The Economic Journal, **112** (*March*), C1–C33. © Royal Economic Society 2002. Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA.

WHY SOCIAL PREFERENCES MATTER – THE IMPACT OF NON-SELFISH MOTIVES ON COMPETITION, COOPERATION AND INCENTIVES

Ernst Fehr and Urs Fischbacher

A substantial number of people exhibit social preferences, which means they are not solely motivated by material self-interest but also care positively or negatively for the material payoffs of relevant reference agents. We show empirically that economists fail to understand fundamental economic questions when they disregard social preferences, in particular, that without taking social preferences into account, it is not possible to understand adequately (i) effects of competition on market outcomes, (ii) laws governing cooperation and collective action, (iii) effects and the determinants of material incentives, (iv) which contracts and property rights arrangements are optimal, and (v) important forces shaping social norms and market failures.

Economics may be called the dismal science because economists routinely make worst case assumptions regarding people's motives. Economic reasoning is typically based on the self-interest hypothesis, ie, on the assumption that *all* people are *exclusively* motivated by their material self-interest. This assumption rules out any heterogeneity with respect to other-regarding, social, preferences. It also contrasts sharply with economists' emphasis on the heterogeneity of people's tastes with regard to consumption activities. Our models typically allow heterogeneous tastes with regard to apples and bananas, with regard to hazardous and safe jobs, etc., but in the important realm of social preferences mainstream economics adheres to the extreme self-interest assumption.

The purpose of this paper is to show that economists fail to understand core questions in economics if they insist on the self-interest hypothesis and rule out heterogeneity in the realm of social preferences. This is so for two reasons. First, during the last decade experimental economists have gathered overwhelming evidence that systematically refutes the self-interest hypothesis and suggests that a substantial fraction of the people exhibit social preferences, in particular, preferences for reciprocal fairness. Second, there is also strong evidence indicating that the deviations from self-interest have a fundamental impact on core issues in economics.

One core question in economics is to understand the workings of competition and the interplay of competition and cooperation in markets and organisations. Other core questions pertain to understanding the conditions for successful collective actions, the prevailing structure of contracts and property rights, and, above all, the workings of material incentives because material incentives constitute the essence of economics. We claim that a satisfactory understanding of these questions is impeded by the self-interest hypothesis. In particular, we provide evidence suggesting that preferences for reciprocal fairness shape the functioning of competition, govern the laws of cooperation and collective action, and have a decisive impact on how material incentives are

constituted and how they function. The evidence also indicates that, by changing the incentives for the selfish types, reciprocal fairness affects the prevailing interaction patterns and the constraints on individual behaviour, ie, the prevailing contracts and institutions.

The structure of our paper is as follows. In Section 1 we briefly describe the most important types of social preferences. We then illustrate the preference for reciprocity by means of two simple one-shot experiments and discuss whether reciprocal behaviour in these experiments can be interpreted as a cognitive mistake, ie, a kind of habit that is learned in the repeated interactions outside the laboratory and inappropriately applied to one-shot situations, or whether reciprocal behaviour is better interpreted as a preference. Section 2 then shows that if one neglects social preferences one cannot understand important effects of competition on market prices. Section 3 deals with cooperation and shows that decisive determinants of cooperation cannot be understood on the basis of the self-interest hypothesis. Section 4 deals with material incentives, contracts and property rights. We present evidence indicating that neither the effects nor the determinants of material incentives can be adequately understood if one neglects social preferences and that the interaction between material incentives and social preferences is likely to have important effects on the optimality of different types of contracts and property rights. Section 5 discusses some problems in the modelling of reciprocity and fairness preferences and to what extent it is possible to mimic preferences for reciprocity by simpler and more tractable models of inequity aversion. Finally, Section 6 concludes the paper.

1. The Nature of Social Preferences

The last 15 years have seen a large number of studies indicating that – in addition to material self-interest – social preferences shape the decisions of a substantial fraction of the people. A person exhibits social preferences if the person not only cares about the material resources allocated to her but also cares about the material resources allocated to relevant reference agents. Depending on the situation, the relevant reference agents may be the colleagues in the firm with whom a person interacts most frequently, or a person's relatives, or the trading partners, or a person's neighbours. It is important to keep in mind that in different domains a person may have different reference agents. In this paper we do not attempt to summarise the empirical evidence on social preferences (for surveys see Fehr and Schmidt (2000) and Sobel (2001)). Instead we are interested in the economic implications of people's social preferences. Before we proceed it is nevertheless useful to mention the quantitatively most important types of social preferences that have been uncovered by the literature.

A particularly important type of social preference is the preference for *reciprocity* or reciprocal fairness, respectively. A reciprocal individual responds to actions that are perceived to be kind in a kind manner, and to actions that are

perceived to be hostile in a hostile manner. Whether an action is perceived as kind or hostile depends on the fairness or unfairness of the consequences and the intention associated with the action. The fairness of the intention, in turn, is determined by the equitability of the payoff distribution, relative to the set of feasible payoff distributions, caused by the action. It is important to emphasise that reciprocity is not driven by the expectation of future material benefit. It is, therefore, fundamentally different from 'cooperative' or 'retaliatory' behaviour in repeated interactions. These behaviours arise because actors expect future material benefits from their actions; in the case of reciprocity, the actor is responding to friendly or hostile actions even if no material gains can be expected. Models of reciprocity have been developed by Rabin (1993), Levine (1998), Falk and Fischbacher (1999), Dufwenberg and Kirchsteiger (1999), Segal and Sobel (1999) as well as Charness and Rabin (2000).[1]

A second type of social preference is *inequity aversion* as modelled in Fehr and Schmidt (1999) or Bolton and Ockenfels (2000). Fehr and Schmidt (1999) assume, eg, that inequity averse persons want to achieve an equitable distribution of material resources. This means that they are altruistic towards other persons, ie, they want to increase the other persons' material payoffs, if the other persons' material payoffs are below an equitable benchmark, but they feel envy, ie, they want to decrease the other persons' payoffs, when the payoffs of the others exceed the equitable level. In many situations reciprocal persons and inequity averse persons behave in similar ways. For example, both reciprocity and inequity aversion imply the desire to reduce the payoff of another person if that person made a decision such that the payoff of the reciprocal or inequity averse person is much lower than the payoff of the other person. Recent evidence (Falk *et al.*, 2000*a*, *b*) suggests, however, that reciprocity is the stronger and quantitatively more important motive.

The similarity in the behaviour of reciprocal and inequity averse persons is due to the fact that both concepts depend in important ways on the notion of a fair or equitable payoff. Since models of inequity aversion are much simpler and more tractable than models of reciprocity it is often convenient to 'mimic' or to 'black box' reciprocal behaviour by inequity aversion (see also Section 5). Some authors (eg, Charness and Rabin, 2000) have also found evidence suggesting that subjects tend to help the least well off. Such behaviour is, however, often not distinguishable from inequity aversion, in particular, non-linear inequity aversion. Recently, Neilson (2000) provided an axiomatic characterisation of a non-linear version of Fehr–Schmidt type inequality aversion.

Reciprocity and inequity aversion are very different from *pure altruism* which constitutes a third type of social preference. Altruism is a form of unconditional kindness; that is, altruism given does not emerge as a response to altruism received (Andreoni, 1989; Andreoni and Miller, forthcoming; Cox *et al.*, 2001). In technical terms altruism means that a person values the material resources allocated to a

[1] Levine's model of reciprocity is, strictly speaking, not based on intentions but on the reciprocation to the other players' preferences. A subject with Levine-type preferences is more altruistic (or less spiteful) towards an altruistic player and more spiteful (or less altruistic) towards a spiteful player. The model thus captures a kind of type-based reciprocity.

relevant reference agent positively. An altruistic person, therefore, never takes an action that decreases the payoff of a reference agent. Yet, as we will see below an important stylised fact concerns people's willingness to punish other people for unfair or hostile actions. Since altruism is a form of unconditional kindness, it cannot explain the phenomenon of conditional cooperation, ie, the fact that many people are willing to increase their voluntary cooperation in response to the co-operation of the other players.

Finally, research has also shown that a fraction of the people exhibits *spiteful or envious* preferences (Falk *et al.*, 2000*a*). A spiteful or envious person *always* values the material payoff of relevant reference agents negatively. The person is, therefore, willing to decrease the material payoff of a reference agent at a personal cost to himself (Kirchsteiger, 1994; Mui, 1995), irrespective of the payoff distribution and irrespective of the reference agent's fair or unfair behaviour. Spiteful choices seem to be quantitatively less important than reciprocal choices. Moreover, spitefulness (as well as altruism) cannot explain why the *same* people often are willing to help others at a personal cost in one situation while they harm other people in other situations (Falk *et al.*, 2000*b*).

Although previous research clearly indicates that many people exhibit social preferences it is important to keep in mind that not everybody exhibits social preferences. In fact, most studies indicate that there is also a substantial fraction of people who behave in a purely *selfish* manner. A key question, therefore, is how the heterogeneity of motives at the individual level can be captured by parsimonious models and how the different individual motivations interact. In the following we concentrate on the existence of reciprocal and selfish types. The reason for this is three-fold. First, empirical evidence suggests that in the domain of payoff-decreasing or punishing behaviour negative reciprocity is the dominant motive (Falk *et al.*, 2000*a*, *b*; Kagel and Wolfe, 2000; Offerman, forthcoming), although other motives like, eg, inequity aversion and envy also play a role. Second, in the domain of helping or rewarding behaviour positive reciprocity seems to be less dominant but it also plays an important role (Cox, 2000; Charness and Rabin, 2000; Falk *et al.*, 2000*b*; Offerman, forthcoming). For reasons of parsimony we will, therefore, neglect the other motives in the following.

Finally, theory as well as empirical evidence suggest that the interaction between reciprocal and selfish types is of first-order importance for many economic questions. The reason for this is that the presence of reciprocal types often changes the material incentives for the selfish types which induces the selfish types to make 'non-selfish' choices. For example, a selfish person is deterred from behaving opportunistically if the person expects to be punished by the reciprocators. Likewise, a selfish person may be induced to behave in a co-operative and helpful manner because she expects the reciprocators to return the favour. Since the presence of reciprocal types changes the pecuniary incentives for the selfish types the reciprocal types often have a big impact on the aggregate outcome in markets and organisations.

1.1. *Positive and Negative Reciprocity: Two Examples*

In the following we illustrate the existence of reciprocity and fairness with the help of two simple experiments.[2] We focus on experiments because in most real life situations it is impossible to isolate the impact of reciprocity and fairness unambiguously. A sceptic may always discount field evidence with the argument that, in the field, the notion of fairness is only used for rhetorical purposes that disguise purely self-interested behaviour in an equilibrium of a repeated game.[3]

Our first illustration concerns the Ultimatum game (UG) that was introduced by Güth *et al.* (1982). In the UG a pair of subjects has to agree on the division of a fixed sum of money. Person *A*, the proposer, can make one proposal of how to divide the amount. Person *B*, the responder, can accept or reject the proposed division. In the case of rejection, both receive nil; in the case of acceptance, the proposal is implemented. Under the standard assumptions that (i) both the proposer and the responder are rational *and* care only about how much money they get and (ii) the proposer knows that the responder is rational and selfish, the subgame perfect equilibrium predicts a rather extreme outcome: The responder accepts *any* positive amount of money and, hence, the proposer gives the responder the smallest money unit, ε, and keeps the rest.

A robust result in the UG, across hundreds of experiments, is that proposals offering the responder less than 20% of the available surplus are rejected with probability 0.4 to 0.6. In addition, the probability of rejection is decreasing in the size of the offer (see, eg, Camerer (in press); Roth (1995) and the references therein). Apparently, many responders do not behave in a self-interested manner. In general, the motive indicated for the rejection of positive, yet 'low', offers is that subjects view them as unfair. A further robust result is that many proposers seem to anticipate that low offers will be rejected with a high probability. This is suggested, for example, by the comparison of the results of Dictator Games (DG) and UGs. In a DG the responder's option to reject is removed – the responder must accept any proposal. Forsythe *et al.* (1994) were the first who compared the offers in UGs and DGs. They report that offers are substantially higher in the UG which suggests that many proposers do apply backwards induction in this game. The offers and rejection rates in the UG are generally quite robust across (developed) cultures, levels of stakes (including \$100–\$400 in the United States and 2–3 months' wages in other countries), and changes in experimental methodology (see Camerer, in press).

[2] In economic experiments human subjects make decisions with real monetary consequences in carefully controlled laboratory settings. In particular, the experimenter can implement one-shot interactions between the subjects so that long-term self-interest can be ruled out as an explanation for what we observe. As we will see, in some experiments the monetary stakes involved are quite high – amounting up to the income of three months' work. In the experiments reviewed below subjects do not know each others' identities, they interact anonymously and, sometimes, even the experimenter cannot observe their *individual* choices. Due to the anonymity conditions the laboratory environment is quite unfavourable to the emergence of reciprocal behaviour. Yet, if we observe reciprocal behaviour under such unfavourable conditions it is even more likely to prevail in non-anonymous interactions between people who know each other.

[3] For evidence suggesting that fairness and reciprocity is important in the field see, eg, Agell and Lundborg (1995), Bewley (1999), Frey and Weck-Hannemann (1984), Frey and Pommerehne (1993), Greenberg (1990), Kahneman *et al.* (1986), Lind and Tyler (1988), Ostrom (1990, 2000), Seidl and Traub (1999), Zajac (1995).

The UG indicates that a sizeable fraction of responders is willing to punish behaviour that is perceived as unfair. In contrast, the Gift exchange game (GEG) indicates that a substantial fraction of the responders are willing to reward actions that are perceived as generous or fair. The GEG has been introduced by Fehr, Kirchsteiger and Riedl (1993). In the GEG the proposer offers a wage $w \in [\underline{w}, \bar{w}]$, $\underline{w} \geqslant 0$, to the responder. The responder can accept or reject w. In the case of rejection both players receive zero payoff; in the case of acceptance the responder has to make a costly 'effort' choice $e \in [\underline{e}, \bar{e}]$, $\underline{e} > 0$. The monetary payoff for the proposer is $x^P = ve - w$ while the responder's payoff is $x^R = w - c(e)$ where v denotes the marginal value of effort for the proposer and $c(e)$ the strictly increasing effort cost schedule.[4] Under the standard assumptions (i) and (ii) above the responder will always choose the lowest feasible effort level $\underline{e}$ and will, in equilibrium, never reject any w. Therefore, the subgame perfect equilibrium predicts that the wage will be set equal to the lowest feasible wage level $\underline{w}$.

The GEG captures a principal–agent relation with highly incomplete contracts in a stylised way. Variants of the GEG have been conducted by several authors.[5] All of these studies report that there is, in general, a strong positive correlation between the mean effort and the offered wage which is consistent with the interpretation that the responders, on the average, reward generous wage offers with generous effort choices. However, as in the case of the UG, there are considerable individual differences among the responders. While there is typically a sizeable fraction of responders (frequently roughly 40%, sometimes more than 50%) who exhibit a reciprocal effort pattern, there is also a substantial fraction of responders who always make purely selfish effort choices or whose choices seem to deviate randomly from the self-interested action. Similar to the UG the regularities in the GEG are quite robust with regard to stake levels. In experiments in which subjects earned on the average between two and three times their monthly incomes the same wage and effort patterns prevailed (Fehr and Tougareva, 1995).

1.2. *One-Shot and Repeated Interactions*

Sometimes it is argued that reciprocal behaviour in anonymous one-shot experiments is due to subjects' inability to adjust properly to one-shot interactions. One idea is that outside the laboratory subjects are typically involved in a network of repeated interactions. It is well known from repeated game theory that in repeated interactions rewarding and punishing may be in the long-run self-interest of an individual. Hence, according to this argument, subjects who routinely interact in the repeated game of life import routines and habits, that are appropriate for repeated interactions, into the laboratory one-shot situation because they do not

[4] In some applications of this game the proposer's payoff was given by $x^P = (v - w)e$. This formulation rules out that proposers can incur losses when they offer generously high wages. Likewise, in some applications of the GEG the responder did not have the option to reject w. Thus, the proposer just sent w while the responder chose an effort level. Under the standard assumptions of rationality and selfishness the subgame perfect equilibrium is, however, not affected by these differences.

[5] See, eg, Fehr, Kirchsteiger and Riedl (1993, 1998), Charness (1996, 2000), Fehr and Falk (1999), Gächter and Falk (forthcoming), Falk *et al.* (1999), Hannan *et al.* (forthcoming) and Brandts and Charness (1999).

understand the strategic differences between one-shot and repeated interactions. Therefore, the observation of reciprocal behaviour in one-shot interactions should not be interpreted as a deviation from self-interest but merely as a form of rule-of-thumb behaviour, ie, as a cognitive failure to distinguish properly between one-shot and repeated interactions.

Our response to this argument is two-fold. First, and most importantly, the argument is refuted by evidence indicating that the vast majority of the subjects understand the strategic differences between one-shot and repeated interactions quite well. Below we discuss the study by Fehr and Fischbacher (2001) that explicitly tested for this. Second, even if the argument were correct, economists would have strong reasons to take the habits and routines that shape people's behaviour into account. The importance of reciprocal behaviour for economics does not depend on whether it is interpreted as a deviation from self-interest or as a form of bounded rationality. Reciprocal behaviour is important because it affects in fundamental ways the functioning of markets, organisations, incentives and collective actions.

In principle it is testable whether people have the ability to distinguish temporary one-shot play from repeated play. Fehr and Fischbacher (2001) investigated this problem in the context of the Ultimatum game and Gächter and Falk (forthcoming) provided evidence for the Gift exchange 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 iterations of the game. In each iteration of the *baseline condition* the proposers knew nothing about the past behaviour 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 behaviour of their current responders, ie, 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 of receiving high offers from the proposers.

If the responders understand that there is a pecuniary payoff from rejecting low offers in the reputation condition one should in general 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 behaviour across conditions. Since the subjects participated in both conditions it was possible to observe behavioural changes at the individual level. It turns out that the vast majority (slightly more than 80%, $N = 72$) of the responders *increase* their acceptance thresholds in the reputation condition relative to the baseline condition.[6]

[6] The remaining subjects, except one, exhibit no significant change in the acceptance threshold. Only one out of 70 subjects exhibits a significant decrease in the threshold relative to the baseline. Note that if a subject places a very high value on fairness the acceptance threshold may already be very high in the baseline condition so that there is little reason to change the threshold in the reputation condition. Identical thresholds across conditions are, therefore, also compatible with a social preference approach. Only a decrease in the acceptance threshold is incompatible with theories of social preferences.

This contradicts the hypothesis that subjects do not understand the difference between one-shot and repeated play.

A plausible alternative hypothesis is that responders face strong emotions when faced with a low offer and that these emotions trigger the rejections. The evolutionary origin of these emotions may be located in the repeated game interactions in our ancestral past and the emotions may, therefore, not be fine-tuned to one-shot interactions (Binmore, 1998). However, for modelling purposes, behaviourally relevant emotions can be captured by appropriate formulations of the utility function. This is exactly what theories of social preferences do.

2. Competition

In this section we illustrate our claim that it is not possible to understand the effects of competition if concerns for fairness and reciprocity are neglected. We will show, in particular, that the self-interest hypothesis hinders economists from understanding important comparative static effects of competition. In addition, we present results indicating that competition may sometimes have no impact on market outcomes because of the presence of reciprocal factors.

2.1. *The Effects of Competition under Exogenous Contract Enforcement*

Consider the following slightly modified ultimatum game. Instead of one responder there are now two competing responders. When the proposer has made his offer the two responders simultaneously accept or reject the offer. If both accept, a random mechanism determines with probability 0.5 which one of the responders will get the offered amount. If only one responder accepts he will receive the offered amount of money. If both responders reject, the proposer and both responders receive nil.

The ultimatum game with responder competition can be interpreted as a market transaction between a seller (proposer) and two competing buyers (responders) who derive the same material payoff from an indivisible good. Moreover, if the parties' pecuniary valuations of the good are public information all involved parties know the surplus. Since there is a known fixed surplus there is no uncertainty regarding the quality of the good provided by the seller. The situation can thus be viewed as a market in which the contract (quality of the good) is enforced exogenously.

If all parties are selfish, competition among the responders does not matter because the proposer is predicted to receive the whole surplus already in the bilateral case. Adding competition to the bilateral ultimatum game has therefore no effect on the power of the proposer. It is also irrelevant whether there are two, three or more competing responders. The self-interest hypothesis thus implies a very counterintuitive result, namely, that increasing the competition among the responders does not affect the share of the surplus that the responders receive. Fischbacher, Fong and Fehr (2001) tested this prediction by conducting ultimatum games with one, two and five responders. To allow for convergence and learning effects, in each experimental session a large group of subjects played the same game for 20 periods. For example, in the case with two responders one third

of the subjects were always in the role of the proposer and two thirds of the subjects were in the role of the responder. In every period the proposers and the responders were randomly re-matched to ensure the one-shot nature of the interactions. All subjects knew that after period 20 the experiment would end. The results of these experiments are presented in Fig. 1.

Fig. 1 shows the responders' average share of the surplus in each period across conditions. In the bilateral case the average share is – except for period 1 – quite close to 40%. Moreover, the share does not change much over time. In the final period the responders still appropriate slightly more than 40% of the surplus. In the case of two responders the situation changes dramatically, however. Already in period 1 the responders' share is reduced by 16.5 percentage points relative to the bilateral case. Moreover, from period 1 till period 12 responder competition induces a further reduction of the share by 14 percentage points (from 35% to 21%) and in the final period the share is even below 20%. Thus, the addition of just one more responder has a dramatic impact on the share of the responders. If we add three additional responders the share goes down even further. From period 3 onwards it is below 20% and comes close to 10% in the second half of the session.[7]

The reason why the responders' share decreases when competition increases is that the rejection probability of the responders declines when there are more

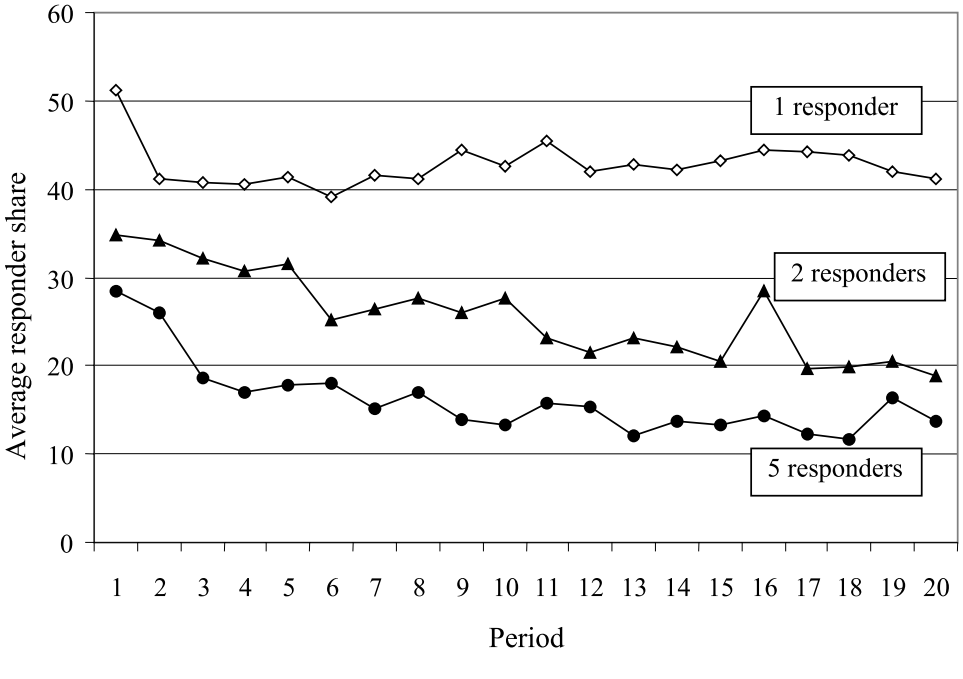

Fig. 1. *Responder Share in the Ultimatum Game with One, Two and Five Competing Responders*
Source: Fischbacher, Fong and Fehr (2001)

[7] In Roth *et al.* (1991) competition led to an even more extreme outcome. However, in their market experiments 9 competing proposers faced only 1 responder and the responder was forced to accept the highest offer.

competing responders. These facts can be parsimoniously explained if one takes the presence of reciprocal or inequity averse responders into account (Falk and Fischbacher, 1999; Fehr and Schmidt, 1999). Recall that reciprocal responders reject low offers in the bilateral ultimatum game because by rejecting they are able to punish the unfair proposers. In the bilateral case they can always ensure this punishment while in the competitive case this is no longer possible. In particular, if one of the other responders accepts a given low offer, it is impossible for a reciprocal responder to punish the proposer. Since there is a substantial fraction of selfish responders, the probability that one of the other responders is selfish is higher the larger the number of competing responders. This means, in turn, that the expected non-pecuniary return from the rejection of a low offer is smaller the larger the number of competing responders. Therefore, reciprocal responders will reject less frequently the larger the number of competing responders.

2.2. *The Effects of Competition under Endogenous Contract Enforcement*

The previous example illustrates that the self-interest model underestimates the power of competition. This example should, however, not make us believe that sufficient competition will in general weaken or remove the impact of fairness on market outcomes, quite the contrary. In the following we will show that the presence of reciprocal individuals may completely nullify the impact of competition on market outcomes. Whether competition does have the effects illustrated in Fig. 1 depends critically on the enforceability of the contracts.

To illustrate this argument consider the double auction experiments conducted by Fehr and Falk (1999). Fehr and Falk deliberately chose the double auction as the trading institution because a large body of research has shown the striking competitive properties of experimental double auctions. In hundreds of such experiments, prices and quantities quickly converged to the competitive equilibrium predicted by standard self-interest theory (see Davis and Holt (1993), for a survey of important results). Therefore, showing that reciprocity renders competition completely powerless in an experimental double auction provides a strong piece of evidence in favour of the importance of reciprocity in markets.

Fehr and Falk use two treatment conditions: a bilateral condition in which competition is completely removed and a competitive condition. In the competitive condition they embed the gift exchange framework into the context of an experimental double auction that is framed in labour market terms.[8] The crucial difference between the competitive condition and the gift exchange game described in Section 1 is that both experimental firms and experimental workers can make wage bids in the interval [20,120] because the workers' reservation wage is 20 and the maximum revenue from a trade is 120. If a bid is accepted, a labour contract is concluded and the worker has to choose the effort level. As in the gift exchange game the workers ('responders') can freely choose any feasible effort

[8] In the meantime the gift exchange game has been framed in goods market terms, labour market terms and in a completely neutral language. The results indicate that there are no framing effects.

level. They have to bear effort costs while the firm ('proposer') benefits from the effort. Thus, the experiment captures a market in which the quality of the good traded ('effort') is not exogenously enforced but is chosen by the workers. Workers may or may not provide the effort level that is expected by the firms.

In the competitive double auction there are 8 firms and 12 workers and each firm can employ at most one worker. A worker who enters into a contract has costs of 20. Therefore, due to the excess supply of labour, the competitive wage level is 20. A double auction lasts for ten periods and a period lasts for three minutes.[9] In contrast to the double auction firms in the bilateral condition are exogenously matched with a worker and if a worker rejects the firm's offer both parties earn nothing. The bilateral condition consists of a series of ten one-shot gift exchange games that are also framed in labour market terms. There are ten firms (proposers) and ten workers (responders). In each of the ten periods each firm is matched with a different worker. Firms have to make an offer to the matched worker in each period. If the worker accepts he has to choose the effort level. As in the competitive condition a worker who accepts a wage offer has costs of 20 and the maximum revenue from a trade is 120.

The self-interest model predicts that in both conditions the workers will only provide the minimum effort so that the firms will pay a wage of 20 or 21 in equilibrium. However, we know already from bilateral ultimatum games that firms (proposers) cannot reap the whole surplus, ie, wages in the bilateral gift exchange game also can be expected to be much higher than predicted by the self-interest model. Moreover, since in the gift exchange game the effort is in general increasing in the wage level firms have an additional reason to offer workers a substantial share of the surplus. The question, therefore, is to what extent competition in the double auction pushes wages below the level in the bilateral condition.

Fig. 2 shows the evolution of wages in both conditions. The figure indicates the startling result that competition has no long run impact on wage formation in this setting. Only at the beginning wages in the double auction are slightly lower than the wages in the bilateral condition but since workers responded to lower wages with lower effort levels firms raised their wages from period four onwards. In the last five periods firms paid slightly higher wages in the double auction; this difference is not significant, however. It is also noteworthy that competition among the workers was extremely intense. In each period many workers offered to work for wages below 30 but firms preferred instead to pay workers on average wages around 60. It was impossible for the workers to get a job by underbidding the going wages because the positive effort–wage relation made it profitable for the firms to pay high, non-competitive, wages.[10]

[9] In each period the same stationary situation is implemented, ie, there are 12 workers, 8 firms, and each worker's reservation wage is 20. In a given period employers and workers can make as many wage bids as they like, as long as they have not yet been signed on. Trading is anonymous. Every worker can accept an offer made by a firm and every firm can accept an offer made by a worker.

[10] A variety of studies have found that one major reason why managers are reluctant to cut wages in a recession is the fear that wage cuts may hamper work performance. Among others, Bewley (1999) reports that managers are afraid that pay cuts 'express hostility to the work force' and will be 'interpreted as an insult'. For similar results see Agell and Lundborg (1995) and Campbell and Kamlani (1997).

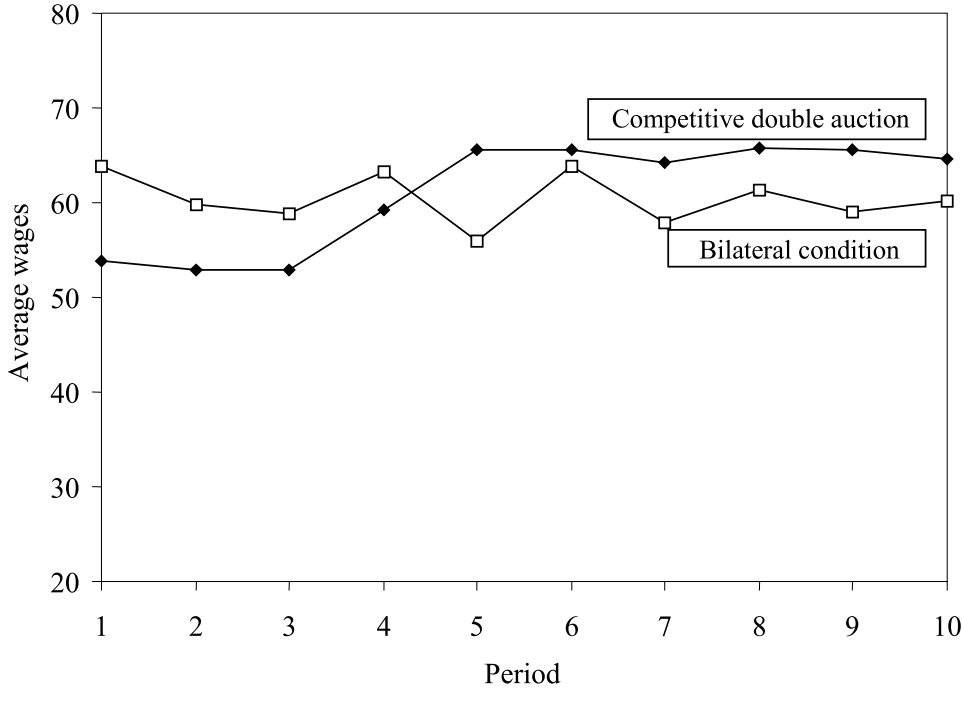

Fig. 2. *Wage Levels in the Competitive Double Auction and the Bilateral Condition*
Source: Fehr and Falk (1999)

The previous evidence indicates that reciprocity severely limits the impact of competition in markets in which effort or quality is not enforced exogenously. It restricts the impact of competition on wages by generating an efficiency wage effect that renders it profitable for the firms to pay non-competitive wages. As is well known such non-competitive wages may in turn cause involuntary unemployment (Akerlof, 1982). In addition, the existence of reciprocal types may endogenously generate a distinction between insiders and outsiders. Firms are, of course, interested in workers who do not exploit every opportunity to shirk, ie, workers who are loyal and who also perform when they are unobserved. Since workers are heterogeneous in this regard and since a worker's type may be difficult to find out, firms will, in general, be reluctant to replace existing workers by new workers even if the new workers would be willing to work for less than the going wage. This protects the existing workforce of firms from outside competition. Thus, reciprocity provides – in addition to the mechanisms discussed by Lindbeck and Snower (1988) – an independent reason for why insiders are protected from outside competition.

Finally, reciprocity may also contribute to the existence of non-competitive wage differentials. In the 1980s and the early 1990s there has been a heated debate about whether inter-industry wage differentials should be interpreted as non-competitive job rents. The debate did not result in a consensus because the results could also be interpreted as reflecting unobservable heterogeneity in working

conditions, and unobservable heterogeneity in skill levels.[11] Laboratory experiments can help resolve some of the open issues because in the laboratory it is possible to rule out heterogeneity in working conditions and skill levels. This was done by Fehr, Gächter and Kirchsteiger (1996) who embedded the gift exchange framework into a competitive market environment in which experimental firms differ according to their profit opportunities. Once a worker has accepted a firm's wage offer and before she makes her effort choice she is informed about the profit opportunity of the firm. This procedure ensures that only the effort decision but not the contract acceptance decision of the worker is affected by the firm's profit opportunity. Both firms and workers know this information revelation procedure in advance. The experiment shows that firms with better profit opportunities pay systematically higher wages and higher job rents. This wage policy is quite rational because for a more profitable firm a given effort increase leads to a larger profit increase. Hence, high-profit firms have a stronger incentive to appeal to the workers' reciprocity by paying high wages.

3. Cooperation

Free-riding incentives are a pervasive phenomenon in social life. Participation in collective actions against a dictatorship or in industrial disputes, collusion among firms in oligopolistic markets, the prevention of negative environmental externalities, workers' effort choices under team-based compensation schemes or the exploitation of a common resource are typical examples. In these cases the free-rider cannot be excluded from the benefits of collective actions or the public good although he does not contribute. In view of the ubiquity of cooperation problems in modern societies it is crucial to understand the forces shaping people's cooperation. In this section we will show that the neglect of reciprocity may induce economists to misunderstand the nature of many cooperation problems completely. As we will see a key to the understanding of cooperation problems is again the interaction between selfish and reciprocal types and how this interaction is shaped by the institutional environment.

3.1. *Conditional Cooperation*

Reciprocity changes the typical cooperation problem for two reasons. First, reciprocal subjects are willing to cooperate if they are sure that the other people who are involved in the cooperation problem will also cooperate. If the others cooperate – despite pecuniary incentives to the contrary – they provide a gift that induces reciprocal subjects to repay the gift, ie, reciprocators are conditionally cooperative. Second, reciprocal subjects are willing to punish free-riders because free-riders exploit the cooperators. Thus, if potential free-riders face reciprocators they have an incentive to cooperate to prevent being punished.

The impact of reciprocity on cooperation can be demonstrated in the context of the Prisoner's Dilemma (PD). To make things transparent consider the following

[11] For the severe difficulties created by unobservable heterogeneity in this context, see Murphy and Topel (1990) and Gibbons and Katz (1992).

situation. Subjects *A* and *B* both possess £10. They can either keep their £10 or they can transfer it to the other person. If they transfer the money the experimenter triples the transferred amount, ie, the recipient receives £30 from the transfer. *A* and *B* have to decide simultaneously whether they keep or whether they transfer the £10. If both subjects transfer their money both earn £30 while if they keep their money both earn only £10. Moreover, irrespective of whether the other subject transfers the money it is always in the self-interest of a subject to keep the £10.[12] The self-interest hypothesis predicts, therefore, that both subjects keep their money. In fact, however, many subjects cooperate in situations like this (see Ledyard, 1995; Dawes, 1980). For example, in one-shot PDs cooperation rates are frequently between 40 and 60%.

In the presence of sufficiently reciprocal subjects cooperative outcomes in the PD can be easily explained because the above game – although a PD in terms of material payoffs – is not a PD in utility payoffs. It is, instead, a coordination game with two pure strategy equilibria. If both subjects are reciprocators and if *A* believes that *B* will cooperate (ie transfer the money), *A* *prefers* to cooperate. The same holds true for *B* if *B* believes that *A* will cooperate. Thus, the strategy combination (cooperate, cooperate) constitutes an equilibrium. Likewise, if both believe that the other person will defect (ie, keep the money), they prefer to defect, too. Therefore, the combination (defect, defect) is also an equilibrium.[13] The fact that the PD in material terms is transformed into a coordination game in the presence of reciprocal players can explain two further facts. It has been shown dozens of times that communication leads to much higher cooperation rates in PDs and other social dilemma games (Sally, 1995).[14] If all subjects were completely selfish this impact of communication would be difficult to explain. If, however, the PD in material terms is in fact a coordination game, communication allows the subjects to coordinate on the superior equilibrium. It has also been shown that cooperation is affected by how the PD is framed. If the PD is framed in 'cooperative' terms, subjects are more likely to cooperate than if it is framed in 'competitive' terms. Since it is likely that the 'cooperative' frame induces more optimistic beliefs about the behaviour of the co-player than the 'competitive' frame subjects are more likely to coordinate on the good equilibrium in the 'cooperative' frame.

If it is indeed the case that in social dilemma situations like the PD the actual preferences of the subjects transform the game into a coordination game, the self-interest hypothesis induces economists to misperceive social dilemma problems fundamentally. In view of the importance of this claim it is, therefore, desirable to have more direct evidence on this. Fischbacher, Gächter and Fehr (2001) and Croson (2000) elicited subjects' willingness to cooperate conditional on the average cooperation of others in the context of 4-person public good games in

[12] This situation mimics a classic exchange problem in the absence of exogenous contract enforcement. *A* would like to have the good that *B* possesses because she values that good more than *B* does, and *vice versa*. Since *A* and *B* cannot write a contract that is enforced by third parties and since both have to send their goods simultaneously to the other person they have a strong incentive to cheat.

[13] For rigorous proofs that reciprocity (or inequity aversion) transform the PD in material terms into a coordination game see Section IV in Fehr and Schmidt (1999).

[14] Social dilemma games are generalised PD games in the following sense: There is a Pareto-superior cooperative outcome that renders everybody strictly better off relative to the Nash equilibrium.

which the dominant strategy for each subject was to free-ride completely. It was in the selfish interest of each subject to free-ride although the socially optimal decision required contributing the whole individual endowment (ie, 20 money units) to the public good. Both studies find considerable evidence for the presence of conditional cooperators. The results of the Fischbacher *et al.* study are presented in Fig. 3 below. The figure shows that 50% of the subjects are willing to increase their contributions to the public good if the other group members' average contribution increases although the pecuniary incentives always implied full free-riding. The behaviour of these subjects is consistent with models of reciprocity (or inequity aversion). The figure also reminds us that a substantial fraction of the subjects (30%) are complete free-riders while 14% exhibit a hump-shaped response. Yet, taken together there are sufficiently many conditional cooperators such that an increase in the other group members' contribution level causes an increase in the contribution of the 'average' individual (see bold line in Fig. 3).

The coexistence of conditional cooperators and selfish subjects has important implications. It implies, eg, that subtle institutional details may cause large behavioural effects. To illustrate this assume that a selfish and a reciprocal subject are matched in the *simultaneous* PD and that the subjects' type is common knowledge. Since the reciprocal type knows that the other player is selfish he knows that the other will always defect. Therefore, the reciprocal type will also defect, ie, (defect, defect) is the unique equilibrium. Now consider the *sequential* PD in which the selfish player first decides whether to cooperate or to defect. Then the reciprocal

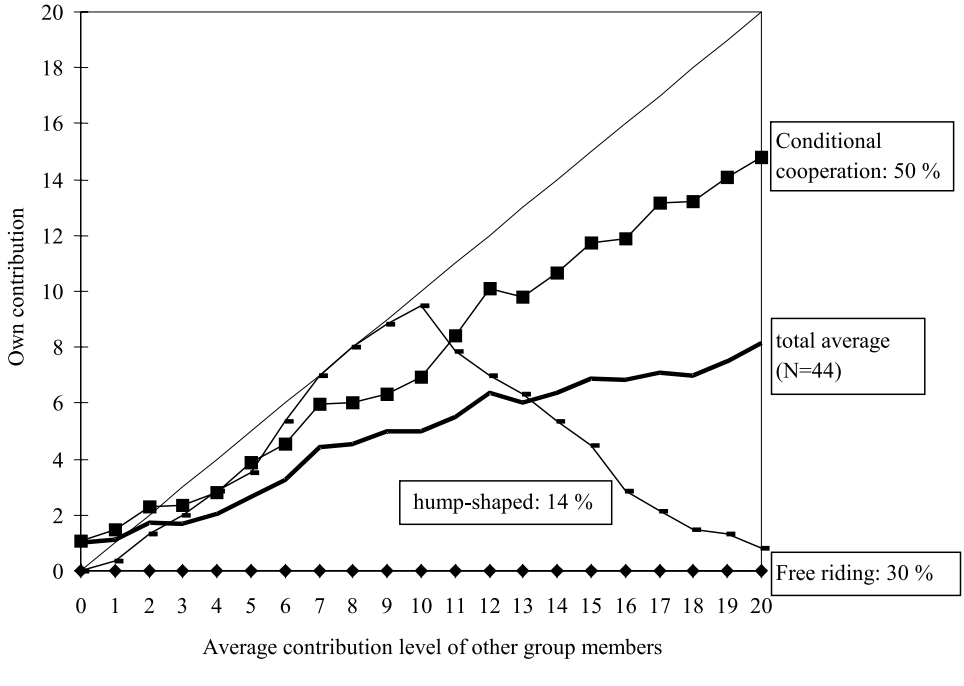

Fig. 3. *Contributions of Individual Subjects as a Function of Other Members' Average Contribution*
Source: Fischbacher, Gächter and Fehr (2001)?

player observes what the first-mover did and chooses his action. In the sequential case the unique equilibrium outcome is that both players cooperate because the reciprocal second-mover will match the choice of the first-mover. This means that the selfish first-mover essentially has the choice between the (cooperate, cooperate) outcome and the (defect, defect) outcome. Since mutual cooperation is better than mutual defection the selfish player will also cooperate. Thus, while in the simultaneous PD the selfish player induces the reciprocal player to defect, in the sequential PD the reciprocal player induces the selfish player to cooperate in equilibrium. This example neatly illustrates how institutional details interact in important ways with the heterogeneity of the population.

Since there are many conditional cooperators the problem of establishing and maintaining cooperation involves the management of people's beliefs. If people believe that the others cooperate to a large extent, cooperation will be higher compared to a situation where they believe that others rarely cooperate. Belief-dependent cooperation can be viewed as a social interaction effect that is relevant in many important domains. For example, if people believe that cheating on taxes, corruption, or abuses of the welfare state are wide-spread, they are themselves more likely to cheat on taxes and are more willing to take bribes or to abuse welfare state institutions. It is therefore important that public policy prevents the initial unravelling of civic duties because, once people start to believe that most others engage in unlawful behaviour, the belief-dependency of individuals' co-operation behaviour may render it very difficult to re-establish lawful behaviour.

In an organisational context the problem of establishing cooperation among the members of the organisation also involves the selection of the 'right' members. A few shirkers in a group of employees may quickly spoil the whole group. Bewley (1999), eg, reports that personnel managers use the possibility of firing workers mainly as a means to remove 'bad characters and incompetents' from the group and not as a threat to discipline the workers. The reason is that explicit threats create a hostile atmosphere and may even reduce the workers' generalised willingness to cooperate with the firm. Managers report that the employees themselves do not want to work together with lazy colleagues because these colleagues do not bear their share of the burden which is viewed as unfair. Therefore, the firing of lazy workers is mainly used to establish internal equity, and to prevent the unravelling of cooperation. This supports the view that conditional cooperation is also important inside firms.

Reciprocity and conditional cooperation are also likely to shape the structure of social policies that aim at helping the poor (Bowles and Gintis, 1998; Wax, 2000; Fong, 2001). The reason is that the political support for policies favouring the poor depends to a large extent on whether the poor are perceived as 'deserving' or as 'undeserving'. If people believe that the poor are poor because they do not *want* to work hard the support for policies that help the poor is weakened because the poor are perceived as undeserving. If, in contrast, people believe that the poor try hard to escape poverty but that for reasons beyond their control they could not make it, the poor are perceived as deserving. This indicates that the extent to which people perceive the poor as deserving is shaped by reciprocity. If the poor exhibit good intentions, ie, they try to contribute to society's output, or

if they are poor for reasons that have nothing to do with their intentions, they are perceived as deserving. In contrast, if the poor are perceived as lacking the will to contribute to society's output, they are perceived as undeserving. This means that social policies that enable the poor to demonstrate their willingness to reciprocate the generosity of society will mobilise greater political support than social policies that do not allow the poor to exhibit their good intentions. Wax (2000) convincingly argues that an important reason for the popularity of President Clinton's 1996 welfare reform initiative was that the initiative appealed to the reciprocity of the people.[15]

3.2. *Cooperation and Punishment*

We argued above that the presence of a selfish subject will induce the reciprocal subject in the simultaneous PD to defect, as well. This proposition also holds more generally in the case of *n*-person social dilemma situations. It can be shown theoretically that even a small minority of selfish subjects induces a majority of reciprocal (or inequity averse) subjects to free-ride in simultaneous social dilemmas (Fehr and Schmidt, 1999, proposition 4). In an experiment with anonymous interactions subjects do not of course know whether the other group members are selfish or reciprocal but if they interact repeatedly over time they may learn the others' types. Therefore, one would expect that over time cooperation will unravel in (finitely repeated) simultaneous social dilemma experiments. This unravelling of cooperation has indeed been observed in dozens of experiments (Ledyard, 1995; Fehr and Schmidt, 1999).

This raises the question of whether there are social mechanisms that can prevent the decay of cooperation. A potentially important mechanism is social ostracism and peer pressure stemming from reciprocal subjects. Recall that reciprocal subjects exhibit a willingness to punish unfair behaviour and it is quite likely that cooperating reciprocators view free-riding as very unfair. To examine the willingness to punish free-riders and the impact of punishment on cooperation Fehr and Gächter (2000*a*) introduced a punishment opportunity into a public goods game. In their game there are two stages. Stage one is the same public goods game as described in Fischbacher, Gächter and Fehr (2001). In particular, in the game at stage one the dominant strategy of each player is to free-ride completely although the socially optimal decision requires contributing the whole endowment to the public good. In stage two, after every player in the group has been informed about the contributions of each group member, each player can assign up to ten punishment points to each of the other group members. The assignment of one punishment point reduces the first-stage income of the punished subject, on the average, by three points but it also reduces the income of the punisher.[16] This kind of punishment mimics an angry

[15] The official title of Clinton's reform initiative – 'Personal Responsibility and Work Opportunity Reconciliation Act' – is telling in this regard.

[16] The written instructions for the subjects do not use value laden terms like, eg, punishment points. Instead the instructions are framed in neutral terms. For example, subjects do not assign 'punishment points' but just 'points' to the other players.

group member scolding a free-rider, or spreading the word so the free-rider is ostracised – there is some cost to the punisher, but a larger cost to the free-rider. Note that since punishment is costly for the punisher, the self-interest hypothesis predicts zero punishment. Moreover, since rational players will anticipate this, the self-interest hypothesis predicts no difference in the contribution behaviour between a public goods game without punishment and the game with a punishment opportunity. In both conditions zero contributions are predicted.

The experimental evidence completely rejects this prediction (see Fig. 4).[17] In contrast to the game without a punishment opportunity, where cooperation declines over time and is close to zero in the final period, the punishment opportunity causes a sharp jump in cooperation (compare period 10 with period 11 in Fig. 4). Moreover, in the punishment condition there is a steady increase in contributions until almost all subjects contribute their whole endowment. This sharp increase occurs because free-riders often get punished, and the less they give, the more likely punishment is. Cooperators feel that free-riders take unfair advantage of them and, as a consequence, they are willing to punish the free-riders. This induces the punished free-riders to increase cooperation in the following periods. A nice feature of this design is that the actual rate of punishment is very low in the last few periods – the mere threat of punishment, and the memory of its sting from past punishments is enough to induce potential free-riders to cooperate.

3.3. *Strategic versus Non-Strategic Punishment*

Peer pressure, social ostracism and, more generally, the cooperation-enhancing punishment of free-riders play a key role in the enforcement of social norms. They are also important in industrial disputes between workers and firms, in team production settings, in the management of common property resources or as an enforcement device for collusion in oligopolistic industries. For example, striking workers often ostracise strike breakers (Francis, 1985) or, under a piece rate system, the violators of production quotas are punished by those who try to maintain effort withholding norms (Roethlisberger and Dickson, 1947; Whyte, 1955).[18] During World War I British men who did not volunteer for the army faced strong public contempt and they were called 'whimps'. Ostrom *et al.* (1994) also report that punishment is frequently imposed on those who use common property resources excessively. They convincingly argue that the successful management of

[17] In the experiments subjects first participate in the game without a punishment opportunity for ten periods. After this they are told that a new experiment takes place. In the new experiment, which lasts again for ten periods, the punishment opportunity is implemented. In both conditions subjects remain in the same group for ten periods and they know that after ten periods the experiment will be over.

[18] Francis' (1985, p. 269) description of social ostracism in the communities of the British miners provides a particularly vivid example. During the 1984 strike of the miners, which lasted for several months, he observed the following: 'To isolate those who supported the ''scab union'', cinemas and shops were boycotted, there were expulsions from football teams, bands and choirs and ''scabs'' were compelled to sing on their own in their chapel services. ''Scabs'' witnessed their own ''death'' in communities which no longer accepted them.'

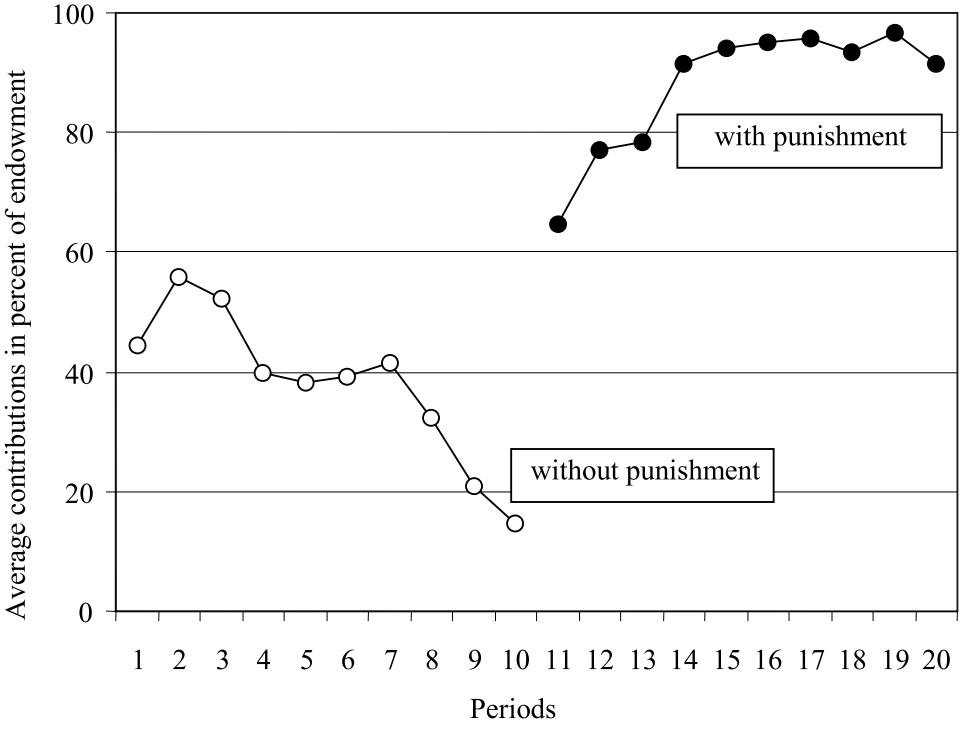

Fig. 4. *Average Contributions to the Public Good*
Source: Fehr and Gächter (2000*a*)

such resources requires institutions that render the excess extraction of common resources visible or easy to detect. This enables the users of the resource to impose sanctions on the wrongdoers.

A further interesting example is provided by Slade (1990) who analysed the behaviour of firms during price wars in oligopolistic industries. She shows that during price wars firms sell their products far below their marginal costs. While this behaviour may be rationalised as part of a complicated punishment strategy in a repeated game involving only self-interested players, it seems more likely that players get angry so that their punishment behaviour is driven by non-selfish forces. Anecdotal evidence from oil company marketers supports this view. According to Slade (personal communication) the marketers stated that they would follow a rival's price cut right down to zero if that rival started a price war. Yet, it is also clear that anecdotal evidence alone, as suggestive as it may be, is not fully convincing.

All of the examples above raise a similar question. To what extent is the punishment observed in the field strategically motivated, ie, caused by the expectation of future material benefit, and to what extent is it due to the mere (non-strategic) desire to punish. Moreover, what are the implications of the existence of non-strategic sanctions over and above what repeated game theory already tells us.

Our answer to these questions is as follows. *First*, repeated game theory tells us in fact very little about the actual behaviour in infinitely repeated interactions because for sufficiently high discount factors there is typically a plethora of equilibria,

including equilibria with no punishment and no cooperation. Thus, at a minimum, the results on punishment-based cooperation show that people do punish and they typically coordinate on cooperative outcomes. *Second*, there are in fact many situations in which interactions are only one-shot or finitely repeated or where people's discount factors are so low that self-interested agents cannot sustain cooperation in equilibrium. The results of Fehr and Gächter (2000*a*) show that in these situations non-strategic punishment is a powerful cooperation-enforcement device. *Third*, if fairness considerations are an important driving force of non-strategic sanctions it is quite likely that strategic sanctioning is also shaped in important ways by fairness concerns. In particular, we believe that many people will forgo the possibility of sanctioning others for purely pecuniary reasons if the sanction is viewed as unfair. They may refrain from sanctioning for intrinsic reasons or because they fear that the sanctioned player will retaliate.[19] Thus, unfair punishments are quite unlikely even if they yield material benefits while fair punishments will occur even if they cause a net decrease in the punisher's payoff. *Finally*, although it is true that due to the ambiguity of most field situations it is not possible to attribute the sanctions to non-pecuniary motives unambiguously, this does not mean that the sanctions are automatically driven by strategic reasons. In fact, we do not know of any rigorous evidence that free-riders are punished for strategic reasons.

The lack of evidence in favour of strategic sanctions led Falk *et al.* (2000*a*) to examine this question. They conducted a public goods experiment with a punishment opportunity in two conditions. In the partner condition three group members stay together for six periods. In the perfect stranger condition the game also lasts for six periods but it is ensured that nobody meets any of the other participants more than once. Thus, in the partner condition subjects can benefit in material terms from their punishments because the punished group members typically raise their contributions in the following periods while in the perfect stranger condition no such benefits can accrue. If there are more sanctions in the partner condition we have evidence in favour of strategic sanctions. The results of this experiment are displayed in Fig. 5.

The figure shows the sanctioning behaviour as a function of the deviation of the contribution of the sanctioned subject from the contribution of the sanctioning subject. It indicates that in the first five periods the sanctioning pattern as well as the strength of the sanctions is very similar in both conditions. The sanctions in the partner condition are only slightly stronger and the difference is not significant. Thus, the bulk of the sanctions already exists when there are no pecuniary benefits from sanctioning so that there is little or no evidence in favour of strategic sanctioning. Moreover, it turns out that in the final (sixth) period of the partner treatment the sanctions are even higher (although insignificantly so) than in the previous five periods of this treatment.[20] Since subjects know in advance that the

[19] Suppose we offer you £100 for hitting a stranger in the face. Even if the stranger had no possibility to hit back most people would probably reject this offer.

[20] A plausible reason for this is that if subjects cooperate successfully for five periods and then some group members try to cheat (free-ride) in the final period, the cooperators may be more angry than when they face free-riding in earlier periods. Being cheated by a 'friend' might make people angrier than being cheated by a 'stranger'.

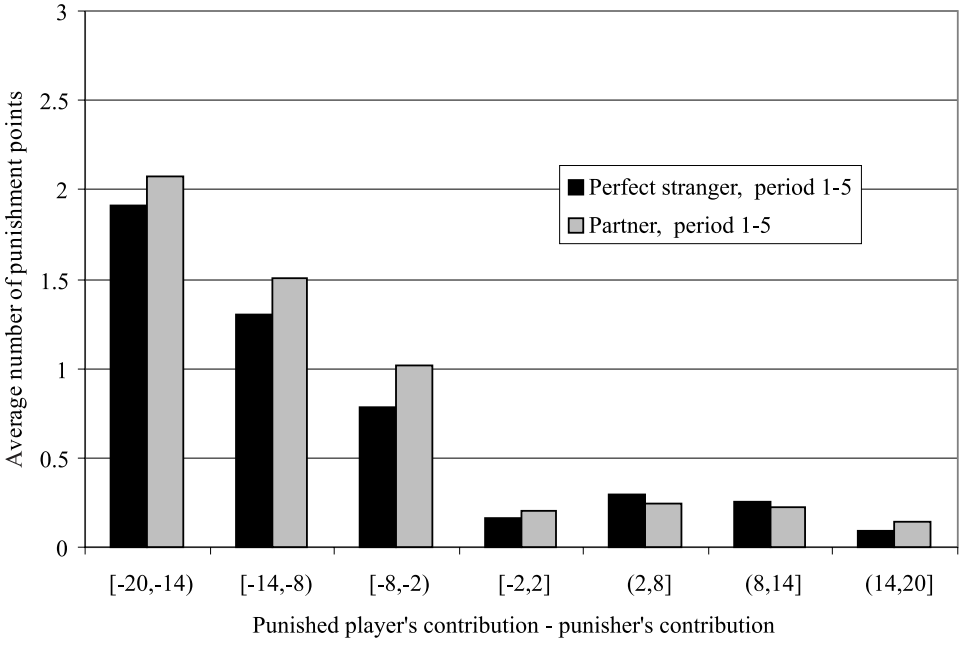

Fig. 5. *Punishment Pattern in the Public Goods Game – Partners versus Perfect Strangers*
Source: Falk *et al.* (2000*a*)

experiment ends after period six this result also indicates a lack of evidence in favour of strategic sanctions. Although we do not regard our experiment as the last word on this question this evidence should remind us that the mere fact that strategic punishments can be part of an equilibrium does not yet mean that strategic punishments will actually occur in the real world or in the laboratory.[21]

In view of the ubiquity of opportunities for free-riding the existence of a substantial amount of non-strategic punishment of free-riders is quite important. It suggests that even in one-shot situations or when the discount factor is low, collusive practices in output and labour markets are much more likely than predicted by the self-interest hypothesis. It also lends support to theories stressing the economic importance of social norms (Lindbeck, 1995; Lindbeck *et al.*, 1999, Dufwenberg and Lundholm, 2001) and to the insider–outsider theory of involuntary unemployment developed by Lindbeck and Snower (1988). This theory is based on the idea that the firm's existing workforce will harass outsiders and will not cooperate with them if the outsiders are employed below the going wage. Our

[21] There is an interesting difference between the ultimatum game experiments with reputation formation discussed in Section 1.2 above and the punishment of free-riders in the partner treatment. Recall that responders' acceptance thresholds were significantly higher in the reputation treatment of the ultimatum game relative to the baseline treatment. In the reputation treatment a responder could acquire an *individual* reputation for being a tough bargainer and he could reap the full benefits of his reputation. In the partner treatment of the public goods game the punishment of free-riders constitutes a second-order public good because all group members benefit from the cooperation-enhancing effect of the punishment. This may be one reason why we observe so little strategic punishment in the partner treatment.

evidence suggests that insiders will harass outsiders even if this is costly for the insiders and yields no material benefits for them.

4. Material Incentives and Property Rights

In this section we show that the neglect of reciprocal fairness prevents the understanding of crucial determinants and effects of material incentives. We will show, in particular, that material incentives may reduce efficiency in situations in which they are predicted to be efficiency-enhancing by the self-interest model. In addition, we show that reciprocity may have strong consequences for the optimal provision of incentives in a moral hazard context. Incentive contracts that are optimal when there are only selfish actors become inferior when some agents care for reciprocity. Conversely, contracts that are doomed to fail when there are only selfish actors provide powerful incentives and become superior when there are also reciprocal players. Finally we provide evidence indicating that reciprocity leads to collective property rights where the self-interest hypothesis predicts individual property rights.

4.1. *Material Incentives may be Harmful*

In the gift exchange game described in Section 1 there are no material incentives to provide non-minimal effort levels. Despite this many responders (workers) put forward non-minimal effort levels in the case of fair wage offers. In reality, material incentives are, of course, also used to induce workers to provide high effort. The question, therefore, arises how explicit performance incentives interact with motivations of fairness and reciprocity. One possibility is that reciprocity gives rise to extra effort on top of what is enforced by material incentives alone. However, it is also possible that explicit incentives may cause a hostile atmosphere of threat and distrust, which reduces any reciprocity-based extra effort. Bewley (1999, p. 431), for example, reports that many 'managers stress that punishment should seldom be used to obtain co-operation' because of the negative effects on work atmosphere.

In a series of experiments Fehr and Gächter (2000*b*) examine this possibility. They implement a baseline gift exchange game with a slight modification. In addition to the wage experimental employers also stipulate a desired effort level. However, the desired effort represents merely cheap talk so that it is not binding for the workers. This means that workers still face no material incentives in this treatment. In addition, they also implement a treatment with explicit performance incentives. This treatment keeps everything constant relative to the baseline treatment except that employers now have the possibility of stipulating a fine, to be paid by the worker to the employer in case of verified shirking. The probability of verification is given by 0.33 and the fine is restricted to an interval between zero and a maximum fine. The maximum fine is fixed at a level such that a selfish risk-neutral worker will choose an effort level of 0.4 when faced with this fine.[22]

[22] To prevent hostility being introduced merely by the use of value laden terms we avoided terms like 'fine', 'performance', etc. Instead we used a rather neutral language like, for example, 'price deduction'.

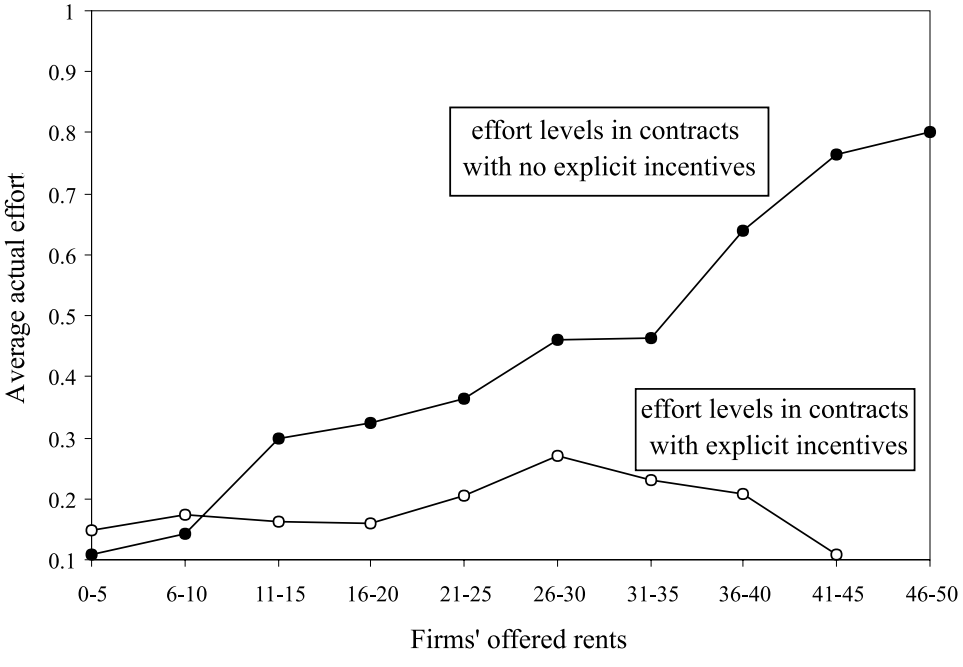

Fig. 6. *Average Effort Levels and Explicit Incentives*
Source: Fehr and Gächter (2000*b*)?

Fig. 6 presents the results of this experiment. The line with the black dots in Fig. 6 shows workers' effort behaviour in the baseline treatment. It depicts the average effort on the vertical axis as a function of the rent offered to the workers. The offered rent is implied by the original contract offer. It is defined as the wage minus the cost of providing the desired effort level. Due to the presence of many reciprocal workers the average effort level is strongly increasing in the offered rent and rises far above the selfish level of $e = 0.1$. The line with the white dots in Fig. 6 shows the relationship of rent to effort in the presence of the explicit performance incentive. Except at the low rent levels, the average effort is *lower* in the presence of the explicit incentives! This result suggests that reciprocity-based effort elicitation and explicit performance incentives may indeed be in conflict with each other. Performance incentives that are perceived as hostile cause hostile responses by the workers. In the context of our incentive treatment this meant that reciprocal workers were no longer willing to provide non-minimal effort levels.[23] Similarly counter productive incentive effects are reported in the studies of Bohnet *et al.* (2001); Benz *et al.* (2001); Evans *et al.* (2001) and Schulze and Frank (2001).

In the experiments of Fehr and Gächter (2000*b*) the average effort taken over all trades, and hence the aggregate monetary surplus, is lower in the incentive treatment than in the baseline treatment. However, employers' profits are higher because in the incentive treatment they rely much less on the 'carrot' of generous

[23] Note that according to this interpretation there is no crowding out of an intrinsic (reciprocal) motivation here. Instead, the preference for reciprocity implies that workers respond in a hostile manner to incentives that are perceived as hostile.

wage offers. Instead, they threaten the workers with the maximal fine in most cases. For the employers the savings in wage costs more than offset the reductions in revenues that are caused by the lower effort in the incentive treatment. However, while the wage savings merely represent a transfer from the workers to the firms, the reduction in effort levels reduces the aggregate surplus. This shows that in the presence of reciprocal types efficiency questions and questions of distribution are inseparable. Since the perceived fairness of the distribution of the gains from trade affects the effort behaviour of the reciprocal types different distributions are associated with different levels of the aggregate gains. Thus, lump-sum transfers between trading parties have efficiency consequences.

4.2. *Reciprocity-Based Incentives versus Explicit Incentives*

Standard principal–agent models predict that contracts should be made contingent on all verifiable measures that are informative with regard to the agent's effort. But in reality, we often observe highly incomplete contracts. For example, as noted earlier, wages are often paid without explicit performance incentives. To this point, the discussion has focused on demonstrating that reciprocity has powerful economic effects in situations where explicit incentives are absent. This section seeks to explore underlying causes for the absence of explicit incentives. Reciprocity plays a twofold role in this context. First, as the previous experiment has shown, certain kinds of explicit incentives have negative side effects because they reduce reciprocity-based voluntary cooperation. Second, it renders contracts that do not rely on explicit incentives more efficient relative to the prediction of the self-interest model because reciprocity itself constitutes a powerful contract enforcement device. Each of these two reasons may induce the principals to prefer contracts without explicit incentives.

To study the impact of reciprocity on contractual choices, Fehr, Klein and Schmidt (2001) conducted an experiment in which principals had the choice between an explicit incentive contract and an implicit contract without explicit incentives. In a typical session of this experiment there are 12 principals and 12 agents who play for ten periods. In each of the ten periods an agent faces a different principal, which ensures that all matches are one-shot. A period consists of three stages. At stage one of a period, the principal has to decide whether to offer the agent an implicit or an explicit contract. The implicit contract specifies a fixed wage and a desired effort level (where effort choices can range from 1 to 10). In addition, the principal can promise a bonus that may be paid after the actual effort has been observed. In the implicit contract, there is no contractual obligation to pay the announced bonus, nor is the agent obliged to choose the desired effort level. The principal is, however, committed to pay the wage. An explicit contract also specifies a binding fixed wage and a desired effort level between 1 and 10. Here, however, the principal can impose a fine on the agent that has to be paid to the principal in the case of verified shirking. Except for one detail the explicit contract is identical to the performance contract discussed in the previous section. The difference concerns the fact that the choice of the explicit contract involves fixed verification cost of 10 units. This reflects the fact that the verification

of effort is, in general, costly. Note that the implicit contract does not require third-party verification of effort. It is only necessary that effort is observable by the principal.[24]

At stage two, the agent observes which contract has been offered and decides whether to accept or reject the offer. If the agent rejects the offer, the game ends and both parties get a payoff of zero. If the agent accepts, the next step is for the agent to choose the actual level of effort.

At stage three, the principal observes the actual effort level. If the principal has offered an implicit contract, the next decision is whether the agent should be awarded the bonus payment. If the principal offered an explicit contract and if the agent's effort falls short of the agreed effort level, a random draw decides with probability 1/3 whether shirking is verifiable, in which case the agent has to pay the fine.

If all players have purely selfish preferences, the analysis of this game is straightforward. A selfish principal would never pay a bonus. Anticipating this, there is no incentive for the agent to spend more than the minimum effort. If the principal chooses the explicit contract, the principal should go for the maximum punishment because this is the best deterrence for potential shirkers. The parameters of the experiment are chosen such that a risk neutral and selfish agent maximises expected utility by choosing an effort level of 4 if faced with the maximum fine. Since the enforceable effort level is 4 under the explicit contract while it is only 1 under an implicit contract, the self-interest model predicts that principals prefer the explicit contract.

The experimental evidence is completely at odds with these predictions. In total, the implicit contract was chosen in 88% of the cases. In view of the relative profitability of the different contracts, the popularity of the implicit contract is not surprising. Those principals choosing the explicit contract made an average loss of 9 tokens per contract, while those preferring the implicit contract made an average profit of 26 tokens per contract. Since the fixed verification cost in the explicit contract was 10 tokens, the explicit contract would have been much less profitable even in the absence of these costs. For both contracts the average income of the agents was roughly 18 tokens. Implicit contracts were more profitable because – contrary to the standard prediction – they induced much higher effort levels. The effort level in the implicit contract was 5.2 on the average (on a scale of 1 to 10), while the effort level in the explicit contract was 2.1 on the average.

How did implicit contracts induce so much higher effort levels than predicted? A major reason is that in the presence of reciprocal principals, the promised bonus does not merely represent cheap talk, because reciprocal principals can – and actually do – condition the bonus payment on the effort level. The average data clearly reflect this impact of the reciprocal types because the actual average bonus rises steeply with the actual effort level. The principal's capability to commit himself to paying a conditional bonus is based on his reciprocal inclinations. Conditional bonus payments, in turn, provide a strong pecuniary incentive for the agents to

[24] Employers are, in general, not free to cut a worker's wage for shirking while they have few legal problems when they refuse to pay a promised bonus.

perform as desired by the principals. Why did explicit contracts induce lower effort levels than predicted? A likely reason is that these contracts are perceived as hostile and even induce negative reciprocity, as shown in the previous section.

One might also conjecture that the preference for implicit contracts in this particular experiment is solely caused by the fact that the explicit contract involves a punishment while the implicit contract involves a reward. Further experiments by Fehr, Klein and Schmidt (2001), however, cast doubt on this explanation. If the above described implicit contract competes with a piece rate contract the vast majority of principals still prefer the implicit contract.

4.3. *Individual versus Joint Ownership*

The impact of preferences for reciprocity on contractual choices suggests that reciprocity may not only cause substantial changes in the functioning of given economic institutions but that it may also have a powerful impact on the selection and formation of institutions. Recent work by Ellingsen and Johannesson (2000) and by Sonnemans *et al.* (2001) indicates, eg, that such preferences have a decisive impact on the efficiency of different property rights and bargaining institutions. This impact on efficiency is also likely to affect the selection of the institutions. To see this in more detail consider the present theory of property rights (Hart, 1995). This theory predicts that joint ownership will in general severely inhibit relations-specific investments so that it emerges only under very restrictive conditions. This may no longer be true in the presence of reciprocal actors who are willing to cooperate if they expect the trading partner to cooperate as well, and who are willing to punish even at a cost to themselves.

To illustrate this point, consider two parties, *A* and *B*, who are engaged in a joint project (a 'firm') to which they have to make some relationship-specific investments today in order to generate a joint surplus in the future. An important question that has received considerable attention in recent years is who should own the firm. In a seminal paper, Grossman and Hart (1986) argue that ownership rights allocate residual rights of control to the physical assets that are required to generate the surplus. For example, if *A* owns the firm, then he will have a stronger bargaining position than *B* in the renegotiation game in which the surplus between the two parties is shared *ex post*, because he can exclude *B* from using the assets which makes *B*'s relationship-specific investment less productive. Grossman and Hart show that there is no ownership structure that implements first best investments, but some ownership structures do better than others and there is a unique second best optimal allocation of ownership rights. They also show that joint ownership is, in general, not optimal.

This result is at odds with the fact that there are many jointly owned companies, partnerships or joint ventures. Furthermore, the argument neglects that reciprocal fairness may be an important enforcement mechanism to induce the involved parties to invest more under joint ownership than otherwise predicted. In order to test this hypothesis, Fehr, Kremhelmer and Schmidt (2001) conducted a series of experiments on the optimal allocation of ownership rights. The experimental game is a simplified version of Grossman and Hart (1986): there are two parties, *A*

and B, who have to make investments, a, $b \in \{1, \ldots, 10\}$, respectively, in order to generate a joint surplus $v(a,b)$. Investments are sequential: B has to invest first, her investment level b is observed by A, who has to invest thereafter. We consider two possible ownership structures: Under A-ownership, A hires B as an employee and pays her a fixed wage w. In this case monetary payoffs are $v(a,b)-w-a$ for A and $w-b$ for B. Under joint ownership, each party gets half of the gross surplus minus her investment cost, ie $0.5v(a,b)-a$ for A and $0.5v(a,b)-b$ for B. The gross profit function has been chosen such that maximal investments are efficient, ie $a^{FB} = b^{FB} = 10$, but if each party gets only 50% of the marginal return of their investments, then it is a dominant strategy for a purely self-interested player to choose the minimum level of investment, $a = b = 1$. Finally, in the first stage of the game, A can decide whether to be the sole owner of the firm and make a wage offer to B, or whether to have joint ownership.

The prediction of the self-interest model is straightforward. Under A-ownership B has no incentive to invest and will choose $b = 1$. On the other hand, A is the residual claimant, so she will invest efficiently. Under joint ownership each party gets only 50% of the marginal return which is not sufficient to induce any investments. Hence in this case B's optimal investment level is unchanged, but A's investment level is reduced to $a = 1$. Thus, A-ownership outperforms joint ownership and A should hire B as an employee.

In the experiments just the opposite happens. Party A chooses joint ownership in more than 80% (187 out of 230) of all observations and gives away 50% of the gross return to B. With joint ownership B-players choose on the average an investment level of 8.9 and A responds with an investment of 6.5 (on the average). On the other hand, if A-ownership is chosen and A hires B as an employee, B's average investment is only 1.3, while all A-players choose an investment level of 10. Furthermore A-players earn much more on the average if they choose joint ownership rather than A-ownership.

These results are inconsistent with the self-interest model, but it is straightforward to explain them with concerns for reciprocity. Under joint ownership the investments are associated with positive externalities and, hence, joint ownership favours positively reciprocal behaviour. If, under joint ownership, B expects A to behave reciprocally even a selfish B-player has a strong incentive to make high investments because this induces the reciprocal As to invest, too. Under A-ownership the incentives for B are quite different because B does not benefit from the investments of A. Hence, the selfish Bs choose the minimal investment under A-ownership. If there is sufficient complementarity between the investments of A and B, the joint surplus is therefore much higher under joint ownership. This makes it profitable for A to choose joint ownership.

5. The Modelling of Reciprocity

The evidence in Sections 2, 3 and 4 indicates that reciprocity has a deep impact on fundamental economic issues. It is an important behavioural force that shapes the functioning of competition, that governs the laws of cooperation, and that also has a decisive impact on how incentives work. Reciprocity creates implicit incentives

and renders some explicit incentives quite inefficient. By changing the incentives for the selfish types it also affects the prevailing interaction patterns and the constraints on individual behaviour, ie, the prevailing contracts and institutions, relative to a world that is exclusively populated by self-interested people.

We believe that – in view of the importance of reciprocity – mainstream economics has much to gain by routinely incorporating concerns for reciprocity into the analysis. This means that – when analysing an economic problem – one should routinely try to derive the implications of the assumption that, in addition to the purely self-interested types, roughly 40–50% of the people exhibit reciprocal preferences. It is obvious that to achieve this a precise mathematical model of reciprocal preferences is desirable. In the past few years several authors have worked on models of reciprocity (Rabin, 1993; Levine, 1998; Dufwenberg and Kirchsteiger, 1999; Falk and Fischbacher, 1999; Segal and Sobel, 1999; Charness and Rabin, 2000). These papers are very useful because they sharpen the notion of reciprocity. However, they also show that it is extremely difficult to build simple and tractable models of reciprocity. The problem is that the explicit modelling of intention-based or type-based reciprocity quickly renders these models very complex and difficult to handle.

The first best solution to the modelling problem would surely be a simple and tractable model of reciprocity. However, since this solution is not available, at least at present, there is also a need for simpler models that mimic the outcomes of reciprocity models in a wide variety of circumstances but that do not explicitly model reciprocity. Models like this have been developed by Fehr and Schmidt (1999) and Bolton and Ockenfels (2000). They are based on the assumption that 'fair' types dislike an inequitable distribution of material resources. The impressive feature of these models is that – although they are much simpler than reciprocity models – they correctly predict the outcome of experiments in a wide variety of games. The model of Fehr and Schmidt, eg, is consistent with the stylised facts in the bilateral ultimatum and gift exchange game reviewed in Section 1, with the market games under exogenous and endogenous contract enforcement presented in Section 2, with the cooperation games with and without a punishment opportunity described in Section 3, and with the contract choice and property rights experiments presented in Section 4. This suggests that in many instances it is possible to capture the behavioural predictions of reciprocity with simpler models of fairness.[25]

However, the black-boxing of reciprocity via simple fairness models must be done with a background knowledge about the limits of these models. Mindless application of these models may lead to wrong predictions as can be illustrated by the experiments of Brandts and Sola (forthcoming) and Falk *et al.* (forthcoming). In the Falk *et al.* paper the rejection rates of the (8/2)-offer (8 for the proposer and

[25] Following Bhaskar (1990), Driscoll and Holden (2001) provide an interesting application of a simple theory of social preferences to an important macroeconomic problem. In the spirit of the models of Loewenstein *et al.* (1989) and Fehr and Schmidt (1999) they assume that workers care disproportionately more about being paid less than other identical workers than they care about being paid more than them. Their model neatly explains inflation persistence which has puzzled macroeconomists for a long time.

2 for the responder) in four different mini-ultimatum games are compared. The games differ only with regard to the available alternative to the (8/2)-offer. In one game the alternative was (5/5), in the second game it was (2/8), in the third game it was (8/2) and in the last game it was (10/0). Note that if the responder cares only about the distribution of payoffs the rejection rate of the (8/2)-offer should be the same in all four games. In fact, however, the rejection rate is monotonically declining across the four games. It is highest in the (5/5)-game, where (5/5) was the alternative, and lowest in the (10/0)-game. A plausible interpretation of this result is that in the (5/5)-game an offer of (8/2) indicates an unfair intention or an unfair type while in the (10/0)-game this is not the case. Thus, if responders punish unfair intentions or unfair types they should exhibit a higher rejection rate in the (5/5)-game. This example indicates that if the set of feasible alternatives changes across situations such that the possibilities for expressing good or bad intentions changes, simple fairness models do not capture important aspects of behaviour. It is, however, interesting that even in these situations a simple model can be useful because the prediction of the model provides hints when intention- or type-based reciprocity is likely to matter. The prediction thus alerts the researcher about the limit of the model. For instance, if (5/5) instead of (10/0) is the alternative to (8/2), the Fehr–Schmidt model predicts that, for reasonable rejection rates, the population of proposers who make the (8/2)-offer is less fair. Thus responders will make different inferences about the type or the intention of the population who made the (8/2)-offer when (5/5) is the alternative compared to when (10/0) is the alternative. This inference induces reciprocal responders to reject the (8/2)-offer more frequently when (5/5) is the alternative.[26]

It is also important to keep in mind that models that have been developed to explain a diverse set of facts in laboratory experiments must be used with care and need perhaps some adaptations when applied to field situations. For example, in the field it is often not possible to determine the relevant reference agents without further empirical analysis while in an experiment the set of the other players in the group is often a good first approximation. Likewise, it does not seem likely that the effort-relevant fairness judgements of a worker are based on a comparison between the worker's income and the income of the top managers of the firm. Instead, the behaviourally relevant comparisons tend to be more local, ie, comparisons with their co-workers or comparisons with the average value of the output they generate.

6. Concluding Remarks

The self-interest hypothesis assumes that all people are exclusively motivated by their material self-interest. This hypothesis is a convenient simplification and there are, no doubt, situations in which almost all people behave as if they were strictly self-

[26] We provide an explanation of the different rejection rates in the Falk *et al.* paper along these lines in the appendix of the working paper version of this paper (http://www.iew.unizh.ch/wp/iewwp084.pdf). Our explanation is based on a modified Fehr–Schmidt model of inequity aversion in which the disutility from disadvantageous inequality is lower if a person faces a subject with a high preference against advantageous inequality. This basically boils down to a type-based model of reciprocity. The model by Falk and Fischbacher (1999) can also explain this evidence.

interested. In particular, for comparative static predictions of aggregate behaviour self-interest models may make empirically correct predictions because models with more complex motivational assumptions predict the same outcome. However, the evidence presented in this paper also shows that fundamental issues in economics cannot be understood on the basis of the self-interest model. The evidence indicates that concerns for fairness and reciprocity are important for bilateral negotiations, for the functioning of markets and incentives, for the structure of property rights and contracts and for the laws governing collective action and cooperation.

Many influential economists, including Smith (1759), Becker (1974), Arrow (1981), North (1990), Samuelson (1993) and Sen (1995), pointed out that people often do care for the well-being of others and that this may have important economic consequences. Yet, so far, these opinions did not have a strong impact on mainstream economics. We believe that this has to do with a strong convention in economics of not explaining puzzling observations by changing assumptions on preferences. Changing preferences is said to open Pandora's box because everything can be explained by assuming the 'right' preferences. In view of this convention it is important to stress that what is proposed here is not an arbitrary and empirically unfounded change in the preference assumptions. The goal of the simple models of fairness discussed above was not to invent a new motive for each different game but to explain a large and diverse set of facts with the same motivational assumptions. Moreover, due to the development of experimental techniques Pandora's box is kept closed because these tools enable us to examine the nature of preferences in a scientifically rigorous way. In fact, there has been much progress and fascinating new insights into the nature of fairness preferences during the past decade. While there is still much to be done this research clearly shows that it is possible to discriminate between theories based on different preference assumptions (see, eg, Falk *et al.* (2000*a*), Charness and Rabin (2000) and Fehr and Schmidt (2000) for surveys). Therefore, in view of the facts, the new theoretical developments, the importance of concerns for reciprocity in many economic domains, and in view of the existence of rigorous experimental techniques that allow us to examine hitherto unsolvable problems in a scientific manner, we believe that it is time to recognise that a substantial fraction of the people is also motivated by reciprocity. People differ not only in their tastes for chocolate and bananas but also along a more fundamental dimension. They differ with regard to their inclination to behave in a selfish or reciprocal manner, and this does have important economic consequences.

University of Zurich

References

Agell, Jonas and Lundborg, Per (1995). 'Theories of pay and unemployment: survey evidence from Swedish manufacturing firms', *Scandinavian Journal of Economics*, vol. 97, pp. 295–308.

Akerlof, George (1982). 'Labor contracts as partial gift exchange', *Quarterly Journal of Economics*, vol. 97, pp. 543–69.

Andreoni, James (1989). 'Giving with impure altruism: applications to charity and Ricardian equivalence', *Journal of Political Economy*, vol. 97, pp. 1447–58.

Andreoni, James and Miller, John (forthcoming). 'Giving according to GARP: an experimental test of the rationality of altruism', *Econometrica*.
Arrow, Kenneth J. (1981). 'Optimal and voluntary income redistribution', in (Steven Rosenfield, ed.), *Economic Welfare and the Economics of Soviet Socialism: Essays in Honor of Abram Bergson*, Cambridge: Cambridge University Press.
Becker, Gary S. (1974). 'A theory of social interactions', *Journal of Political Economy*, vol. 82, pp. 1063–93.
Benz, M., Fehr, E. and Frey, B. (2001). Multitasking and explicit incentives, Working Paper, University of Zürich.
Bewley, Truman (1999). *Why Wages Don't Fall During a Recession*, Harvard: Harvard University Press.
Bhaskar, V. (1990). 'Wage Relativities and the natural range of unemployment', ECONOMIC JOURNAL 100, 60–6.
Binmore, Ken (1998). *Game Theory and the Social Contract: Just Playing*, Cambridge, MA: MIT Press.
Bohnet, Iris, Frey, Bruno and Huck, Steffen (2001). More order with less law: on contract enforcement, trust and crowding. *American Political Science Review*, vol. 95(1), pp. 131–44.
Bolton, Gary E. and Ockenfels, Axel (2000). 'A theory of equity, reciprocity and competition', *American Economic Review*, vol. 100, pp. 166–93.
Bowles, Samuel, and Gintis, Herbert (1998). 'Is equity passé', *Boston Review*, vol. 23(6), December/January, 1998–9.
Brandts, Jordi, and Charness, Gary (1999). 'Gift-exchange with excess supply and excess demand', mimeo, Pompeu Fabra, Barcelona.
Brandts, Jordi, and Sola, Carles (forthcoming). 'Reference points and negative reciprocity in simple sequential games', *Games and Economic Behavior*.
Camerer, Colin F. (in press). *Behavioral Game Theory*, Princeton: Princeton University Press.
Campbell, Carl M. and Kamlani, Kunal S. (1997). 'The reasons for wage rigidity: evidence from a survey of firms', *Quarterly Journal of Economics*, vol. 112, pp. 759–89.
Charness, Gary (1996). 'Attribution and reciprocity in a labor market: an experimental investigation', mimeo, University of California at Berkeley.
Charness, Gary (2000). 'Responsibility and effort in an experimental labor market', *Journal of Economic Behavior and Organization*, vol. 42, pp. 375–84.
Charness, Gary, and Rabin, Matthew (2000). 'Social preferences: some simple tests and a new model', mimeo, University of California at Berkeley.
Cox, Jim (2000). 'Trust and reciprocity: implications of game triads and social contexts', mimeo, University of Arizona at Tucson.
Cox, Jim, Sadiraj Klarita and Sadiraj, Vjollca (2001). 'A theory of fairness and competition without inequality aversion', mimeo.
Croson, Rachel (2000). 'Theories of altruism and reciprocity: evidence from linear public goods games', Discussion Paper, Wharton School, University of Pennsylvania.
Davis, Douglas D. and Holt, Charles A. (1993). *Experimental Economics*, Princeton: Princeton University Press.
Dawes, Robin (1980). 'Social dilemmas', *Annual Review of Psychology*, vol. 31, pp. 169–93.
Driscoll, John C. and Steiner Holden (2001). 'Fair Treatment and Inflation Persistence', Working Paper, University of Oslo.
Dufwenberg, Martin and Lundholm, Martin (2001). 'Social norms and moral hazard', ECONOMIC JOURNAL, vol. 111, pp. 506–25.
Dufwenberg, Martin and Kirchsteiger, Georg (1999). 'A theory of sequential reciprocity', Discussion Paper. Center, Tilburg University.
Ellingsen, Tore and Johannesson, Magnus (2000). 'Is there a hold up problem?' Stockholm School of Economics, Working Paper Series in Economics and Finance, No. 357.
Ellingsen, Tore and Johannesson, Magnus (2001). 'Sunk costs, fairness and disagreement', Mimeo Stockholm School of Economics.
Evans, J.H. Hannan, R.L. Krishnan, R. and Moser, D.V. (2001). 'Honesty in Managerial Reporting', *The Accounting Review*, vol. 76, 4, forthcoming.
Falk, Armin, Fehr, Ernst and Fischbacher, Urs (2000*a*). 'Informal sanctions', Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 59.
Falk, Armin, Fehr, Ernst and Fischbacher, Urs (2000*b*). 'Testing theories of fairness – intentions matter', Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 63.
Falk, Armin, Fehr, Ernst and Fischbacher, Urs (forthcoming). 'On the nature of fair behaviour', *Economic Inquiry*.
Falk, Armin and Fischbacher, Urs (1999). 'A theory of reciprocity', Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 6.
Falk, Armin, Gächter, Simon and Kovács, Judith (1999). 'Intrinsic motivation and extrinsic incentives in a repeated game with incomplete contracts', *Journal of Economic Psychology*, vol. 20 (3), pp. 251–84.

Fehr, Ernst and Falk, Armi (1999). 'Wage rigidity in a competitive incomplete contract market', *Journal of Political Economy*, vol. 107, pp. 106–34.
Fehr, Ernst and Fischbacher, Urs (2001). 'Reputation and retaliation', mimeo, University of Zürich.
Fehr, Ernst and Gächter, Simon (2000*a*). 'Cooperation and punishment in public goods experiments', *American Economic Review*, vol. 90, pp. 980–94.
Fehr, Ernst and Gächter, Simon (2000*b*). 'Do incentive contracts crowd out voluntary contribution?' Working paper 34, Institute for Empirical Research in Economics, University of Zurich.
Fehr, Ernst, Gächter, Simon and Kirchsteiger, Georg (1996). 'Reciprocal fairness and noncompensating wage differentials', *Journal of Institutional and Theoretical Economics*, vol. 152 (4), pp. 608–40.
Fehr, Ernst, Kirchsteiger, Georg and Riedl, Arno (1993). 'Does fairness prevent market clearing? An experimental investigation', *Quarterly Journal of Economics*, vol. 108, pp. 437–60.
Fehr, Ernst, Kirchsteiger, Georg and Riedl, Arno (1998). 'Gift exchange and reciprocity in competitive experimental markets', *European Economic Review*, vol. 42, pp. 1–34.
Fehr, Ernst, Klein, Alexander and Schmidt, Klaus M. (2001). 'Endogenous incomplete contracts', Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 72.
Fehr, Ernst, Kremhelmer, Susanne and Schmidt, Klaus (2001). 'Fairness and the optimal allocation of property rights', mimeo, University of Munich.
Fehr, Ernst and Schmidt, Klaus M. (1999). 'A theory of fairness, competition and co-operation', *Quarterly Journal of Economics*, vol. 114, pp. 817–68.
Fehr, Ernst and Schmidt, Klaus M. (2000). 'Theories of fairness and reciprocity – evidence and economic applications', forthcoming in (M. Dewatripont, L. P. Hansen, S. Turnovsky), *Advances in Economic Theory, Eighth World Congress of the Econometric Society*, Cambridge: Cambridge University Press.
Fehr, Ernst and Tougareva, Elena (1995). 'Do high monetary stakes remove reciprocal fairness? Experimental evidence from Russia', mimeo. Institute for Empirical Economic Research, University of Zurich.
Fischbacher, Urs, Fong, Christina and Fehr, Ernst (2001). 'Competition and fairness', mimeo. Institute for Empirical Economic Research, University of Zurich.
Fischbacher, Urs, Gächter Simon and Fehr, Ernst (2001). 'Are people conditionally cooperative? Evidence from a public goods experiment', *Economics Letters*, vol. 71, pp. 397–404.
Fong, Christina (2001). ''Social preferences, self-interest, and the demand for redistribution', *Journal of Public Economics*, vol. 82, pp. 225–46.
Forsythe, Robert L., Horowitz, Joel, Savin, N. E. and Sefton, Martin (1994). 'Fairness in simple bargaining games', *Games and Economic Behavior*, vol. 6, pp. 347–69.
Francis, Hywel (1985). 'The law, oral tradition and the mining community', *Journal of Law and Society*, vol. 12, pp. 267–71.
Frey, Bruno S. and Pommerehne, Werner W. (1993). 'On the fairness of pricing – an empirical survey among the general population', *Journal of Economic Behavior and Organization*, vol. 20, pp. 295–307.
Frey, Bruno and Weck-Hannemann, Hannelore (1984). 'The hidden economy as an ''unobserved'' variable', *European Economic Review*, vol. 26, pp. 33–53.
Gächter, Simon and Falk, Armin (forthcoming). 'Reputation and reciprocity: consequences for the labour relation', *Scandinavian Journal of Economics*.
Gibbons, Robert and Katz, Lawrence (1992). 'Does unmeasured ability explain inter-industry wage differentials?' *Review of Economic Studies*, vol. 59, pp. 515–35.
Greenberg, Jerald (1990). 'Employee theft as a reaction to underpayment inequity: the hidden cost of pay cuts', *Journal of Applied Psychology*, vol. 75, pp. 561–68.
Grossman, Sanford and Hart, Oliver (1986). 'The costs and benefits of ownership: a theory of vertical and lateral integration', *Journal of Political Economy*, vol. 94(1), pp. 691–719.
Güth, Werner, Schmittberger, Rolf and Schwarze, Bernd (1982). 'An experimental analysis of ultimatium bargaining', *Journal of Economic Behavior and Organization*, vol. 3, pp. 367–88.
Hannan, Lynn, Kagel, John and Moser, Donald (forthcoming). 'Partial gift exchange in experimental labor markets: impact of subject population differences, productivity differences and effort requests on behavior', *Journal of Labor Economics*.
Hart, Oliver (1995). *Firms, Contracts, and Financial Structure*. Oxford: Clarendon Press.
Kagel, John and Wolfe, Katherine (2000). 'Tests of difference aversion to explain anomalies in simple bargaining games', mimeo. Ohio State University.
Kahneman, Daniel, Knetsch, Jack L. and Thaler, Richard (1986). 'Fairness as a constraint on profit seeking: entitlements in the market', *American Economic Review*, vol. 76, pp. 728–41.
Kirchsteiger, Georg (1994). 'The role of envy in ultimatum games', *Journal of Economic Behavior and Organization*, vol. 25, pp. 373–89.
Ledyard, John (1995). 'Public goods: a survey of experimental research', Ch. 2 in (Alvin Roth and John Kagel, eds.), *Handbook of Experimental Economics*. Princeton: Princeton University Press.

Levine, David (1998). 'Modeling altruism and spitefulness in experiments', *Review of Economic Dynamics*, vol. 1, pp. 593–622.
Lind, Allan and Tyler, Tom (1988). *The Social Psychology of Procedural Justice*. New York and London: Plenum Press.
Lindbeck, Assar and Snower, Dennis J. (1988). 'Cooperation, harassment, and involuntary unemployment: an insider–outsider approach', *American Economic Review*, vol. 78 (1), pp. 167–89.
Lindbeck, Assar (1995). 'Welfare-State Disincentives with Endogenous Habits and Norms'. *Scandinavian Journal of Economics*, vol. 97(4), pp. 477–94.
Lindbeck, Assar, Nyberg, S. and Weibull, J. (1999). 'Social Norms and Economic Incentives in the Welfare State', *Quarterly Journal of Economics*, vol. 114(1), pp. 1–35.
Loewenstein, George F., Leigh Thompson, and Max H. Bazerman (1989). 'Social utility and Decision Making in Interpersonal Contexts', *Journal of Personality and Social Psychology*, vol. 57, pp. 426–41.
Mui, Vai-Lam (1995). 'The economics of envy', *Journal of Economic Behavior and Organization*, vol. 26, pp. 311–36.
Murphy, Kevin M. and Topel, Robert H. (1990). 'Efficiency wages reconsidered: theory and evidence', in (Y. Weiss and G. Fishelson, eds.), *Advances in the Theory and Measurement of Unemployment*, London: Macmillan.
Neilson, William S. (2000). 'An axiomatic characterization of the Fehr–Schmidt model of inequity aversion', working paper, Texas A&M University.
North, Douglass C. (1990). *Institutions, Institutional Change and Economic Performance*, Cambridge: Cambridge University Press.
Offerman, Theo (forthcoming). 'Hurting hurts more than helping helps', *European Economic Review*.
Ostrom, Elinor (1990). *Governing the Commons – the Evolution of Institutions for Collective Action*, New York: Cambridge University Press.
Ostrom, Elinor (2000). 'Collective action and the evolution of social norms', *Journal of Economic Perspectives*, vol. 14, pp. 137–58.
Ostrom, Elinor, Gardner, Roy and Walker, James (1994). *Rules, Games, and Common Pool Resources*, Michigan: The University of Michigan Press.
Rabin, Matthew (1993). 'Incorporating fairness into game theory and economics', *American Economic Review*, vol. 83(5), pp. 1281–302.
Roethlisberger, Fritz F. and Dickson, William J. (1947). *Management and the Worker: an account of a research program conducted by the Western Electric Company, Hawthorne Works, Chicago*, Cambridge, MA: Harvard University Press.
Roth, Alvin E. (1995). 'Bargaining experiments', in (J. Kagel and A. Roth, eds.), *Handbook of Experimental Economics*, Princeton, Princeton University Press.
Roth, Alvin E., Prasnikar, Vesna, Okuno-Fujiwara, Masahiro and Zamir, Shmuel (1991). 'Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study', *American Economic Review*, vol. 81, pp. 1068–95.
Samuelson, Paul A. (1993). 'Altruism as a problem involving group versus individual selection in economics and biology', *American Economic Review*, vol. 83, pp. 143–8.
Sally, David (1995). 'Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992', *Rationality and Society*, vol. 7(1), pp. 58–92.
Schulze, G.G. and Frank, B. (2001). Deterrence versus Intrinsic Motiration: Experimental Evidence on the Determinants of Corruptability. University of Konstanz, Department of Economics Discussion Paper, Series 1, No. 303.
Segal, Uzi and Sobel, Joel (1999). 'Tit for tat: foundations of preferences for reciprocity in strategic settings', mimeo, University of California at San Diego.
Seidl, Christian and Traub, Stefan (1999). 'Taxpayers' attitudes, behavior, and perceptions of fairness in taxation', mimeo, Institut für Finanzwissenschaft und Sozialpolitik, University of Kiel.
Sen, Amartya (1995). 'Moral codes and economic success', in (C. S. Britten and A. Hamlin, eds.), *Market Capitalism and Moral Values*, Aldershot: Edward Elgar.
Slade, Margret (1990). 'Strategic pricing models and interpretation of price-war data', *European Economic Review*, vol. 34, pp. 524–37.
Smith, Adam (1759), reprinted 1982. *The Theory of Moral Sentiments*. Indianapolis: Liberty Fund.
Sobel, Joel (2001). 'Social preferences and reciprocity', mimeo, University of California San Diego.
Sonnemans, Joep, Oosterbeek, Hessel and Sloof, Randolph, 2001. 'On the Relation Between Asset Ownership and Specific Investments', ECONOMIC JOURNAL 111, 781–820.
Wax, Amy L. (2000). 'Rethinking welfare rights: reciprocity norms, reactive attitudes, and the political economy of welfare reform', *Law and Contemporary Problems*, vol. 63 (1–2), pp. 257–98.
Whyte, William (1955). *Money and Motivation*, New York: Harper and Brothers.
Zajac, Edward (1995). *Political Economy of Fairness*, Cambridge, MA: MIT Press.