IZA DP No. 1635

Driving Forces Behind Informal Sanctions

Armin Falk
Ernst Fehr
Urs Fischbacher

June 2005

Driving Forces Behind
Informal Sanctions

Armin Falk
IZA Bonn
and University of Bonn

Ernst Fehr
University of Zurich
and IZA Bonn

Urs Fischbacher
University of Zurich

ABSTRACT

Driving Forces Behind Informal Sanctions[1]

This paper investigates the driving forces behind informal sanctions in cooperation games and the extent to which theories of fairness and reciprocity capture these forces. We find that cooperators' punishment is almost exclusively targeted towards the defectors but the latter also impose a considerable amount of spiteful punishment on the cooperators. However, spiteful punishment vanishes if the punishers can no longer affect the payoff differences between themselves and the punished individual, whereas the cooperators even increase the resources devoted to punishment in this case. Our data also discriminate between different fairness principles. Fairness theories that are based on the assumption that players compare their own payoff to the group's average or the group's total payoff cannot explain the fact that cooperators target their punishment at the defectors. Fairness theories assuming that players aim to minimize payoff inequalities cannot explain the fact that cooperators punish defectors even if payoff inequalities cannot be reduced. Therefore, retaliation, i.e., the desire to harm those who committed unfair acts, seems to be the most important motive behind fairness-driven informal sanctions.

JEL Classification: A13, D63, D23, C92, K42

Keywords: sanctioning, cooperation, social norm, reciprocity, fairness, spitefulness

Corresponding author:

Armin Falk
IZA and University of Bonn
Schaumburg-Lippe-Strasse 7/9
D-53113 Bonn
Germany
Email: falk@iza.org

1. INTRODUCTION

This paper examines the driving forces behind informal sanctions. We term sanctions to be informal if they are not imposed by formal, legal bodies but by private parties who punish other peoples' observed behaviors. Informal sanctions are important because the bulk of peoples' daily interactions is not governed by explicit, enforceable contracts but by implicit agreements and social norms. Informal sanctions typically enforce these agreements and norms (Francis 1985, Ostrom 1990, Hechter and Opp 2001, Knez and Simester 2001). In the light of previous evidence (Güth et al. 1982, Roth 1995, Fehr and Gächter 2000) it is not the question anymore *whether* there is informal sanctioning. The problem, which is not yet understood, however, is *why* people sanction and, in particular, why they sanction others' cooperative or defective behaviors. For this reason, we examined the motivational forces behind informal sanctions in a series of cooperation experiments. We conducted several three player prisoners' dilemma (PD) experiments with direct sanctioning opportunities, i.e., players had the option of sanctioning other group members after being informed of the latters' choices in the PD. We deliberately focused on games with more than two players because most past research examines rejection behavior in two player bargaining games, while social norms clearly extend beyond the context of bilateral interactions.[2] In addition, the examination of bilateral interactions cannot answer important questions regarding the nature of the fairness principles that drive informal sanctions. Thus, our limited understanding of the driving forces behind sanctions in multilateral interactions also implies a limited understanding of the relevant fairness principles.

Our experiments yield the following findings. First, if we average over all treatments, 63 percent of the subjects (N = 315) cooperate and 37 percent defect. Second, when cooperators punish, they almost exclusively penalize defectors. Third, when viewed through the lens of fairness theories, the share of defectors who punish is surprisingly large. These defectors

impose roughly equal sanctions on cooperators and on (other) defectors. Fourth, the cooperators' sanctions are quantitatively much more important than those of the defectors because the percentage of cooperators who punish is much larger than the percentage of punishing defectors and cooperators' sanctions are much more severe at the individual level.

The pattern of individual sanctions is also related to a fifth result which we find particularly interesting. We implemented a high and a low sanction condition in our experiments. Every money unit spent on punishment reduced the punished individuals' payoff by 2.5 or $3^{1/3}$ money units in the high sanction condition, while the same expenditure only reduced the punished individual's payoff by one money unit in the low sanction condition. Thus, sanctioning was not associated with a change in the payoff difference between the punisher and the punished subject in the low sanction condition. We find that the sanctions of individual cooperators exceed those of individual defectors by a factor of almost three in the high sanction condition. Moreover, defectors' sanctions vanish completely in the low sanction condition, whereas the cooperators spend about 2.5 times *more* money on punishing defectors in the low than in the high sanction condition. These results suggest that cooperators have a very strong motive for sanctioning and that the desire to increase the payoff difference between the defector and the punished individual drives defectors' spiteful sanctions.

The observed punishment patterns also have implications for the validity of different theories of fairness (Rabin 1993, Levine 1998, Fehr and Schmidt 1999, Bolton and Ockenfels 2000, Dufwenberg and Kirchsteiger 2004, Falk and Fischbacher forthcoming). One implication is that – with the exception of Levine's (1998) approach – all fairness theories have difficulties in explaining spiteful sanctions. In addition, our data help us answer two important questions regarding cooperators' fairness principles. First, can the cooperators' sanctions be explained by fairness preferences which neglect the payoffs of individual group members and focus, instead, on the comparison of a player's own payoff with an aggregate measure of the group's payoff,

such as the group's average or total payoff? Second, can the cooperators' sanctions be explained by approaches which assume that players want to minimize payoff inequalities, or can these sanctions be better explained by the motive to retaliate, i.e., the motive to harm those who acted unfairly?

2. THE EXPERIMENTAL DESIGN

We conducted several PD experiments with sanctioning opportunities in order to examine the driving forces behind informal sanctions. In the first stage of these experiments, subjects decided simultaneously whether to cooperate or defect. In the second stage, every subject was informed about the other PD players' individual decisions in stage 1. Each subject could then punish the other players by assigning them deduction points. For convenience, we sometimes denote these experiments in the following as PDs; the reader should keep in mind, however, that a sanctioning opportunity existed in all experiments.

213 subjects participated in a three player one-shot PD in the first two treatments. The payoff consequences of the cooperation decisions in the PD are presented in Table I. The table demonstrates the free rider incentive at the cooperation stage. Player i's dominant action is to defect, and his payoff increases the more other players cooperate. If all three players in a group defect, each player earns 20, while if all cooperate, each earns 36. At the second stage, each player can reduce the payoff of either or both of the other two players by a maximum of 25 tokens. If player i assigns p_{ij} deduction points to j, $j's$ payoff is reduced by fp_{ij} ($f \geq 1$) and $i's$ payoff is reduced by p_{ij}. Thus, punishment is costly for both the punishing and the punished player.

Table I about here

The treatments differ with regard to the parameter *f*. In the *low-sanction* treatment, *f* was set equal to 1. This treatment is useful because both theories of inequity aversion (Fehr and Schmidt 1999, Bolton and Ockenfels 2000) as well as Levine's (1998) theory predict that no punishment will take place if $f \leq 1$, while intention based reciprocity theories (Rabin, 1993, Dufwenberg and Kirchsteiger 2004, Falk and Fischbacher forthcoming) are consistent with punishment, even if it is equally costly or more costly for the punisher than for the punished. In addition, the low sanction treatment also enables us to study the nature of spiteful punishment in more detail.

Punishment was more effective in the *high-sanction* condition. The effectiveness of sanctions in the high-sanction treatment varied depending on whether *i* punished a cooperator or a defector. If *i* assigned one deduction point to a cooperator, the cooperator's payoff was reduced by 3 1/3 tokens (*f* = 3 1/3), whereas if *i* assigned a point to a defector, the defector's payoff was only reduced by 2.5 tokens (*f* = 2.5). In other words, the punishment of cooperators yielded the same payoff reduction at a lower cost. The main reason for this design feature was that fairness approaches, which neglect individual group members' payoffs and focus exclusively on the relation between a player's own payoff and the group's total or average payoff, predict that punishment should be exclusively targeted towards the cooperators if punishing cooperators is cheaper than punishing defectors. This contrasts sharply with most other fairness theories which predict that even if punishing cooperators is cheaper, sanctioning should be exclusively targeted towards the defectors.

Each group member made a *contingent* sanctioning decision at the punishment stage before being informed of the other players' decision at the cooperation stage. That is, each member assigned points to another member both for the case that the other member cooperated and that the other defected. This has the advantage of allowing us to collect much more information about subjects' sanctioning behavior. The contingent sanctioning decisions were

made with the help of four decision screens. On each screen, i had to indicate whether he wanted to assign deduction points to no other player in the group, or to one or both of the other players and if so, how many points. The four screens represented the four possible choice combinations of the other two players: both of the other players defected, both cooperated, the second player cooperated while the third defected, and vice versa. After all three players had made their contingent sanctioning decisions, they were informed of the other players' behaviors at the cooperation stage and how much they themselves had been sanctioned. The entire procedure for making decisions at stage two was carefully explained to the subjects in the instructions. We also asked several hypothetical questions at the end of the instructions to check subjects' comprehension of the procedure. The experiment began after all subjects had solved all questions successfully.

In principle, it is possible that the elicitation of contingent responses (i.e., the "strategy method") induces different behaviors relative to a situation where the subjects face given, known cooperation decisions (i.e., the "specific response method"). However, Cason and Mui (1998) and Brandts and Charness (2000) report evidence indicating that contingent responses and the specific response method do not elicit different behaviors. Nevertheless, we also conducted the high sanction treatment with the specific response method to check for potential artifacts caused by the strategy method in our context. In this control experiment 102 subjects participated. They made their sanctioning decisions in the punishment stage of each period by responding to the actual cooperate/defect decisions of the players in the cooperation stage.

In all treatments, subjects did not know the personal identities of their interaction partners, and all interactions between the subjects were anonymous.[3] No subject participated in more than one treatment. The cooperation decision was always framed in terms of investments into a project. The punishment decision was framed as the assignment of deduction points to

the other group members. We used this frame to avoid value laden terms such as "punishment" or "sanction".

The subjects in all experiments described in this paper were undergraduate or graduate students from the University of Zurich or the Swiss Federal Institute of Technology in Zurich. All experiments were programmed in z-Tree (Fischbacher 1999) and conducted in the computer laboratory of the Institute for Empirical Research in Economics. The sessions with the strategy method lasted roughly 35 minutes and subjects earned CHF 25 ($\approx$ $18.5) on average. The sessions with the specific response method also lasted 35 minutes and average earning per subject also was CHF 24.6 ($\approx$ $18).

3. RESULTS

In this section, we first show how cooperators and defectors punish other cooperators and defectors (Section 3.1.). We then interpret these punishment patterns in light of preferences for fairness and spitefulness (Section 3.2.). Next, we present the results regarding the robustness of spiteful preferences when the specific response method is used to determine punishment (Section 3.3.). Finally, we discuss the implications of our results for the relevance of different fairness principles (Section 3.4.).

3.1. *Who sanctions whom?*

120 subjects participated in the high-sanction condition (f = 2.5 or 3.33) of our one-shot PD. 61 percent of them cooperated, and the remaining subjects defected. 93 subjects participated in the low-sanction treatment (f = 1) of the one-shot PD and 51 percent of them cooperated while the others defected. Figures 1a and 1b provide a first indication of the general pattern of sanctions. Figure 1a depicts the average sanctions in the high-sanction treatment and Figure 1b does so in

the low sanction treatment. Note that Figure 1 depicts the severity of the sanctions imposed on other players from the viewpoint of the subject who punishes. The figure does not directly show the average payoff reductions the sanctioned subjects experience. Figure 1a (see bar corresponding to C → D) shows, for example, that a cooperator reduces a defector's average income by 5.7 tokens. Therefore, if there are two cooperators in a group and one defector, the defector's income is reduced by 11.4 tokens on average.

<center>Figures 1a – 1c about here</center>

Figures 1a and 1b exhibit several remarkable features. *First* of all, cooperators almost exclusively tend to punish defectors. A non-parametric Wilcoxon signed rank test rejects the null hypothesis that cooperators punish defectors and other cooperators with the same probability ($z = 6.64$, $p < 0.001$).

Second, the defectors also punish in the high sanction treatment, but they tend to punish both defectors and cooperators. This is indeed surprising when viewed through the lens of fairness theories. Moreover, the strength of the sanctions imposed on other defectors is almost the same as that of the sanctions imposed on cooperators, and defectors are equally likely to punish cooperators and defectors. The null hypothesis that defectors sanction other defectors and cooperators with the same probability cannot be rejected (Wilcoxon signed rank test, $z = 1.34$, $p = 0.179$). *Third*, the figures show that cooperators impose by far the strongest sanctions on defectors.

Fourth, only one type of punishment occurs in the low-sanction treatment: cooperators punish the defectors. All other types of punishment are virtually non-existent. Another interesting feature of Figure 1b is that the severity of cooperators' sanctions almost equals that in Figure 1a. Since the payoff reduction per assigned deduction point, *f*, was much lower in the low-sanction treatment than in the high-sanction treatment, the punishing cooperators spent

much more money on sanctioning in the low-sanction treatment. This indicates that the cooperators' motives behind the sanctions are very strong. Figure 2 further illustrates the similarity of the cooperators' punishment pattern across conditions, by showing the distribution of payoff reductions that individual cooperators imposed on average on the defectors. In both conditions, 60 – 70% of the cooperators punish defectors. Most punishers impose payoff reductions below 10, but some even chose the maximal reduction of 25.

The share of people in the different sanctioning categories is important for assessing the relevance of different motives behind the sanctions. Therefore, Table II shows the percentage of cooperators and defectors who sanction in the different conditions. In the high-sanction condition, for example, 68.5 percent of the cooperators (50 out of 73 subjects) sanctioned defectors and 40.4 percent of the defectors (19 out of 47) sanctioned other defectors. 34 percent of the defectors (16 out of 47) sanctioned the cooperators. There is thus a surprisingly large share of defectors who punish in the high sanction condition. However the share of cooperators who punish is even higher. In contrast, only 2.2 percent of the defectors (1 out of 46) but 59.6 percent of the cooperators (28 out of 47) punish in the low-sanction treatment.

Table II about here

3. 2. *Fairness versus Spite*

To what extent do spitefulness and fairness considerations drive the sanctions depicted in Figures 1a and 1b and in Table II? Spiteful sanctions are those that occur because the sanctioning subject values the payoff of the sanctioned subject negatively, irrespective of whether the sanctioned subject behaved fairly or unfairly, and irrespective of the distribution of pre-sanction payoffs. A general form of spiteful preferences (Kirchsteiger 1994, Mui Vai Lam 1995) can be represented by $U_i(\pi_1,..., \pi_n)$ where $(\pi_1,..., \pi_n)$ denotes the vector of material payoffs in an n-player game and U_i obeys $\partial U_i/\partial \pi_i > 0$ and $\partial U_i/\partial \pi_j < 0$ for all $j \neq i$ and all

vectors $(\pi_1,...,\pi_n)$. A particular form of spitefulness prevails if a subject prefers inequality, i.e., if i's utility is increasing in $(\pi_i - \pi_j)$. There are two reasons why spitefulness cannot explain the cooperators' sanctions. First, spiteful subjects are indifferent with regard to the punishment target because they value both the cooperators' and the defectors' payoffs negatively. Therefore, they sanction both defectors as well as cooperators. Yet, cooperators sanctioned the defectors almost exclusively. Second, subjects with spiteful preferences are unlikely to cooperate because cooperation means increasing the payoff of the other group members at the expense of one's own material payoff. However, the behavior of those subjects who cooperate *and* punish the defectors is consistent with fairness approaches. For example, the fairness models of Dufwenberg and Kirchsteiger (2004), or Falk and Fischbacher (forthcoming) can rationalize this punishment pattern because defection in the PD is viewed as an unkind act which triggers the sanctions. This suggests (see Table II) that more than two thirds of the cooperators in the high-sanction treatment punished in response to the violation of fairness principles while roughly 60 percent of the cooperators sanctioned for this reason in the low sanction treatment.

However, Table II also indicates that a large percentage of the defectors punished in the high sanction treatment. It seems difficult, if not impossible, to reconcile the defectors' sanctioning behavior with any reasonable notion of fairness. After all, the defectors benefit from the cooperation of the others but refuse to bear "their" part of the cost in the investment project. Not only do these defectors benefit from the project without sharing the burden, they also sanction cooperators and other defectors. Therefore, the defectors' sanctions point towards motivational forces that most fairness theories have neglected so far.[4] It is difficult to rationalize the defectors' punishment of the cooperators except by assuming some form of spitefulness. In fact, a comparison of Figure 1a with Figure 1b suggests that a particular form of spitefulness drives the defectors' sanctions. Recall that a sanctioning defector can increase

the payoff difference between himself and the sanctioned subject in the high-sanction condition because $f > 1$. In the low-sanction condition, however, any sanction leaves the payoff difference between the sanctioning and the sanctioned subject unchanged because $f = 1$. It is, therefore, striking that the percentage of defectors who punish is rather high in the high-sanction condition but almost zero in the low-sanction condition. Only one defector (out of 46) punishes in the low-sanction condition. This pattern is consistent with spitefulness in the form of a desire to increase payoff differences between the punishing and the punished subject.

Are fairness driven sanctions or spiteful sanctions more important in our PD? There are two reasons why fairness-driven sanctions are more important. First, the sanctions that the cooperators impose on the defectors are much stronger than those the defectors impose on others. Second, the sanctioning cooperators prevail numerically over the sanctioning defectors. A non-parametric Fisher exact test shows that the fraction of cooperators who sanction the defectors *only* is significantly higher ($p < 0.001$) than the fraction of defectors who sanction other players.

3. 3. *A Robustness Check*

In this section, we report the results of the control treatment where the specific response method determined the sanctioning behavior. This treatment is designed to check whether the strategy method used in the first two treatments is responsible for some of the main results documented in the previous sections. In particular, the surprisingly large share of spiteful sanctions observed in Figure 1a and Table II motivates this robustness check.

79 subjects cooperated and 23 defected in the control treatment. A comparison of Figures 1a and 1c illustrates that the punishment pattern under the specific response method is qualitatively similar to that observed under the strategy method: cooperators' punishment of defectors is by far the most important sanctioning category, but defectors also impose

considerable sanctions on other players. However, there is a striking quantitative difference in the severity of the sanctions cooperators impose on defectors: the cooperators' sanctions are almost twice as high under the specific response method as under the strategy method – a highly significant difference (Mann Whitney test, $p = 0.018$). The sanctions the defectors impose on other defectors under the specific response method are also higher than under the strategy method, but this difference is not significant (Mann Whitney test, $p = 0.293$). Furthermore, there are no significant differences across elicitation methods with regard to the remaining sanctioning categories ($C \rightarrow C$, $D \rightarrow C$).

Table II provides further information on the impact of the different elicitation methods (numbers in parentheses correspond to the specific response method). The last column of the table shows that the percentages of cooperators who punish is similar under both methods, namely roughly 70 percent for cooperators who punish defectors and 7 percent for cooperators who punish other cooperators. This indicates that the strong increase in the severity of the cooperators' sanctions on defectors under the specific response method is not driven by a change in the share of cooperators who punish. Instead, those cooperators who punish impose stronger sanctions.

The percentage of defectors who punish other defectors is 50 percent under the specific response method and 40.4 percent under the strategy method. The corresponding percentages for defectors who punish cooperators are 17.4 percent and 34.0 percent, respectively. Thus, the percentage of defectors who punish cooperators is lower under the specific response method but the severity of the sanctions in this punishment category does not differ from that observed under the strategy method (compare the bars for $D \rightarrow C$ in Figures 1a and 1c).

Taken together, the evidence in this section suggests that – despite some important quantitative differences – the specific response method generates similar qualitative

punishment patterns to the strategy method: (i) cooperators almost exclusively punish defectors; (ii) a significant fraction of the defectors punishes other players; (iii) cooperators impose by far the strongest sanctions on defectors.

3. 4. *The Relevance of different Fairness Principles*

The results of the previous sections suggest that the fairness motive is the central motive behind informal sanctions, although spiteful sanctions were also surprisingly frequent. An understanding of the nature of the fairness principles that drive the sanctions is thus important. We turn to this question next by discussing two fairness principles. The first principle relates to the question whether the other players' *individual* payoffs provide the empirically correct input for a player's fairness preferences or whether a summary indicator of the *group's* overall payoff such as the group's average or total payoff, constitutes the correct basis for a player's fairness preferences. The second principle relates to the question whether fairness driven sanctioning is motivated by the desire to retaliate, i.e., because the punishers want to harm those who committed unfair acts or whether the punishers want to minimize payoff inequities. These two questions are closely related to the different fairness models. In several fairness theories (e.g., Levine 1998, Fehr and Schmidt 1999, Falk and Fischbacher forthcoming), a player's own material payoff and the other players' *individual* material payoffs provide the basis for a players' fairness preference. This contrasts with the approach by Bolton and Ockenfels (2000) which assumes that the comparison between a player's own material payoff and the group's total payoff affects a player's preferences. A key feature of this approach is that the other players' individual material payoffs play no role for the construction of fairness preferences.[5] The answer to the second question separates the inequity aversion models (Fehr and Schmidt 1999, Bolton and Ockenfels 2000), which assume that players aim to minimize

unfair payoff inequalities, from fairness models that are based on the retaliation motive (Rabin 1993, Dufwenberg and Kirchsteiger 2004, Falk and Fischbacher forthcoming).

The pattern of punishment in our high sanction treatment can be used to answer the first question. If punishment is solely driven by a player's comparison of his own payoff with the total or the average payoff of the group, respectively, the player is indifferent between punishing a cooperator and punishing a defector if the costs of punishing them are identical. The reason is that both the player's own material payoff as well as the group's total or average payoff is affected in the same way regardless of whether a cooperator or a defector is punished. Thus, if the costs of punishing cooperators and defectors are identical, the approach by Bolton and Ockenfels predicts that cooperators and defectors are equally likely to be punished because the punishment target will be more or less determined randomly. Moreover, if punishing cooperators is cheaper than doing so to defectors, as is the case in our high sanction treatment, a punisher will always prefer to punish the cooperators because this is the cheaper way of affecting the total payoff. The data in our high sanction treatment sharply contradicts this prediction, however. In fact, the cooperators' sanctions were predominantly imposed on the defectors. 45 of the 73 cooperators punished the defectors exclusively, and 5 punished the other cooperators as well as the defectors. In addition, the defectors sanctioned cooperators and other defectors at the same rate although Bolton and Ockenfels predict preferential punishment of cooperators (see Figure 1a). Taken together, these data suggest that fairness models that disregard the other players' individual payoffs and rely, instead, on a comparison with the group's total or average payoff, respectively, fail to explain decisive aspects of informal sanctions.

To answer the second question raised above, recall that a cooperator cannot reduce the payoff inequalities in the low sanction treatment, regardless of whether the inequalities are measured in terms of individual payoff differences, as in Fehr and Schmidt (1999), or whether

they are measured in terms of the deviation of a player's relative payoff share from the fair relative share, as in Bolton and Ockenfels (2000). Thus, theories of inequality aversion cannot explain why cooperators punish in the low sanction condition. Despite this, 59.6 percent of the cooperators punished the defectors (see Table II). This suggests that the *desire to retaliate,* instead of the motive of reducing unfair payoff inequalities, seems to be the driving force of these sanctions. It is also remarkable that those cooperators who punish in the low-sanction condition impose on average almost the same payoff-reduction (8.0 tokens) as the cooperators who punish in the high-sanction condition (8.3 tokens). The small difference across conditions is not significant (Mann Whitney test, $p = 0.395$). Note that this means that a punishing cooperator in the low-sanction condition spends roughly 2.5 times more money on the punishment of the defectors than does a punishing cooperator in the high-sanction condition. Thus, the retaliation motive behind the punishment of defectors in the low-sanction condition seems to be rather strong.

4. CONCLUSIONS

The willingness to sanction norm violations and non-cooperative behavior is crucial for the maintenance of social order. Such sanctions sustain the viability of a myriad of informal agreements in markets, organizations, families, and neighborhoods. In this paper, we examined the sanctioning motives in the context of a cooperation problem.

Our findings indicate that the most important category of sanctions – in terms of the percentage of individuals involved and in terms of the severity of the sanctions – are those which cooperators impose on defectors. We also find an unexpectedly large fraction of individuals who defect but nevertheless punish other group members. This spiteful punishment is somewhat reduced if the specific response method is used to determine sanctioning behavior. Interestingly, spiteful punishment by defectors vanishes completely if the defectors can no

longer increase the payoff difference between themselves and the punished individual. This result contrasts sharply with the punishment patterns of the cooperators; the latter strongly increase their expenditures for punishment if the impact of a given investment in punishment causes a lower payoff reduction for the punished individual. Our results suggest two principles that should be taken seriously in theories of fairness. First, fairness preferences should be based on individualized payoff information. Second, they should take the motive to retaliate seriously. These principles are suggested by the failure of fairness theories that neglect individual payoff information and focus exclusively on the relation of a player's material payoff to the group's total or average payoff to explain the predominant form of punishment – that of cooperators punishing defectors. In addition, fairness theories that are exclusively based on the idea that players want to minimize payoff inequalities fail to explain a considerable share of the cooperators' sanctions.

Institute for the Study of Labour (IZA), P.O. Box 7240, 53072 Germany, falk@iza.org, http://www.iza.org/home/falk

and

Institute for Empirical Research in Economics, University of Zürich, Blümlisalpstrasse 10, CH-8006 Zürich, fehr@iew.unizh.ch, http://www.iew.unizh.ch/home/fehr/; fiba@iew.unizh.ch, http://www.iew.unizh.ch/home/fischbacher/.

REFERENCES

BOLTON, G. E., AND A. OCKENFELS (2000): "A Theory of Equity, Reciprocity and Competition," *American Economic Review*, XC, 166-193.

BRANDTS, J., AND G. CHARNESS (2000): "Hot versus Cold: Sequential Responses and Preference Stability in Experimental Games," *Experimental Economics*, 2, 227-238.

CASON, T., AND V. MUI (1998): "Social Influence in the Sequential Dictator Game," *Journal of Mathematical Psychology*, 42, 248-265.

DUFWENBERG, M., AND G. KIRCHSTEIGER (2004): "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, 47, 268-298.

FALK, A., AND U. FISCHBACHER (forthcoming): "A Theory of Reciprocity", Games and Economic Behavior, forthcoming.

FEHR, E., AND S. GÄCHTER (2000): "Cooperation and Punishment in Public Goods Experiments", *American Economic Review* 90, 980-994.

FEHR, E. AND K. SCHMIDT (1999): "A Theory of Fairness, Competition and Cooperation", *Quarterly Journal of Economics*, CXIV, 817-851.

FISCHBACHER, U. (1999): "Z-tree. Zurich Toolbox for Readymade Economic Experiments – Experimenter's Manual" Working Paper No. 21, Institute for Empirical Research in Economics, University of Zurich.

FRANCIS, H. (1985): "The Law, Oral Tradition and the Mining Community", *Journal of Law and Society,* 12, 267-271.

GÜTH, W., R. SCHMITTBERGER, AND B. SCHWARZE (1982): "An Experimental Analysis of Ultimatum Bargaining", *Journal of Economic Behavior and Organization,* 3, 367-88.

GÜTH, W., AND E. VAN DAMME (1998): "Information, Strategic Behavior and Fairness in Ultimatum Bargaining: An Experimental Study", *Journal of Mathematical Psychology,* 42, 227-247.

HECHTER, M., AND K. D. OPP (2001): *Social Norms*. New York: Russell Sage Foundation.

KAGEL, J., AND K. WOLFE (2001): "Tests of Fairness Models based on Equity Considerations in a Three-Person Ultimatum Game," *Experimental Economics*, 4, 213 – 219.

KIRCHSTEIGER, G. (1994): "The Role of Envy in Ultimatum Games", *Journal of Economic Behavior and Organization*, 25, 373-389.

KNEZ, M., AND D. SIMESTER (2001): "Firm-wide Incentives and Mutual Monitoring at Continental Airlines", *Journal of Labor Economics,* 19:4, 743-772.

LEVINE, D. K. (1998): "Modeling Altruism and Spitefulness in Experiments", *Review of Economic Dynamics,* 1, 593-622.

MUI, V. (1995): "The Economics of Envy", *Journal of Economic Behavior and Organization*, 26, 311-336.

OSTROM, E. (1990): *Governing the Commons – The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

RABIN, M. (1993): "Incorporating Fairness into Game Theory and Economics", *American Economic Review,* 83, 1281-1302.

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

FOOTNOTES

---

[1] This paper is part of the Research Priority Program of the University of Zurich on "The Foundations of Human Social Behavior: Altruism versus Egoism". Financial support by the EU Research Network ENABLE is also gratefully acknowledged. We also want to thank seminar participants in the Behavioral Economics Seminar at Harvard University, at Pompeu Fabra University in Barcelona, and at the MacArthur Preferences Network Meeting for helpful comments and discussions.

[2] Exceptions are the experiments by Güth and van Damme (1998) and by Kagel and Wolfe (2001), where three players participated in each of these experiments. However, the third player was completely passive. He could not make any choice, but the actions of the other two players affected his payoff. In our three person experiments all players have a non-trivial strategy set. In particular, each player can sanction any other player.

[3] The instructions for the low sanction treatment can be found in a discussion paper version of this paper (http://www.iew.unizh.ch/wp/iewwp059.pdf). The same basic instructions, suitably modified to capture the variations across conditions, were used in all other treatments.

[4] An exception is the model by Levine (1998), which explicitly models spiteful types and altruistic types in the same framework.

[5] Formally, Bolton and Ockenfels assume that a player's utility function is given by $U_i(\pi_1,..., \pi_n) = U_i(\pi_i, \sigma_i)$. $(\pi_1,..., \pi_n)$ denotes the vector of material payoffs, n the number of players and the relative share, σ_i, is defined as $\pi_i/(n\pi_a)$ where π_a denotes the average payoff in the group and $n\pi_a$ the group's total payoff. U_i is nondecreasing in π_i, increasing in σ_i if $\sigma_i < 1/n$ and decreasing in σ_i if $\sigma_i > 1/n$.

TABLE I

PAYOFFS TO PLAYER i IN THE 3-PLAYER PRISONER'S DILEMMA

|	Both other players defect	One of the other two players cooperates	Both other players cooperate
Player i defects	20	32	44
Player i cooperates	12	24	36

TABLE II

PERCENTAGE OF COOPERATORS AND DEFECTORS WHO PUNISH IN THE HIGH-SANCTION AND THE LOW-SANCTION CONDITION

	Sanctioning subject is a			
	defector		cooperator	
Sanctioned subject is a defector	low-sanc. cond. 2.2	high-sanc. cond. 40.4 (50.0)	low-sanc. cond. 59.6	high-sanc. cond. 68.5 (70.3)
Sanctioned subject is a cooperator	low-sanc. cond. 0.0	high-sanc. cond. 34.0 (17.4)	low-sanc. cond. 0.0	high-sanc. cond. 6.8 (6.6)

Remark: Numbers in parentheses relate to the control treatment with the specific response method. All other numbers relate to the high or the low sanction treatment with the strategy method. 73 subjects cooperate and 47 defect in the high sanction condition with the strategy method whereas in the low sanction conditions 47 subjects cooperate and 46 defect. In the control treatment with the specific response method 79 subjects cooperate and 23 defect.

FIGURE 1 – AVERAGE PAYOFF REDUCTION ACROSS DIFFERENT TREATMENTS
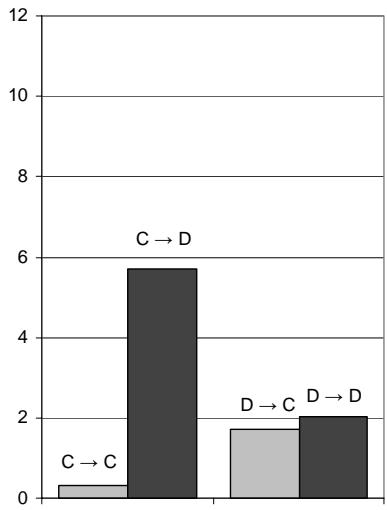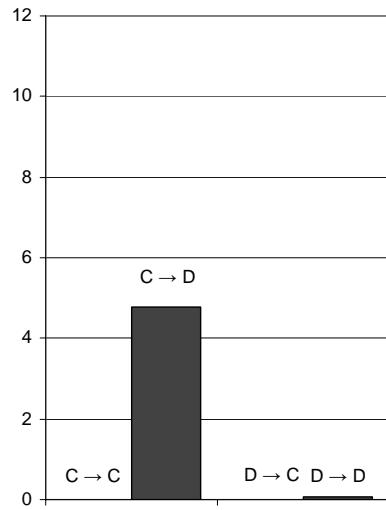
FIGURE 1a	FIGURE 1b	FIGURE 1c
HIGH SANCTION TREATMENT (STRATEGY METHOD)	LOW SANCTION TREATMENT (STRATEGY METHOD)	HIGH SANCTION TREATMENT (SPECIFIC RESPONSE METHOD)
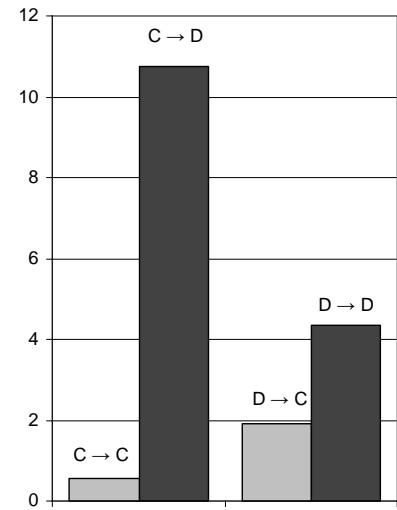
Remark: C → C represents the situation in which a cooperator faces another cooperator, C → D is the situation in which a cooperator faces a defector; D → C and D → D can be interpreted analogously.
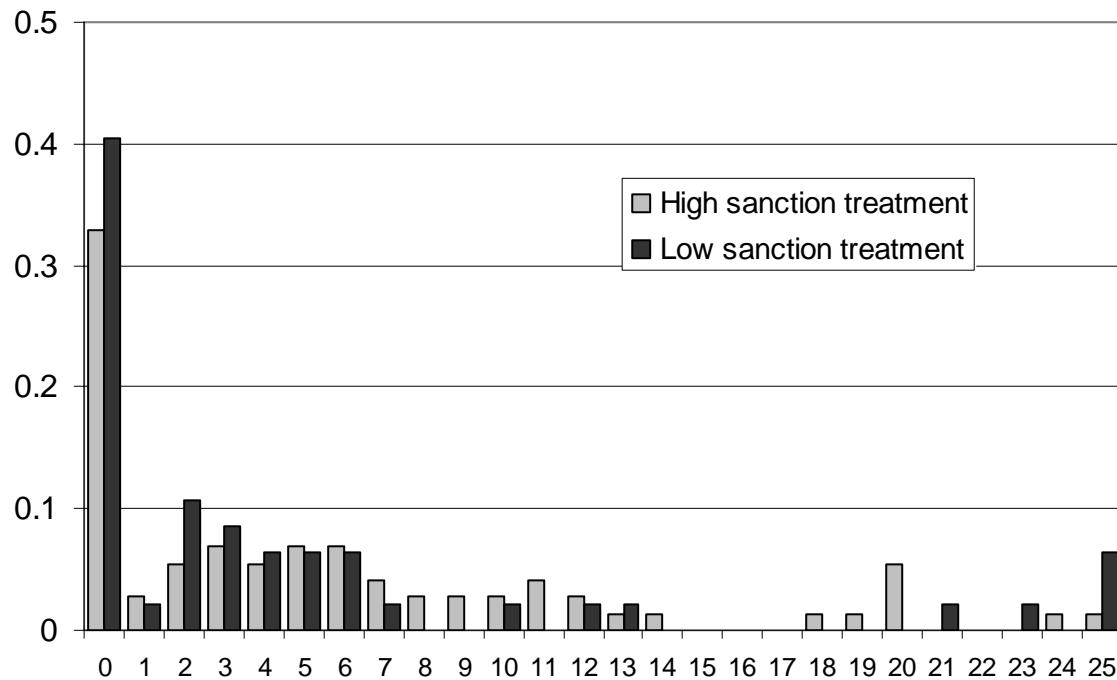
FIGURE 2 – RELATIVE FREQUENCY OF AVERAGE PAYOFF REDUCTIONS THAT ARE IMPOSED BY INDIVIDUAL COOPERATORS ON DEFECTORS

FIGURE 1 – AVERAGE PAYOFF REDUCTION ACROSS DIFFERENT TREATMENTS

FIGURE 2 – RELATIVE FREQUENCY OF AVERAGE PAYOFF REDUCTIONS THAT ARE IMPOSED BY

INDIVIDUAL COOPERATORS ON DEFECTORS