The final publication is available at http://dx.doi.org/10.4284/0038-4038-2011.033.

Do People Accurately Anticipate Sanctions?

Raúl López Pérez and Hubert J. Kiss*

Abstract: We provide lab data from four different games that allow us to study whether people have accurate expectations regarding monetary sanctions (punishment/reward) and non-monetary sanctions (disapproval/approval). Although the strength of the sanction is always predicted with some error (particularly in the case of monetary sanctions), we observe that (i) most subjects anticipate correctly the sign of the average sanction, (ii) expectations co-vary with sanctions, (iii) the average expectation is very often not significantly different than the average actual sanction, and (iv) the errors exhibit no systematic bias, except in those situations where rewards are frequent. In this line, we find some evidence that punishment is better anticipated than rewards.

Keywords: Approval; disapproval; expectations; monetary sanctions; non-monetary sanctions; punishment; rewards; social norms.

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

* López Pérez: Department of Economic Analysis, Universidad Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain. E-mail address: raul.lopez@uam.es. Kiss: Department of Economics, Eötvös Loránd University, Lágymányosi Campus, 1117 Budapest, Hungary. E-mail address: hubert.kiss@tatk.elte.hu. Kiss is also an Affiliate Fellow at CERGE-EI, Prague. We are grateful to the editor, Laura Razzolini, two anonymous referees, Eli Spiegelman, and participants at IMEBE 2011 for useful comments and suggestions. In addition, María García provided helpful research assistance. We also gratefully acknowledge financial support from the Spanish Ministry of Education through the research project ECO2008-00510.

1. Introduction

Social researchers (Akerlof 1980; Elster 1989; Fehr and Fischbacher 2004) often stress that punishment and reward are important to foster cooperation and enforce social norms.[1] In addition, they also point out that both monetary and non-monetary sanctions play a role in this respect.[2] A burgeoning experimental literature attests to these statements. On one hand, part of this literature shows that the availability of *monetary* punishment and reward promotes cooperation, generosity, and fairness (Güth et al. 1982; Ostrom et al. 1992; Roth et al. 1995; Fehr and Gächter 2000; Andreoni et al. 2003; Falk et al. 2005; Sefton et al. 2007; Vyrastekova and van Soest 2008). On the other hand, several laboratory studies show that even *non-monetary* punishment and reward can foster pro-social behaviors (Masclet et al. 2003; Rege and Telle 2004; Noussair and Tucker 2005; Ellingsen and Johannesson 2008; Xiao and Houser 2009, López-Pérez and Vorsatz 2010).

Why does behavior change when sanctions become feasible? One potential reason is that people *expect* to be sanctioned for certain choices and change their behavior to avoid negative sanctions and seek out positive ones. In light of this, it is natural to ask whether people have accurate expectations regarding sanctions, both monetary and non-monetary. This paper addresses this question experimentally.[3] Our research question may shed light on several social phenomena. As a first illustration, consider cooperation and compliance with social norms when sanctions are possible. For any particular sanction to be effective in promoting compliance with a social norm or an etiquette rule, it seems necessary that the potential recipients of the sanction anticipate it. A restaurant customer, for instance, may leave a generous tip if the quality of the service is good. Yet this positive sanction cannot have any effect ex ante on the waiter's service unless he correctly anticipates their potential occurrence and the kind of service that will foster them – e.g., opening a bottle of wine in front of the customers. As a second illustration of the importance of accurate expectations, consider the ultimatum game. In this well-known bargaining game, a first mover proposes a division of some money between herself and a second mover, who can accept or reject the proposal. The division is implemented if it is accepted, whereas both players receive zero if it is rejected. Abundant experimental evidence, starting from Güth

[1] As in the classic study by Coleman (1990), we will use the term sanctions as an umbrella term to refer both to punishment (negative sanctions) and rewards (positive sanctions).

[2] Examples of non-monetary punishment include disapproval, humiliation, insults, shaming and ostracism, while social approval, honors, and praise are examples of non-monetary rewards.

[3] Our focus is on expectations in one-shot games, not on whether people learn and improve their expectations with time in repeated games. In any case, note that both questions are probably related: If the initial expectations affect behavior in the present period, this might in turn affect the future expectations and behavior via a learning process.

et al. (1982), shows that proposals are often rejected if they give little money to the second mover (even if the amount is strictly positive), which can be interpreted as a negative sanction of an unfair offer. Whatever the reasons, the evidence clearly suggests that bargaining impasses are more likely if bargainers under-estimate the potential negative reactions of the other party.

To study accuracy, our experiment uses four games and two treatments (with monetary and non-monetary sanctions, respectively). In each game, we elicit one player's beliefs regarding a co-player's sanctioning behavior. The four games are selected so as to induce different motivations for punishment and reward. Since our design is within-subjects, we can investigate whether players are able to anticipate the differences in these motivations across games. In addition, our two treatments allow us to investigate the expectations about each type of sanctions, to check whether people predict one better than the other, and infer whether they expect differences across treatments.

Our key results for both treatments are the following. In those situations where the average actual sanction is significantly different from zero and hence has a definite sign (i.e., non-null), *first*, subjects tend to anticipate that sign – i.e., they anticipate whether punishment or rewards are prevalent. *Second*, expectations co-vary with sanctions in most situations in both treatments, suggesting that people foresee the differences across games and treatments. *Third*, expectations are not significantly different on average from sanctions in a large majority of the situations, suggesting that people are not systematically biased in their expectations. *Fourth*, the average error is never larger than 35% (27%) of the maximum possible error in any game of the treatment with monetary (non-monetary) sanctions. Arguably, therefore, relative errors are moderate, particularly with regard to non-monetary sanctions. In fact, *fifth*, we observe that average errors are often significantly smaller in the non-monetary treatment. Hence, it seems that people anticipate better non-monetary sanctions. Finally, we find some evidence that people predict punishment better than rewards, as most exceptions to the previous results occur in those situations where rewards are frequent. This evidence suggests a relative advantage of punishment in providing incentives.

In the rest of the paper, we first describe and motivate our experimental design and procedures in section 2. Section 3 presents and discusses our results in both treatments. Section 4 concludes with a brief review of some related literature on beliefs, a discussion of some implications of our data, and some ideas for further research.

2. Experimental Design and Procedures

Our experimental design consists of a Monetary (M) and a Non-Monetary (NM) treatment. Each subject participated only in one treatment. We describe first the M-treatment. Participants played four two-player games, each with a two-stage structure. In the first stage, one player (the *decider*, or A) chooses between two allocations of money for herself and another player (the *sanctioner*, or B).[4] Table 1 shows the two (decider, sanctioner) payoff allocations available in each game (called left and right; in what follows, we denote by 1L the left-hand allocation in game 1, by 2R the right-hand allocation in game 2, and so on). Payoffs are presented in points, at the exchange rate 10 points = 1 Euro; we motivate the selected payoff constellations later.

Table 1 about here

In the second stage, the sanctioner can impose a *monetary* punishment or reward on the decider, at a *fixed* cost of five points for the sanctioner.[5] More precisely, the sanctioner can either increase or decrease the decider's payoff by up to 100 points, but for that she must pay five points from her allocation share.[6] If the sanctioner does not want to affect the decider's balance, no points are deducted from her own. As an illustration, suppose that the decider chooses allocation (250, 100) in game 1. If the sanctioner decides not to pay the five points, this allocation is implemented. If she pays the five points fee, however, she can choose a "point score" $s \in [-100, 100]$ so that the decider's payoff in the game is $250 + s$, while the sanctioner gets a payoff of $100-5 = 95$. For simplicity, s had to be a multiple of 10. Note that our simple technology of punishment/reward probably facilitates the already complex task of predicting sanctions. Since there is still little evidence on expectations about sanctions, a simple technology seems convenient for our research question.

The NM-treatment is identical to the M-treatment, with one important difference: In the second stage of each game, the sanctioner cannot affect the monetary payoff of the co-player, but approve or disapprove her choice –i.e., she can reward or punish the decider in a *non-monetary* manner. As in the

[4] In the subjects' instructions, the decider was always called 'A', and the sanctioner was referred to as 'B'.

[5] We avoided the terms "punishment" and "reward" in the instructions.

[6] This small fee prevents random choices from selfish players, which could theoretically occur if sanctioning was not costly and would complicate the interpretation of our results.

M-treatment, the sanctioner must pay a fixed cost of five points for that. If she pays the five points, she can send an "evaluation score" $s \in [-100, 100]$ to the decider expressing either approval or disapproval of the decider's choice. To achieve common knowledge in this respect, the instructions explicitly used the words "approval" for positive scores and "disapproval" for negative scores. Hence, an evaluation score of $s = +100$ means maximal approval, and $s = -100$ means maximal disapproval. As an illustration of the payoff structure, suppose that the decider chooses allocation (250, 100) in game 1. Since the sanctioner cannot affect the decider's balance in this treatment, the latter is then sure to get a payoff of 250. With respect to the sanctioner, she would get a payoff of $100 - 5 = 95$ if she sends an evaluation score and a payoff of 100 otherwise. We note that our approval/disapproval technology is an adaptation of that used by Masclet et al. (2003).

The experimental procedures were identical in both treatments. We conducted seven sessions at the Universidad Autónoma de Madrid, and a total of 92 subjects participated in the M-treatment (four sessions), and 84 in the NM-treatment (three sessions). Subjects were students from different disciplines (9.6 percent came from the faculty of economics) and not students of the experimenters. Further, the share of female participants was 54 percent in the M-treatment and 69 percent in the NM-treatment. Before the start of each session, we distributed instruction and decision sheets (dependent on role) in a classroom, leaving enough space between seats to ensure anonymity. Then the subjects entered the room. The sheets were initially covered and the subjects could freely choose their seat; in that manner, we assigned them to be either a decider or a sanctioner. Subjects could read the instructions at their own pace and questions were answered in private. Before proceeding with their decisions, participants had to fill out control questions to make sure that they understood the rules.

In each treatment, subjects played the same four games in the same role and with the same anonymous co-player. To prevent income effects, we paid subjects for their decisions in just one game, randomly selected at the end of the experiment. Further, no subject was informed of her counterpart's actual choice in any game in order to avoid repeated game effects and changes of mood which would severely complicate the data analysis (e.g., the mood of a sanctioner could change depending on the decider's choice in a preceding game). Therefore, we employed the *strategy method* to elicit the decisions of the sanctioners, i.e., they indicated in each allocation of each game whether they wanted to pay the five points fee, and if this was the case, we asked them which score $s \in [-100, 100]$ they wanted to assign to their co-player. Finally, we presented all four games on the same decision sheet to prevent potential order effects that could appear if the games were presented one by one. It is worth noting that

some of these design features might affect behavior and thus possibly expectations. For instance, sanctioners in our setting could attempt to play in a consistent manner across games, potentially generating 'spillovers' from one game to another.[7] Hence some prudence is warranted when extrapolating our results to other settings, like one in which subjects play only one game. In any case, we believe that our within-subjects design provides strong evidence on whether people anticipate changes in sanction patterns across games.

In both treatments, we elicited the deciders' expectations with regard to the sanctioners' behavior. More precisely, we asked two questions for each allocation: (i) The percentage of sanctioners who would pay the 5 points fee at that allocation, and (ii) the average score assigned by those sanctioners. Letting p^e and s^e respectively denote the expectations (i) and (ii) of a generic decider at one allocation, we henceforth refer to the product $p^e \cdot s^e$ of these two numbers as the *expectation* by that decider at that allocation. Subjects were not paid for answering questions (i) and (ii). This has one potential shortcoming but also some advantages. The shortcoming is that it cannot be completely ruled out that some subjects report falsely their expectations, maybe choosing randomly. Paying for accuracy, in contrast, would arguably *reduce* any incentive to misreport. Yet we doubt that many subjects chose randomly, as it would be extremely difficult to make sense of our data in that case (we provide specific examples later). Alternatively, subjects might respond sincerely but reflect less on their answers if they are not paid for accuracy, possibly augmenting the variance of the expectations – Gächter and Renner (2010) discuss this issue.[8] Yet we do not view a casual level of reflection as a shortcoming (see the second advantage below). On the positive side, we find at least three advantages. First, our data facilitate comparison with previous studies. For instance, our results are arguably relevant to understand behavior in the numerous experiments on punishment and reward where subjects are not paid for accuracy (see

[7] As another example, the strategy method might induce different behavior than the direct-response method where participants know the choice made by the co-player – e.g., Casari and Cason (2009). Yet 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).

[8] There is still an ongoing debate on whether incentivized belief elicitation increases accuracy. For example, Gächter and Renner (2010) elicit subjects' beliefs in a 10-times repeated public good game, and compare incentivized and non-incentivized elicitation. They find no differences across treatments in the absolute average error in period 1, whereas the error is significantly smaller in the incentivized treatment in the remaining periods. Hence, their data suggests that incentives affect the way in which people's beliefs change, but not the initial beliefs. Provided that this result can be generalized, it would be clearly relevant in our games, as they are one-shot.

Gächter and Renner, 2010 for some examples). Second, we believe that our data facilitate extrapolation to behavior out of the lab where, generally, people are not paid by others to form accurate beliefs about punishment and reward. If we provided incentives and they affected accuracy, in contrast, it would not be so clear that our results are relevant to understand behavior out of the lab. Finally, paying for the beliefs would change the expected distribution of payoffs, and that should theoretically affect the behavior of those sanctioners with preferences like inequity-aversion (they should condition their behavior on their expectations that the decider has accurate expectations), complicating the interpretation of our results.

After subjects made their decisions in the four games, they answered a brief questionnaire. Then we collected their decision sheets and selected the payoff-relevant game. Subjects were paid privately, and earned on average 18.3 Euros in the M-treatment and 20.6 Euros in the NM-treatment. All deciders were informed at the time of payment about the score (if any) sent by his/her sanctioner at the payoff-relevant allocation.[9] Each session lasted approximately 60 minutes.[10]

We finish this section with a discussion of our payoff constellation. The goal was to induce variability in sanctions across allocations. In this manner, we can analyze whether the deciders anticipate the varying behavior of the sanctioners, and whether some situations are more prone to error. Based on several experimental studies (Leibbrandt and López-Pérez 2010a survey this literature), we expected *inequity-aversion* and *reciprocity* to be key motivators for *monetary* sanctions in our games. To induce variability, therefore, we selected allocations in which either (i) both motives predict punishment/reward, (ii) only one of these motives predicts punishment/reward, or (iii) none of them predicts either punishment or reward. Table 2 indicates the allocations of the four games in which inequity-aversion and reciprocity predict the occurrence of punishment or reward. To understand these predictions, note that models of inequity-aversion (Fehr and Schmidt 1999; Bolton and Ockenfels 2000) predict punishment if the decider has a larger payoff than the sanctioner, and reward if the decider has a smaller payoff; whereas reciprocity models like Dufwenberg and Kirchsteiger (2004) – based on Rabin

[9] Subjects knew in advance that the experimenter was going to observe their payoff-relevant decision, which may have affected their choices (e.g., Cox and Deck, 2005) and perhaps their beliefs. Hence, some caution is also warranted when extrapolating our results to other settings with an increased social distance.

[10] The instructions for the deciders in the M-treatment and the NM-treatment, translated from Spanish to English, are available in a web appendix at https://sites.google.com/site/kisshub/home/research/webappendix-southern-economic-journal. We also provide one example of a decision sheet for the M-treatment, to illustrate how beliefs were elicited.

(1993) – predict punishment if the sanctioner is harmed by the decider's choice (i.e., in our games if the decider chooses the allocation where the sanctioner's payoff is smallest), and reward if she is helped.[11]

Table 2 about here

Our payoff constellation is also convenient for the NM-treatment, as we again expected some variability in sanctions across allocations. We must first clarify, however, that the models of inequity-aversion and reciprocity listed above cannot explain the occurrence of non-monetary sanctions, as these behaviors are costly in our design and cannot reduce inequity or be used to reciprocate (in a *material* manner). Yet models like Holländer (1990) and Kandel and Lazear (1992) predict that sanctioners approve (disapprove) choices that help (harm) them. Hence, they predict non-monetary reward/punishment in the same allocations where models of reciprocity predict *monetary* reward/punishment in the M-treatment (see Table 2), thus generating differences in sanctions across allocations. Our design allows us to investigate if the deciders foresee these potential variations.

3. Results

In this section, we compare *sanctions* and *expectations*, that is, the actual sanctions by the sanctioners (i.e., the scores s chosen by them at each allocation; note that we set s = 0 when a sanctioner does not pay the 5 points fee), and the expectations in this regard of the deciders, respectively. We perform this comparison for every allocation of each treatment. Our first aim is to investigate whether the deciders are able to anticipate the sign and strength of the *average sanction* in each allocation of the four games (Section 3.1), and the size of the errors committed (3.2). While these two initial sections have a descriptive nature, we afterwards explore potential determinants of the sanctions by the sanctioners *and* the prediction errors by the deciders, suggesting one encompassing explanation and offering some evidence (3.3). Finally, we discuss the choices of the deciders in both treatments (3.4).

[11] These predictions hold for a very large range of the models' parameters. The reader may consult Leibbrandt and López-Pérez (2010a) for a more detailed discussion. For simplicity, the predictions in Table 2 only specify the sign of the predicted sanction, not its strength – e.g., inequity aversion predicts a sanction of -55 points in allocation 1R.

3.1 Expectations of the deciders: Accuracy

Do the deciders accurately estimate the average sanction at each allocation? To answer this question, we first study whether they at least correctly anticipate the *sign* of the average sanction, and later move a step further and analyze whether they accurately anticipate its *strength*.

Studying the sign is an important first step, as a punishment (negative sign) is qualitatively very different from a reward (positive sign). Table 3 reports for each allocation in both treatments the percentage of deciders who expect either a positive or a negative sanction (the remaining deciders expect a null sanction). The variable employed is the expectation by each decider at each allocation. Recall from Section 2 that this variable represents the average sanction expected by each decider at each allocation, as it equals the corresponding product $p^e \cdot s^e$: the percentage of sanctioners who are expected to pay the 5 points fee at that allocation multiplied by the average expected score from those players. For instance, 21.7 and 69.6 percent of the deciders report an expectation with positive and negative sign, respectively, at allocation 1L in the M-treatment. Now, the sign of the average sanction at allocation 1L happens to be negative (Table 4 below reports the precise average sanction at each allocation of each treatment). Consequently, the percentage of deciders who correctly anticipate the sanction sign at this allocation equals 69.6 (i.e., the sign of their expectation coincides with the sign of the corresponding average sanction), as we indicate in bold in the table.

Table 3 about here

Our focus is on those allocations where the average sanction is significantly different from zero at the 10 percent level: Since the average sanction has a very clear sign (either positive or negative) at these allocations, it makes sense to ask if subjects can anticipate it. With respect to the M-treatment, the average sanction is significant at the 10 percent level at allocations 1L, 2L, 3L and 4R (Mann-Whitney test, in fact $p < 0.01$ always). As one can confirm from Table 3, more than two thirds of the subjects correctly anticipate the sign in these allocations, and a one-sample test of proportion rejects the hypothesis that these figures are equal to 50 percent ($p < 0.01$ always). In the NM-treatment, in turn, the average sanction is significantly different from zero at the 10 percent level at all allocations except 1R, and we observe that most subjects anticipate correctly the sign of the average sanction at all these allocations except 3R, where only 36.6 percent of the subjects properly forecast a positive sanction. Furthermore, the proportion of correct sign predictions is significantly higher than 0.5 at the 10 percent

significance level at allocations 1L, 2L, 2R, 3L, and 4R ($p < 0.07$ always). We summarize our findings for both treatments as follows:

Result 1: In 4 (7) allocations of the M-treatment (NM-treatment), the average actual sanction is significantly different from zero and hence has a clear sign. In almost all these allocations, the fraction of deciders who accurately anticipate the corresponding sign is larger (mostly significantly) than 50 percent. The only exception is allocation 3R of the NM-treatment, where rewards are prevalent but most deciders fail to anticipate it.

We now investigate if, on average, deciders correctly predict the strength of the average sanction. For each allocation of each treatment, Table 4 reports the average expectation across all deciders – e.g., the average decider expects a sanction of -26.47 points at allocation 1L of the M-treatment –, and the average actual sanction by the sanctioners. Table 4 also indicates the median and standard deviation of each variable.[12]

Table 4 about here

We can see that the average expectation is similar to the average sanction in most allocations of both treatments, suggesting that the average decider anticipates the average sanction rather accurately.[13] Indeed, the two-sided Mann-Whitney test almost never rejects at the 10 percent level the hypothesis that the average expectation and sanction at each allocation are equal (see Section W3 of the web appendix for the precise p-values of the tests). The only exceptions occur at allocations 3L (p-value < 0.1) and 4R (p-value < 0.01) of the M-treatment, and allocation 3R of the NM-treatment (p-value < 0.01)

Figure 1 uses a box-and-whisker plot to offer more disaggregated evidence on the subjects' accuracy in the M-treatment. The darker (lighter) boxes correspond to the sanctions (expectations) at each allocation. Each box represents the inter-quartile range of the corresponding variable, while the upper/lower whiskers (that is, the adjacent dashed lines) extend to the maximum/minimum value of the corresponding variable that is within the 1.5 inter-quartile range from both quartiles. For instance, at allocation 2L the highest sanction within that range is 100, and the highest expectation is 50. The crosses represent outlying values. Further, the black circles (thick lines) indicate averages (medians) of each

[12] For further detail, consult the cumulative distribution functions of the expectations and the sanctions at each allocation of each treatment, available in the web appendix. See also section W4 in that appendix for additional data on sanctions.

[13] This is one signal that most subjects reported their expectations in a non-random manner. Another signal is that expectations vary considerably across allocations.

variable; the precise figures appear in Table 4. Note as well that the allocations have been ordered on the X-axis from those with the lowest average sanction (on the left) to those with the highest average sanction (on the right). We remark that this ordering coincides with the ordering according to the median, although this last ordering is less informative due to the fact that at 4 allocations the median equals zero. In turn, Figure 2 is the analogue of Figure 1 for the NM-treatment; allocations have been ordered as in Figure 1 to facilitate comparison.

Figure 1 about here

Two things stand out in these figures. First, expectations co-vary with sanctions. One indication of this is that the ranking of the allocations according to the average sanction roughly coincides with the ranking according to the average expectation: Subjects tend to anticipate which allocations will be most highly punished or rewarded. Second, the variation in the expectations (lighter boxes) is smaller than that of the sanctions, a somehow natural point given that the deciders' expectations focus on the average value of the sanctions: unless deciders are extremely inaccurate and heterogeneous in this regard, one would expect their expectations to vary less than the observations from the whole distribution. For further evidence, we use Levene's test to assess whether the variances of the expectations and the sanctions are equal.[14] In the M-treatment, the null hypothesis that the variances are equal is rejected at the 10 percent significance level at all allocations except 3L and 4L. In the NM-treatment, in turn, the null hypothesis is rejected at the 1 percent level at any allocation. In a similar line, the Kolmogorov-Smirnov test rejects at 10 percent significance level the hypothesis that expectations and sanctions come from the same distribution at all allocations, except 1R, 2R, and 4L of the M-treatment (p-value > 0.12), and allocation 4L of the NM-treatment (p-value > 0.16).

Figure 2 about here

We summarize our findings for both treatments as follows:

Result 2: The average expectation is not significantly different than the average sanction at most allocations in the M-treatment (in 6 out of 8 cases) and the NM-treatment (in 7 out of 8 cases). In the

[14] We use Levene's test because it is robust to non-normality of the error distribution.

only exceptions, rewards are frequent and strong. Further, expectations are less dispersed around the mean than actual sanctions.

Result 2 relates to the question of whether punishment is better anticipated than reward, which is important for understanding the effectiveness of each type of sanction: for instance, punishment could be more effective in changing behavior if the potential receiver anticipates it better than rewards. In this respect, several phenomena suggest that subjects tend to anticipate the 'stick' better than the 'carrot' in our games. First, average expectations and sanctions in the M-treatment are significantly different only at allocations 3L and 4R, the two allocations where the average sanction reaches its highest positive level.[15] Second, average expectations and sanctions in the NM-treatment present a significant difference only at allocation 3R, and again rewards are strong at this allocation. In our games, consequently, a *necessary* condition for average expectations and sanctions not to coincide at one allocation seems to be that rewards are strong at that allocation. As Result 1 indicates, third, allocation 3R in the NM-treatment is the only one in which most subjects fail to anticipate the sign of the (significant) sanction. While all this evidence is preliminary and further experimental tests are warranted, we provide below additional support for the idea that the 'stick' is better anticipated than the 'carrot', particularly with respect to monetary sanctions.

So far we have mostly compared average expectations and sanctions. Yet expectations are defined as the product $p^e \cdot s^e$ of the expected percentage of sanctioners paying the fee to sanction, and the expected sanction from them. In a more disaggregated analysis, therefore, it makes sense to compare the average values of p^e and s^e to their actual average values. In this manner, we investigate if the average decider fails to anticipate any of these variables. Table A in the appendix shows the percentages and the sanctions that subjects expect, together with the average actual counterparts.

3.2 The prediction error

The previous section showed that, *on average*, subjects predict sanctions rather accurately at most allocations in both treatments. On the individual level, nevertheless, deciders frequently estimate the strength of the sanction with some error. To show some initial evidence in this respect, we define a decider's *absolute error* at one allocation as the absolute difference between the expectation by that decider at that allocation and the corresponding average sanction. Table 5 depicts the *average* absolute

[15] In fact, these are the only allocations where the average sanction is significantly higher than zero (hence, deciders are strongly rewarded), the p-values being less than 0.01 in both cases (Mann-Whitney test).

error at each allocation of each treatment; this is 35.61 at allocation 1L in the M-treatment, for instance.[16] We observe that this average absolute error is high in many allocations and in fact significantly higher than zero in all allocations of both treatments (Mann-Whitney test, p-value < 0.01).

Table 5 about here

While the absolute error provides some insights, it has also some shortcomings when making comparisons across allocations. As an illustration, it is different to commit an absolute error of 20 units when the average sanction is 100 than when it is zero. In the first prediction problem, one can commit an absolute error as large as 200 (i.e., if one predicts -100), so that an error of 20 seems low. In the second case, the largest possible absolute error is 100, and an error of 20 appears to be more substantial. To provide a normalized measure of the error that takes into account this problem, we define a subject's *relative error* at one allocation as her absolute error divided by the largest absolute error possible – observe that the largest absolute error at any allocation equals the average sanction at that allocation plus 100, as the previous examples illustrate. So defined, the relative error takes values between zero and 1; low values suggest that a subject did rather well in anticipating sanctions. In Table 5, we present the average relative error at each allocation as a percentage – i.e., multiplied by 100.

We make a number of remarks on Table 5. First, the absolute and the relative error are correlated within each treatment, which is not surprising given that the relative errors are just a normalized version of their absolute counterparts. Second, the average relative errors are never larger than 35 % in the M-treatment, and never larger than 27% in the NM-treatment. Hence, subjects commit errors, but of a limited size in average, particularly in the NM-treatment.

This latter point merits some attention: Do deciders commit larger mistakes in predicting monetary sanctions? The relative errors allow for a meaningful comparison across treatments. In this respect, the Mann-Whitney test rejects equality of the relative error at the 1 percent level at all allocations except 2L (p-value > 0.19), 2R (p-value > 0.20), and 4L (p-value > 0.09). It follows that the relative error is significantly lower in the NM-treatment for almost all remaining allocations, a signal that the deciders have more problems to foresee monetary sanctions in our games. Note that the only

[16] To put this figure in context, consider an allocation where the average sanction is equal to s and a hypothetical individual who tries to predict s by randomly picking a number between -100 and 100. One can show that the average absolute error would then be equal to $50 + s^2/200$. In allocation 1L, for instance, s = -26.47 so that the average error should therefore be 53.5. This figure is larger than the average absolute error in this allocation (35.6), and in fact this happens in all allocations.

exception to this statement occurs in allocation 3R, where relative errors are significantly higher in the NM-treatment.

Another relevant question is whether errors are higher at allocations where rewards are prevalent. Again, some evidence suggests this to be the case, particularly in the M-treatment. In effect, the relative error reaches the highest and the third highest value at allocations 3L and 4R, which are rewarded most. In contrast, Table 5 shows that in allocations 1L and 2L, where punishment is strongest, the relative error is the second and third lowest – yet we note that there is no significant difference between the error at 1L and 3L. The results are not so clear in the NM-treatment. Punishment is high at allocation 2L and yet the relative error is higher than in allocation 4R, where rewards are strong.

We consider now whether the subjects' errors are systematic. In this respect, the analysis in Section 3.1 showed that the average expectation is not significantly different than the corresponding average sanction at most allocations in both treatments (the only exceptions, recall, are allocations 3L and 4R in the M-treatment, and allocation 3R of the NM-treatment). This fact suggests that subjects are not systematically biased in their expectations and that at most allocations the errors caused by overestimation of the average sanction by some subjects compensate the errors caused by underestimation by other subjects.

To check more carefully for this possibility, we consider a signed version of the relative error. More precisely, a decider's *signed relative error* at one allocation is defined as the relative error, but not in absolute value. The sign of this variable shows the direction of the error, the positive/negative sign indicating over/underestimation. If the distribution of deciders' signed relative errors at one allocation has a non-zero median, we infer that deciders are biased in their expectations at that allocation. In this respect, the Wilcoxon signed-rank test rejects the hypothesis that the median is zero at the 5 per cent level in allocations 3L, 3R, and 4R of the M-treatment (p-values < 0.05, 0.01, and 0.01, respectively), and 1R, 2R, 3R and 4R of the NM-treatment (p-values < 0.05, 0.01, 0.01, and 0.05, respectively). These results suggest that subjects commit systematic errors at seven out of 16 allocations. Since the average of the signed relative error at all seven allocations happens to have a negative sign, furthermore, it follows that subjects systematically underestimate the average sanction at these allocations. We also note that the average sanction is always positive in these seven allocations. Apparently, therefore, subjects tend to underestimate where rewards are prevalent, but not when punishment is prevalent.

We summarize our findings for both treatments as follows:

Result 3: In absolute terms, subjects tend to commit errors of a significant size. In comparison with the maximum possible error, however, the relative errors are moderate, particularly in the NM-treatment. When considering the sign of the error, we observe systematic biases at those allocations where rewards are relatively frequent. In all these allocations, subjects tend to underestimate the average sanction.

3.3 Discussion

In this section we offer one possible explanation of our data, gathering together three questions: (i) how do sanctioners choose sanctions? (ii) how do deciders form expectations of sanctions? and (iii) why are these expectations sometimes inaccurate? We also explore and shed some evidence on question (iii) with the help of a regression analysis.

The first part of our explanation considers question (i) above, that is, the determinants of the sanctions. This point has been analyzed in detail by Leibbrandt and López-Pérez (2010a), an accompanying paper focused on the sanctioners' behavior in our experiment.[17] To prevent duplication, therefore, we just briefly summarize their key results, starting with the M-treatment. Although many factors are at play in these games, Leibbrandt and López-Pérez (2010a) reckon that both inequity aversion and reciprocity are the crucial forces behind monetary punishment and rewards. This result is in line with previous studies stressing the importance of both forces (e.g., Fehr and Fischbacher, 2004; Falk et al., 2005), and can be illustrated with the help of Figure 3, a box-and-whisker plot combining information from Figures 1 and 2 on the actual sanctions at each allocation in each treatment. The darker/lighter boxes correspond to the inter-quartile range of sanctions in the M/NM-treatment, respectively. Note that the allocations are ordered on the X-axis from those with the lowest average sanction *in the M-treatment* (on the left) to those with the highest average sanction in the same treatment (on the right).

Figure 3 is consistent with the idea that inequity aversion and reciprocity are both important to account for the occurrence of *monetary* sanctions. In effect, the figure shows that the allocations where the average sanction is lowest are 2L and 1L, and these are the only allocations in which both inequity-aversion and reciprocity predict punishment (see Table 2). Further, the average sanction reaches the highest positive level at allocations 3L and 4R, where both forces predict rewards. Finally, the average

[17] We complement Leibbrandt and López-Pérez (2010a) by studying whether the deciders are able to predict the sanctioners' average behavior.

sanction reaches an intermediate level at the remaining allocations, where inequity-aversion and reciprocity predict different things.

Figure 3 about here

With respect to the NM-treatment, Leibbrandt and López-Pérez (2010a) find that both harm (help) and *strict* payoff equality are crucial to explain the occurrence of non-monetary punishment (rewards). In this line, Figure 3 shows that the average non-monetary sanction is negative *if and only if* the sanctioner is harmed and strict payoff equality is not achieved (i.e., in allocations 1L, 2L, and 4L), and it is positive *if and only if* the sanctioner is helped and/or strict payoff equality is achieved (i.e., in allocations 1R, 2R, 3L, 3R, and 4R).

One possible interpretation of the previous evidence is that the sanctioners are heterogeneous. With respect to monetary sanctions, for instance, there could be some types who are purely inequity-averse, others who are reciprocal, others who are motivated by a mix of these two previous forces, and some who are selfish and hence never punish or reward in our games.[18] When anticipating a sanction, therefore, the deciders should take into account the frequency of each type, the precise utility function of each one, etc. Let us then assume that the deciders have some priors in this respect, but possibly not totally accurate so that they predict sanctions with some error. Since we have games in each treatment in which the main forces driving sanctions make different predictions, we can explore which of these forces are better predicted, and hence some potential determinants of the prediction error. For this we use a regression approach in which the dependent variable is the (absolute) error committed by each decider at each allocation (recall that this variable is strongly correlated with the relative error). As independent variables, we use individual socio-demographic variables collected from our questionnaire (gender, political ideology, and religiosity), session dummies, and several key variables that refer to the forces mentioned above.

Assume for instance that many deciders in the M-treatment fail to accurately forecast an increase in sanctions due to increasing payoff differences, as predicted by inequity aversion. Since this prediction failure might depend on whether the other player is advantaged or disadvantaged, we

[18] This idea is consistent with the results from the classification analysis in Leibbrandt and López-Pérez (2010b), who also find some other motives apart of inequity-aversion and reciprocity apparently shaping monetary sanctions. In effect, there are small fractions of subjects who punish in a spiteful or maybe competitive manner, and apparently altruistic subjects who tend to reward always. For simplicity, our analysis here does not consider these types, as they play an arguably minor role.

distinguish two variables in this respect. The first one is called *paydiff-* and is defined as the difference between the decider's and the sanctioner's payoff in the corresponding allocation, provided that the sanctioner has a smaller payoff (otherwise, it takes value zero). For example, this variable equals 50 in allocation 1R, and 0 in allocation 3L. In turn, variable *paydiff+* is analogously defined for the case when the sanctioner has a larger payoff than the decider – e.g., it takes value 0 and 100 in allocations 1R and 3L, respectively. In addition, we consider a dummy variable called *strict*, which takes value 1 if players' payoffs at the corresponding allocation are identical (this variable is potentially important in the NM-treatment, as we have mentioned before).

Players might also fail to anticipate that reciprocity is important to explain sanctions. For this reason, we introduce a dummy variable *Rec+*, taking value 1 when reciprocity (Dufwenberg and Kirchsteiger, 2004) predicts rewards, and value 0 otherwise (i.e., if reciprocity predicts punishment; see Table 2 for further clarification). For instance, this variable equals 0 at allocation 2L, and 1 at allocation 4R. Finally, we consider two additional variables. One of them (*spiteful type*) is a dummy variable that takes value 1 if the corresponding decider chose allocation 2L. Since this choice harms the sanctioner and provides no benefit to the decider, it signals some negative predisposition towards the other player. We conjecture that this might have an effect on the prediction error: As these types are potentially least collaborative, they might be more likely to misreport their expectations or choose them randomly. Alternatively, if these types tend to believe that many others are like them (False consensus; consult Offerman, 2002; Altmann et al., 2008; Blanco et al., 2009; or Gächter and Renner, 2010; for discussion and additional references), this could also lead to significant mistakes. Pursuing this kind of idea further, we also consider a dummy variable (IA type) that takes value 1 if the corresponding decider chooses allocation 4L, a choice which could be a signal of inequity-aversion.

Table 6 presents the results of four OLS regressions. The first two regressions refer to the M-treatment. Regression (1) considers all variables, while regression (2) does not consider the socio-demographic variables, or those which are not significant at the 10 percent level in model (1). Regression (3) considers all variables in the NM-treatment, and regression (4) excludes those which are never significant at the 10 percent significance level. We focus in what follows on regressions (2) and (4). Before, however, some caveats are in order. First of all, the residuals are not normally distributed, so that the results of hypothesis testing and the resulting p-values may not be valid. Second, OLS requires that the residuals be identically and independently distributed. In this respect, the Durbin-Watson d-statistic indicates that residuals are not correlated across observations – to carry out the test, we

considered all observations separately; the values of the statistic fall between 1.72 and 2.23 in all regressions. Third, one should take the usual care in interpreting our results, because any observed associations may result from a spurious correlation, or even an inverse causal relationship.[19] To the best of our knowledge, finally, we are the first to study the determinants of errors in predicting sanctions. For all these reasons, we stress that our regression results should be seen as preliminary and subject to further replication – in this vein, the web appendix shows the results of an alternative regression in which the main independent variables are allocation dummies.

Table 6 about here

In the regression (2) for the M-treatment, we observe that the independent variables paydiff- and paydiff+ are highly significant. The estimated coefficients tell us that for each 100 points difference, the prediction error increases in 9.4 and 9.9 points, respectively. This suggests that subjects' anticipation of punishment/reward tend to be less accurate when the payoff differences grow larger. Similarly, the variable Rec+ is also significant, so if reciprocity predicts reward/punishment, the prediction error increases/decreases by almost 8 points. This apparent failure to anticipate the role played by reciprocity in explaining monetary rewards is in line with our previous evidence that subjects anticipate punishment better than rewards. We also observe that the players classified as spiteful make significantly higher errors than the other players.

With respect to regression (4), we observe again that the payoff difference between the players is significantly correlated with the prediction error, as it was in the M-treatment. In contrast to the M-treatment, however, people do not commit significantly larger mistakes when reciprocity predicts rewards, but they do when the allocation is strictly egalitarian (variable *strict*). In fact, the absolute error increases in almost 23 points at those allocations where both players get identical payoffs. This suggests that deciders do not anticipate perfectly that strict equality has a significant role on the occurrence of non-monetary sanctions. We also find significant evidence that those deciders classified as inequity-

[19] For example, we will see that the 'spiteful' people choosing 2L commit very large errors. For the reasons cited above, one could argue that choosing 2L signals a certain preference or motivation, which in turn translates into large errors. Yet the other way around could be argued as well: Choice 2L might be the result of a severe inability to anticipate how the co-player will react (and not of any preference for allocation 2L), so that the errors are somehow the antecedent of the action.

averse (i.e., those choosing allocation 4L) make lower errors. Contrary to what happened in the M-treatment, intriguingly, we do not find that the subjects classified as spiteful make larger errors.[20]

3.4 Deciders' expectations and allocation choices

Our focus in this paper is on the accuracy of the deciders' beliefs. Yet some readers might be interested in whether there is any relation between beliefs and behavior (consult Table B in Appendix A for a summary of the deciders' choices). For instance, do deciders choose the allocation of the game with the highest expected payoff, given their beliefs? The answer is mostly affirmative. Since sanctions can affect the decider's payoff only in the monetary treatment, we focus most of our discussion here on that treatment. In effect, we observe there that most deciders tend to choose the allocation with the highest expected payoff, where this expected payoff is computed as the sum of the allocation payoff and the sanction expected by the corresponding player. This happens in games 1, 2, 3 and 4 for respectively 70, 91, 91, and 93 percent of the deciders. This suggests that beliefs are meaningful and taken into account when choosing.[21]

However, one should understand that the previous percentages are only a partial indication of the relation between beliefs and behavior, for several reasons. A rational decider need not choose necessarily the allocation with the highest expected payoff, but take other motives like inequity-aversion into account. This can explain why in game 1, for instance, some deciders choose the more egalitarian allocation 1R even if they do not expect that choice to be payoff-maximizing. The data from the NM-treatment provides even clearer evidence on this point. The frequency of deciders who choose the payoff-maximizing allocation in the NM-treatment can be directly computed from table B in the appendix, and we can see that not all players choose so in any game. As a final remark, it may be noted that the mere elicitation of beliefs could have primed the deciders when making their choices, so that any relation between beliefs and choices in our experiment must be considered with some caution.

4. Conclusion

This paper analyzes whether people accurately anticipate sanctions, an important question in order to explore compliance with social norms. In this respect, we must distinguish between average and individual results. On one hand, we find that the average expectation is not significantly different from

[20] This suggests that the reason why these subjects commit larger mistakes in the M-treatment is not because they choose in a random, careless manner. Otherwise, one would expect a similar result in the NM-treatment.

[21] Another signal again that most subjects did not report their expectations in a random manner.

the average sanction at most allocations. This is remarkable in itself because it is one indication that subjects do not fail in a systematic manner –i.e., they are not biased. The only exceptions to this statement are those allocations where rewards dominate, one example of an apparent tendency to better anticipate the 'stick' than the 'carrot'. On the other hand, we observe at the individual level that subjects predict the sign of the sanction correctly most of the time (the only exception to this being again one allocation where rewards are frequent), although they commit some errors when predicting the strength. Yet these errors are arguably moderate in relative terms, in particular in the case of non-monetary sanctions. Finally, we also find that subjects tend to forecast non-monetary sanctions with a smaller error than monetary ones. Since monetary sanctions are arguably not so frequently used out of the lab as non-monetary ones, one might speculate that subjects are more familiar with the latter, although further and more focused research should clarify this point.

To our knowledge, the question of whether people anticipate accurately sanctions has not received systematic attention in any previous study in experimental economics. Yet our paper contributes to a literature on the broader field of beliefs – consult Palfrey and Wang (2009) and Blanco et al. (2010) for surveys. One example of this literature is Costa-Gomes and Weizsäcker (2008), who study one-shot interactions and find that subjects fail to best respond to their own stated beliefs in almost half of the cases– in contrast, our analysis in Section 3.4 suggests that deciders in our games best responded most often.[22] As another example, subjects in Huck and Weizsäcker (2002) have to estimate the choices of other subjects in a set of binary lottery-choice tasks. Subjects almost always correctly predict which lottery is most frequently chosen, but the elicited frequencies were systematically biased. Our paper is also related to a literature with a long history in the social sciences, stressing how groups of independently-deciding individuals come up with an accurate *average* prediction. For instance, Galton (1907) found that in a weight-judging competition at a county fair the median and average guesses were fairly accurate, and more recently, Griffiths and Tenenbaum (2006) show that people are able to predict remarkably well the duration or extent of everyday phenomena (e.g. life span, movie run time). In any case, subjects in all previous studies do not have to anticipate sanctions from others. In contrast, a few papers consider games with punishment and/or rewards and elicit beliefs about sanctions. Thus, Suleiman (1996) elicits proposers' beliefs in a modified ultimatum game, finding that they are able to predict their counterparts' replies to some degree. Further, Offerman (2002) considers a sequential game

[22] The difference may be due to the fact that Costa-Gomes and Weizsäcker (2008) used 3x3 games which possibly are more complicated than our games.

where the second mover can punish or reward the first mover, or simply leave her payoff unchanged. In the treatment most relevant for us, the first mover's choices and then her probabilistic beliefs about the second mover's choice are elicited in an incentive-compatible manner. In contrast to our evidence, Offerman reports that first movers have substantial biases in predicting *both* punishment and reward.

The evidence provided here has potential implications for understanding previous experimental evidence on the effectiveness on sanctions. For instance, at each allocation we observe some fraction of subjects who anticipate the wrong sign of the sanction, thus showing very inaccurate expectations. In this line, a failure to predict sanctions is very likely the reason why some proposers in the ultimatum game make low offers, which are often rejected (Güth et al. 1982; Forsythe et al. 1994). Something similar might occur in public good games with a punishment stage (Fehr and Gächter 2000), where we observe some subjects who contribute little in the first period but then increase their contributions after being punished. In addition, our finding that the average subject predicts better the 'stick' than the 'carrot' might partially account for the relative ineffectiveness of rewards observed in studies like Sefton et al. (2007) for public good games, Andreoni et al. (2003) for the ultimatum game, and Houser et al. (2008) for the trust game.

Our study is a first attempt to systematically study players' expectations about monetary and non-monetary sanctions and may encourage more research on this topic. To start, additional studies with different games are required to check that our results extend to other situations – as we have noted, Offerman (2002) offers one setting where they do not. As another example, although our evidence shows that the average subject predicts rather well the average sanction in most allocations, we also observe substantial heterogeneity in accuracy. Studying the sources of this heterogeneity seems a promising endeavour, with potential applications to issues like crime. In a study about shoplifting, for instance, Kraut (1976) shows that those who shoplifted generally underestimated both the risk of apprehension and the severity of the sanction. Finally, another possible research question appears when both types of sanctions are available: Do people correctly anticipate when each one is used by the co-players?

References

Altmann, Steffen, Dohmen, Thomas, and Wibral, Matthias. (2008) "Do the Reciprocal Trust Less?", *Economics Letters*, 99, pp. 454-57.

Andreoni, James, Harbaugh, William, and Vesterlund, Lise. (2003) "The Carrot or Stick: Reward, Punishment and Cooperation", *American Economic Review*, 93(3), pp. 893-902.

Akerlof, George A. (1980) "A Theory of Social Custom, of which Unemployment may be One Consequence", *Quarterly Journal of Economics*, 94, pp. 749-775.

Blanco, **Mariana**, **Engelmann**, **Dirk**; **Koch**, **Alexander K.**, **and Normann, Hans-Theo**. (2009) "Preferences and Beliefs in a Sequential Social Dilemma: A Within-Subjects Analysis", IZA Discussion Paper no. 4624.

Blanco, **Mariana**, **Engelmann**, **Dirk**; **Koch**, **Alexander K.**, **and Normann, Hans-Theo**. (2010) "Belief Elicitation in Experiments: Is there a Hedging Problem?", *Experimental Economics*, 13(3), pp. 364-377

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

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

Casari, Marco, and Cason, Timothy N. (2009) "The Strategy Method Lowers Measured Trustworthy Behavior", *Economics Letters*, 103, pp. 157–159.

Coleman, James S. (1990) "*Foundations of Social Theory*", Cambridge, MA: Harvard University Press.

Costa-Gomes, Miguel A., and Weizsäcker, Georg. (2008) "Stated Beliefs and Play in Normal-Form Games", *Review of Economic Studies*, 75, pp. 729-762.

Cox, James C., and Deck, Cary A. (2005) "On the Nature of Reciprocal Motives", *Economic Inquiry*, 43, pp. 623–635.

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

Ellingsen, Tore, and Magnus Johannesson. (2008) "Anticipated Verbal Feedback Induces Pro-social Behavior", *Evolution and Human Behavior*, 29, pp. 100–105.

Elster, Jon. (1989) "Social Norms and Economic Theory", *Journal of Economic Perspectives*, 3(4), pp. 99-117.

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

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

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

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

Forsythe, Robert, Horowitz, Joel L., Savin, N. E., and Sefton, Martin. (1994) "Fairness in Simple Bargaining Experiments", *Games and Economic Behavior*, 6, pp. 347-369.

Galton, Francis. (1907) "Vox Populi", *Nature*, 75, pp. 450-451.

Gächter, Simon, and Renner, Elke. (2010) **"**The Effects of (Incentivized) Belief Elicitation in Public Good Experiments**",** *Experimental Economics*, 13, pp. 364–77.

Griffiths, Thomas L., and Tenenbaum, Joshua B. (2006) "Optimal Predictions in Everyday Cognition", *Psychological Science,* 17, pp. 767-773.

Güth, Werner, Schmittberger, Rolf, and Schwarze, Bernd. (1982) "An Experimental Analysis of Ultimatum Bargaining", *Journal of Economic Behavior and Organization*, 3, pp. 367-388.

Holländer, Heinz. (1990) "A Social Exchange Approach to Voluntary Cooperation", *American Economic Review*, 80, pp. 1157-1167.

Houser, Daniel, Xiao, Erte, McCabe, Kevin, and Smith, Vernon. (2008) "When Punishment Fails: Research on Sanctions, Intentions and Non-Cooperation", *Games and Economic Behavior*, 62, pp. 509-532.

Huck, Steffen, and Weizsäcker, Georg. (2002) "Do players correctly estimate what others do? Evidence of conservatism in beliefs", *Journal of Economic Behavior and Organization*, 47, pp. 71-85

Kandel, Eugene, and Lazear, Edward. (1992) "Peer Pressure and Partnerships", *Journal of Political Economy*, 100, pp. 801–817.

Kraut, Robert E. (1976) "Deterrent and Definitional Influences on Shoplifting", *Social Problems*, 23(3), pp. 358-369.

Leibbrandt, Andreas, and López-Pérez, Raúl. (2010a) "Different Carrots and Different Sticks: A Comparison of Monetary and Non-monetary Rewards and Punishment", mimeo.

Leibbrandt, Andreas, and López-Pérez, Raúl. (2010b) "Is the Magic in the Mix? Individual Heterogeneity in Punishment and Reward", mimeo.

López-Pérez, Raúl, and Vorsatz, Marc. (2010) "On Approval and Disapproval: Theory and Experiments", *Journal of Economic Psychology*, 31, pp. 527-541.

Masclet, David, Noussair, Charles, Tucker, Steven, and Villeval, Marie-Claire. (2003) "Monetary and Non-monetary Punishment in the Voluntary Contributions Mechanism", *American Economic Review*, 93, pp. 366–380.

Noussair, Charles, and Tucker, Steven. (2005) "Combining Monetary and Social Sanctions to Promote Cooperation", *Economic Inquiry*, 43, pp. 649–660.

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

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

Palfrey, Thomas R., and Wang, Stephanie W. (2009) "On eliciting beliefs in strategic games", *Journal of Economic Behavior & Organization*, 71, pp. 98-109.

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

Rege, Mari, and Telle, Kjetil. (2004) "The Impact of Social Approval and Framing on Cooperation in Public Good Situations", *Journal of Public Economics*, 2004, 88, pp. 1625–1644.

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

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

Suleiman, Ramzi. (1996) "Expectations and Fairness in a Modified Ultimatum Game" *Journal of Economic Psychology*, 17, pp. 531-554.

Vyrastekova Jana, and van Soest, Daan. (2008) "On the (in)effectiveness of Rewards in Sustaining Cooperation", *Experimental Economics*, 11, pp. 53–65.

Xiao, Erte, and Houser, Daniel. (2009) "Avoiding the Sharp Tongue: Anticipated Written Messages Promote Fair Economic Exchange", *Journal of Economic Psychology*, 30(3), pp. 393-404.

Tables and figures:

TABLE 1—THE ALLOCATIONS IN THE 4 GAMES		
Game	Allocation	
	Left	Right
1	(250, 100)	(200, 150)
2	(250, 100)	(250, 250)
3	(100, 200)	(150, 150)
4	(100, 200)	(100, 300)

TABLE 2—PREDICTIONS OF MONETARY PUNISHMENT/REWARD

Game	A's and B's Allocation Share			Predictions Left		Predictions Right	
	Left		Right	Punishment	Reward	Punishment	Reward
1	(250, 100)	vs.	(200, 150)	IA, RP	----	IA	RP
2	(250, 100)	vs.	(250, 250)	IA, RP	----	----	RP
3	(100, 200)	vs.	(150, 150)	----	IA, RP	RP	----
4	(100, 200)	vs.	(100, 300)	RP	IA	----	IA, RP

Notation: IA = Inequity aversion, RP = Reciprocity.

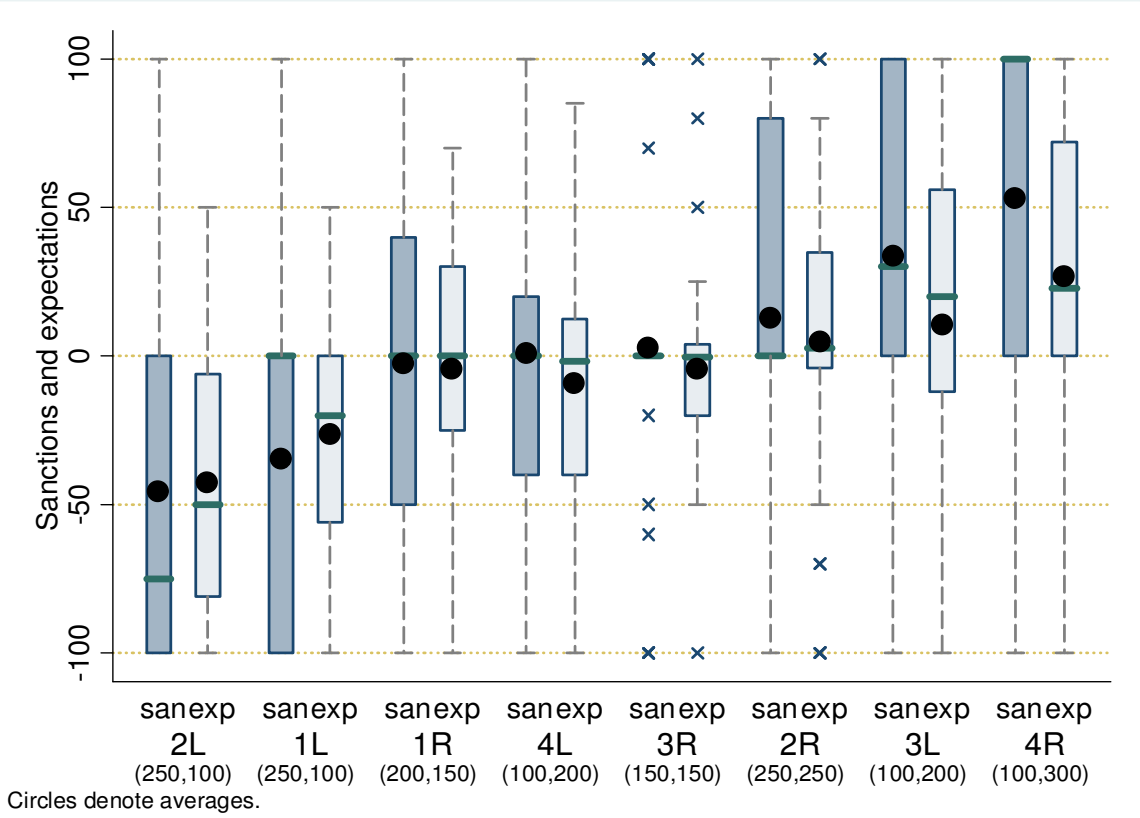
TABLE 3—Percentage of deciders expecting a positive or a negative sanction at each allocation

		Game 1		Game 2		Game 3		Game 4	
Allocation		(250,100)	(200,150)	(250,100)	(250,250)	(100,200)	(150,150)	(100,200)	(100,300)
Monetary treatment (N=46)	% expecting sanction > 0	21.7	43.5	10.9	52.2	**67.4**	34.8	36.9	**67.4**
	% expecting sanction < 0	**69.6**	43.5	**76.1**	34.8	26.1	52.2	52.2	21.7
Non-monetary treatment (N=42)	% expecting sanction > 0	2.4	59.5	0	**61.9**	**64.3**	**36.6**	19.1	**78.6**
	% expecting sanction < 0	**85.7**	14.3	**83.3**	0	4.7	30.1	**59.5**	0

Note: The bold number at some allocations indicates those deciders who correctly forecast the sign of the average sanction (if the average sanction is non-significant at that allocation, we select no number). In allocation 1L, for instance, we select the deciders expecting a negative sanction, as the actual sanction is significantly negative there.

TABLE 4— EXPECTATIONS AND SANCTIONS AT EACH ALLOCATION: SUMMARY

		Game 1				Game 2				Game 3				Game 4			
Allocation		(250, 100)		(200, 150)		(250, 100)		(250, 250)		(100, 200)		(150, 150)		(100, 200)		(100, 300)	
		exp.	sanc.	exp.	sanc.	exp.	sanc.	exp.	sanc.	exp.	sanc.	exp.	sanc.	exp.	sanc.	exp.	sanc.
Monetary treatment (N=46)	Average	-26.5	-34.6	-4.3	-2.2	-42.8	-45.2	4.9	12.6	10.5	33.7	-4.1	3	-9.3	0.6	26.6	52.8
	Median	-20	0	0	0	-50	-75	2.5	0	20	30	-0.4	0	-1.7	0	22.9	100
	Standard deviation	41.5	66.1	40.8	63.8	43.6	61.9	45	60.5	56.3	68.1	30.9	49.7	45	65.1	51.5	57.9
Non-monetary treatment (N=42)	Average	-24.9	-26.4	10.7	12.6	-42.5	-39.8	17.6	41.7	18.5	19.8	0.9	33.8	-13.74	-14.8	30.3	39.8
	Median	-18	0	5.5	0	-33	-25	7.5	0	10	0	0	0	-7.5	0	25	35
	Standard deviation	25	46.6	17.9	47.7	35.9	52	25.5	50.3	29.4	56	21.4	50.3	30.7	61.9	30.1	56.7



Circles denote averages.

Figure 1: Distribution of actual sanctions and expectations at each allocation (M-treatment)

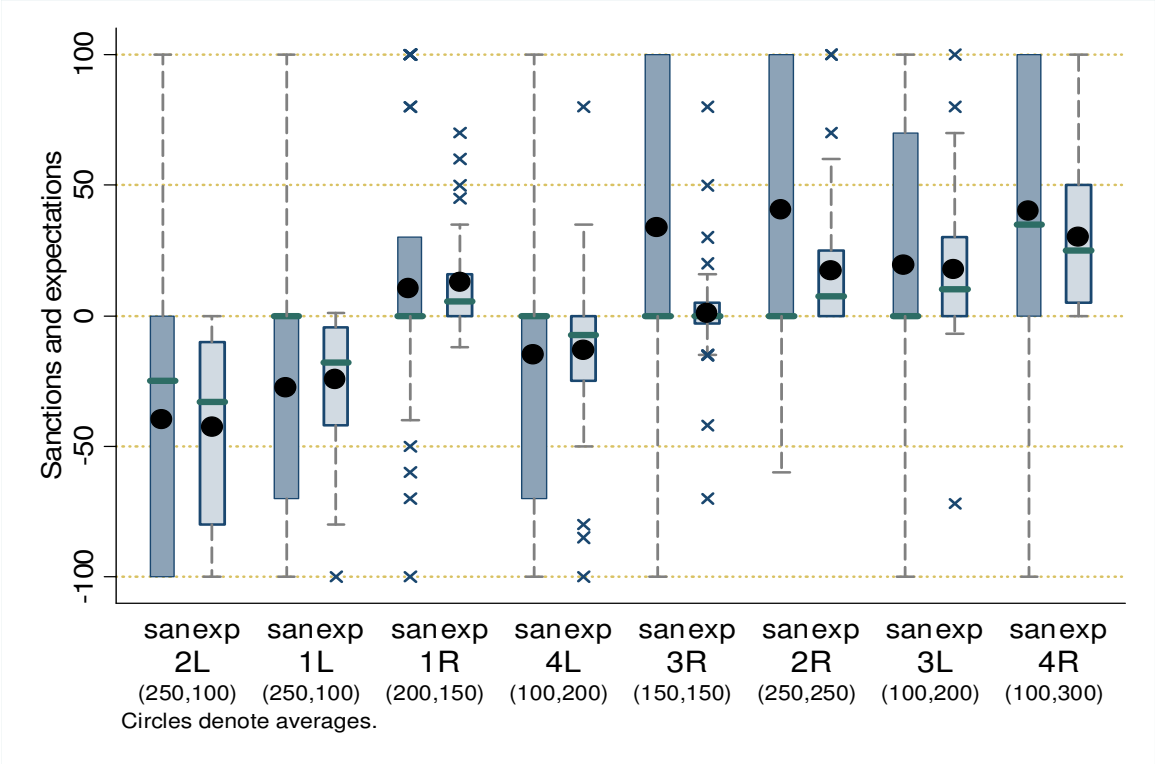


Figure 2: Distribution of actual sanctions and expectations at each allocation (NM-treatment)

TABLE 5—AVERAGE ERRORS IN EXPECTATIONS AT EACH ALLOCATION

		Game 1		Game 2		Game 3		Game 4	
Allocation		(250,100)	(200,150)	(250,100)	(250,250)	(100,200)	(150,150)	(100,200)	(100,300)
M-treatment (N=46)	Absolute error	35.6	30.5	38.6	32.2	46	19.6	33.4	46.4
	Relative error	26.5	29.8	26.5	28.5	34.4	19	33.2	30.4
NM-treatment (N=42)	Absolute error	20.2	13.5	31.9	32.5	21.7	35.9	21.2	26.9
	Relative error	16	12	22.9	23	18.1	26.8	18.5	19.3

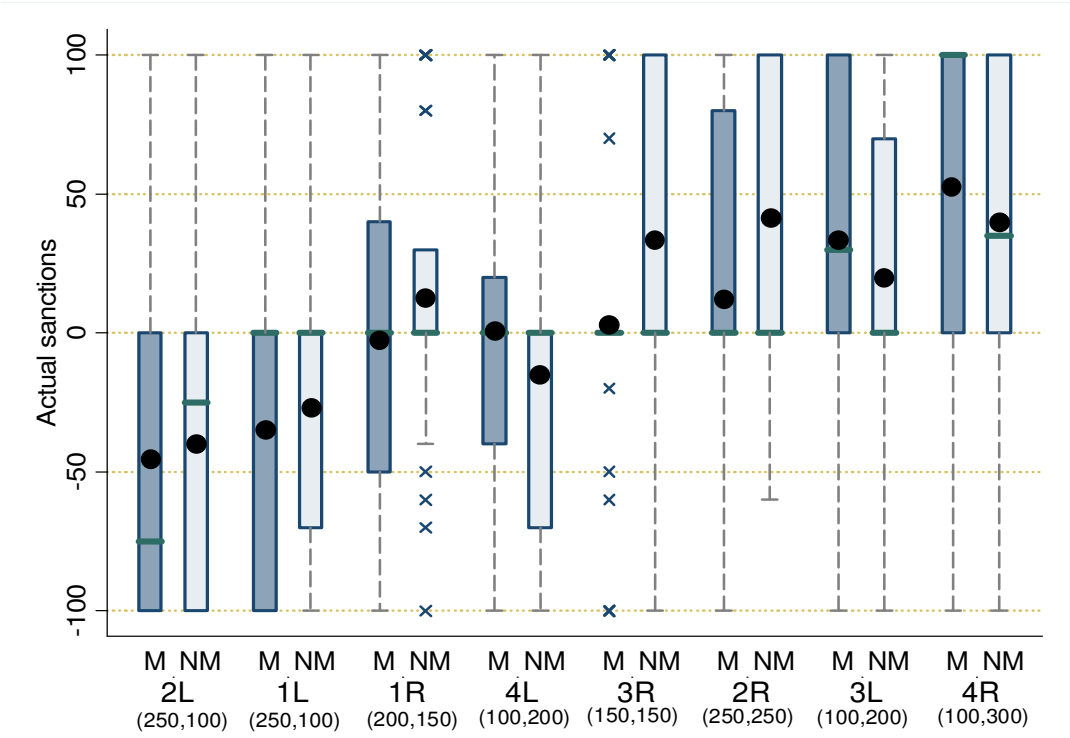


Figure 3: Distribution of sanctions at each allocation in the M and NM-treatment

TABLE 6: RESULTS OF REGRESSION ANALYSIS

Dependent Variable: Absolute prediction error

Variables	Monetary treatment		Non-monetary treatment	
	(1)	(2)	(3)	(4)
Female dummy	-5.921 (4.497)	----	-2.916 (2.639)	----
Ideology	0.260 (1.140)	----	-0.962 (0.790)	----
Religiosity	-0.431 (0.845)	----	0.791* (0.403)	0.701* (0.394)
Paydiff -	0.068 (0.053)	0.094*** (0.031)	0.080** (0.032)	0.095*** (0.027)
Paydiff +	0.104*** (0.035)	0.099*** (0.022)	0.087*** (0.020)	0.088*** (0.019)
Rec. +	6.142 (4.010)	7.885*** (2.770)	-2.555 (2.541)	----
Spiteful type	33.938*** (10.961)	32.567*** (10.624)	3.557 (2.770)	----
Strict	-3.433 (5.824)	----	21.960*** (3.453)	22.962*** (3.180)
IA type	-3.877 (4.308)	----	-6.472* (2.569)	-5.528** (2.361)
Constant	24.572** (10.045)	16.561** (6.350)	20.029*** (6.968)	10.302** (4.194)
N	368	368	320	328
R-squared	0.22	0.21	0.19	0.19

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered on individual level are presented in parentheses. Two subjects in the NM treatment did not answer the question about their political ideology, and one of these two did not reveal the degree of religiosity. Therefore, in regression (3) instead of the possible 336 observations we have only 320, while in regression (4) we have 328. We used dummies to control for sessions. They proved non-significant in all cases, except in regression 4 where in session 2 the errors were significantly higher (by 6.689 units) than in session 3.

Appendix A: Additional data

We provide here more disaggregated data on the deciders' expectations, and use the one-sample Wilcoxon signed-rank test to assess whether the sample median of the expected percentage/strength differs significantly from the actual counterparts. Several things stand out in Table A. First, it is noticeable that the expected percentage is lower than the actual percentage at all allocations in both treatments. Further, many of these differences in the average percentages are significant, particularly in the NM-treatment. Regarding the strength, subjects in the M-treatment anticipate reasonably well the actual average sanction from the sanctioners, while these predictions are rather inaccurate in the NM-treatment (except at allocations 1L and 4L). It seems that in the non-monetary treatment the underestimation of the percentage and the overestimation of the strength overall resulted in an accurate prediction of the average sanction.

TABLE A — EXPECTATIONS AND SANCTIONS AT EACH ALLOCATION: DISAGGREGATED DATA

		Game 1		Game 2		Game 3		Game 4	
Allocation		(250,100)	(200,150)	(250,100)	(250,250)	(100,200)	(150,150)	(100,200)	(100,300)
M-treatment (N=46)	Percentage (exp)	48.2	45.7	54.9	38.5	56.8	29.3	43.9	53.2
	Percentage (act)	58.7	65.2	67.4	45.7	67.4	30.4	56.5	69.6
	Signed rank test	**	***	**	_	*	_	***	**
	Strength (exp)	-49.1	-3.2	-67.5	11.9	26.2	-11.1	-14.2	44.1
	Strength (act)	-58.9	-3.3	-67.1	27.6	50.0	10.0	1.2	75.9
	Signed rank test	-	-	**	-	-	*	*	**
NM-treatment (N=42)	Percentage (exp)	35	17.9	46.8	19.3	24.2	20.7	30	32.3
	Percentage (act)	42.9	42.9	57.1	50	54.8	45.2	54.8	59.5
	Signed rank test	**	***	*	***	***	***	***	***
	Strength (exp)	-63.7	53.8	-81.9	87.1	65.2	16.4	-27.6	90.0
	Strength (act)	-61.7	29.4	-69.6	83.3	36.1	74.7	-27.0	66.8
	Signed rank test	-	***	***	***	***	***	-	***

Note: Strength (act) refers to the average choice s among those sanctioners who paid the 5 points fee. Stars indicate significant differences between the expected and the actual values indicated above the stars, according to the following scale: * < 0.1; ** < 0.05; *** < 0.01. Symbol – appears when differences are not significant at 10 % level.

TABLE B—Deciders' Choices							
Game	Allocation			percentage (M-treatment)		percentage (NM-treatment)	
	Left		Right	Left	Right	Left	Right
1	(250,100)	vs.	(200,150)	60.9	39.1	78.6	21.4
2	(250,100)	vs.	(250,250)	10.9	89.1	9.5	90.5
3	(100,200)	vs.	(150,150)	30.4	69.6	4.8	95.2
4	(100,200)	vs.	(100,300)	26.1	73.9	23.8	76.2

Web appendix **Not to be published**

Instructions for Participant A (Monetary Treatment)

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

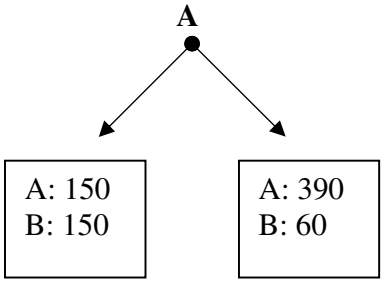
10 points = 1 Euro

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 A participant. Further, you will be anonymously matched with a type B participant** (in what follows, we call him/her B)**.** 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 A, and B will take decisions in four scenarios, all of them with a two-stage structure. In the first stage of each scenario, you have to decide between two allocations of points for you and B. In the hypothetical example of the figure, the left-hand allocation gives 150 points to you and 150 points to B. The right-hand allocation gives 390 points to you and 60 points to B.



Remember: 10 points = 1 Euro.

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

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

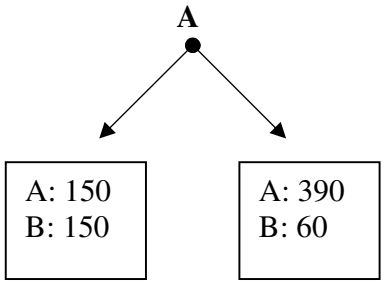
Example 2: Suppose that you choose the right-hand allocation in the previously illustrated scenario and that B then decides to spend the 5 points and assigns to you -30 points. Then you would have a balance of 390-30 = 360, and B would have 60-5 = 55 points.

Important: When deciding, B will not know the allocation actually chosen by you in any scenario. For this reason, B will indicate her decision for any possible choice by you at any scenario. Following with the example of the figure, B 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 B's choice in the payment-relevant game (without, of course, revealing B's identity).

<u>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.</u>

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)? _____

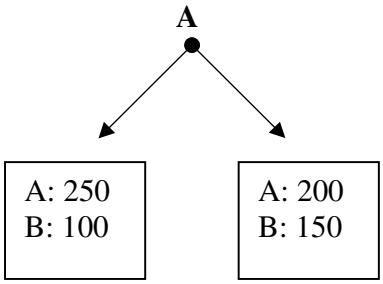
In addition:

- Will you know any of the decisions taken by B before you have made your decision in all four scenarios? Yes No
- Will B know any of your effective decisions before B 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 B ever affect your balance without spending 5 points? Yes No

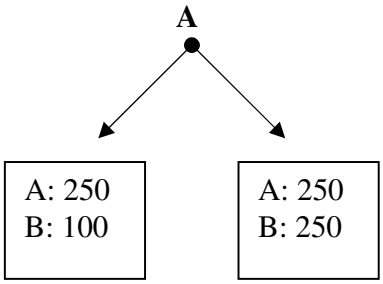
Decisions of a type-A participant

The 4 scenarios

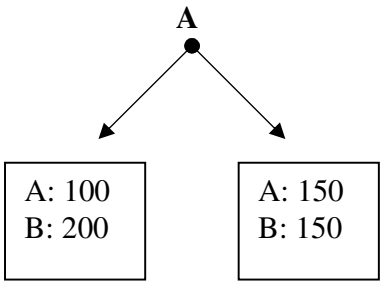
For your information, we present here the point allocations available in each of the 4 scenarios. In the next sheets, you can take your decisions in each scenario.



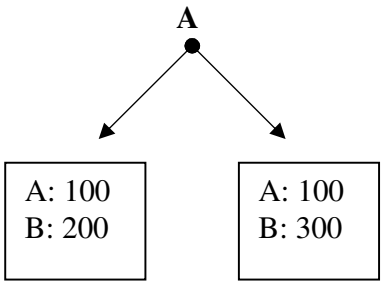
Scenario 1



Scenario 2



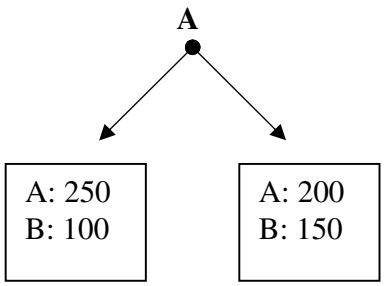
Scenario 3



Scenario 4

Note: In the next sheets, you can take your decisions in any order as you wish (that is, you do not need to start deciding in scenario 1). Until we collect your decision form, moreover, you can always change your decision in any scenario if you decide so (to facilitate this, you can initially use a pencil; write down your final decision with a pen, though).

Scenario 1



Recall: 10 points = 1 Euro

The point allocation that I choose in this scenario is (select it with a circle):

- A: 250, B: 100
- A: 200, B: 150

Independently of your previous choice, we kindly ask you to make a series of estimations (your answers here will not affect your final payoff):

- What is the percentage of participants B that will pay the 5 points if A chooses (250, 100)? _____ (this must be a number between 0 and 100, both included)
- In the previous case, how many points (in average) will these B-participants assign to the A-participant? _____ (this must be a number between -100 and 100, both included)
- What is the percentage of participants B that will pay the 5 points if A chooses (200, 150)? _____ (this must be a number between 0 and 100, both included)
- In the previous case, how many points (in average) will these B-participants assign to the A-participant? _____ (this must be a number between -100 and 100, both included)

Instructions for Participant A (Non-Monetary Treatment)

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

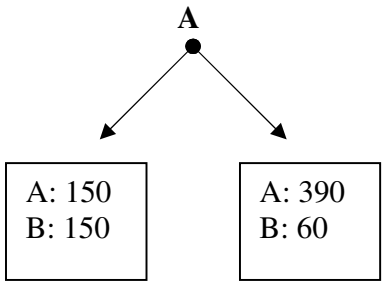
10 points = 1 Euro

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 A participant. Further, you will be anonymously matched with a type B participant** (in what follows, we call him/her B)**.** 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, that is, no participant will ever know which participant made which choice.

Description of the Experiment

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



Remember: 10 points = 1 Euro.

In the second stage of each scenario, B cannot affect your balance, but can approve or disapprove your prior choice. For this, B must pay 5 points. If B pays the 5 points, B can then assign an evaluation score between -100 and +100 to you. A negative score indicates that B disapproves your choice (-100 is maximum disapproval), while a positive score indicates that B approves your choice (+100 is maximum approval). We note again that, whatever its sign, this score will not affect your balance. If B chooses not to pay the 5 points, B cannot assign a score to you

Example 1: Suppose that you choose the left-hand allocation in the previously illustrated scenario and that B then decides to spend the 5 points and assign a score of +60 to you. That means that B approve your choice with intensity equal to 60 out of 100. Note also that your balance is unchanged (you get 150 points), whereas B would get 150 – 5 = 145 points.

Example 2: Suppose that you choose the right-hand allocation in the previously illustrated scenario and that B then decides to spend the 5 points and assign a score of -30 to you. That means that B disapproves your choice with intensity equal to 30 out of 100. Note also that your balance is unchanged (you get 390 points), whereas B would get 60 – 5 = 55 points.

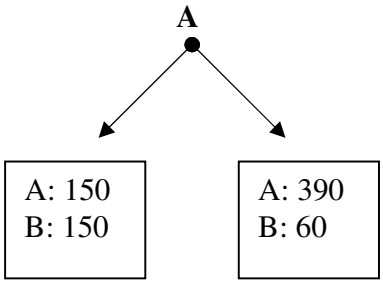
Important: When deciding, B will not know the allocation actually chosen by you in any scenario. For this reason, B will indicate her decision for any possible choice by you at any scenario. Following with the example of the figure, B should answer four questions: (1) Would you pay the 5 points if A had chosen (150, 150)?, (2) in case you pay the 5 points, what score (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 everyone will be paid in private and that we will inform you in that moment about the evaluation score that B assigned to you in the payment-relevant game (without, of course, revealing B's identity).

<u>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.</u>

In the hypothetical example of the figure, suppose that A chooses allocation (A: 150, B: 150) and that B decides to pay the 5 points, and assigns then a score +100 to A. In this case:

- What would be A's final balance? _____
- Does B approve or disapprove A's choice? _____



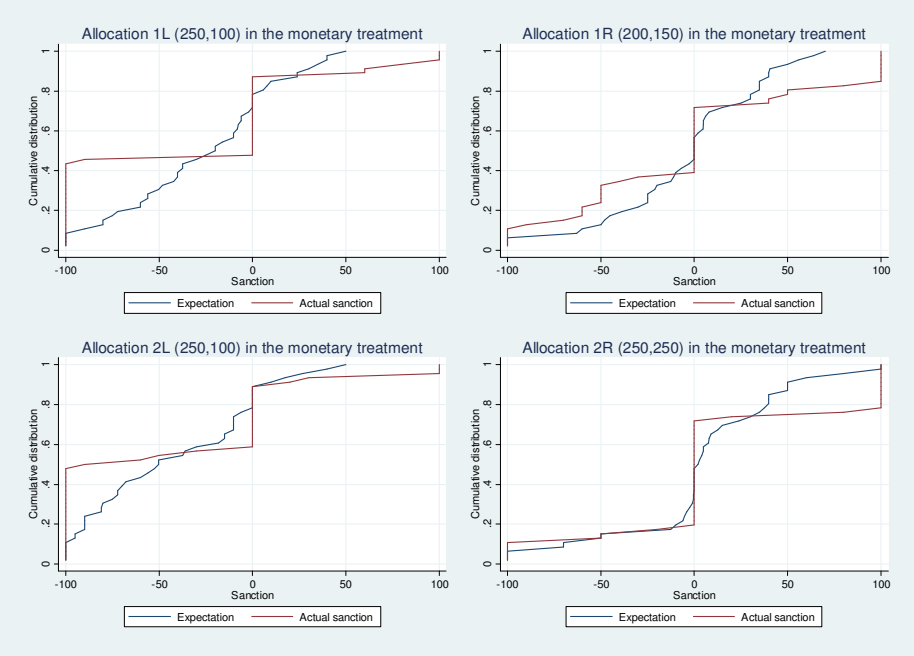
Suppose now that A chooses allocation (A: 390, B: 60) and that B decides not to pay the 5 points.

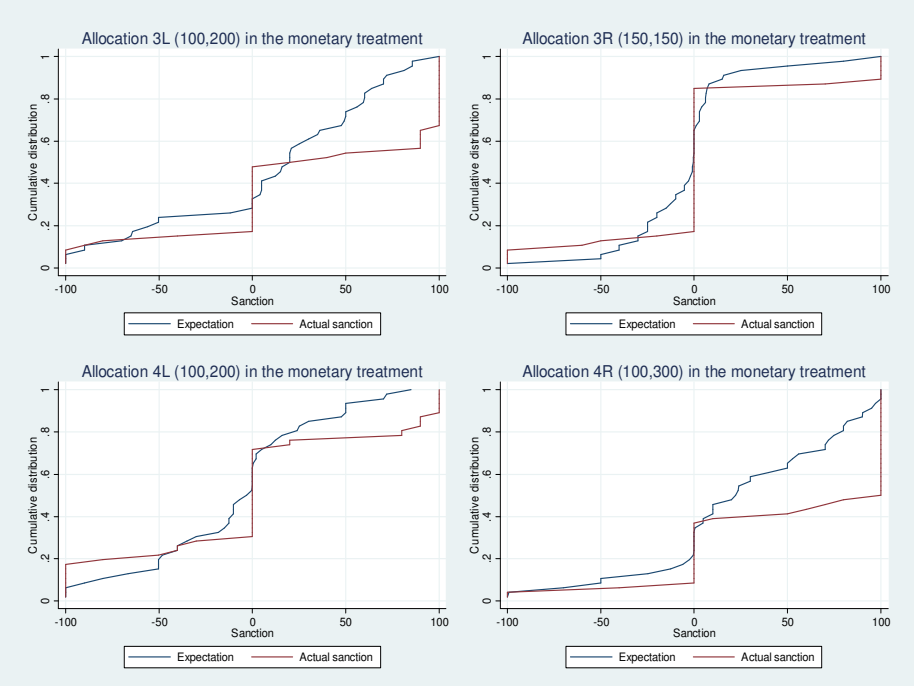
- What would be A's final balance then? _____
- What would be B's final balance then? _____

In addition:

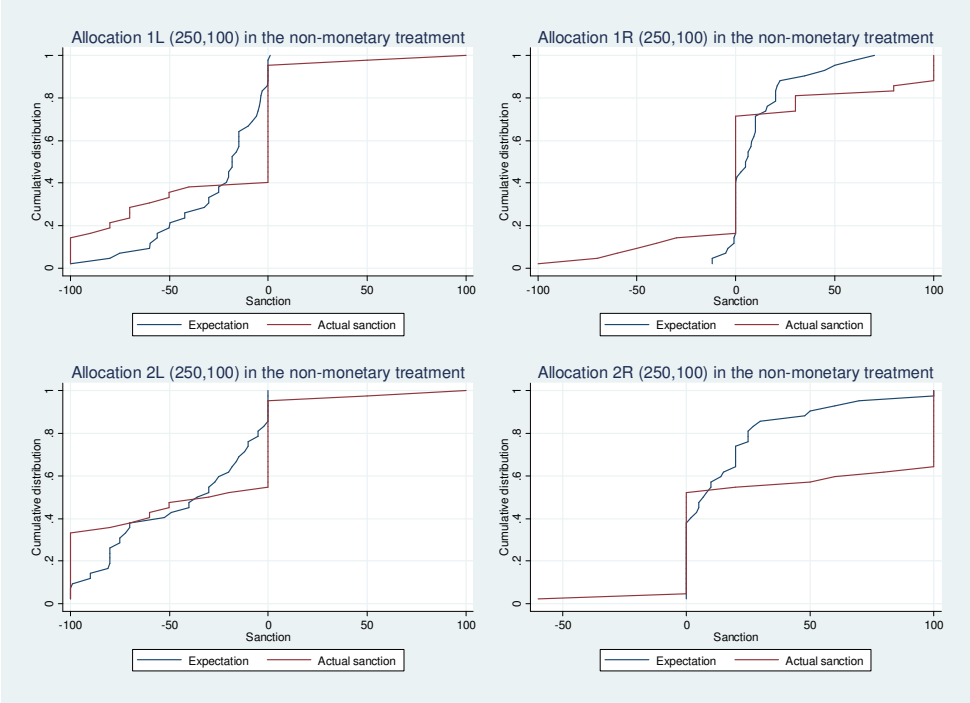
- Will you know any of the decisions taken by B before you have made your decision in all four scenarios? Yes No
- Will B know any of your decisions before she/he has made her/his decision in all four scenarios? Yes No
- How many scenarios has this experiment? _____ How many scenarios will be relevant for your payment? _____
- Can B ever affect your balance? Yes No

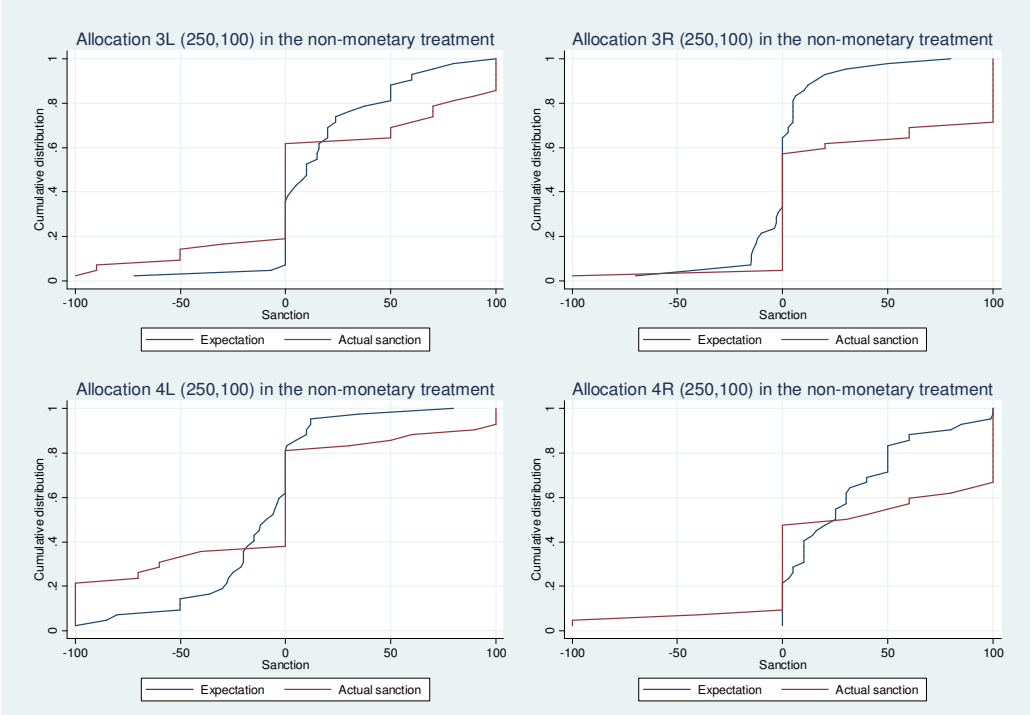
W1. Cumulative distribution functions of the expectations and the actual sanctions at each allocation of the monetary treatment





W2. Cumulative distribution functions of the expectations and the actual sanctions at each allocation of the non-monetary treatment





W3. Test results: P-values

Allocation		Game 1		Game 2		Game 3		Game 4	
		(250,100)	(200,150)	(250,100)	(250,250)	(100,200)	(150,150)	(100,200)	(100,300)
M-treatment (N=46)	MW	0.4674	0.7589	0.3008	0.9111	0.0985	0.2601	0.3959	0.0067
	Levene	0.0000	0.0132	0.0010	0.0555	0.0355	0.1949	0.1194	0.0844
	KS	0.007	0.284	0.003	0.128	0.000	0.003	0.128	0.000
NM-treatment (N=42)	MW	0.1170	0.2570	0.6261	0.4158	0.4445	0.0024	0.7322	0.6981
	Levene	0.0000	0.0001	0.0010	0.0000	0.0000	0.0000	0.0001	0.0000
	KS	0.000	0.035	0.035	0.009	0.097	0.009	0.159	0.035

We report the p-values for the following tests:

(Mann-Whitney, MW) H0 for each allocation: average expectation = average sanction.

(Levene) H0 for each allocation: variance of expectations = variance of sanctions.

(Kolmogorov-Smirnov; KS) H0 for each allocation: expectations and sanctions come from the same distribution.

W4. Disaggregated data on the sanctioners' behavior

We provide here a more detailed picture of punishment and reward by the sanctioners in our two treatments. Table B1 shows for each allocation of each game in the M-treatment, the frequency of sanctioners who invested five points to punish, the average score s chosen by those players who punished, and the same figures with respect to rewards. In game 1, for instance, we observe that 45.7 percent of the sanctioners punish at the left-hand allocation (250/100), and that the average punishment is s = -99.5. The corresponding numbers for rewards in the same allocation are 13 percent and s = 83.3.

TABLE B1—Punishment and reward in the M-treatment: Frequency and strength

Game	Allocation			Monetary punishment				Monetary reward			
				Left		Right		Left		Right	
	Left		*Right*	%	average	%	average	%	average	%	average
1	(250,100)	vs.	(200,150)	45.7	99.5	37	68.2	13	83.3	28.3	81.5
2	(250,100)	vs.	(250,250)	56.5	93.5	17.4	77.5	10.9	70	28.3	92.3
3	(100,200)	vs.	(150,150)	15.2	87.1	15.2	75.1	52.2	90	15.2	95.7
4	(100,200)	vs.	(100,300)	28.3	80	6.5	80	28.3	82.3	63	92.1

Table B2 is the analogue of B1 for the NM-treatment. It presents for each allocation of each game, (a) the percentage of sanctioners who invested five points to punish in a non-monetary manner (i.e., disapprove), (b) average disapproval among those players who disapproved, and the same figures with respect to non-monetary reward (i.e., approval). In game 1, for instance, we observe that 38.1 percent of the sanctioners disapprove the choice of allocation 1L, and that the average disapproval is of s = -78.8. The corresponding numbers for approval in the same allocation are 4.8 percent and s = 75.

TABLE B2—Disapproval and approval in the NM-treatment: Frequency and strength

Game	Allocation			Disapproval				Approval			
				Left		Right		Left		Right	
	Left		*Right*	%	average	%	average	%	average	%	average
1	(250,100)	vs.	(200,150)	38.1	78.8	14.3	58.3	4.8	75	28.6	73.3
2	(250,100)	vs.	(250,250)	52.4	82.7	2.4	60	4.8	75	47.6	90.5
3	(100,200)	vs.	(150,150)	16.7	65.7	2.4	100	38.1	80.6	42.9	84.4
4	(100,200)	vs.	(100,300)	35.7	83.3	7.1	80	20.1	78.8	52.4	86.8

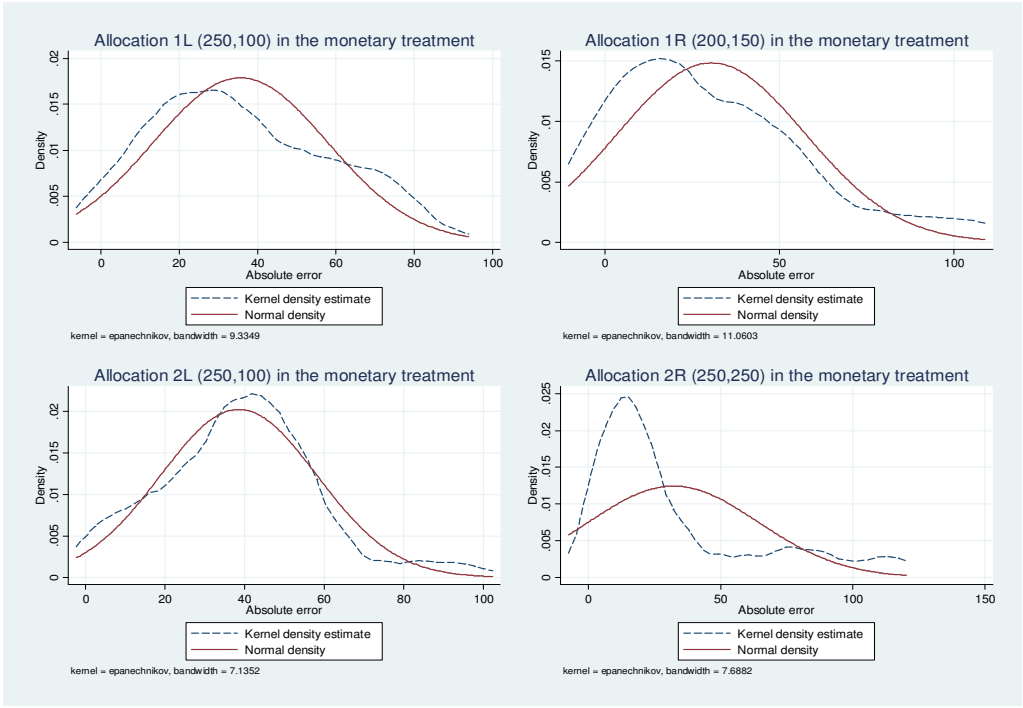
W5. Alternative regressions

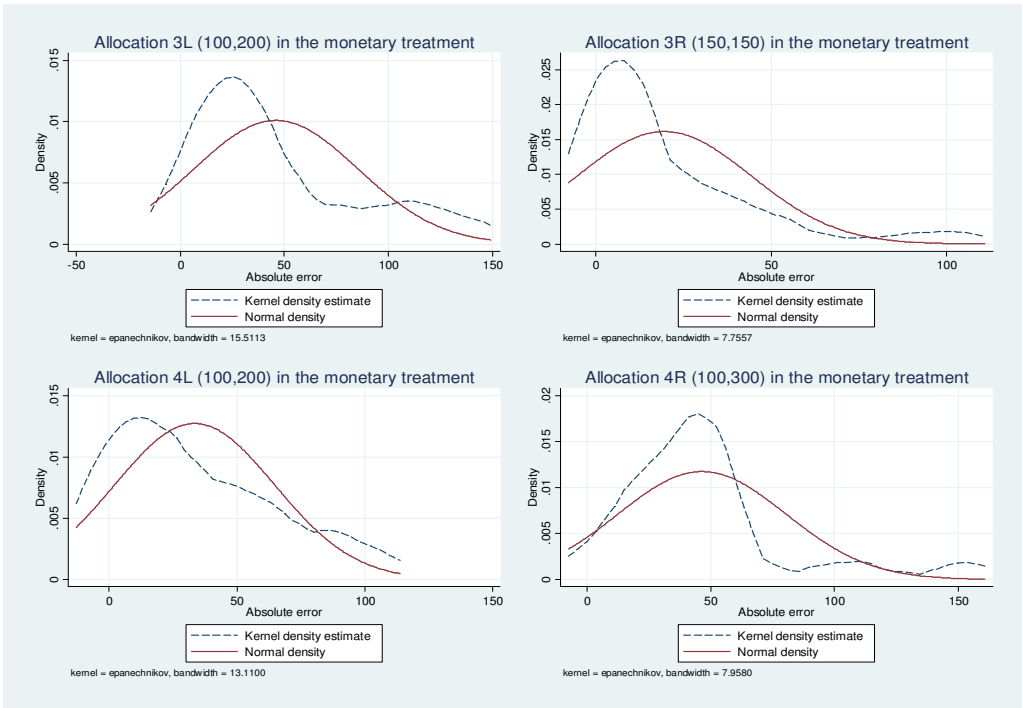
For informative purposes, we include here the results from two regressions (one for each treatment) using alternative independent variables than those used in section 3.3. In particular, we include one dummy variable for each allocation except one (to prevent multicollinearity). The non-included allocation is that with the lowest average absolute error in each treatment, that is, allocation 3R in the M-treatment, and 1R in the NM-treatment.

RESULTS OF REGRESSION ANALYSIS Dependent Variable: Absolute prediction error				
Variables	Monetary treatment		Non-Monetary treatment	
	Coefficient	Standard error	Coefficient	Standard error
Female dummy	-10.560**	5.248	-3.087	2.549
Ideology	0.694	1.377	-1.024	0.815
Religiosity	-0.841	0.775	0.796*	0.401
Spiteful type	25.304**	11.162	4.084	2.664
Pessimism	2.164**	1.072	1.134	1.702
IA type	-0.210	6.100	-7.996*	4.205
All1L	17.802*	9.083	6.222*	3.232
All1R	12.693	8.918	----	
All2L	20.753**	8.593	18.055***	3.308
All2R	13.078**	5.881	19.065***	2.371
All3L	36.775***	9.139	8.561**	3.368
All3R	----		22.119***	3.006
All4L	24.193***	8.734	7.908*	4.212
All4R	37.191***	9.055	13.849***	3.236
Session1	9.405	7.748	-0.569	2.267
Session2	0.164	5.829	4.015	3.797
Session3	7.963	8.461	----	
Constant	8.391	9.829	18.180***	(4.233)
N	368		320	
R-squared	0.25		0.23	

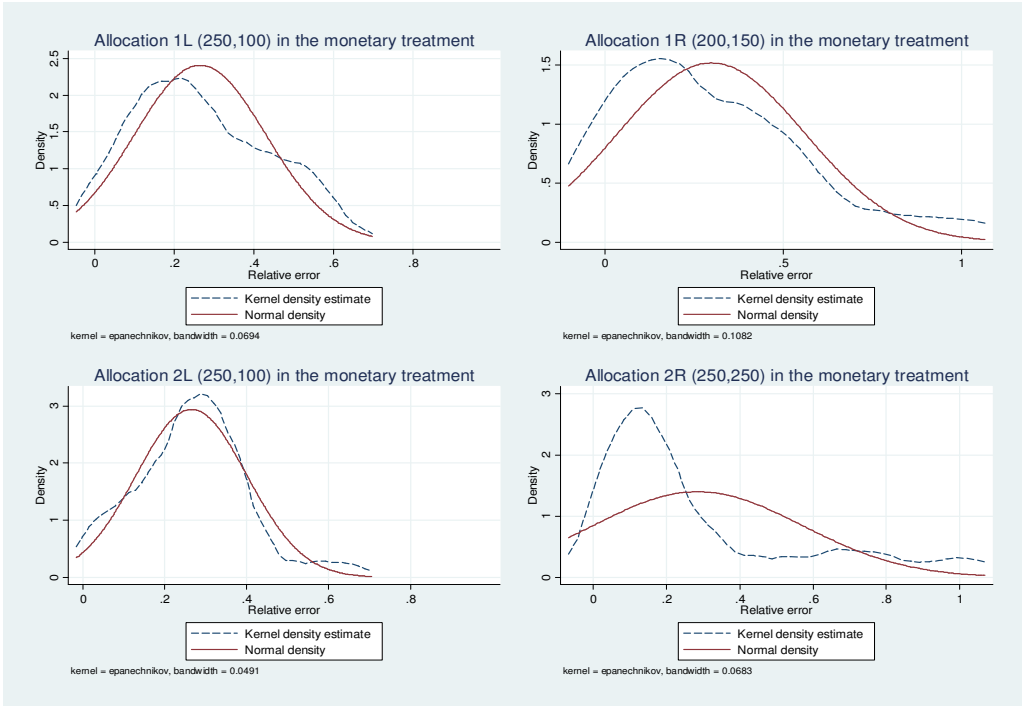
Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered on individual level. Two subjects in the NM-treatment did not answer the question about their political ideology, and one of these two did not reveal the degree of religiosity.

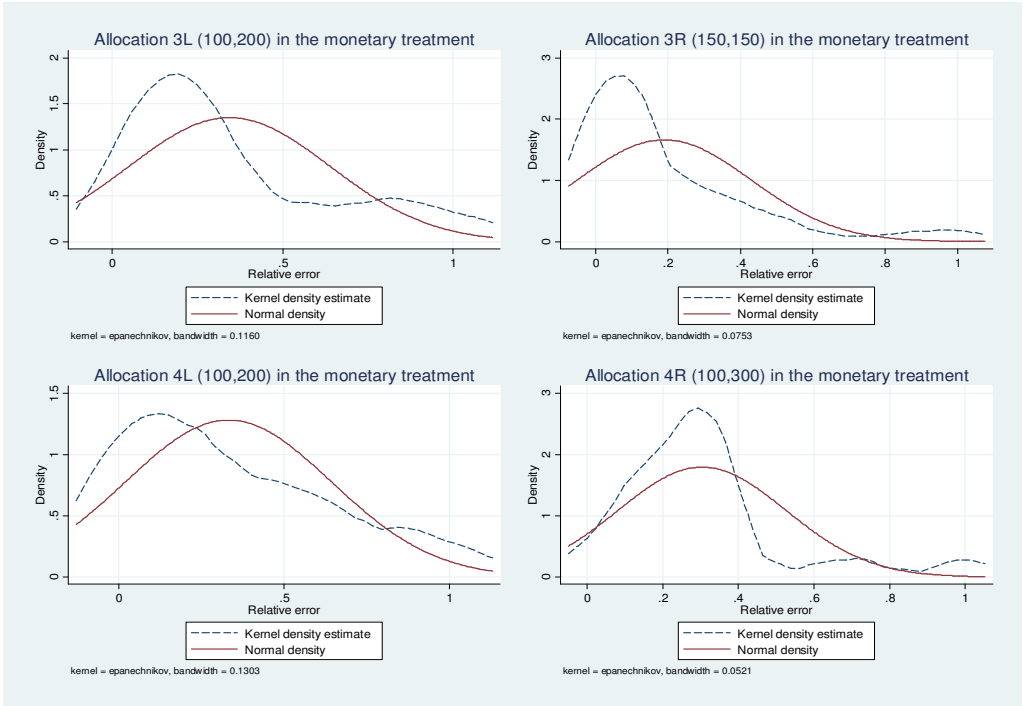
W6. Density functions of the errors: Absolute error in the monetary treatment



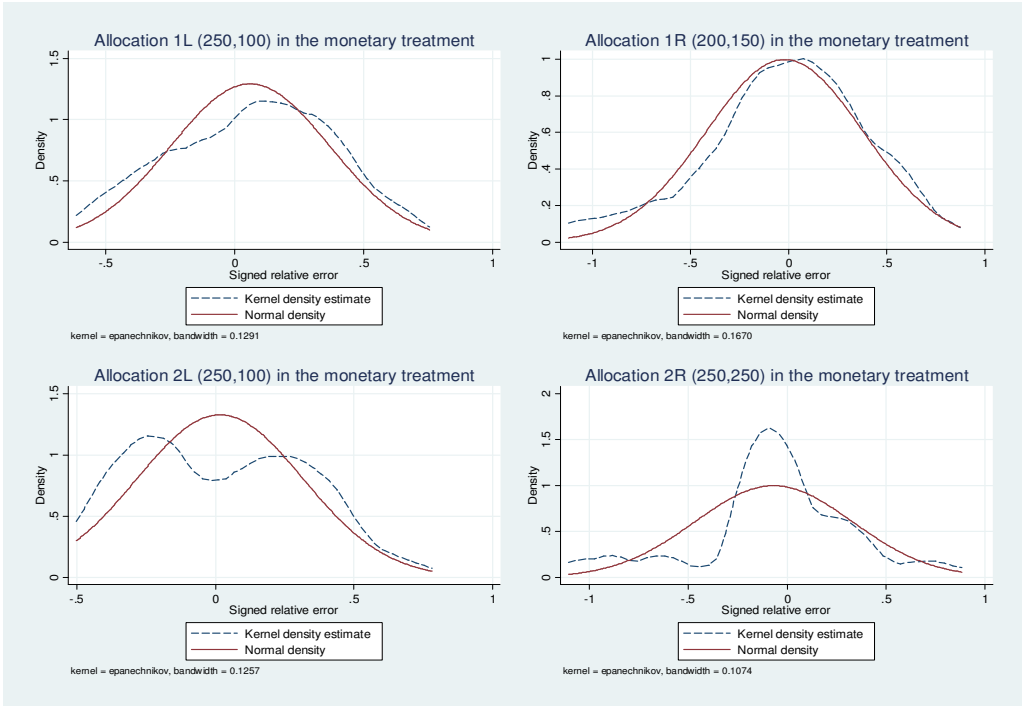


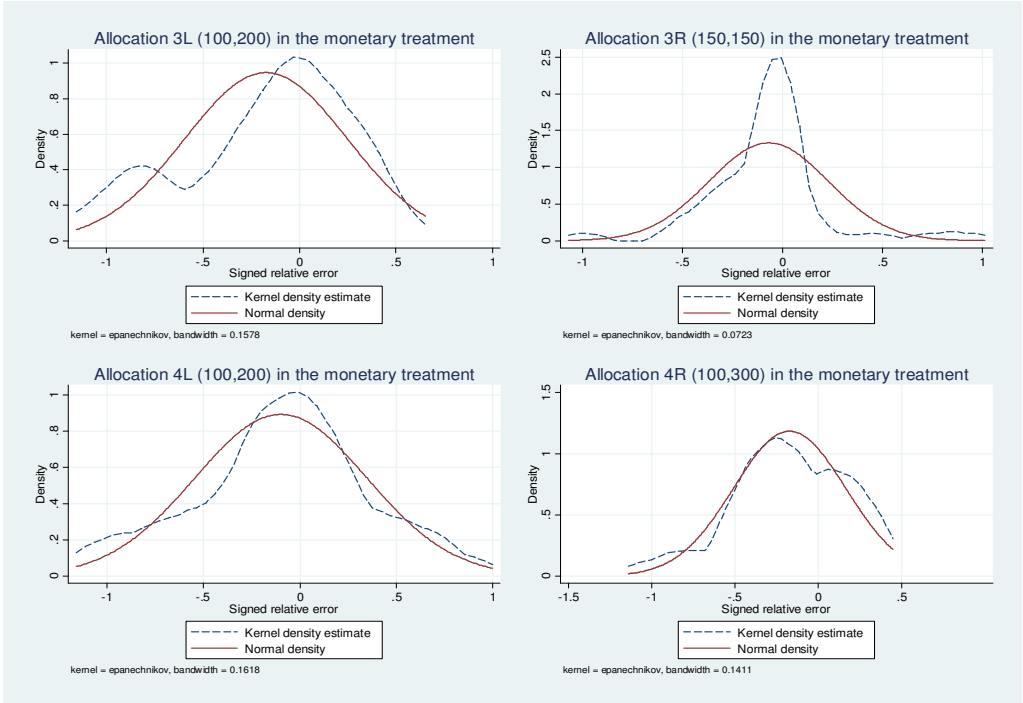
Relative error in the monetary treatment





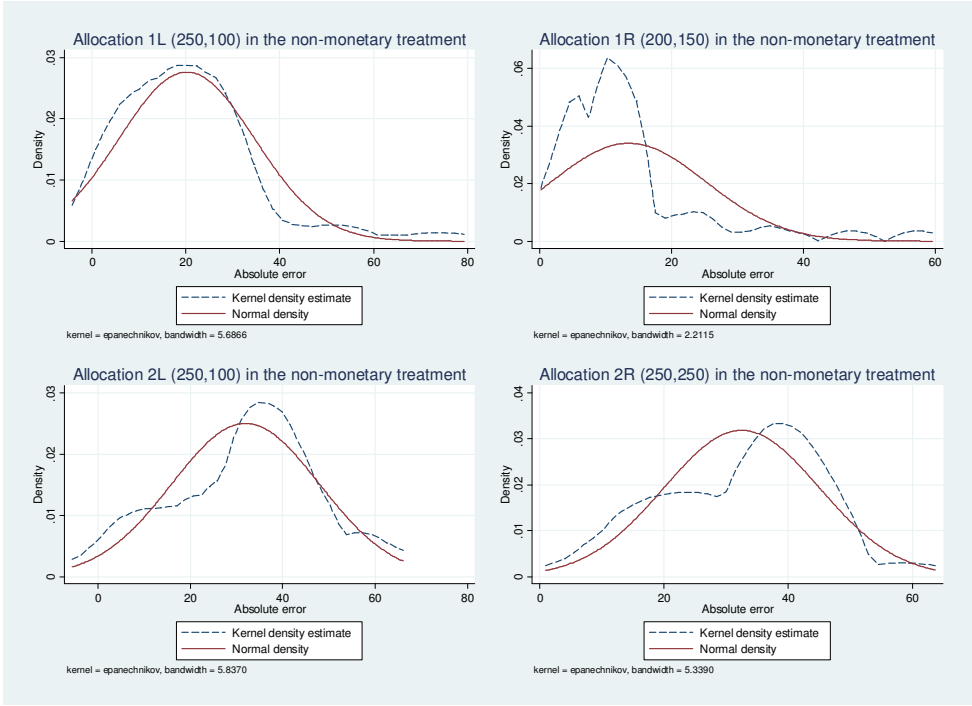
Signed relative error in the M-treatment

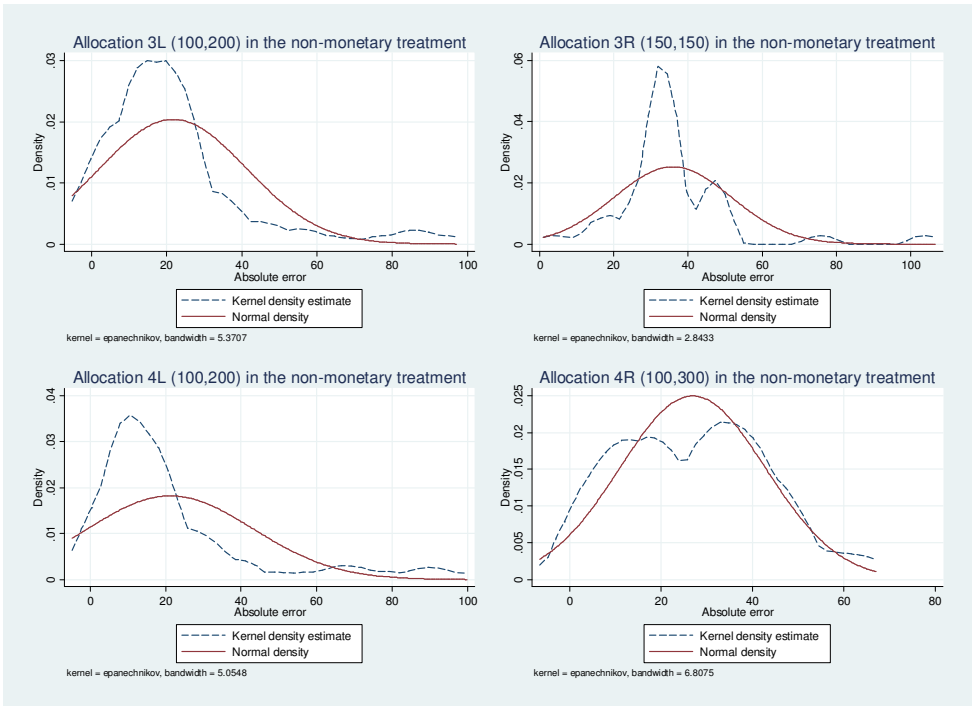




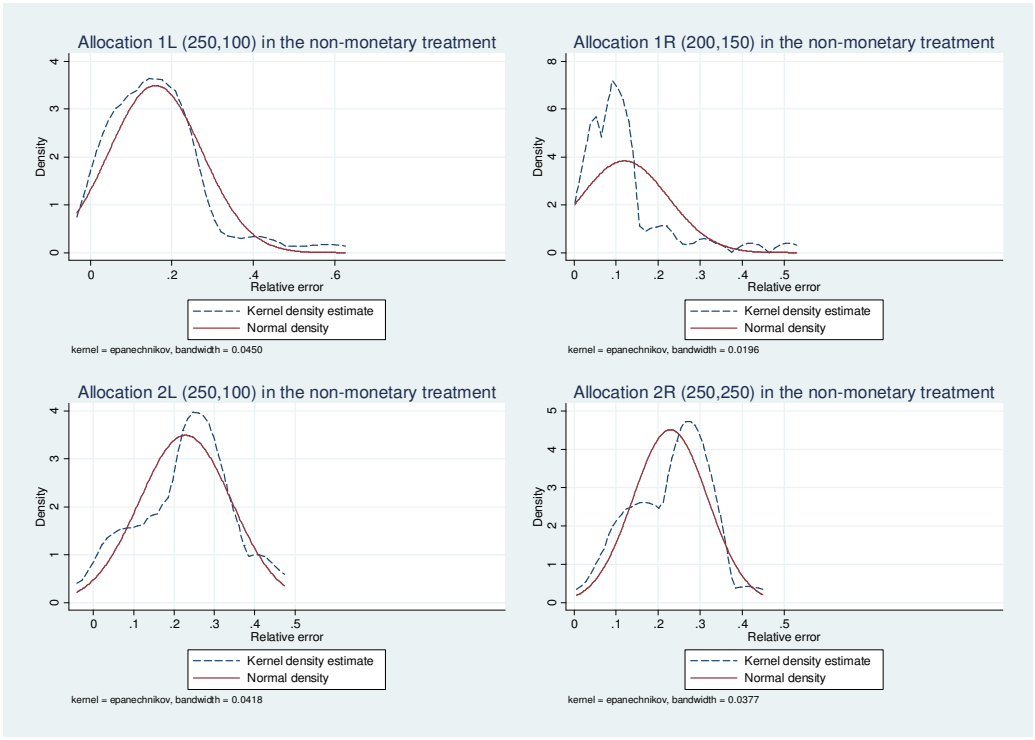
W7. Density functions of the errors: Non-Monetary treatment

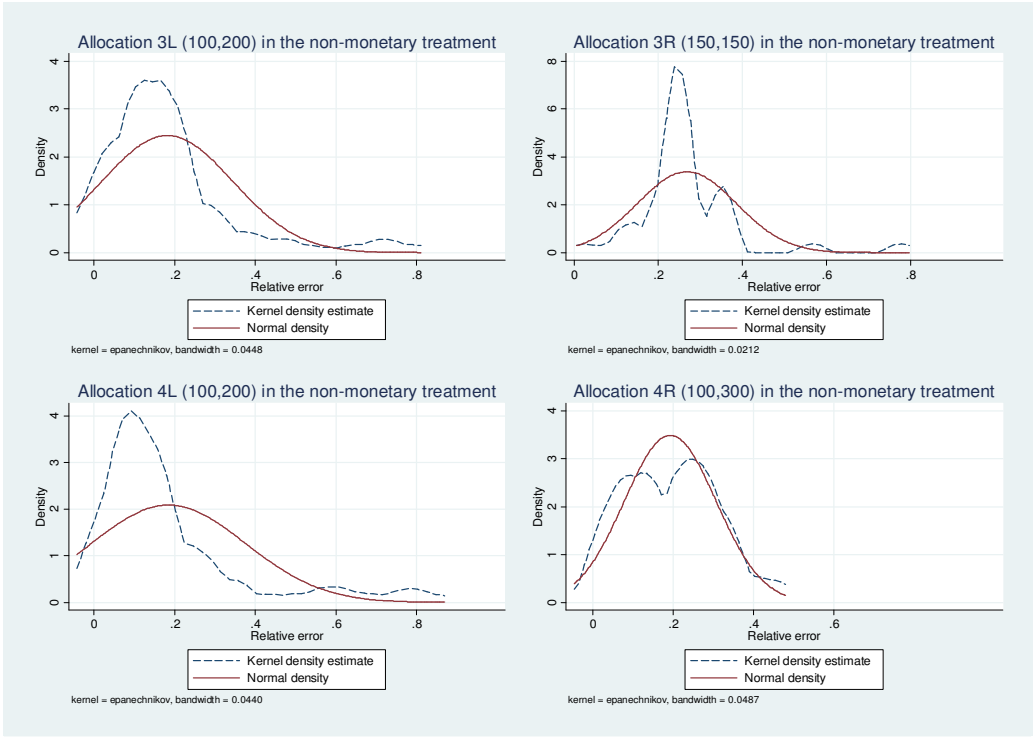
Absolute error in the NM-treatment





Relative error in the NM-treatment





Signed relative error in the NM-treatment