible, as no subject was informed about her counterpart's actual choice in any game. This feature of our design also prevents changes of mood which may complicate the data analysis (i.e., the mood of a B-player could change depending on the A-player's choice in a preceding game). Since players received no feedback, we employed the strategy method to elicit the decisions of the B-players, i.e., they indicated for both allocations in each of the four games whether they wanted to pay five points to affect their counterpart's balance, and if this was the case, they had to decide which score $s \in [-100, 100]$ they wanted to assign to their co-player.[3]

After subjects made their decisions in the four games, they answered a brief questionnaire. Then we collected their decision sheets and selected only one game randomly for payment in order to prevent possible income effects. This and all other features of the experiment were common knowledge to the subjects. Subjects were paid in private, and earned on average 18.3 Euros. Each session lasted approximately 60 minutes.

[3] One additional advantage of the strategy method is that it maximizes the amount of statistical data gathered. In principle, the strategy method may induce different behaviors than the specific response method, where participants know the choice made by the other player. However, Falk et al. (2005) investigate this issue and find no differences in subjects' punishment patterns, although the strength of punishment seems to be somewhat lower overall when using the strategy method. In addition, Brandts and Charness (2009) review the experimental studies that use both methods and find no treatment differences in most of them. Moreover, they find that differences are particularly unlikely in experiments in which players make numerous choices (as in ours).

3. Punishment and Rewards: Theory and Evidence

This section studies whether and how individuals differ in their motivations behind costly punishment and rewarding. To organize our analysis, we first consider five prominent utility models that provide different rationales for costly punishment and reward, and derive their predictions for our games (3.1). We then briefly report some aggregate data across individuals (3.2) before analyzing behavior on the individual level and, with the help of a classification analysis, investigating the distribution of motivations across subjects (3.3).

3.1 Theories

We start by noting that **selfish** players never punish or reward in our games because these behaviors are costly and moreover reputation effects are ruled out by the fact that players are provided no feedback about the co-player's choices. As a result, punishment and reward can only be explained by non-selfish motivations. We consider five utility models.

The first two models predict *un*conditional behavior. To start, models of **spitefulness** (Kirchsteiger, 1994; Levine, 1998) assume that a player's utility depends negatively on the co-player's material payoff.[4] Formally, if u_i and x_i denote a player's utility and material payoff, respectively, we have $u_i = x_i + \sigma \cdot x_j$ ($i \neq j$; $\sigma \leq 0$). In our games, a spiteful B should punish A maximally (i.e., 100 points) at all allocations if $|\sigma|$ is larger than 0.05.[5] In contrast, theories of **altruism** (Andreoni and Miller, 2002; see also the model of quasi-maximin preferences in Charness and Rabin, 2002) assume $\sigma \geq 0$ in the previous utility function and hence predict that an altruistic B should reward A maximally at all allocations if $\sigma > 0.05$. Any spiteful or altruistic player with $|\sigma| < 0.05$ will never pay the 5 points fee, thus behaving like a self-interested player.

Consider now models of **reciprocity**. The pioneering model is Rabin (1993), which applies to any two-player, normal form game. Dufwenberg and Kirchsteiger (2004) extend Rabin's approach to a large class of n-player extensive form games; given the sequential nature of our games, we focus on this model here. Table 2 summarizes the predictions by this model in our games for a reciprocal enough player B (more precisely, for a B-player with a reciprocity parameter Y large enough; see Appendix II for a more detailed explanation of these predictions). The table also includes in parenthesis the precise score s predicted by the

[4] Levine (1998) assumes the existence of other types of players. However, the empirical relevance of his model relies on the existence of a significant proportion of spiteful types.

[5] Such behavior increases the utility of a spiteful player in $100 \cdot |\sigma|$, at a cost of 5 –i.e., the fee. Hence, it is optimal only if $|\sigma| > 0.05$.

theory (a negative score indicates punishment, while a positive one indicates reward). Intuitively, B punishes if A harms her, that is, if A chooses the allocation (x_A, x_B) of the game with the smallest x_B; while B rewards if A helps her, that is, if A chooses the allocation (x_A, x_B) of the game with the largest x_B. In game 1 (250/100 vs. 200/150), for example, B punishes A if she chooses allocation (250/100), and rewards her if she chooses (200/150). Since the cost of punishment/reward is fixed, it follows also that a B-player who punishes/rewards will do it with the maximal strength (i.e., -/+ 100 points).

TABLE 2—PREDICTIONS OF COSTLY PUNISHMENT/REWARD

	Allocation			Predictions *Left*	Predictions *Right*
Game	*Left*		*Right*		
1	(250,100)	vs.	(200,150)	C, IA, R (0, -100, -100)	C, IA, R (≤ -60, -60, +100)
2	(250,100)	vs.	(250,250)	C, IA, R (0, -100, -100)	C, IA, R (≤ -10, 0, +100)
3	(100,200)	vs.	(150,150)	C, IA, R (0, +90, +100)	C, IA, R (≤ -10, 0, -100)
4	(100,200)	vs.	(100,300)	C, IA, R (0, +90, -100)	C, IA, R (0, +100, +100)

The following notation is used: C = Competitive, IA = Inequity aversion, R = Reciprocity. In parenthesis, we depict the precise score s predicted by each theory at the corresponding allocation –e-g-, in the right-hand allocation of game 1, C predicts any score smaller or equal than -60, IA predicts punishment of -60, and reciprocity a reward of +100. The predictions of inequity-aversion and competitiveness take into account that subjects could choose only scores multiples of 10.

Models of **inequity-aversion** (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) assume that there are agents who dislike inequity of payoffs and are willing to spend money to reduce it. For instance, Fehr and Schmidt (1999) posit the following utility function in two-player games:

$$u_i = x_i - \alpha \cdot \max\{x_j - x_i, 0\} - \beta \cdot \max\{x_i - x_j, 0\},$$

where α, $\beta > 0$ and $\alpha > \beta$. Table 2 indicates the predictions of this model in our games for large enough α's and β's (see Appendix II for a more detailed explanation). Intuitively, an

inequity-averse B-player punishes A at *any* allocation in which A gets a higher payoff, provided that α is sufficiently large. Since additional units of punishment are costless, B-players punish so as to reduce inequity as much as possible, taking into account that only multiples of 10 were allowed when punishing and rewarding. In the left-hand allocation of game 1 (250/100 vs. 200/150), for example, B should reduce A's payoff by 100 points if her α is larger than 0.1. Analogously, an inequity-averse B-player rewards A at *any* allocation in which A gets a lower payoff, provided that her β is large enough. The strength of the reward should be such that the payoff distance is maximally reduced. In the left-hand allocation of game 3 (100/200 vs. 150/150), for instance, B should increase A's payoff by 90 points if her β is larger than 0.06. Note that a B-player will never punish or reward if both her α and β are zero or sufficiently close to zero. Since punishing/rewarding is so cheap in our design, however, it is worthy to stress that the values of α and β required for punishment/reward are very low.

Other theories of inequity aversion make parallel predictions in our games. Bolton and Ockenfels (2000) assume that individuals dislike inequity with respect to their *relative* payoff -not the absolute one, as Fehr and Schmidt (1999) posit. However, this distinction is immaterial in two-player games. Finally, Falk and Fischbacher (2006) incorporate inequity aversion and reciprocal concerns in their model. Since punishment/reward is triggered when a player believes that she will get a smaller/larger payoff than her co-player, however, this model predicts the occurrence of punishment and reward in the same allocations as Fehr and Schmidt (1999), provided that beliefs and parameters are chosen conveniently.

The last motivation we consider is **competitiveness** (as suggested in Levine, 1998). A competitive player never rewards, but she punishes the co-player if that allows her to become the player with the highest material payoff. In game 1 (250/100 vs. 200/150), for instance, a sufficiently competitive B-player should punish A if the latter chooses 200/150 (if B reduces A's payoff in at least 60 points, B gets a larger payoff than A), but not if she chooses 250/100. While the amount of punishment should be at least enough to achieve an advantage, any additional amount is also predicted, as it has no cost. In allocation 200/150 of game 1, therefore, any amount of punishment equal or above 60 is consistent with competitiveness. Table 2 includes the predictions of the model in each allocation of our four games.

Note that table 2 does not depict the predictions by altruism and spitefulness, as they are trivial (always reward maximally and always punish maximally, respectively). Finally, we make two remarks: (i) Our selection of games allows us to discriminate between any pair of

theories in at least 4 allocations, and (ii) any two theories share predictions in some allocations; as we clarify later, this is crucial to study more complex motivations than the five cited above (e.g., a model predicting punishment/reward if it is predicted by *both* IA and R, as in the left-hand allocation of game 2, where both predict punishment).

3.2 Experimental Results: Aggregate Overview

Among the 368 decisions made by the B-subjects (8 decisions for each of the 46 B-subjects), we find that 29.9 percent of them were to reward ($s > 0$), 27.7 percent to punish ($s < 0$), and the remaining 42.4 percent were decisions to neither reward nor punish, i.e. not to pay the fee. Table 3 shows for each allocation of each game: (a) the percentage of B-subjects who spent the five point fee to punish, and (b) the average strength of the sanction among those players who punished. For instance, we observe that 45.7 percent of the B-players punish the co-player A if she chooses the left-hand allocation in game 1 (250/100 vs. 200/150); the average score assigned by these B-players in this case is $s = -99.5$.

TABLE 3—OVERVIEW OF PUNISHMENT

Game	Allocation			Frequency		Average strength	
	Left		Right	Left	Right	Left	Right
1	(250,100)	vs.	(200,150)	45.7	37	99.5	68.2
2	(250,100)	vs.	(250,250)	56.5	17.4	93.5	77.5
3	(100,200)	vs.	(150,150)	15.2	15.2	87.1	75.1
4	(100,200)	vs.	(100,300)	28.3	6.5	80	80

Table 4 presents the same data as Table 3 but with respect to rewards. Further, we note that the behavior of the A-players, which is not the focus of our study, is summarized in Table A in Appendix III. In addition, Table B in the same appendix displays the choices made by each B-subject in each game.

TABLE 4—OVERVIEW OF REWARD

Game	Allocation			Frequency		Average strength	
	Left		Right	Left	Right	Left	Right
1	(250,100)	vs.	(200,150)	13	28.3	83.3	81.5
2	(250,100)	vs.	(250,250)	10.9	28.3	70	92.3
3	(100,200)	vs.	(150,150)	52.2	15.2	90	95.7
4	(100,200)	vs.	(100,300)	28.3	63	82.3	92.1

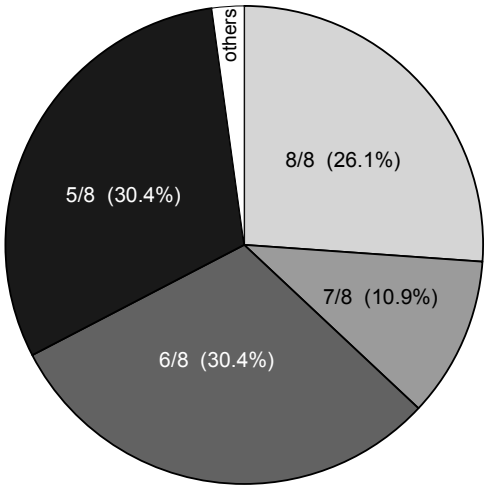
We point out that the aggregate data already suggests that subjects act in heterogeneous manners. For instance, while 8.7 percent of the B-subjects never punish or reward, we find some subjects who reward (6.5 percent) or punish (2.2 percent) in each allocation of the four games. Further, 15.2 percent never punish but sometimes reward, 23.9 percent never reward but sometimes punish, and the remaining fraction of subjects (43.5 percent) both punish and reward in at least one allocation. In the next section we focus on the individual data, and study heterogeneity in more detail.

3.3 Individual Analysis of Punishment and Reward

In this section, we analyze the behavior on the individual level and classify subjects according to the theory that best fits their behavior. We focus first on the occurrence of punishment and reward, leaving the analysis of their intensity for later – i.e., we study which theory predicts best *the sign* of the scores s ($s > 0$; $s < 0$; $s = 0$) chosen by each individual, not their specific values. Apart from greatly facilitating the exposition, there are several other reasons for this. To start, we can evaluate not only the theories cited in section 3.1, but also combinations and variations of these theories. Since these variations do not correspond to existing formal models, it seems wiser to study first if they can at least predict correctly the occurrence of punishment/reward. Further, we believe that studying which theory predicts best the sign of the scores chosen by an individual is worthy in itself, even if it is a less exigent task than predicting the point choices. In any way, we will see that theories that predict best the sign of the scores chosen by an individual (i.e., occurrence) also happen to predict with rather similar success the exact values of the scores (i.e., intensity).

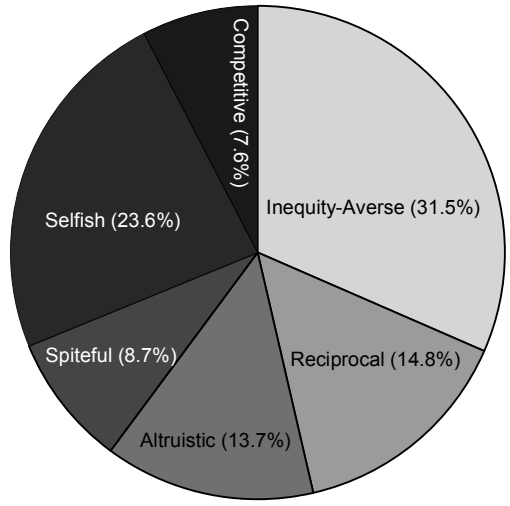
We start with two main findings: (a) For most subjects, there is no theory among those cited in section 3.1 that perfectly fits their actual behavior, and (b) if subjects are classified according to the theory that they follow most consistently, we find substantial heterogeneity. Figure 1 clarifies point (a) and illustrates the extent to which individuals consistently follow one of the six motivations mentioned in section 3.1. Considering for each player the theory that she follows most consistently, we categorize a player as "8/8" if she acts as predicted by that theory in all eight allocations, as "7/8" if she acts in seven out of the eight allocations as predicted by the theory, and so on. We can see that only 26.1 percent of the players behave in a manner *perfectly* consistent with one of the theories, i.e. they are categorized as 8/8. Indeed, most players follow their most consistent motivation in an imperfect manner (60.8 percent are classified as 5/8 and 6/8). For the majority of individuals, therefore, even the theory that best accounts for their behavior must allow for some error.

Fig. 1: How consistently do players follow one motivation?



The next pie graph (figure 2) illustrates which of the six motivations each individual most consistently follows. For example, if a player follows inequity-aversion with 5/8, and all other motivations with 4/8 she is included in the fraction of inequity-averse players.[6] The graph highlights two things. First, there are players for all six motivations. Second, there is no single motivation to which more than 31.5 percent are assigned (the maximum corresponding to inequity-aversion). Thus, this figure suggests that there is considerable individual heterogeneity in motivations.

Fig. 2: Motivation of Players in Games



Since individuals apparently differ in the motivation and often do not follow this motivation in a perfectly consistent manner, we apply the procedure from El-Gamal and Grether (1995) to classify subjects into types and determine whether heterogeneity

[6] If a player follows k motivations in an equally consistent manner (and the remaining motivations less consistently), we add 1/k players to each of the k motivations when computing their corresponding fractions.

significantly improves our understanding of punishment and reward. For this, we first define a *behavioral rule* as a vector of strategies for player B, specifying one strategy for each of our games. As there are four games and player B has two information sets in each game, a behavioral rule is a vector of eight moves. The moves specify whether the B-player punishes or rewards (or does nothing) at the corresponding information set, without specifying the strength of punishment or reward. Thus defined, there are already 3^8 possible behavioral rules in our games; for brevity, we focus on rules based on the theories presented in section 3.1, that is, (1) *selfish*, which predicts no punishment or reward at any allocation; (2) *altruism*, which predicts rewards at all allocations; (3) *spite*, which predicts punishment at all allocations; (4) *reciprocity*; (5) *inequity-aversion*; and (6) *competitive*. The last three rules can be obtained from Table 2. In addition, we also consider some variants of these six rules, which we introduce later in this section and in Appendix IV.

In the simplest scenario, we posit that all players punish and reward according to the same rule R, but also that they deviate from R at each allocation with probability $\varepsilon > 0$. This probability suggests how well rule R fits the data and can be estimated by maximum likelihood –simply by computing the percentage of actual choices that are unexplained by R.[7] Proceeding in this manner for any rule, we can find the best single rule –i.e., that one with the minimal ε. In a more complex version of the classification procedure, we can allow for individual heterogeneity by assuming that players follow different rules. For any combination of k rules (i.e. a *model*), we can classify subjects in types by assigning each subject to the rule that best fits her behavior, and given this assignment, we can compute the corresponding ε in an analogous manner as before. This estimated ε can be used to compare different models, find the model of two, three, etc. rules that best accounts for the behavior in our games, and thus provide an accurate picture of individual heterogeneity. Importantly, this procedure prevents the multicollinearity problems that could appear in a classical regression analysis if the rules were treated as independent variables and allows appropriate inferences even when testing all possible rules –no matter how similar their predictions are– at the same time.

Table 5 indicates the best single rule and the best models with two, three and four rules, together with the percentage of subjects assigned to each rule in each model, and the estimated error rate.[8] We first stress the following: (1) The best single rule is inequity-aversion, with an error rate of 49.1%; (2) the best model with two rules includes inequity-

[7] A more detailed explanation of the classification procedure is provided in --- (2009).
[8] Table 5 only reports error rates for the best models. In appendix IV, we discuss the error rates of some other models.

aversion and selfishness; (3) with three rules, the best model includes altruism, inequity-aversion, and selfishness; (4) the best model with four rules includes altruism, inequity-aversion, selfishness, and spite; this model has an error rate of 0.258, which means that it can explain almost 75 percent of the choices of the average player.

TABLE 5— RESULTS OF CLASSIFICATION PROCEDURE

Number of rules	Best model	Percentage of subjects (N = 46)	Error rate	ML-test (p-value)
1	Inequity-aversion.	100%	0.491	
2	Inequity-aversion; selfish.	63%, 37%	0.367	> 200 (<0.001)
3	Inequity-aversion; selfish; altruism.	43.5%, 34.8%, 21.7%	0.298	116,99 (<0.001)
4	Inequity-aversion; selfish; altruism; spite.	37.7%, 28.3%, 21.4%, 12,7%	0.258	61.03 (<0.1)

The table shows that models with more rules have considerably lower error rates, providing further evidence that individuals are heterogeneous in the way they punish/reward. To infer whether heterogeneous models significantly improve the error rate, table 5 also presents the results from maximum-likelihood (ML) tests. Each of these tests indicates whether the net improvement obtained with the corresponding model is significant.[9] In this respect, we can see that a model with three rules explains significantly better our evidence than a model with two rules (which in turn explains better than a model with one rule). In contrast, a model with four rules only explains behavior marginally better (at the 10% level) than the model with three rules. In any case, our results clearly indicate that heterogeneous models *significantly* improve our understanding of punishment and reward.

[9] To clarify this, note first that the optimal model with k rules is a restriction of the model with k+1 rules (if any), because the k original rules are also included in the model with k+1 rules (i.e., the models are "nested"). With this in mind, let $\lambda = L_R/L_U$ denote the likelihood ratio, where L_R and L_U are respectively the values of the likelihood function for the restricted model and the more complex model. Since the statistic $-2\cdot Ln(\lambda)$ is asymptotically distributed as a chi-squared with degrees of freedom equal to the number of restrictions imposed (see Greene, 1991), we can reject the restricted model if we obtain very large values for $-2\cdot Ln(\lambda)$, as they indicate in turn very small values of λ. As a final clarification, observe that in a model with k rules we must determine for each subject the rule that he/she follows and those that he/she does not follow. This means k parameters for each subject (each one corresponding to one of the rules in the model); these parameters can take either value zero (the subject does not follow the corresponding rule) or 1 (the subject follows the rule). When we move from a model with k+1 rules to another model with k rules, we therefore impose n = 46 restrictions, as the parameters corresponding to one rule are restricted to take value zero for every subject.

Not finding the reciprocity rule selected in table 5, one may think that few individuals are motivated by reciprocity and that it is a relatively less important motivation than inequity-aversion. For instance, figure 2 is implicitly based on a model with our main six rules (error rate = 0.215) and shows that reciprocal players, although they constitute the third biggest group, are also much less frequent than inequity-averse types. Further, the reciprocity rule has an error rate of 0.6 when considered alone, considerably higher than inequity-aversion (0.49). Yet some points suggest that reciprocity is as important as inequity-aversion. First, comparing the best model with two rules (i.e., inequity-aversion + selfish; $\varepsilon = 36.7\%$) to other models with two rules, we note that (reciprocity + selfish; $\varepsilon = 38\%$) is the model with the second lowest error rate.[10] Second, comparing the best model with three rules (i.e., altruism + inequity-aversion + selfish; $\varepsilon = 29.8\%$) to alternative models with three rules, the model (inequity-aversion + reciprocity + selfish; $\varepsilon = 31.5\%$) has the second lowest error rate.

How can we reconcile this apparently conflicting evidence? One potential explanation is that a significant fraction of the players are motivated by both reciprocity *and* inequity-aversion. These players might follow (with some error) the following behavioral rule (which we denote RIA, from reciprocity and inequity-aversion): (i) Reward the co-player if she has helped me and has a lower payoff than me, and (ii) punish if the co-player has harmed me and has a higher payoff than me. In other words, this rule predicts punishment/reward when *both* the inequity-aversion and the reciprocity rule predict punishment/reward (see table 2 for further clarification).

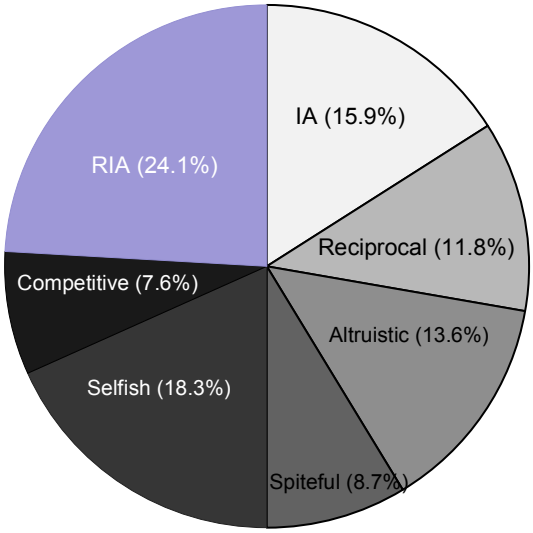
The evidence from our games is in line with this explanation. Consider for instance figure 3, which is identical to figure 2 except that RIA is included.[11] We observe that RIA is assigned to 24.1 percent of the players, i.e. to considerably more than to any other rule. In addition, if we include the RIA rule in our classification analysis, we find that it predicts behavior better than any of our six basic rules. Considered in isolation, the RIA rule has the lowest error rate (0.476). It also appears in the best models with two rules (RIA + selfish; error rate = 0.38), three rules (RIA + selfish + competitive; $\varepsilon = 0.296$) and four rules (RIA +

[10] Observe that the marginal improvement in the error rate when introducing the selfish rule is much higher for reciprocity (from 0.6 to 0.38) than for inequity-aversion (from 0.49 to 0.367). This suggests that inequity-aversion alone explains better than reciprocity alone because the former shares more predictions with the selfish rule (inequity-aversion predicts no punishment/reward when the allocation is strictly egalitarian; this is not the case for reciprocity). As a result, inequity-aversion accounts better not only for the behavior of the inequity-averse types, but also for the behavior of the selfish types.

[11] Table B in Appendix III indicates the type of each subject according to this analysis.

selfish + altruism + spite; ε = 0.252).[12] Note also that the model with all seven rules illustrated in figure 3 has an error rate of 0.185.

Fig. 3: Types in Games (7 motivations)



If we compare figures 2 and 3, we find that the RIA rule takes the largest share from the inequity-aversion rule, which drops from 31.5 to 15.9 percent. The share of reciprocity in contrast is less affected (drops from 14.8 percent to 11.8 percent). Hence, the existence of the RIA types qualifies the importance of inequity-aversion. As further evidence for the RIA rule, we note that RIA can partially explain some of our prior contradictory findings like (i) the model (reciprocity + selfish) is almost as accurate as the model (inequity-aversion + selfish), and (ii) reciprocity does not appear among the more accurate models in table 5. More precisely, finding (ii) is probably due to two factors. First, reciprocity shares four out of eight predictions with inequity-aversion, and therefore it is relatively less complementary to inequity-aversion than other rules (altruism and spite share 3 out of 8, and competitiveness 1 out of 8 predictions). Second, the classification procedure categorizes the RIA players as predominantly inequity-averse, and not as predominantly reciprocal.[13]

In summary, our previous analysis and figure 3 suggest that the majority of the subjects are motivated by inequity-aversion and/or reciprocity, but in different manners. We can see these subjects as displayed along a segment, depending on the relative weight that

[12] These models do not include the purely inequity-averse and reciprocal types, most probably because their behavior is more akin to the behavior of the RIA types than the behavior of the other types. As a result, the models improve their accuracy mostly by including the types most different to the RIA types.

[13] Observe from table 2 that the RIA rule shares more predictions in our games with the inequity-aversion rule than with reciprocity: In the right-hand allocations of games 2 and 3, RIA predicts neither punishment nor reward (as inequity-aversion), whereas reciprocity respectively predicts reward and punishment. As a result, our analysis in table 5 (which does not consider the RIA rule) classifies the RIA types as inequity-averse.

they assign to inequity-aversion with respect to reciprocity. On the extremes of the segment, 15.9 percent of the players can be classified as purely inequity averse, and 11.8 percent as purely reciprocal. In the midpoint, 24.1 percent of the players take into account both motivations, in the manner suggested by the RIA rule.

We dedicate the remainder of this section to consider the intensity of punishment and rewards. In particular, we investigate if the best theory assigned to each individual predicts with comparable accuracy the occurrence of punishment/reward (i.e., the sign of the score s chosen at each allocation), *and* its strength or intensity (i.e., the specific value of s). If this is the case, our previous analysis in terms of occurrence also applies in terms of strength. We organize the analysis by considering each of the seven rules that appear in figure 3 (except the selfish rule, as this rule never predicts occurrence). First, among all subjects classified as purely inequity-averse, the theory predicts correctly the sign for 68 out of 88 choices. We observe that 46 of these 68 hits (67.6%) coincide exactly with the point prediction by inequity-aversion. If we do consider not only these choices, but also those which fall in a 10-points interval around the corresponding prediction (e.g., this is a case of an actual choice s = 90 when the prediction is s = 100; note that 10 points is the minimum error possible), we observe that 58 of the 68 hits (85.3%) happen to be in this class. Arguably, therefore, inequity-aversion does not only accurately predict the sign, but also the intensity of punishment and rewards among these players. Second, among all subjects classified as purely reciprocal, the theory predicts correctly the sign of 48 out of their 64 choices. Among these 48 hits, 31 choices (64.6%) coincide exactly with the point prediction by reciprocity, and 34 choices (70.8%) are within the 10-points interval. Third, altruism predicts correctly the sign of 45 of the 56 choices of the subjects classified as altruists. Further, 86.7% percent of these hits coincide exactly with the point prediction, and 88.9% err by at most 10 points. Fourth, spite theory correctly forecasts the occurrence of 25 out of 32 of the choices of the subjects classified as spiteful, and 76% of those choices are equal to -100, that is, the corresponding point prediction (80% are in the 10-points interval). Fifth, both the occurrence and the strength of 26 out of the 32 choices (81.2%) made by the competitive subjects are correctly predicted by competitiveness.

Finally, we consider the RIA rule. It happens that this rule correctly forecasts the sign of 104 of the 136 choices made by the RIA subjects. If we assume that the RIA rule makes the same point prediction as the reciprocity rule always when RIA predicts punishment/reward, we find that almost all hits (95.2%) coincide with the corresponding point prediction (99% happen to be in the 10-points interval).

In summary, our analysis shows that the theory or type assigned to each subject does not only predict rather successfully the sign of her choices, but also its *exact* value or strength. It follows that heterogeneity is not only important to understand the occurrence of punishment/reward, but also its intensity.

4. Reconciling the Experimental Literature on Punishment

In this section we argue that the conjecture that approximately half of the individuals are either purely inequity-averse, purely reciprocal, or RIA may reconcile some of the contradicting evidence in the experimental literature on costly punishment. We formulate testable predictions based on this conjecture and test for their validity by reviewing other experimental findings on punishment. This exercise is not only interesting because it could help to organize important results in the existing experimental literature, but also because it suggests that our results are robust and can be extended to other games and studies.

Predictions for Punishment: We predict that punishment towards a co-player in one-shot games should be (a) most pronounced in a situation in which the co-player is richer and has harmed the punisher. In effect, in this case the RIA, purely reciprocal, and purely inequity-averse players should punish (provided of course that the cost of punishment is sufficiently low). In comparison, punishment occurs less frequently towards (b1) a richer co-player who has not harmed the punisher (only purely inequity-averse players punish), and (b2) a co-player who has harmed the punisher but is not richer (only purely reciprocal players punish). Finally, (c) little punishment occurs if there was no harm and the person who can be punished is not richer (only the smaller fractions of spiteful and competitive players may punish).

Evidence: Prediction (a) is consistent with punishment behavior in public goods games with a punishment stage as most punishment comes from the contributors and is targeted towards the non-contributors who as a result of their harming behavior are richer than the contributors (Ostrom et al. 1992; Fehr and Gächter, 2000). Falk et al. (2005) find similar patterns of punishment in social dilemma games even if the punishment technology is such that one player alone cannot reduce inequity with another player by sanctioning her (because it costs one unit to reduce the co-player's payoff in one unit; see also Sefton et al., 2007). Although

the inequity-averse types should not punish in this case, this evidence is compatible with our conjecture on heterogeneity, as some RIA and purely reciprocal players might punish.[14]

In turn, prediction (b1) is consistent with the evidence from Falk et al. (2003; see also Brandts and Solá, 2001), who study several mini-ultimatum games and show that the rejection rate of the (proposer, responder) offer (8, 2) depends on the alternative offer. While rejection rates are significantly lower if choice (8, 2) does not harm the responder (i.e., the responder's payoff at the alternative offer is lower or equal than 2), there is some rejection of offer (8, 2) even in these cases. The existence of the inequity-averse types can also explain why some players punish a richer co-player even if she is passive and hence could make no harm (Zizzo, 2003; Dawes et al. 2007, Johnson et al., 2009; reference, 2009), or why unaffected "third parties" in reference (2009) direct their sanctions mostly towards richer co-players – see Fehr and Fischbacher (2004) as well.

Furthermore, prediction (b1) is in line with several papers that report an effect of intentionality on punishing behavior (Blount, 1995; Offerman, 2002; Falk et al, 2008). In the context of the ultimatum game, for instance, Blount (1995) finds that subjects are significantly less likely to reject an offer if it is randomly selected by a computer than if it is intentionally chosen by the co-player, but also that some rejections occurred even when the computer determined the offer. This pattern is consistent with our prediction as inequity-averse types punish in both cases whereas RIA and reciprocal types should punish only other players who intentionally harmed them. To finish, we note that prediction (c) is also in line with several papers. For instance, the previously cited study by Falk et al (2005) consider one treatment in which punishment is relatively cheap, and report that only 13 percent of the non-cooperators reduced the payoff of the (poorer) cooperators (see also Fehr and Gächter, 2000).

Finally, we note that similar predictions with respect to rewards can be tested using experiments. We predict that the frequency of reward towards an active co-player in one-shot games should be (a) most pronounced in a situation in which the co-player is poorer and moreover helped the punisher (RIA, purely reciprocal and purely inequity-averse players reward, provided that the cost of reward is sufficiently low). In comparison, (b1) reward occurs less frequently towards a poorer co-player who has not helped the punisher (only purely inequity-averse players reward), and (b2) towards a co-player who has helped the punisher but is richer (only purely reciprocal players reward). Finally, (c) little reward occurs

[14] Yet note that punishment should occur less frequently as its price increases, as the evidence from Anderson and Putterman (2006) and Carpenter (2007) indicates. Hence, it seems plausible that the most price-sensitive RIA and reciprocal types should not punish if the effectiveness of punishment is 1:1 as in Falk et al. (2005).

if there was no help and the person who can be rewarded is richer (only the small fraction of altruistic players may reward).

5. Conclusion

Our findings indicate that models with multiple types significantly improve the understanding of costly punishment and rewards. We find that not only both inequity-aversion and reciprocity should be taken into account, but also the intersection between the two: Some subjects seem primarily motivated by inequity-aversion, others primarily by reciprocity, and a large fraction by a combination of both motivations. In addition, we also find fractions of altruistic, competitive, selfish, and spiteful players. We hope that these findings may help to further develop recent models of social preferences and reconcile seemingly contradicting findings in the related experimental literature.

We end with ideas for further experimental research. To start, one could change some parameters in our experimental design to study different issues. For instance, one can vary the price of punishment/reward and investigate how this affects the different player types (e.g., are inequity-averse types as sensitive to the price as reciprocal types?). It may be also interesting to introduce communication between players and study how this affects the behavior of the different types. For instance, some types of B-players may use communication to approve or disapprove the A-player for her choice, and abstain in that case from punishing or rewarding in a monetary manner (see Xiao and Houser, 2005). Other types, however, may not regard communication as a substitute for monetary punishment and hence continue punishing.

References

Abbink, Klaus; Irlenbusch, Bernd and Renner, Elke. "The Moonlighting Game –An Experimental Study on Reciprocity and Retribution", *Journal of Economic Behavior and Organization*, 2000, 42, pp. 265-277.

Andreoni, James and Miller, John. "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism", *Econometrica,* 2002, pp. 737-753.

Anderson, Christopher and Putterman, Louis. "Do Non-Strategic Sanctions Obey the Law of Demand? The Demand for Punishment in the Voluntary Contribution Mechanism", *Games and Economic Behavior*, 2006, 54, pp.1-24.

Blount, Sally. "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences", *Organizational Behavior and Human Decision Process*, 1995, 63, pp. 131-144.

Bolton, Gary and Ockenfels, Axel. "ERC: A Theory of Equity, Reciprocity, and Competition", *American Economic Review*, 2000, 90(1), pp. 166-93.

Brandts, Jordi and Charness, Gary. "The Strategy versus the Direct-response Method: A Survey of Experimental Comparisons", mimeo, 2009.

Brandts, Jordi and Solá, Carles. "Reference Points and Negative Reciprocity in Simple Sequential Games", *Games and Economic Behavior*, 2001, 36, pp.138-157.

Carpenter, Jeffrey. "The Demand for Punishment", *Journal of Economic Behavior and Organization*, 2007, 62(4), pp. 522-542.

Charness, Gary and Rabin, Matthew. "Understanding Social Preferences with Simple Tests", *Quarterly Journal of Economics*, 2002, 117, pp. 817-869.

Charness, Gary, "Attribution and Reciprocity in an Experimental Labor Market", *Journal of Labor Economics*, 2004, 22(3), pp. 665-688.

Cox, James C.. "How to Identify Trust and Reciprocity", *Games and Economic Behavior*, 2004, 46, pp. 260–281.

Cox, James C.; Friedman, Daniel and Gjerstad, Steven. "A Tractable Model of Reciprocity and Fairness", *Games and Economic Behavior*, 2007, 59, pp. 17-45.

Dawes, Christopher; Fowler, James H.; Johnson, Tim; McElreath, Richard and Smirnov, Oleg. "Egalitarian Motives in Humans", *Nature*, 2007, 446, pp. 794-796.

Dufwenberg, Martin and Kirchsteiger, Georg. "A Theory of Sequential Reciprocity", *Games and Economic Behavior*, 2004, 47, pp. 268-98.

El-Gamal, Mahmoud and David Grether. "Are People Bayesian? Uncovering Behavioral Strategies", *Journal of the American Statistical Association*, 1995, 90(432), pp. 1137-1145.

Falk, Armin; Fehr, Ernst and Fischbacher, Urs. "On the Nature of Fair Behavior", *Economic Inquiry*, 2003, 41(1), pp. 20-26.

Falk, Armin; Fehr, Ernst and Fischbacher, Urs. "Driving Forces behind Informal Sanctions", *Econometrica*, 2005, 7(6), pp. 2017-30.

Falk, Armin; Fehr, Ernst and Fischbacher, Urs. "Testing Theories of Fairness – Intentions Matter", *Games and Economic Behavior*, 2008, 62, pp. 287-303.

Falk, Armin and Fischbacher, Urs. "A Theory of Reciprocity", *Games and Economic Behavior*, 2006, 54, pp. 293-315.

Fehr, Ernst and Fischbacher, Urs. "Third Party Punishment and Social Norms", *Evolution and Human Behavior*, 2004, 25, 63-87.

Fehr, Ernst and Gächter, Simon. "Cooperation and Punishment in Public Goods Experiments", *American Economic Review*, 2000, 90, pp. 980-994.

Fehr, Ernst and Gächter, Simon. "Altruistic Punishment in Humans", *Nature*, 2002, 415, pp. 137-140.

Fehr, Ernst; Kirchsteiger, Georg and Riedl, Arno. "Does Fairness prevent Market Clearing?", *Quarterly Journal of Economics*, 1993, 108, pp. 437-460.

Fehr, Ernst and Schmidt, Klaus. "A Theory of Fairness, Competition and Cooperation", *Quarterly Journal of Economics*, 1999, 114(3), pp. 817-68.

Greene, William H. "Econometric Analysis", 1991, New York: Macmillan Publishing.

Güth, Werner. "On Ultimatum Bargaining Experiments: A Personal Review", *Journal of Economic Behavior and Organization*, 1995, 27, pp. 329-344.

Herrmann, Benedikt; Thöni, Christian and Gächter, Simon. "Antisocial Punishment across Societies**",** *Science*, 2008, 319, pp. 1362 – 1367.

Johnson, Tim; Dawes, Christopher T.; Fowler, James H.; McElreath, Richard and Smirnov, Oleg. "The Role of Egalitarian Motives in Altruistic Punishment", *Economics Letters*, 2009, 102, pp. 192-194.

Kirchsteiger, Georg. "The Role of Envy in Ultimatum Games", *Journal of Economic Behavior and Organization*, 1994, 25(3), 373-389.

Ledyard, John. "Public Goods: A Survey of Experimental Research", in J. Kagel and A. Roth (Eds.), *Handbook of Experimental Economics*, 1995, Princeton, Princeton University Press.

Levine, David K. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics*, 1998, 1, pp. 593-622.

Nikiforakis, Nikos. "Punishment and Counter-punishment in Public Good Games: Can We Really Govern Ourselves?" *Journal of Public Economics.* 2008, 92(1-2), 91-112.

Offerman, Theo. "Hurting Hurts more than Helping Helps", *European Economic Review*, 2002, 46, pp. 1423-1437.

Ostrom, Elinor; Walker, James and Gardner, Roy. "Covenants with and without a Sword: Self-Governance is Possible", *American Political Science Review*, 1992, 86(2), 404-417.

Rabin, Matthew. "Incorporating Fairness into Game Theory and Economics", *American Economic Review*, 1993, 83(5), pp. 1281-1302.

Roth, Alvin E. "Bargaining Experiments", in J. Kagel and A. Roth (eds.): *Handbook of Experimental Economics*, 1995, Princeton, Princeton University Press.

Sefton, Martin; Shupp, Robert and Walker, James. "The Effect of Rewards and Sanctions in Provision of Public Goods", *Economic Inquiry*, 2007, 45(4), pp. 671–690.

Zizzo, Daniel J. "Money Burning and Rank Egalitarianism with Random Dictators", *Economics Letters*, 2003, 81, pp. 263-66.

Xiao, Erte and Houser, Daniel. "Emotion Expression in Human Punishment Behavior", *Proceedings of the National Academy of Science,* 2005, 102(20), pp. 7398-7401.

Appendix I: Instructions for the B-players.

General Instructions for Participant B

Welcome to this experiment on decision making. At the end of the experiment, you will be paid some money; the precise amount will depend on your decisions and the decisions of another participant. During the experiment we always speak of points; note that

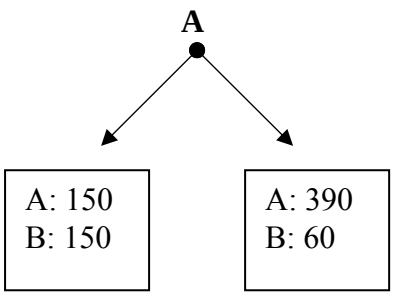
10 points = 1 Euro

Please, do not talk to any other participant during the experiment. If you do not follow this rule we will have to exclude you from the experiment and you will not earn any money. If you have questions, please raise your hand and we will attend you.

There are two types of participants in this experiment: A and B. There is the same number of participants of each type. Previously, the instructor has distributed in a random manner the same number of instructions for each type across the room. Given your seat choice, **you are a type B participant. Further, you will be anonymously matched with a type A participant** (in what follows, we call him/her A)**.** You will never know the type of any other participant, nor will any other participant get to know your type. The decisions in this experiment are anonymous. This means no participant will ever know which participant made which choice.

Description of the Experiment

You, as player B, and A will take decisions in four scenarios, all of them with a two-stage structure. In the first stage of each scenario, A has to decide between two allocations of points for A and you. In the hypothetical example of the figure, the left-hand allocation gives 150 points to A and 150 points to you. The right-hand allocation gives 390 points to A and 60 points to you.



Remember: 10 points = 1 Euro.

In the second stage of each scenario, you can affect the balance of A. For this, you must pay previously 5 points. If you pay the 5 points, you can then assign to A any amount of points between -100 and +100. This amount will decrease or increase the balance of A by the same amount. If you choose not to pay the 5 points, you cannot assign any points to A so that the allocation chosen by A is implemented.

Example 1: Suppose that A chooses the left-hand allocation in the previously illustrated scenario and that you decide then to spend the 5 points and assign +60 points to A. Then A would have a balance of 150 + 60 = 210, and you would get 150 – 5 = 145 points.

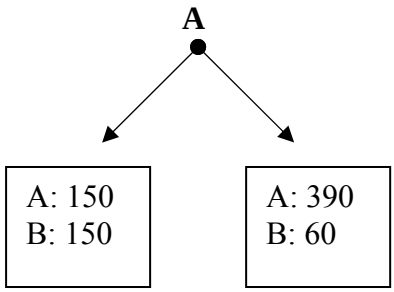
Example 2: Suppose that A chooses the right-hand allocation in the previously illustrated scenario and that you decide then to spend the 5 points and assign -30 points to A. Then A would have a balance of 390-30 = 360, and you would have 60-5 = 55 points.

Important: When deciding, you will not know the allocation actually chosen by A in any scenario. For this reason, you will indicate your decision for any possible choice by A at any scenario. Following with the example of the figure, you should answer four questions: (1) Would you pay the 5 points if A had chosen (150, 150)?, (2) in case you pay the 5 points, what amount of points (between -100 and +100) would you assign then to A?, (3) and (4) the same questions if A had chosen (390, 60).

After all participants have taken their decisions in the four scenarios and answered a brief questionnaire, the instructor will collect your form. Afterwards, one scenario will be chosen randomly (with the roll of a die). This is important because any participant will be paid only for her/his final point score in that scenario (the instructor will divide that score by 10). To finish, note that you will be paid in private and that we will inform you in that moment about A's choice in the payment-relevant game (without, of course, revealing A's identity).

Before we proceed with the experiment, please answer the following control questions. Raise your hand after that so that we can verify that the answers are correct.

In the hypothetical example of the figure, assume the following: (a) B decides to pay the 5 points if A had chosen allocation (A: 150, B: 150), and assigns then +100 points to A, (b) B decides not to pay the 5 points if A had chosen allocation (A: 390, B: 60).



Taking into account all this, answer the following questions,

- What would be the final point score of A if she/he chooses (A: 150, B: 150)? _____
- What would be the final point score of B if A chooses (A: 150, B: 150)? _____
- What would be the final point score of A if she/he chooses (A: 390, B: 60)? _____
- What would be the final point score of B if A chooses (A: 390, B: 60)? _____

In addition:

- Will you know any of the decisions taken by A before you have made your decision in all four scenarios? Yes No
- Will A know any of your decisions before she/he has made her/his decision in all four scenarios? Yes No
- How many scenarios has this experiment? _____ How many scenarios will be relevant for your payment? _____
- Can you ever affect the balance of A without spending 5 points? Yes No

Appendix II: Concise explanation of the models' predictions in our games

We first sum up the basic concepts and hypothesis behind the model of **reciprocity** by Dufwenberg and Kirchsteiger (2004) –many of these ideas also apply for the model by Rabin (1993). Given player A's expectations about the distribution of monetary payoffs to be reached in one game, let (x_{Ae}, x_{Be}) denote B's beliefs about A's expectations –i.e., B's second-order beliefs. From B's point of view, A's intentions are kind (unkind) to B if x_{Be} is larger (smaller) than B's "equitable" payoff -i.e., the average of the maximum and minimum of B's payoffs within the set of allocations that B deems attainable by means of Pareto-efficient strategies. Given this, a player's utility function consists of the sum of her monetary payoff and a reciprocity component. In this respect, it suffices to note that B's reciprocity component takes positive values if she punishes (rewards) A whenever she believes that A's intentions are unkind (kind) to her.

To derive precise predictions in any of our games, observe first that B will always believe that A's intentions are unkind (kind) if A chooses the allocation (x_A, x_B) of the game with the smallest (largest) x_B –i.e., if A harms (helps) B. In effect, since the punishment/reward technology is such that B must pay only 5 points to punish or reward, B gets always less (more) than the "equitable" payoff if A harms (helps) her, no matter whether B uses the technology afterwards. If player B is sufficiently reciprocal, therefore, she should punish A in case she harms B, and otherwise reward her. More precisely, players in Dufwenberg and Kirchsteiger (2004) can be heterogeneous, as a parameter Y measures the sensitivity to reciprocity concerns. Then one can show that any reciprocal player with Y larger than 0.003 should punish and reward in the manner cited.[1] Since the cost of punishment/reward is fixed, moreover, it follows that a B-player who finds it optimal to punish/reward will do it with the maximal strength (i.e., -/+ 100 points).

With respect to the model of **inequity-aversion** by Fehr and Schmidt (1999), recall that it posits the following utility function in two-player games:

$$u_i = x_i - \alpha \cdot \max\{x_j - x_i, 0\} - \beta \cdot \max\{x_i - x_j, 0\},$$

where α, $\beta > 0$, $1 > \beta$, and $\alpha > \beta$. In our games, an inequity-averse B-player should punish A at *any* allocation in which A gets a higher payoff, provided that α is large enough. Since

[1] A proof of this result can be requested from the authors. A reciprocal player with Y < 0.003 should exhibit a different pattern of reward and punishment. For instance, she should never punish or reward if Y is sufficiently close to zero. If Y is sufficiently close but smaller than 0.003, she should punish (reward) *only if* the size of harm (help) is high. Our analysis indicates that this kind of behavior is non-relevant.

additional units of punishment are costless in our design, B-players should punish so as to reduce the payoff distance as much as possible –for this, remember that subjects in our experiment were restricted for simplicity to choose levels of punishment/reward which are multiples of 10. In the right-hand allocation of game 1 (250/100 vs. 200/150), therefore, B should reduce A's payoff by 60 points if her α is larger than $\alpha > 0.1 \cdot (1 + \beta)$. This behavior leads to the distribution 140/145, which is optimal because $145 - \beta \cdot (145 - 140) > 150 - \alpha \cdot (200 - 150)$ for any $\alpha > 0.1 \cdot (1 + \beta)$.[2] The argument is similar in the other allocations in which punishment is predicted, and even smaller values of α predict punishment there (any $\alpha > 0.1$). Note in any case that, since β is smaller than 1 by assumption, expression $0.1 \cdot (1 + \beta)$ can never be larger than 0.2.

With respect to rewards, an inequity-averse B-player should reward A at *any* allocation in which A gets a lower payoff, provided that her β is large enough. Further, the reward should be maximal or almost maximal (90 points). This latter behavior is for instance optimal in the left-hand allocation of game 3 (100/200 vs. 150/150), as it leads to the allocation 190/195, which is optimal because $195 - \beta \cdot (195 - 190) > 200 - \beta \cdot (200\text{-}100)$ if $\beta > 0.06$. Observe that a reward of 90 is better in this allocation than one of 100, which creates a disadvantage for the B-player. Note finally that a B-player will never punish or reward if both her α and β are zero or sufficiently close to zero.[3]

[2] Note that a punishment of 60 is better than a punishment of 50 because, although both achieve the same distance in payoffs in this allocation, the former punishment creates advantageous inequity, which is more liked than the disadvantageous inequity created by the punishment of 50.

[3] A player whose α is smaller but sufficiently close to 0.1, in contrast, will punish when the disadvantageous payoff distance is larger or equal than 100 –as in the left-hand allocation of game 1 (250/100 vs. 200/150)-, but not otherwise. Something analogous occurs with the rewarding behavior if β is smaller but sufficiently close to 0.06. While the existence of this kind of behavior could be analyzed with the help of our classification procedure, we have chosen not to do that in detail for two reasons: (a) The values of α and β that make this behavior possible seem rather marginal; e.g. the distribution of types suggested by Fehr and Schmidt (1999) never considers such low values for α and β, and (b) when we apply the classification analysis to explore a similar behavior (see Appendix IV), we achieve no substantial improvement in the understanding of the data.

Appendix III: Aggregated choices of the A-players and individual choices by the B-players.

TABLE A—A-players' Behavior

Game	Allocation			Frequency	
	Left		Right	Left	Right
1	(250,100)	vs.	(200,150)	60.9%	39.1%
2	(250,100)	vs.	(250,250)	10.9%	89.1%
3	(100,200)	vs.	(150,150)	30.4%	69.6%
4	(100,200)	vs.	(100,300)	26.1%	73.9%

TABLE B—B-players' choices and types

subject	Type	Scores (s) chosen at each allocation by the corresponding subject							
		1L	1R	2L	2R	3L	3R	4L	4R
1	IA/RIA/SE	-100	-50	-100	0	0	0	0	0
2	RE	-100	100	-100	100	100	100	-100	100
3	SE/CO	0	-60	0	100	0	0	0	100
4	RIA	-100	0	-100	0	0	0	-100	100
5	RIA	-100	0	-100	0	0	0	0	100
6	SE	0	-50	0	0	40	0	0	0
7	CO	0	-60	0	-20	0	-20	0	0
8	RIA	-100	0	-100	0	-90	0	-100	100
9	RIA	-100	0	-100	0	100	0	-100	100
10	SE	0	0	0	0	0	0	0	0
11	RE	0	50	-100	100	100	70	-40	100
12	CO	-100	-100	-100	-100	0	-100	0	0
13	RIA	-100	0	-100	0	100	0	0	100
14	IA/RIA	0	0	-90	0	90	0	80	70
15	RE/AL/RIA/SE	0	100	0	100	100	0	0	100
16	SP	0	-70	0	-50	-40	0	-40	-40
17	CO	0	-100	0	-100	-100	-100	0	0
18	RIA/SE	0	0	0	100	100	0	0	100
19	SP	0	-100	0	-100	-100	0	-100	-100
20	AL	100	100	100	100	100	100	100	100
21	SE	0	-100	0	0	0	0	0	0
22	IA/RE/RIA	-90	40	-100	0	100	0	-100	100
23	RIA/SE	-100	0	-100	0	0	0	0	0
24	IA	-100	-50	-100	0	100	0	100	100
25	IA/RE/RIA	-100	50	-100	0	90	0	-30	100
26	IA/RIA	-100	0	-100	0	100	0	100	100

27	AL	80	100	-50	100	100	0	90	100
28	SP	-100	-100	-100	-100	-100	-100	-100	-100
29	SP	-100	-90	-100	-100	-100	-100	-100	0
30	RE/RIA	-100	100	-100	100	100	0	0	100
31	RE	60	80	-30	80	-80	0	-50	10
32	IA	-100	-60	-100	0	90	0	90	100
33	RIA	-100	0	-100	0	0	0	0	100
34	AL	0	100	0	100	100	100	100	100
35	SE	0	0	0	0	0	0	0	0
36	IA	0	-40	0	-50	20	0	20	60
37	IA	0	-50	-60	0	90	0	90	0
38	SE	0	0	0	0	0	0	0	0
39	AL	100	100	100	100	100	100	100	100
40	AL	100	100	100	100	100	100	100	100
41	IA	0	-30	20	0	50	0	20	50
42	RE	-100	40	-100	20	90	-50	-80	100
43	RIA	-100	0	-100	0	100	-60	0	100
44	SE	0	0	0	0	0	0	0	0
45	AL	60	0	30	0	0	100	80	80
46	IA/RIA/SE	-100	-50	-100	0	0	0	0	0

Note: 1L denotes the left-hand allocation in game 1, 2R denotes the right-hand allocation in game 2, and so on. For the types, the following notation is used: AL = Altruistic, C = Competitive, IA = Inequity averse, R = Reciprocal, SE = Selfish, SP = Spiteful. RIA refers to the RIA rule.

Appendix IV: Additional results from the classification analysis

Table 5 only reports error rates for the best models, but is silent for the others. Here we discuss the error rates of other models. In this manner, we can study the relative performance of each of these models.

Altruism & Spite: If we posit that all subjects follow the altruism rule, the error rate equals 0.70. Theories of altruism, however, allow for some parametric heterogeneity (see section 3.1), which implies that subjects with a large σ follow the altruism rule while others follow the selfish rule. This model with two rules (altruism + selfish) has an error rate of 0.42. In turn, the spite rule has an error rate of 0.72 while the model (selfish + spite) has one of 0.48.

Competitiveness: Under the assumption that all subjects follow the competitive rule, the error rate is equal to 0.69. If we introduce some parametric heterogeneity, a model with two rules (competitive + selfish) has an error rate of 0.52.

Inequity-aversion: As indicated in table 5, the error rate of the inequity-aversion rule is 0.49. Note however that theories of inequity aversion like Fehr and Schmidt (1999) assume some parametric heterogeneity so that players with $\alpha > 0.1 \cdot (1 + \beta)$ and $\beta > 0.06$ (see Appendix II) should follow the inequity-averse rule, while players with α and β sufficiently close to zero should follow the selfish rule. In this respect, the model of two rules (inequity-aversion + selfish) has an error rate of approximately 0.37, as also reported in table 5. We can allow for further parametric heterogeneity by introducing an additional "*envy*" rule predicting punishment as the inequity-aversion rule, but no reward (this corresponds for instance to the case $\beta = 0$ but $\alpha > 0.1 \cdot (1 + \beta)$; see footnote 3 in Appendix II in this respect). It follows that models like (envy + selfish), (envy + inequity-aversion), and (envy + inequity-aversion + selfish) have error rates of 0.46, 0.39, 0.33 respectively. By adding this envy rule, therefore, we do not lower the error rate in comparison to the best models with two or three rules.

We finally note that, for brevity, we do not report all the results from our classification analysis here. For instance, we have also considered variations of the RIA rule, like a *reciprocity-equity* rule which coincides with the reciprocity rule except that it predicts no punishment at those allocations with strictly equal payoffs. These slight variations from the RIA also perform rather well, which is again evidence in favor of the idea that some people take into account both equity and reciprocal considerations when punishing/rewarding.