Strong Reciprocity, Human Cooperation and the Enforcement of Social Norms[1]

Ernst Fehr
University of Zürich

Urs Fischbacher
University of Zürich

Simon Gächter
University of St. Gallen

Published in: HUMAN NATURE 13(2002): 1-25

Abstract: This paper provides strong evidence challenging the self-interest assumption that dominates the behavioral sciences and much evolutionary thinking. The evidence indicates that many people have a tendency to voluntarily cooperate, if treated fairly, and to punish non-cooperators. We call this behavioral propensity 'strong reciprocity' and show empirically that it can lead to almost universal cooperation in circumstances in which purely self-interested behavior would cause a complete breakdown of cooperation. In addition, we show that people are willing to punish those who behaved unfairly towards a third person or who defected in a Prisoner's Dilemma game with a third person. This suggests that strong reciprocity is a powerful device for the enforcement of social norms like, e.g., food-sharing norms or collective action norms. Strong Reciprocity cannot be rationalized as an adaptive trait by the leading evolutionary theories of human cooperation, i.e., by kin selection theory, reciprocal altruism theory, indirect reciprocity theory and costly signaling theory. However, multi-level selection theories and theories of cultural evolution are consistent with strong reciprocity.

Keywords: Strong Reciprocity, Punishment, Evolution, Human Cooperation, Social Norms

[1]This paper is part of a research project on strong reciprocity financed by the Network on Economic Environments and the Evolution of Individual Preferences and Social Norms of the MacArthur Foundation.

Ernst Fehr is Professor in Economics at the University of Zürich in Switzerland. He is on the editorial board of the Quarterly Journal of Economics, Games and Economic Behavior, the European Economic Review, the Journal of Socio-Economics and Experimental Economics. Fehr studies the interplay between social preferences, social norms and strategic interactions.

Urs Fischbacher has a position at the Institute for Empirical Research in Economics at the University of Zurich. He received his PhD in Mathematics at the University of Zurich. His research focuses on social preferences, the economics of social interactions and game theoretic models of strong reciprocity.

Simon Gaechter is a Professor in Economics at the University of St. Gallen in Switzerland. His main research is on problems of incentive systems, contract enforcement, voluntary cooperation, and social norms.

1. Introduction

A key fact about human society is the ubiquity of material incentives to cheat on implicit or explicit cooperative agreements. In any kind of social or economic exchange situation between two or more individuals in which not all aspects of the exchange are determined by enforceable contracts, there are material incentives to cheat the exchange partners. Even in modern human societies with a large cooperative infrastructure in the form of laws, impartial courts and the police, the material incentive to cheat on cooperative agreements is probably the rule rather than the exception. This is so because, in general, not all obligations that arise in the various contingencies of exchange situations can be unambiguously formulated and subjected to a binding contract.[2] Therefore, by reneging on the implicit or unenforceable obligations a party can always improve its material payoff relative to a situation where it meets its obligations. Of course, in pre-modern societies that lack a cooperative infrastructure, cheating incentives are even more prevalent. It is probably true that in more than 90 percent of human history no cooperative infrastructure, as mentioned above, existed.

Another key fact about human society is that despite these incentives to cheat many "nonbinding" agreements among non-kin occur and are kept. Since cooperation regularly also takes place among non-kin, genetical kinship theory (Hamilton 1964) cannot readily account for this fact. One possibility to account for the manifest cooperation among non-kin is to recognize that many social interactions take place repeatedly. Evolutionary theorists, for example, have shown that natural selection can favor reciprocally cooperative behavior in bilateral interactions when the chances to interact repeatedly with the same individual in the future are sufficiently high (Trivers 1971; Axelrod and Hamilton 1981). Since cheating, i.e., not reciprocating a cooperative act, can be deterred by the withdrawal of future cooperation, it is in the long run interest of organisms not to cheat. Therefore, in bilateral repeated interactions reciprocal cooperation can be an evolutionary stable outcome. In a similar spirit, game theorists have shown that, when the chances for repeated interactions are sufficiently high, rational egoists, i.e., rational actors that are solely interested in their own material well-being, can establish an equilibrium with full

[2] There is a large economic literature that tries to provide microfoundations for the existence of incomplete contracts. The empirical fact that many agreements contain an element of incompleteness is, however, undisputed. A prominent example is, of course, the labor contract.

cooperation despite the existence of short-run cheating incentives (Friedman 1971, Fudenberg and Maskin 1986). The reason for this is that cheating has not only short run benefits but may also have long run costs depending on the nature of the equilibrium that is played. If the players play a cooperative equilibrium the implicit or explicit threat to withdraw future cooperation from the cheaters deters cheating and, as a consequence, cooperation can be sustained by self-interested rational actors in this equilibrium. However, in multilateral n-person interactions, which are typical for human societies, the prospects for sustaining cooperation in an evolutionary equilibrium by individual threats of withdrawing future cooperation are quite limited. Boyd and Richerson (1988) have shown that for reasonable group sizes this mechanism for sustaining cooperation does not work.

In this paper we will provide strong evidence in favor of another, distinct, cooperation-enhancing force that has so far been largely neglected. We call this force '**strong reciprocity'** (see also Gintis 2000; Bowles and Gintis 2001). A person is a strong reciprocator if she is willing (i) to sacrifice resources to be kind to those who are being kind (= strong positive reciprocity) and (ii) to sacrifice resources to punish those who are being unkind (= strong negative reciprocity). The essential feature of strong reciprocity is a willingness to sacrifice resources for rewarding fair and punishing unfair behavior *even if this is costly and provides neither present nor future material rewards for the reciprocator*. Whether an action is perceived as fair or unfair depends on the distributional consequences of the action relative to a neutral reference action (Rabin 1993). We will show that there exist many people who exhibit strong reciprocity and whose existence greatly improves the prospects for cooperation. We provide evidence that strong reciprocity can give rise to almost maximal cooperation in circumstances in which the standard repeated interaction approach predicts no cooperation at all. However, we also provide evidence indicating that there are social structures in which the interaction between reciprocally fair persons and purely selfish persons induces the majority of people to cheat. This highlights the importance of social structures for the achievement of stable cooperation. In addition, we show that strong reciprocity also is a powerful norm enforcement device. Therefore, strong reciprocity may help explain the enforcement of food-sharing norms and norms that prescribe participation in collective actions.

It is important to distinguish strong reciprocity from terms like 'reciprocal altruism' and 'altruism'. An altruistic actor is unconditionally kind, i.e. the kindness of her behavior does not depend on the other actors' behavior. A reciprocally altruistic actor, in contrast, conditions her behavior on the previous behavior of the other actor. Yet, while a reciprocally altruistic actor is willing to help another actor although this involves short run costs, she does this only because she expects long-term net benefits. The distinction between strong reciprocity, altruism and reciprocal altruism can most easily be illustrated in the context of a *sequential* Prisoners' Dilemma (PD) that is played *exactly once*. In a sequential PD player A first decides whether to defect or to cooperate. Then player B observes player A's action after which she decides to defect or to cooperate. To be specific, let the material payoffs for (A, B) be (5, 5) if both cooperate, (2, 2) if both defect, (0, 7) if A cooperates and B defects and (7, 0) if A defects and B cooperates. If player B is an altruist she never defects even if player A defected. Altruism, as we define it here, is thus tantamount to *unconditional* kindness. In contrast, if player B is a strong reciprocator she defects if A defected and cooperates if A cooperated because she is willing to sacrifice resources to reward a behavior that is perceived as kind. A cooperative act by player A, despite the material incentive to cheat, is a prime example of such kindness. The kindness of a strong reciprocator is thus *conditional* on the perceived kindness of the other player. Reciprocal altruism, as it is frequently used in evolutionary biology, also differs fundamentally from strong reciprocity because a reciprocal altruist only cooperates if there are future returns from cooperation. Thus a reciprocally altruistic player B will always defect in a sequential *one-shot* PD. Since a reciprocal altruist performs altruistic actions only if the total material returns exceed the total material costs we do not use this term in the rest of the paper. Instead, we use the term 'selfish' for this motivation.

II. Evolutionary Theory and Strong Reciprocity

The fact that many people retaliate or cooperate when it is costly and provides no material rewards raises the question of why people behave in this way. We believe that the answer to this question ultimately must be sought in the evolutionary conditions of the human species that caused a propensity for strongly reciprocal behavior among a significant fraction of the population. Our evidence suggests that strong reciprocity cannot be explained by the motives that are invoked by the major prevailing evolutionary theories of altruism and cooperation. We will

argue that our evidence is incompatible with the motives invoked by kin selection theory (Hamilton 1964), by reciprocal altruism theory (Trivers 1971; Axelrod and Hamilton 1981), by the theory of indirect reciprocity (Alexander 1987; Nowak and Sigmund 1998) and by costly signaling theory (Zahavi and Zahavi 1997; Gintis, Smith, Bowles 2001). The puzzling question to be solved by evolutionary theory is, therefore, how strong reciprocity could survive in human evolution. This question is important because – as our experiments show – the presence of strong reciprocators greatly increases and stabilizes human cooperation.

Since the claim that all major evolutionary theories of altruism and cooperation cannot account for strong reciprocity is quite provocative, it is worthwhile to prevent any misunderstanding regarding what we mean by this claim. Our argument is that the observed experimental behaviors cannot be rationalized as adaptive behaviors by these evolutionary models. This means that if one assumes the conditions of the experiment (in particular, that strangers interact anonymously with each other in one shot situations), these theories predict that strongly reciprocal behavior cannot prevail in an evolutionary *equilibrium*. Or put differently, from the viewpoint of these theories the observed experimental behaviors and the underlying motives must be classified as maladaptive. In view of the robustness and the frequency of strong reciprocity across different cultures (see, e.g., Henrich et al. 2001) this seems quite unsatisfactory. A particularly unsatisfactory maladaption story is the argument that experimental subjects are not capable of distinguishing between repeated interactions and one-shot interactions. As a consequence, so the story goes, subjects tend to inappropriately apply heuristics and habits in experimental one-shot interactions (i.e., they take revenge or they reward helping behavior), that are only adaptive in a repeated interaction context but not in the one-shot context. In the final part of the paper, where we discuss the proximate mechanisms behind strong reciprocity, we show that this argument is refuted by the data. The evidence suggests that subjects are well aware of the difference between one-shot and repeated interactions because they behave quite differently in these two conditions.

Recently, Gintis (2000) developed an evolutionary model of strong reciprocity.[3] His model is based on the plausible idea that in the relevant evolutionary environment human groups faced

[3] For different evolutionary models of strong reciprocity see Bowles and Gintis (2001) and Sethi and Somananthan (2001a, 2001b).

extinction threats (wars, famines, environmental catastrophes) with a positive probability. When groups face such extinction threats neither reciprocal altruism nor indirect reciprocity can sustain the necessary cooperation that helps the groups to survive the situation because the shadow of the future is too weak. Kin-selection also does not work here because in most human groups membership is not restricted to relatives but is also open to non-kin members. However, groups with disproportionately many strong reciprocators are much better able to survive these threats. Hence, within-group selection creates evolutionary pressures against strong reciprocity because strong reciprocators engage in individually costly behaviors that benefit the whole group. In contrast, between-group selection favors strong reciprocity because groups with disproportionately many strong reciprocators are better able to survive. The consequence of these two evolutionary forces is that in equilibrium strong reciprocators and purely selfish humans coexist. Another model that is capable of explaining punishment in one-shot situations is the cultural evolutionary model of Henrich and Boyd (2001) that is based on the notion of conformist transmission. Henrich and Boyd show that only a very small amount of conformist transmission can stabilize one-shot punishments in an evolutionary equilibrium. It is also worthwhile to point out that multi-level selection theories of altruism, as the one by Sober and Wilson (1998), are compatible with strong reciprocity.

III. Evidence and Consequences of Strong Reciprocity

III.1. The Enforcement of 'Nonbinding' Agreements

In many bilateral one-shot encounters in real life people behave in a reciprocally fair way. A good example is the exchange between a taxi driver and his passenger in a big city (Basu 1984). On the surface this example represents a trivial economic exchange. A closer look, however, shows that this exchange frequently is similar to a sequentially played one-shot PD because the probability of repeated interactions is extremely low and the taxi driver first has to decide whether to cooperate (drive) or not. Ex post, once the taxi driver has provided his service, it would frequently be very easy for the passenger to escape without payment unless the passenger expects that the taxi driver incurs the cost of chasing him. Yet, for the taxi driver the cost of chasing a non-paying passenger are very often much higher than the returns. In these situations a selfish taxi driver will never

chase the passenger. Thus, if all taxi drivers were purely selfish, passengers could often escape without paying the bill. This example shows that even in seemingly trivial exchanges the enforcement of the obligation of at least one party is often not guaranteed so that the contract is in an important sense incomplete. However, despite the incentive to leave without paying, most passengers reciprocate the service by paying the bill. In our view there are two major reasons for this: (i) Many people are indeed inclined to exhibit strong positive reciprocity, i.e., they pay the bill even if they could escape at low cost without paying. (ii) Many taxi drivers would be extremely upset if the passenger tried to go away without paying and, as a consequence, they would be willing to bear considerable costs to catch and punish the cheater. Therefore, if potential cheaters anticipate this, most of them are probably better off by not cheating. This example indicates how the combined effects of strong positive and negative reciprocity contributes to the enforcement of sequential exchanges that are beneficial for both parties.

In addition to real world examples, there also is evidence for strong reciprocity between anonymously interacting trading partners from well-controlled laboratory experiments (Fehr, Kirchsteiger and Riedl 1993 and 1998; Berg, Dickhaut and McCabe 1995).[4] In these experiments subjects could earn real money according to their decisions and the rules of the experiment. To preserve the one-shot character of the experiment subjects were never informed about the identity of their exchange partner. Moreover, experimental procedures also ensured that no individual subject could ever gain a reputation for being, for example, cooperative. Exchange partners were located in different rooms. These features of the experiment ensured that the exchange really took place between anonymous strangers. In the following we illustrate the regularities of strongly reciprocal behavior on the basis of one of these experiments conducted at the University of Zurich (Fehr, Gächter and Kirchsteiger 1997).

In the experiment a subject in the role of an employer (or buyer) can make a job offer (or contract offer for one unit of a good) to the group of subjects in the role of workers (or sellers). Each worker can potentially accept the offer. There are more workers than employers to induce competition among the workers. A job offer consists of a *binding* wage offer (or price offer) w and a *nonbinding* 'desired effort level' (or 'desired quality level') $\hat{e}$. If one of the workers accepts

an offer ($w,\hat{e}$) she has to determine the *actual* effort level e. In the experiment the choice of an effort level is represented by the choice of a number. The higher the chosen number the higher is the effort and the higher are the monetary effort costs borne by the worker. The desired and the actual effort levels have to be in the interval $[e_{min}, e_{max}] \equiv [0.1, 1]$ and the wage offer has to be in the interval $[0, 100]$. The higher e the larger is the material payoff for the employer but the higher are also the costs $c(e)$ of providing e for the worker. Material payoffs from an exchange are given by $\Pi_f = 100e - w$ for the employer and $\Pi_w = w - c(e)$ for the worker. A party who does not manage to trade earns zero. The effort costs are given by the following table.

Table 1: *effort levels and costs of effort*

effort e	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
costs of effort c(e)	0	1	2	4	6	8	10	12	15	18

Note that since $\hat{e}$ is non-binding the worker can choose any e in the interval $[0.1, 1]$, i.e., in particular $e < \hat{e}$, without being sanctioned. It is obvious that, since $c(e)$ is strictly increasing in e, a selfish worker will always choose $e = e_{min} = 0.1$. Therefore, a rational and selfish employer, who believes that there are only selfish workers, will never offer a wage above $w = 1$. This is so because the employer knows that the workers will incur no effort costs and, being selfish, will accept a wage offer of $w = 1$. At $w = 1$ the trading worker earns $\Pi_w = 1$ which is more than if the worker does not trade.

In sharp contrast to the predictions based on the selfishness assumption we observe the following regularities: The vast majority of contract offers imply, if accepted and met, a much larger profit than $\Pi_w = 1$ for the worker. On average the profit implied by the offer, which is defined by $w - c(\hat{e})$, is 35 units. Moreover, the higher $\hat{e}$ the higher is the profit $w - c(\hat{e})$ offered to the worker (see Figure 1). Note that this means that employers do not just compensate the workers for the higher effort costs but that they increase the *profit* of the workers if they desire a higher effort level, i.e., they share the increase in the total returns that are created by higher effort

[4] In all experiments discussed in this paper completely anonymous strangers, who never learn the identities of their interaction partners, interact with each other. It is also not possible to acquire individual reputations for behaving in

levels. This indicates that employers appeal to the strong reciprocity of the workers by being more generous the more costly the desired effort choice for the worker is. Workers, in turn, exhibit a considerable amount of strong reciprocity (see Figure 1). Although the actual average effort is below the desired average effort it is, in general, clearly above e_{min}. Moreover, there is a strong positive correlation between the generosity of the offer, i.e., the level of $w - c(\hat{e})$, and the actual average effort.

Insert Figure 1 here

The results depicted in Figure 1 indicate that *on average* the subjects have a propensity for strong positive reciprocity. However, the average can be somewhat misleading because it hides the presence of purely selfish subjects. In addition to the studies cited above there are by now also several other studies indicating that subjects are heterogeneous with regard to their propensity to exhibit strong reciprocity (Bolle 1998; Fehr and Falk 1999; McCabe, Rassenti and Smith 1998; Charness 2000; McCabe, Rigdon and Smith 2000; Abbink, Irlenbusch and Renner 2000; Gächter and Falk 2001). As in the above experiment, subjects in these studies always interact anonymously with each other and reciprocal behavior is costly in terms of real money for the reciprocator. Taken together, the fraction of subjects showing strong positive reciprocity is rarely below 40 and sometimes 60 percent whereas the fraction of selfish subjects is also often between 40 and 60 percent. Moreover, these frequencies of strong positive reciprocity are observed in such diverse countries as Austria, Germany, Hungary, the Netherlands, Switzerland, Russia and the U.S. It is also worthwhile to stress that *strong positive reciprocity is not diminished if the monetary stake size is rather high.* In the experiments conducted by Fehr and Tougareva (1995) in Moscow subjects earned on average the monetary income of ten weeks in an experiment that lasted for two hours. The monthly median income of subjects was US $17 while in the experiment they earned on average US $45. The impact of strong reciprocity also does not vanish if the experimental design ensures that the experimenter cannot observe individual decisions but only aggregate decisions (Bolton and Zwick 1995; Berg, Dickhaut, and McCabe 1995; Abbink, Irlenbusch and Renner 2000).

---

particular ways in these experiments.

In an extension of the above experiment we examined the impact of giving the employers the option of responding reciprocally to the worker's choice of *e*. We achieved this by giving the employer the opportunity to reward or punish the worker after she observes the actual effort.[5] By spending one experimental dollar on reward the employer could *increase* the worker's payoff by 2.5 experimental dollars, and by spending one dollar on punishment the employer could *decrease* the worker's payoff by 2.5 dollars. Employers could spend up to 10 experimental dollars on punishment or on rewarding their worker. The important feature of this design is that if there are only selfish employers they will never reward or punish a worker because both rewarding and punishing is costly for the employer. Therefore, in case that there are only selfish employers there is no reason why the opportunity for rewarding/punishing workers should affect workers' effort choice relative to the situation where no such opportunity exists. However, if a worker expects her employer to be a strong reciprocator it is likely he will provide higher effort levels in the presence of a reward/punishment opportunity. This is so because strongly reciprocal employers are likely to reward the provision ($e = \hat{e}$) or the overprovision ($e > \hat{e}$) of effort and to punish the underprovision ($e < \hat{e}$). This is in fact exactly what one observes, on average. If there is underprovision of effort employers punish in 68 percent of the cases and the average investment into punishment is 7 dollars. If there is overprovision employers reward in 70 percent of these cases and the average investment into rewarding is also 7 dollars. If workers exactly meet the desired effort employers still reward in 41 percent of the cases and the average investment into rewarding is 4.5 dollars.

We also elicited workers' expectations about the reward and punishment choices of their employers. Hence, we are able to check whether workers anticipate employers' strong reciprocity. It turns out that in case of underprovision workers expect to be punished in 54 percent of the cases and the expected average investment into punishment is 4 dollars. In case of overprovision they expect to receive a reward in 98 percent of the cases with an expected average investment of 6.5 dollars. As a result of these expectations workers choose much higher effort levels when employers have a reward/punishment opportunity. The presence of this opportunity decreases underprovision from 83 percent to 26 percent of the trades, increases exact provision of $\hat{e}$ from 14

[5] It is important to stress that in the experimental instructions the terms „rewarding" and „punishing" have never been used. The instructions were framed in neutral terms to avoid experimenter demand effects. The same holds for other experiments discussed in this paper.

to 36 percent and increases overprovision from 3 to 38 percent of the trades. The average effort level is increased from $e = 0.37$ to $e = 0.65$ so that the gap between desired and actual effort levels almost vanishes. An important consequence of this increase in average effort is that the aggregate monetary payoff increases by 40 percent - even if one takes the payoff reductions that result from actual punishments into account. Thus, the reward/punishment opportunity increases the total pie that becomes available for the trading parties considerably.

The evidence presented above confirms that strong reciprocity substantially contributes to the enforcement of cooperative agreements in bilateral sequential exchanges. The power of strong reciprocity derives from the fact that it provides incentives for the potential cheaters to behave cooperatively or to limit at least their degree of non-cooperation. In the above experiments, for example, even purely selfish employers have an incentive to make a cooperative first move, i.e., to make a generous job offer, if they expect sufficiently many workers to behave in a strongly reciprocal manner. Similarly, even purely selfish workers have an incentive to provide a high effort in case of a reward/punishment opportunity if they expect sufficiently many employers to be strong reciprocators.

Note that the strongly reciprocal behavior in the previous experiments cannot be explained by the major prevailing theories of altruism and cooperation. Since subjects know that they are strangers to each other kin selection theory does not apply. Since the interaction is one-shot, there are no future material rewards from present cooperation or retaliation so that reciprocal altruism theory does not apply either. Since subjects interact anonymously with each other and can, hence, develop no individual reputation for being cooperative or retaliatory, the theory of indirect reciprocity does not apply. Finally, anonymity also ensures that cooperation cannot be used as a costly signal for unobservable traits, i.e., costly signaling theory does not apply either. It is worthwhile to stress that the same arguments can be made with regard to the other experiments discussed in this paper. Because in the experiments strangers interacted just once and anonymously with each other the major prevailing evolutionary theories cannot rationalize the observed experimental behaviors.

III.2. Punishment in Bilateral Bargaining Situations

There are many real life examples of the desire to take revenge and to retaliate in response to harmful and unfair acts. One important example is that people frequently break off bargaining with opponents that try to squeeze them. This can be nicely illustrated by so-called ultimatum bargaining experiments (Güth, Schmittberger and Schwarze 1982; Camerer and Thaler 1995; Roth 1995). In the ultimatum game two subjects have to agree on the division of a fixed sum of money. Person A, the Proposer, can make exactly one proposal of how to divide the amount. Then person B, the Responder, can accept or reject the proposed division. In the case of rejection both receive nothing whereas in the case of acceptance the proposal is implemented. A robust result in this experiment is that proposals that give the Responder *positive* shares below 20 percent of the available sum are rejected with a very high probability. This shows that Responders do not behave in a self-interest maximizing manner. In general, the motive indicated for the rejection of positive, yet ”low”, offers is that subjects view them as unfair. As in the case of positive reciprocity, it is worthwhile to mention that strong negative reciprocity is observed in a wide variety of cultures, and that rather high monetary stakes do not change or have only a minor impact on these experimental results. By now there are literally hundreds of studies of one-shot ultimatum games. Rejections of positive offers are observed in Israel, Japan, many European countries, Russia, Indonesia and the US. For an early comparison across countries see Roth, Prasnikar, Okuno-Fujiwara and Zamir (1991). In the study of Cameron (1999) the amount to be divided by the Indonesian subjects represented the income of three months for them. Other studies with relatively high stakes are Hoffman, McCabe and Smith (1996) where US $ 100 had to be divided by US-students, and Slonim and Roth (1998).

III.3. Multilateral Cooperation and Punishment Opportunities

The previously discussed studies involve bilateral relations. In view of the ubiquity of n-person interactions in human evolution and everyday life an important question is, however, how people behave in n-person situations. One question that is particularly important is how the selfish types and the strongly reciprocal types affect one another in these situations. Or put differently: What are the interaction structures which enable the selfish types to induce the strong reciprocators to behave non-cooperatively and what are the structures that enable the strong reciprocators to force

or induce the selfish types to behave cooperatively? In view of the fact that strong reciprocators are willing to punish unfair behavior it seems likely that the presence or absence of punishment opportunities is crucial here. To illustrate the argument, consider the example of a *simultaneously* played *one-shot* PD, in which a purely selfish subject is matched with a strong reciprocator. If the reciprocal subject knows that she faces a selfish subject, she will defect because she knows that the selfish subject will *always* defect. Consider now a slightly different game in which both players have the opportunity to punish the other player after they could observe the other player's action. Assume further that the *punishment is costly for the punisher,* which ensures that a purely selfish subject will never punish. A cooperating strong reciprocator is, however, willing to punish a defecting subject because the defection is likely to be viewed as very unfair. Therefore, if the selfish subject anticipates that a defection will be punished she has an incentive to cooperate. This suggests that in the presence of punishment opportunities strong reciprocators can force the selfish types to cooperate whereas in the absence of punishment opportunities the selfish types induce the reciprocal types to defect, too. This argument has been generalized and rigorously proven by Fehr and Schmidt (1999). Fehr and Schmidt show that in an n-person public goods game with a heterogeneous population of players, full defection by everybody is likely to be the unique equilibrium in the game without punishment while full cooperation can be an equilibrium in the game with punishment.

The following public goods game, which is essentially a generalized n-person PD, has been used to examine the empirical validity of this conjecture (Fehr and Gächter 2000). In a group of four anonymously interacting subjects each subject is endowed with 20 tokens. Subjects decide *simultaneously* how many tokens to keep for themselves and how many tokens to invest in a common project. For each token that is privately kept a subject earns exactly one token. Yet, for each token a subject invests into the project each of the four subjects earns 0.4 tokens. Thus, the overall *private* return for investing one additional token into the project is -1 + 0.4 = -0.6 tokens while the overall social return is -1 + 4(0.4) = +0.6 tokens. This means that, irrespective of how much the other three subjects contribute to the project, it is always better for a subject to keep all tokens privately. Therefore, if all subjects are purely selfish they will all keep all their tokens privately. Yet, if all fully defect, i.e., keep all their tokens privately, each earns only 20 tokens while if all invest their total token endowment each subjects earns 0.4(80) = 32 tokens. In the *no-*

punishment condition the same group of subjects plays this game for ten periods. At the end of each period they are informed about the contributions of the other three group members. In the *punishment condition* subjects also play the above game for ten periods. In addition to their investment decision, they can also assign punishment points to each of the other group members at the end of each period after they have been informed about the others' contributions. The costs of punishment for the punisher are a convex and increasing function of the total number of punishment points assigned to the others. Each subject can assign up to ten punishment points to each of the other group members. Assigning ten points to another member costs the punisher 30 tokens, assigning no points costs the punisher nothing and assigning an intermediate amount of punishment points costs an intermediate amount of tokens. For each *received* punishment point the monetary income of the punished subject is reduced by 10 percent. A reduction of 10 percent implies, on average, an income reduction between 2 and 3 tokens. The experiment ensures that group members cannot trace the history of individual investments or individual punishments of a particular subject in a group. It is therefore impossible to gain an individual reputation for being (non)cooperative or for being a punisher.

Fehr and Gächter (2000) also conducted punishment and no-punishment conditions in which the group composition was randomly changed in each of the ten periods. In these experiments there is a large group of 24 people and in each of the ten periods new four-person groups are randomly formed. When the group composition is random in every period the probability of meeting the same group members again in future periods is very small. Moreover, even if some of the group members in later periods have already been met in one of the previous periods subjects do not know this because it is not possible to identify the other members in the group. Thus, the random group design essentially constitutes a situation in which strangers anonymously interact in a series of one-shot games.

In the random group design the predictions are quite straightforward if all subjects are selfish and are known to be selfish. Since in each period the group members essentially play a one-shot game the subjects will never punish because punishment is costly for them and yields no future benefits. While it may be possible that punishment increases the cooperation of the punished subject in future periods, for the punisher the probability of gaining from this is very low due to the low probability of meeting the punished group member again. Therefore,

punishing other subjects makes no sense for a selfish individual. Yet, if there is no punishment it also makes no sense for a selfish subject to cooperate because any form of cooperation causes a reduction in the material payoff. Thus, both in the punishment condition as well as in the no-punishment condition of the random group design no cooperation should occur if all subjects are purely selfish. However, if the selfish subjects expect the presence of strong reciprocators in the group, i.e., if they expect to be punished in case of free-riding with a sufficiently high probability, the selfish types have an incentive to cooperate. Hence, in the presence of strong reciprocators we expect quite different cooperation levels across the punishment and the no-punishment condition. In particular, we expect less cooperation in the no-punishment design.

It is also important to notice that in the stable group design we have the same predictions compared to the random group design if it is *common knowledge that all subjects are rational and selfish money maximizers*. In fact, under this assumption, we should observe exactly the same investment behavior in both the punishment and the no-punishment condition, namely *no investment at all in all periods*. The no-investment prediction is most transparent for period ten. Since all subjects know that the experiment ends in period ten their best private choice in the *no-punishment* condition is to invest nothing. In the punishment condition their best choice at the punishment stage in period ten is to not punish at all because punishment is costly. Yet, since rational egoists anticipate that nobody will punish, the presence of the punishment stage does not change the behavioral incentives at the investment stage of period ten. Therefore, in the *punishment condition* also nobody will invest in period ten. Since rational egoists will anticipate this outcome for period ten, they know that their actions in period nine do not affect the decisions in period ten. Therefore, punishing in period nine makes no sense for selfish players and, as a consequence, full defection at the investment stage of period nine is again their best choice. This backward induction argument can be repeated until period one so that full defection and no punishment is predicted to occur for all ten periods of the punishment treatment. The same backward induction logic also predicts, of course, defection in all periods of the no-punishment treatment.[6]

[6] If rationality and selfishness are not common knowledge there exist other equilibria in which there is cooperation and punishment during the early periods. However, for the final periods the no cooperation - no punishment prediction still holds.

The presence of strong reciprocators will again change the predictions substantially because if a subject is punished for free-riding in period $t < 10$, it knows for sure that the punisher is also part of the group in the next period. Hence, the punished subject has a much stronger incentive to cooperate in the stable group design compared to the random group design. As a consequence, cooperation rates should be higher in the stable group design compared to the random group design.

In sharp contrast to the prediction of zero punishment, subjects punish very often both in the stable group design as well as in the random group design. Figure 2 illustrates the punishment behavior in both designs. It depicts the average punishment imposed on a player as a function of the deviation of the investment of the punished player from the average investment of the other group members. The numbers above the bars denote the relative frequency of observations that correspond to each bar. The stable group design is denoted "Partner" design while the random group design is denoted "Stranger" design in Figure 2. A remarkable feature of Figure 2 is that the strength of the punishment in the Stranger design is almost as high as in the Partner design. For example, if a group member invests between 14 and 8 tokens less than the other group members into the project his income is reduced by roughly 55 percent in the Partner design and by 50 percent in the Stranger design. Moreover, punishment follows a clear pattern. The big majority of punishments are imposed on the defectors and executed by the cooperators. The punishment imposed on a subject is the higher the more the subject's contribution falls short of the average contribution of the other three group members. The positive relation between received punishment and the negative deviation from others' contributions is highly significant while there is no relation between positive deviations and the received punishment. It is also worthwhile to mention that the punishment of sub-average investments also prevails in period ten.

Insert Figure 2 here

What is the impact of this punishment pattern on investment behavior? It turns out that contribution rates differ dramatically between the two conditions. Figure 3 shows the time trend of average investments in the stable group design while Figure 4 shows the trend in the random group design. Note that in both designs the same subjects first participated in the no-punishment condition and then they participated in the punishment condition. Fehr and Gächter (2000) also

reversed the order in which subjects participated in the two designs. The results are almost identical to the results shown in Figures 3 and 4. A remarkable feature of Figures 3 and 4 is that in both designs cooperation breaks down in the no-punishment condition. Initially cooperation is relatively high but over time it unravels and in the final period the absolute majority of the subjects invests nothing and the rest of the subjects invest very little. This supports our view that in the absence of an explicit punishment opportunity the selfish types induce the reciprocal types to behave non-cooperatively, too. Non-cooperation is the only way how the reciprocal types can at least implicitly punish the defectors in their groups. However, if the strong reciprocators are given the opportunity to directly target their punishments towards the individual defectors the pattern of cooperation is very different. Both in the stable group design as well as in the random group design average investments even increase over time. In the stable group design the increase in average investments is much larger and eventually almost full cooperation is achieved.

Insert Figure 3 here

Insert Figure 4 here

The very high cooperation in the punishment conditions represents an unambiguous rejection of the standard repeated interaction approach while it is consistent with the strong reciprocity approach. Moreover, the big difference in cooperation rates across punishment and no-punishment conditions indeed suggests that in the presence of punishment opportunities the strong reciprocators can force the selfish types to cooperate while in the absence of such opportunities the selfish types induce the strong reciprocators to defect, too. Thus, interaction structures that have theoretically identical implications if there are only selfish types generate fundamentally different behavioral patterns in the presence of strong reciprocators.

III.4. Strong Reciprocity as a Norm Enforcement Device

Many small scale societies are characterized by extensive food-sharing. A simple game to examine whether food sharing is a social norm that is enforced by social sanctions has been conducted by Fehr and Fischbacher (2001a). The game is called “third party punishment game” and has three players. The game between player A and player B is just a dictator game. Player A receives an endowment of 100 tokens of which he can transfer any amount to player B, the

Recipient. Player B has no endowment and no choice to make. Player C has an endowment of 50 tokens and observes the transfer of player A. After this player C can assign punishment points to player A. For each punishment point assigned to player A player C has costs of 1 token and player A has costs of 3 token. Since punishment is costly a self-interested player C will never punish. However, if there is a sharing norm player C may well punish player A if A gives too little.

In fact, in the above experiments player As are never punished if they transferred 50 or more tokens to player B. If they transferred less than 50 tokens the punishment was the stronger the less player A transferred. In case that player A transferred nothing she received on average 9 punishment points from player C, i.e. the payoff of player A was reduced by 27 tokens. This means that in this three-person game it was still beneficial, from a selfish point of view, for player A to give nothing compared to an equal split, say. If there is more than one player C, who can punish player A, this may, however, no longer be the case.

Another interesting question is to what extent cooperation norms are sustained through the punishment of free-riders by **third parties**. We have already seen that in the public goods experiment with punishment strikingly high cooperation rates can be enforced through punishment. In this game each investment (i.e., contribution to the public good) increases the payoff of each group member by 0.4. Thus, if a group member free-rides instead of cooperating she directly reduces the other group members' payoff. In real life there are, however, many situations in which free-riding has a very low, indeed almost imperceptible, impact on the payoff of particular other individuals. The question then is, whether these individuals nevertheless help enforcing a social norm of cooperation. In case they do a society greatly magnifies its capability of enforcing social norms because every member of a society acts as a potential "policemen".

It is relatively easy to construct cooperation games with punishment opportunities for third (unaffected) parties. Fehr and Fischbacher (2001a), e.g., have conducted PDs in which a member of the two-person group, who played the PD, observes a member of some other group, who also played the PD. Then the member of the first group can punish the member of the second group. Thus, each member could punish and could be punished by somebody outside the own two-person group. It was ensured that reciprocal punishment was not possible, i.e. if subject A could punish subject B, subject B could not punish A but only some third subject C. It turns out that the

punishment by third parties is surprisingly strong. It is only slightly weaker than second party (within group) punishment.

IV. Proximate Mechanisms behind Strong Reciprocity

Within economics, the leading explanation for the patterns of results described above is that agents have social preferences (or "social utility") which take into account the payoffs and perhaps intentions of others. Roughly speaking, social preference theories assume that people have preferences for how money is allocated (which may depend on who the other player is, or how the allocation came about). From a theoretical viewpoint such preferences are not fundamentally different from preferences for food, the present versus the future, how close one's house is to work, and so forth.[7] In recent years several theories of social preferences have been developed (Rabin 1993; Dufwenberg and Kirchsteiger 1999; Falk and Fischbacher 1999; Fehr and Schmidt 1999; Bolton and Ockenfels 2000; Charness and Rabin 2000). Some of these theories are capable of correctly predicting the bulk of the previously described evidence. For example, the theories of Falk and Fischbacher (1999) and Fehr and Schmidt (1999) predict that positive and negative reciprocity help enforcing nonbonding agreements, that negative reciprocity leads to the rejection of very unequal offers in the ultimatum game, that free-riders are punished in n-person cooperation games and that third parties punish low transfers in the dictator game and defection in other groups in the PD.

It is important to stress that social preference theories only capture proximate mechanisms driving the observed behaviors. They do not aim at explaining the ultimate sources of strong reciprocity. Cultural anthropologists and evolutionary psychologists have sought to explain the origin of strong reciprocity. One idea is that in the environment of evolutionary adaptation (EEA) or ancestral past, people mostly engaged in repeated games with people they knew. Evolution

[7] A different interpretation is that people have rules they obey about what to do—such as, share money equally if you haven't earned it (which leads to equal-split offers in the ultimatum game) (Güth 1995). A problem with the rule-based approach is that subjects **do** change their behavior in response to changes in payoffs, in predictable ways. For example, when the incremental payoff from defecting against a cooperator in a Prisonners' Dilemma is higher, people defect more often. When players' private benefits from the public good are higher, they contribute more. When the responder in the ultimatum game can no longer reject an offer the proposers behave more selfishly on average. Any rule-based account must explain why the rules are bent by incentives, and such a theory will probably end up looking like a theory of social preferences which explicitly weighs self-interest against other dimensions.

created specialized cognitive heuristics for playing repeated games efficiently. It is well-known in game theory that behavior which is optimal for a self-interested actor in a one-period game with a stranger - such as defecting or free riding, accepting all ultimatum offers - is not always optimal in repeated games with partners. In a repeated ultimatum game, for example, it pays to reject offers to build up a reputation for being hard to push around, which leads to more generous offers in the future. In the unnatural habitat view, subjects cannot "turn off" the habitual behavior shaped by repeated-game life in the EEA when they play single games with strangers in the lab.

The unnatural habitat theory assumes the *absence* of a module or cognitive heuristic which could have evolved but did not - the capacity to distinguish temporary one-shot play from repeated play. If subjects had this ability they would behave appropriately in the one-shot game. In principle it is testable whether people have the ability to distinguish temporary one-shot play from repeated play. For example, in Fehr and Gächter (2000) it is shown that cooperation rates are generally lower in public good games with as well as without punishment when the group composition changes randomly in every period relative to the case were the group composition is constant across all ten periods. This fact suggests that, on average, subjects can distinguish between one-shot and repeated interactions because when the group composition changes randomly the probability of meeting the same people again in future periods is much lower. However, a fully satisfactory test of subjects' capacity to distinguish one-shot from repeated interactions requires that the same subjects participate in both conditions so that we can examine behavioral changes across conditions at the individual level. Fehr and Fischbacher (2001b) did this in the context of the ultimatum game.

Fehr and Fischbacher conducted a series of ten ultimatum games in two different conditions. In both conditions subjects played against a different opponent in each of the ten periods of the game. In each period of the baseline condition the Proposers knew nothing about the past behavior of their current Responders. Thus, the Responders could not build up a reputation for being "tough" in this condition. In contrast, in the reputation condition the Proposers knew the full history of the behavior of their current Responders, i.e., the Responders could build up a reputation for being "tough". In the reputation condition a reputation for rejecting low offers is, of course, valuable because it increases the likelihood to receive high offers from the Proposers in future periods.

If the Responders understand that there is a pecuniary payoff from rejecting low offers in the reputation condition one should observe higher acceptance thresholds in this condition. This is the prediction of the social preferences approach that assumes that subjects derive utility from both their own pecuniary payoff and a fair payoff distribution. If, in contrast, subjects do not understand the logic of reputation formation and apply the same habits or cognitive heuristics to both conditions one should observe no systematic differences in Responder behavior across conditions. Since the subjects participated in both conditions it was possible to observe behavioral changes at the individual level. It turns out that the vast majority (slightly more than 80 percent) of the Responders increase their acceptance thresholds in the reputation condition relative to the baseline condition. Moreover, there is not a single subject that reduces the acceptance threshold in the reputation condition relative to the baseline in a statistically significant way.[8] This contradicts the hypothesis that subjects do not understand the difference between one-shot and repeated play.

The above experiment informs us about the proximate mechanism that drives Responder behavior in the ultimatum game. Whatever the exact proximate mechanisms will turn out to be, a hypothesis that is based on the story that subjects do not really understand the difference between one-shot and repeated play seems to be wrong. A plausible alternative hypothesis is that Responders face strong emotions when faced with a low offer and that these emotions trigger the rejections. For modeling purposes, behaviorally relevant emotions can be captured by appropriate formulations of the utility function. This is exactly what theories of social preferences do.

V. Concluding remarks

The empirical evidence shows that many people have inclinations to exhibit strongly reciprocal behavior. Strong reciprocity cannot be rationalized as an adaptive trait by the major prevailing evolutionary theories. The motives invoked by kin selection theory, by the theories of reciprocal altruism and indirect reciprocity, and by costly signaling theory can account for strong reciprocity. However, more recent evolutionary models like the ones by Gintis (2000), Bowles and Gintis

[8] Note that constant acceptance thresholds across conditions are consistent with a social preferences approach while a reduction in the acceptance threshold in the reputation condition would be inconsistent with this approach. If, e.g., a

(2001), Henrich and Boyd (2001), Sethi and Somananthan (2001a, 2001b), and multi-level selections theories as the one proposed by Sober and Wilson (1998), provide plausible evolutionary explanations of the punishment aspect of strong reciprocity.

Strong reciprocity constitutes a powerful constraint for potential cheaters that can generate almost universal cooperation in situations in which purely selfish behavior would cause a complete breakdown of cooperation. Moreover, our results on third party punishment indicate that strong reciprocity is a or perhaps *the* key force in the enforcement of social norms. Once the presence of strongly reciprocal actors is taken into account food-sharing norms and collective actions norms are easy to explain. Strong reciprocity derives its power to fundamentally affect the aggregate outcomes of social interactions from the fact that it often changes the incentives for the selfish types in the population. In sequential interactions, for example, strong reciprocity constitutes an important cooperation incentive for purely self-interested first movers. However, as the example of the simultaneously played PD shows, there are also interaction structures in which the selfish types induce the strongly reciprocal types to behave in a very non-cooperative manner. This means that for social scientists it is very important to examine social interactions according to the objective possibilities of the selfish and the reciprocal types to affect the other type's behavior. In general, our experimental results show that the existence of strong reciprocators greatly improves the prospects for cooperation and norm enforcement. At a methodological level our results indicate that all scientists who are interested in the evolution and the structure of human behavior have much to gain from the application of experimental methods. The recent successful experiments conducted by Henrich (2000) and others (see Henrich et al. 2001) suggest that the scientific returns from experimentation are particularly high in those disciplines (e.g., anthropology) where experimentation was absent or rare in the past.

---

subject rejects already every offer below the equal split in the baseline condition then this subject will in general not increase the acceptance threshold in the reputation condition.

References Cited

Abbink, K., B. Irlenbusch, and E. Renner

2000 The Moonlighting Game – An Experimental Study on Reciprocity and Retribution", Journal of Economic Behavior and Organization 42: 265-277.

Alexander, R. D.

1987 The Biology of Moral Systems. Hawthorne, New York: Aldine De Gruyter.

Axelrod, R., W. D. Hamilton

1981 The Evolution of Cooperation. Science 211: 1390-1396.

Basu, Kaushik

1984 The Less Developed Economy. Oxford: Oxford University Press.

Berg, Joyce, John Dickhaut, and Kevin McCabe

1995 Trust, Reciprocity and Social History, Games and Economic Behavior 10: 122-142.

Bolle Friedel

1998 Rewarding Trust: An Experimental Study, Theory and Decision 45: 85 - 100.

Bolton, Gary and Rami Zwick

1995 Anonymity versus Punishment in Ultimatum Bargaining, Games and Economic Behavior 10: 95-121.

Bolton, Gary E. and Axel Ockenfels

2000 A Theory of Equity, Reciprocity and Competition, American Economic Review 100: 166-193.

Bowles, Sam and Herbert Gintis

2001 The Evolution of Strong Reciprocity, Discussion Paper, University of Massachusetts at Amherst.

Boyd, Robert, and Peter J. Richerson

1988 The Evolution of Reciprocity in Sizable Groups, Journal of Theoretical Biology 132(3): 337-56.

Camerer, Colin F. and Richard H. Thaler,

1995 Ultimatums, Dictators and Manners, Journal of Economic Perspectives 9: 209-19.

Cameron, Lisa A.

1999 Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia, *Economic-Inquiry* 37(1): 47-59.

Charness, Gary

2000 Responsibility and Effort in an Experimental Labor Market, Journal of Economic Behavior and Organization 42: 375-384.

Charness, Gary, and Matthew Rabin

2000 Social Preferences: Some Simple Tests and a New Model." Mimeo, University of California at Berkeley.

Dufwenberg, Martin and Kirchsteiger, Georg

1998 A Theory of Sequential Reciprocity, Discussion Paper, CentER, Tilburg University.

Falk, Armin and Urs Fischbacher

1999 A Theory of Reciprocity, Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 6.

Fehr, Ernst and Armin Falk

1999 Wage Rigidity in a Competitive Incomplete Contract Market, Journal of Political Economy 107: 106-134.

Fehr, Ernst and Urs Fischbacher

2001a Third Party Punishment, mimeo, Institute for Empirical Research in Economics, University of Zürich.

Fehr, Ernst and Urs Fischbacher

2001b Retaliation and Reputation, mimeo, Institute for Empirical Research in Economics, University of Zürich.

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl

1993 Does Fairness prevent Market Clearing? An Experimental Investigation, Quarterly Journal of Economics 108: 437-460.

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl

1998 Gift Exchange and Reciprocity in Competitive Experimental Markets, European Economic Review 42: 1-34.

Fehr, Ernst, and Simon Gächter

2000 Cooperation and Punishment in Public Goods Experiments, American Economic Review 90: 980-994.

Fehr, Ernst, Simon Gächter and Georg Kirchsteiger

1997 Reciprocity as a Contract Enforcement Device, Econometrica 65: 833-860.

Fehr, Ernst and Klaus M. Schmidt

1999 A Theory of Fairness, Competition and Co-operation, Quarterly Journal of Economics 114: 817-868.

Fehr, Ernst and Elena Tougareva

1995 Do High Monetary Stakes Remove Reciprocal Fairness? Experimental Evidence from Russia, *Mimeo*. Institute for Empirical Economic Research, University of Zurich.

Friedman James

1971 A Noncooperative Equilibrium for Supergames, Review of Economic Studies 38: 1 – 12.

Fudenberg Drew and Eric Maskin

1986 The Folk Theorem in Repeated Games with Discounting or with Incomplete Information, Econometrica 54: 533-556.

Gintis, Herbert

2000 Strong Reciprocity and Human Sociality, Journal of Theoretical Biology 206: 169-179.

Gintis, Herbert, Eric Smith and Sam Bowles

2001 Costly Signaling and Cooperation, Journal of Theoretical Biology, forthcoming.

Gächter, Simon and Armin Falk

2001 Reputation and Reciprocity - Consequences for the Labour Relation, Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 19. and Scandinavian Journal of Economics, forthcoming.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze

1982 An Experimental Analysis of Ultimatium Bargaining, Journal of Economic Behavior and Organization 3: 367-88.

Güth, Werner

1995 On the Construction of Preferred Choices – The Case of Ultimatum Proposals, Discussion Paper Economic Series No. 59, Humboldt University Berlin.

Hamilton, William D.

1964 Genetical Evolution of Social Behavior I, II, Journal of Theoretical Biology 7(1): 1-52.

Henrich Joe

Does Culture Matter in Economic Behavior – Ultimatum Game Experiments among the Machiguenga of the Peruvian Amazon, American Economic Review 90: 973-979.

J. Henrich and R. Boyd

2001 Why People Punish Defectors: Weak Conformist Transmission can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas. Journal of Theoretical Biology 208: 79–89.

Henrich J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis and R. McElreath

2001 In Search of Homo Economicus – Behavioral Experiments in 15 Small-Scale Societies, American Economic Review 91: 73-79.

Hoffman, Elisabeth, Kevin McCabe, and Vernon Smith, 1996

On Expectations and Monetary Stakes in Ultimatum Games, International Journal of Game Theory 25: 289-301.

Levine, D.

1998 Modeling Altruism and Spitefulness in Experiments, Review of Economic Dynamics 1: 593-622.

McCabe, Kevin A., Stephen J. Rassenti, Vernon L. Smith

1998 Reciprocity, Trust, and Payoff Privacy in Extensive Form Bargaining, Games and Economic Behavior 24: 10-24.

McCabe, K. A., Mary L. Rigdon and V. L. Smith

2000 Positive Reciprocity and Intentions in Trust Games, mimeo, University of Arizona at Tucson, October 2000.

Nowak Martin and Karl Sigmund

1998 Evolution of Indirect Reciprocity by Image Scoring, Nature 393: 573-577

Rabin, Matthew

1993 Incorporating Fairness into Game Theory and Economics, American Economic Review 83: 1281-1302.

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir

1991 Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study, American Economic Review 81: 1068-95.

Roth, Alvin E.

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

Sethi, R. and E. Somananthan

2001a Understanding Reciprocity, Journal of Economic Behavior and Organization, forthcoming.

Sethi, R. and E. Somananthan

2001b Norm Compliance and Reciprocity, Discussion Paper, Department of Economics, Columbia University.

Slonim, Robert, and Alvin E. Roth

1998 Financial Incentives and Learning in Ultimatum and Market Games: An Experiment in the Slovak Republic, Econometrica 65: 569-596.

Trivers, R. L.

1971 The Evolution of Reciprocal Altruism, Quarterly Review of Biology 46: 35-57.

Sober, Elliott and David Sloan Wilson

1998 Unto Others – The Evolution and Psychology of Unselfish Behavior, Cambridge: Harvard University Press.

Zahavi, Amotz and Avishay Zahavi

1997 The Handicap Principle: A Missing Piece of Darwin’s Puzzle, New York: Oxford University Press.

Figure 1: Relation of desired effort and actual effort to the profit offered to the worker (N=141)

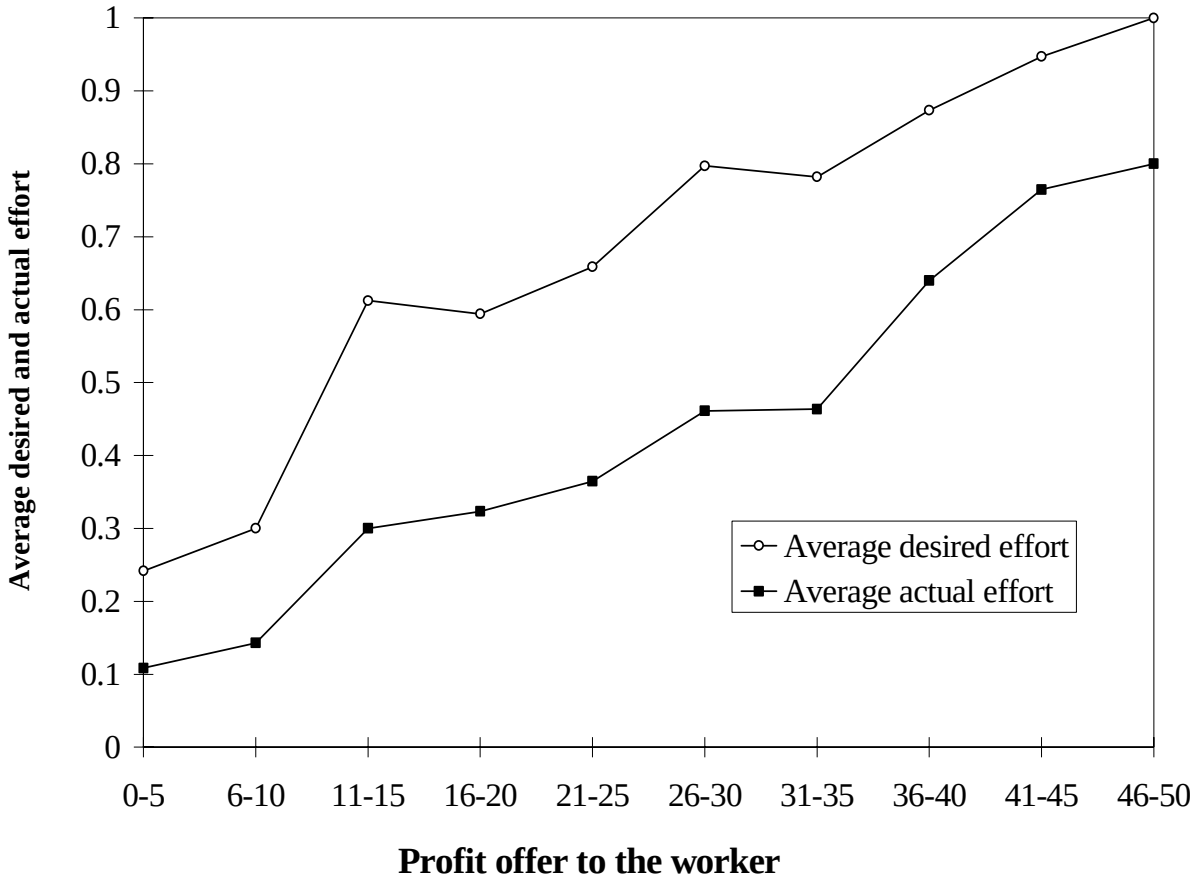

Figure 2: Received punishment points as a function of the deviation from others' average investment (10 partner-groups, 18 stranger-groups)

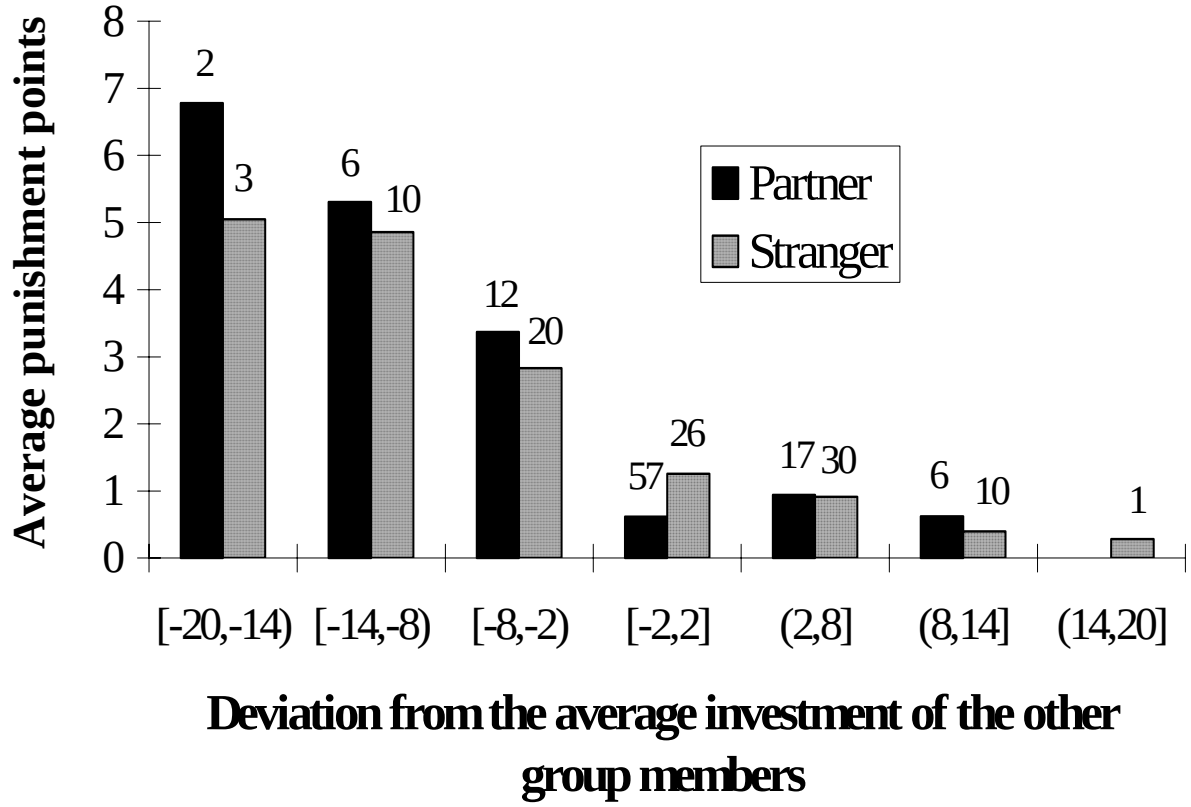

Figure 3: Average investments over time in public good games with stable groups (10 groups)

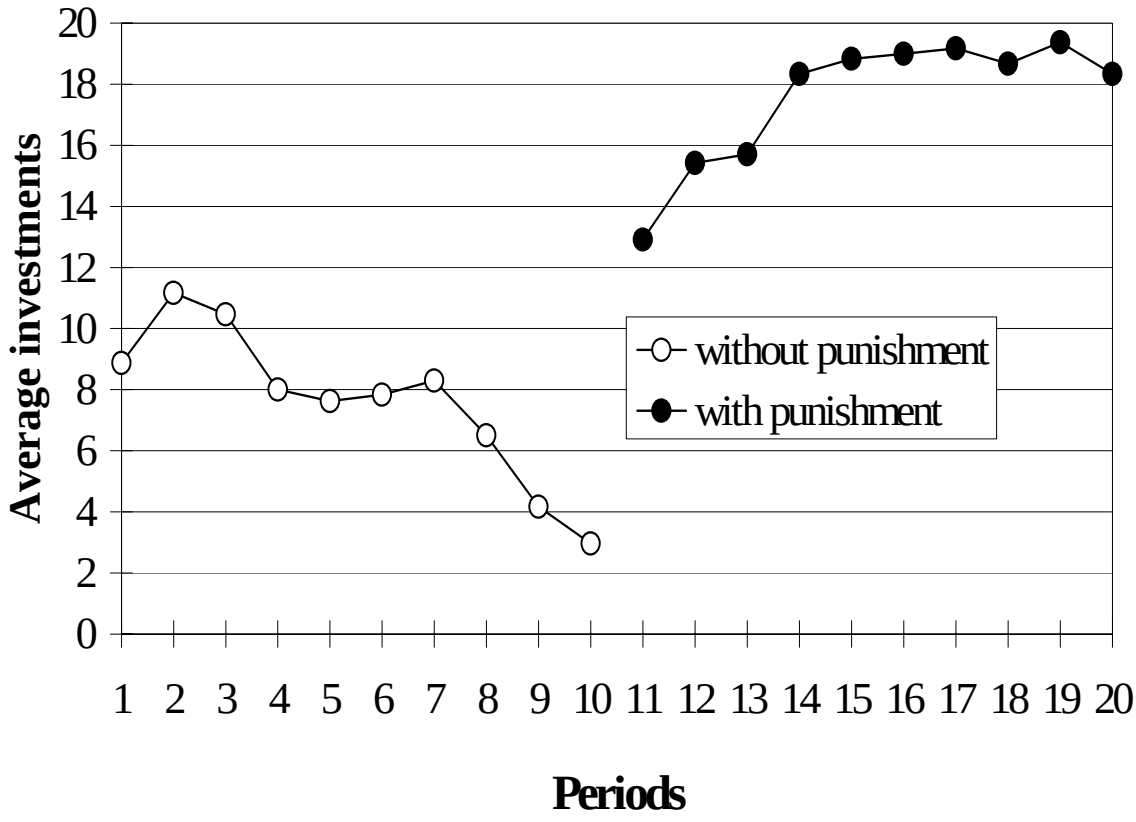

Figure 4: Average investments over time in public good games with random groups (18 groups)