A THEORY OF FAIRNESS, COMPETITION, AND COOPERATION*

ERNST FEHR AND KLAUS M. SCHMIDT

There is strong evidence that people exploit their bargaining power in competitive markets but not in bilateral bargaining situations. There is also strong evidence that people exploit free-riding opportunities in voluntary cooperation games. Yet, when they are given the opportunity to punish free riders, stable cooperation is maintained, although punishment is costly for those who punish. This paper asks whether there is a *simple common principle* that can explain this puzzling evidence. We show that if some people care about equity the puzzles can be resolved. It turns out that the economic environment determines whether the fair types or the selfish types dominate equilibrium behavior.

I. INTRODUCTION

Almost all economic models assume that *all* people are *exclusively* pursuing their material self-interest and do not care about "social" goals per se. This may be true for some (maybe many) people, but it is certainly not true for everybody. By now we have substantial evidence suggesting that fairness motives affect the behavior of many people. The empirical results of Kahneman, Knetsch, and Thaler [1986], for example, indicate that customers have strong feelings about the fairness of firms' short-run pricing decisions which may explain why some firms do not fully exploit their monopoly power. There is also a lot of evidence suggesting that firms' wage setting is constrained by workers' views about what constitutes a fair wage [Blinder and Choi 1990; Agell and Lundborg 1995; Bewley 1995; Campbell and Kamlani 1997]. According to these studies, a major reason for firms' refusal to cut wages in a recession is the fear that workers will perceive pay cuts as unfair which in turn is expected to affect work morale adversely. There are also many well-controlled bilateral bargaining experiments which indicate that a nonnegligible fraction of the

* We would like to thank seminar participants at the Universities of Bonn and Berlin, Harvard, Princeton, and Oxford Universities, the European Summer Symposium on Economic Theory 1997 at Gerzensée (Switzerland), and the ESA conference in Mannheim for helpful comments and suggestions. We are particularly grateful to three excellent referees and to Drew Fudenberg and John Kagel for their insightful comments. The first author also gratefully acknowledges support from the Swiss National Science Foundation (project number 1214-05100.97) and the Network on the Evolution of Preferences and Social Norms of the MacArthur Foundation. The second author acknowledges financial support by the German Science Foundation through grant SCHM 119614-1.

The Quarterly Journal of Economics, August 1999

subjects do not care *solely* about material payoffs [Güth and Tietz, 1990; Roth 1995; Camerer and Thaler 1995]. However, there is also evidence that seems to suggest that fairness considerations are rather unimportant. For example, in competitive experimental markets with complete contracts, in which a well-defined homogeneous good is traded, *almost all* subjects behave as if they are only interested in their material payoff. Even if the competitive equilibrium implies an extremely uneven distribution of the gains from trade, equilibrium is reached within a few periods [Smith and Williams 1990; Roth, Prasnikar, Okuno-Fujiwara, and Zamir 1991; Kachelmeier and Shehata 1992; Güth, Marchand, and Rulliere 1997].

There is similarly conflicting evidence with regard to cooperation. Reality provides many examples indicating that people are more cooperative than is assumed in the standard self-interest model. Well-known examples are that many people vote, pay their taxes honestly, participate in unions and protest movements, or work hard in teams even when the pecuniary incentives go in the opposite direction.[1] This is also shown in laboratory experiments [Dawes and Thaler 1988; Ledyard 1995]. Under some conditions it has even been shown that subjects achieve nearly full cooperation, although the self-interest model predicts complete defection [Isaac and Walker 1988, 1991; Ostrom and Walker 1991; Fehr and Gächter 1996].[2] However, as we will see in more detail in Section IV, there are also those conditions under which a vast majority of subjects completely defect as predicted by the self-interest model.

There is thus a bewildering variety of evidence. Some pieces of evidence suggest that many people are driven by fairness considerations, other pieces indicate that virtually all people behave as if completely selfish, and still other types of evidence suggest that cooperation motives are crucial. In this paper we ask whether this conflicting evidence can be explained by a *single simple* model. Our answer to this question is affirmative if one is willing to assume that, in addition to purely self-interested people, there are a *fraction* of people who are also motivated by fairness considerations. No other deviations from the standard

1. On voting see Mueller [1989]. Skinner and Slemroad [1985] argue that the standard self-interest model substantially underpredicts the number of honest taxpayers. Successful team production in, e.g., Japanese-managed auto factories in North America is described in Rehder [1990]. Whyte [1955] discusses how workers establish "production norms" under piece-rate systems.

2. Isaac and Walker and Ostrom and Walker allow for cheap talk, while in Fehr and Gächter subjects could punish each other at some cost.

economic approach are necessary to account for the evidence. In particular, we do not relax the rationality assumption.[3]

We model fairness as self-centered inequity aversion. Inequity aversion means that people resist inequitable outcomes; i.e., they are willing to give up some material payoff to move in the direction of more equitable outcomes. Inequity aversion is self-centered if people do not care per se about inequity that exists among other people but are only interested in the fairness of their own material payoff relative to the payoff of others. We show that in the presence of some inequity-averse people "fair" and "cooperative" as well as "competitive" and "noncooperative" behavioral patterns can be explained in a coherent framework. A main insight of our examination is that the heterogeneity of preferences interacts in important ways with the economic environment. We show, in particular, that the economic environment determines the preference type that is decisive for the prevailing behavior in equilibrium. This means, for example, that under certain competitive conditions a single purely selfish player can induce a large number of extremely inequity-averse players to behave in a completely selfish manner, too. Likewise, under certain conditions for the provision of a public good, a single selfish player is capable of inducing all other players to contribute nothing to the public good, although the others may care a lot about equity. We also show, however, that there are circumstances in which the existence of a few inequity-averse players creates incentives for a majority of purely selfish types to contribute to the public good. Moreover, the existence of inequity-averse types may also induce selfish types to pay wages above the competitive level. This reveals that, in the presence of heterogeneous preferences, the economic environment has a whole new dimension of effects.[4]

There are a few other papers that formalize the notion of fairness.[5] In particular, Rabin [1993] argues that people want to be nice to those who treat them fairly and want to punish those who hurt them. According to Rabin, an action is perceived as fair if

3. This differentiates our model from learning models (e.g., Roth and Erev [1995]) that relax the rationality assumption but maintain the assumption that all players are only interested in their own material payoff. The issue of learning is further discussed in Section VII below.

4. Our paper is, therefore, motivated by a concern similar to the papers by Haltiwanger and Waldman [1985] and Russell and Thaler [1985]. While these authors examine the conditions under which nonrational or quasi-rational types affect equilibrium outcomes, we analyze the conditions under which fair types affect the equilibrium.

5. Section VIII deals with them in more detail.

the *intention* that is behind the action is kind, and as unfair if the intention is hostile. The kindness or the hostility of the intention, in turn, depends on the equitability of the payoff distribution induced by the action. Thus, Rabin's model, as our model, is based on the notion of an equitable outcome. In contrast to our model, however, Rabin models the role of intentions explicitly. We acknowledge that intentions do play an important role and that it is desirable to model them explicitly. However, the explicit modeling of intentions comes at a cost because it requires the adoption of psychological game theory that is much more difficult to apply than standard game theory. In fact, Rabin's model is restricted to two-person normal form games, which means that very important classes of games, like, e.g., market games and n-person public good games cannot be analyzed. Since a major focus of this paper is the role of fairness in competitive environments and the analysis of n-person cooperation games, we chose not to model intentions explicitly. This has the advantage of keeping the model simple and tractable. We would like to stress, however, that—although we do not model intentions explicitly—it is possible to capture intentions implicitly by our formulation of fairness preferences. We deal with this issue in Section VIII.

The rest of the paper is organized as followed. In Section II we present our model of inequity aversion. Section III applies this model to bilateral bargaining and market games. In Section IV cooperation games with and without punishments are considered. In Section V we show that, on the basis of plausible assumptions about preference parameters, the majority of individual choices in ultimatum *and* market *and* cooperation games considered in the previous sections are consistent with the predictions of our model. Section VI deals with the dictator game and with gift exchange games. In Section VII we discuss potential extensions and objections to our model. Section VIII compares our model with alternative approaches in the literature. Section IX concludes.

II. A Simple Model of Inequity Aversion

An individual is inequity averse if he dislikes outcomes that are perceived as inequitable. This definition raises, of course, the difficult question of how individuals measure or perceive the fairness of outcomes. Fairness judgments are inevitably based on a kind of neutral reference outcome. The reference outcome that is used to evaluate a given situation is itself the product of compli-

cated social comparison processes. In social psychology [Festinger 1954; Stouffer 1949; Homans 1961; Adams 1963] and sociology [Davis 1959; Pollis 1968; Runciman 1966] the relevance of social comparison processes has been emphasized for a long time. One key insight of this literature is that *relative* material payoffs affect people's well-being and behavior. As we will see below, without the assumption that at least for some people relative payoffs matter, it is difficult, if not impossible, to make sense of the empirical regularities observed in many experiments. There is, moreover, direct empirical evidence for the importance of relative payoffs. Agell and Lundborg [1995] and Bewley [1998], for example, show that relative payoff considerations constitute an important constraint for the internal wage structure of firms. In addition, Clark and Oswald [1996] show that comparison incomes have a significant impact on overall job satisfaction. They construct a comparison income level for a random sample of roughly 10,000 British individuals by computing a standard earnings equation. This earnings equation determines the predicted or expected wage of an individual with given socioeconomic characteristics. Then they examine the impact of this comparison wage on overall job satisfaction. Their main result is that—holding other things constant—the comparison income has a large and significantly negative impact on overall job satisfaction.

Strong evidence for the importance of relative payoffs is also provided by Loewenstein, Thompson, and Bazerman [1989]. These authors asked subjects to ordinally rank outcomes that differ in the distribution of payoffs between the subject and a comparison person. On the basis of these ordinal rankings, the authors estimate how *relative* material payoffs enter the person's utility function. The results show that subjects exhibit a strong and robust aversion against disadvantageous inequality: for a given own income x_i, subjects rank outcomes in which a comparison person earns more than x_i substantially lower than an outcome with equal material payoffs. Many subjects also exhibit an aversion to advantageous inequality although this effect seems to be significantly weaker than the aversion to disadvantageous inequality.

The determination of the relevant reference group and the relevant reference outcome for a given class of individuals is ultimately an empirical question. The social context, the saliency of particular agents, and the social proximity among individuals are all likely to influence reference groups and outcomes. Because

in the following we restrict attention to individual behavior in economic experiments, we have to make assumptions about reference groups and outcomes that are likely to prevail in this context. In the laboratory it is usually much simpler to define what is perceived as an equitable allocation by the subjects. The subjects enter the laboratory as equals, they do not know anything about each other, and they are allocated to different roles in the experiment at random. Thus, it is natural to assume that the reference group is simply the set of subjects playing against each other and that the reference point, i.e., the equitable outcome, is given by the egalitarian outcome.

More precisely, we assume the following. First, in addition to purely selfish subjects, there are subjects who dislike inequitable outcomes. They experience inequity if they are worse off in material terms than the other players in the experiment, and they also feel inequity if they are better off. Second, however, we assume that, in general, subjects suffer more from inequity that is to their material disadvantage than from inequity that is to their material advantage. Formally, consider a set of n players indexed by $i \in \{1, \ldots, n\}$, and let $x = x_1, \ldots, x_n$ denote the vector of monetary payoffs. The utility function of player $i \in \{1, \ldots, n\}$ is given by

$$U_i(x) = x_i - \alpha_i \frac{1}{n-1} \sum_{j \neq i} \max\{x_j - x_i, 0\} - \beta_i \frac{1}{n-1} \sum_{j \neq i} \max\{x_i - x_j, 0\}, \tag{1}$$

where we assume that $\beta_i \leq \alpha_i$ and $0 \leq \beta_i < 1$. In the two-player case (1) simplifies to

$$U_i(x) = x_i - \alpha_i \max\{x_j - x_i, 0\} - \beta_i \max\{x_i - x_j, 0\}, \quad i \neq j. \tag{2}$$

The second term in (1) or (2) measures the utility loss from disadvantageous inequality, while the third term measures the loss from advantageous inequality. Figure I illustrates the utility of player i as a function of x_j for a given income x_i. Given his own monetary payoff x_i, player i's utility function obtains a maximum at $x_j = x_i$. The utility loss from disadvantageous inequality ($x_j > x_i$) is larger than the utility loss if player i is better off than player j ($x_j < x_i$).[6]

6. In all experiments considered in this paper, the monetary payoff functions of all subjects were common knowledge. Note that for inequity aversion to be

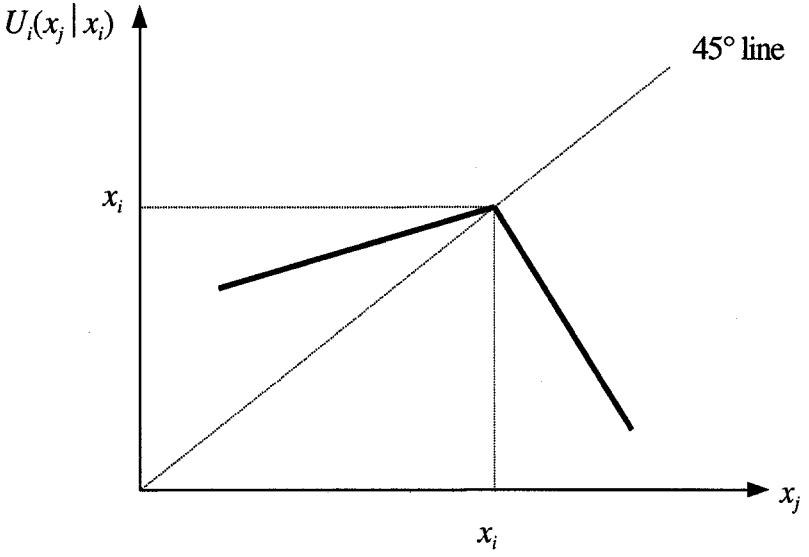

FIGURE I
Preferences with Inequity Aversion

To evaluate the implications of this utility function, let us start with the two-player case. For simplicity, we assume that the utility function is linear in inequality aversion as well as in x_i. This implies that the marginal rate of substitution between monetary income and inequality is constant. This may not be fully realistic, but we will show that surprisingly many experimental observations that seem to contradict each other can be explained on the basis of this very simple utility function already. However, we will also see that some observations in dictator experiments suggest that there are a nonnegligible fraction of people who exhibit nonlinear inequality aversion in the domain of advantageous inequality (see Section VI below).

Furthermore, the assumption $\alpha_i \geq \beta_i$ captures the idea that a player suffers more from inequality that is to his disadvantage. The above-mentioned paper by Loewenstein, Thompson, and

behaviorally important it is not necessary for subjects to be informed about the final monetary payoffs of the other subjects. As long as subjects' material payoff functions are common knowledge, they can compute the distributional implications of any (expected) strategy profile; i.e., inequity aversion can affect their decisions.

Bazerman [1989] provides strong evidence that this assumption is, in general, valid. Note that $\alpha_i \geq \beta_i$ essentially means that a subject is loss averse in social comparisons: negative deviations from the reference outcome count more than positive deviations. There is a large literature indicating the relevance of loss aversion in other domains (e.g., Tversky and Kahneman [1991]). Hence, it seems natural that loss aversion also affects social comparisons.

We also assume that $0 \leq \beta_i < 1$. $\beta_i \geq 0$ means that we rule out the existence of subjects who like to be better off than others. We impose this assumption here, although we believe that there are subjects with $\beta_i < 0$.[7] The reason is that in the context of the experiments we consider individuals with $\beta_i < 0$ have virtually no impact on equilibrium behavior. This is in itself an interesting insight that will be discussed extensively in Section VII. To interpret the restriction $\beta_i < 1$, suppose that player i has a higher monetary payoff than player j. In this case $\beta_i = 0.5$ implies that player i is just indifferent between keeping one dollar to himself and giving this dollar to player j. If $\beta_i = 1$, then player i is prepared to throw away one dollar in order to reduce his advantage relative to player j which seems very implausible. This is why we do not consider the case $\beta_i \geq 1$. On the other hand, there is no justification to put an upper bound on α_i. To see this, suppose that player i has a lower monetary payoff than player j. In this case player i is prepared to give up one dollar of his own monetary payoff if this reduces the payoff of his opponent by $(1 + \alpha_i)/\alpha_i$ dollars. For example, if $\alpha_i = 4$, then player i is willing to give up one dollar if this reduces the payoff of his opponent by 1.25 dollars. We will see that observable behavior in bargaining and public good games suggests that there are at least some individuals with such high α's.

If there are $n > 2$ players, player i compares his income with all other $n - 1$ players. In this case the disutility from inequality has been normalized by dividing the second and third term by $n - 1$. This normalization is necessary to make sure that the relative impact of inequality aversion on player i's total payoff is independent of the number of players. Furthermore, we assume for simplicity that the disutility from inequality is self-centered in the sense that player i compares himself with each of the other

7. For the role of status seeking and envy, see Frank [1985] and Banerjee [1990].

players, but he does not care per se about inequalities within the group of his opponents.

III. Fairness, Retaliation, and Competition: Ultimatum and Market Games

In this section we apply our model to a well-known simple bargaining game—the ultimatum game—and to simple market games in which one side of the market competes for an indivisible good. As we will see below, a considerable body of experimental evidence indicates that in the ultimatum game the gains from trade are shared relatively equally while in market games very unequal distributions are frequently observed. Hence, any alternative to the standard self-interest model faces the challenge to explain both "fair" outcomes in the ultimatum game and "competitive" and rather "unfair" outcomes in market games.

A. The Ultimatum Game

In an ultimatum game a proposer and a responder bargain about the distribution of a surplus of fixed size. Without loss of generality we normalize the bargaining surplus to one. The responder's share is denoted by s and the proposer's share by $1 - s$. The bargaining rules stipulate that the proposer offers a share $s \in [0,1]$ to the responder. The responder can accept or reject s. In case of acceptance the proposer receives a (normalized) monetary payoff $x_1 = 1 - s$, while the responder receives $x_2 = s$. In case of a rejection both players receive a monetary return of zero. The self-interest model predicts that the responder accepts any $s \in (0,1]$ and is indifferent between accepting and rejecting $s = 0$. Therefore, there is a unique subgame perfect equilibrium in which the proposer offers $s = 0$, which is accepted by the responder.[8]

By now there are numerous experimental studies from different countries, with different stake sizes and different experimental procedures, that clearly refute this prediction (for overviews

8. Given that the proposer can choose s continuously, any offer $s > 0$ cannot be an equilibrium offer since there always exists an s' with $0 < s' < s$ which is also accepted by the responder and yields a strictly higher payoff to the proposer. Furthermore, it cannot be an equilibrium that the proposer offers $s = 0$ which is rejected by the responder with positive probability. In this case the proposer would do better by slightly raising his price—in which case the responder would accept with probability 1. Hence, the only subgame perfect equilibrium is that the proposer offers $s = 0$ which is accepted by the responder. If there is a smallest money unit ε, then there exists a second subgame perfect equilibrium in which the responder accepts any $s \in [\varepsilon,1]$ and rejects, $s = 0$ while the proposer offers ε.

see Thaler [1988], Güth and Tietz [1990], Camerer and Thaler [1995], and Roth [1995]). The following regularities can be considered as robust facts (see Table I). (i) There are virtually no offers above 0.5. (ii) The vast majority of offers in almost any study is in the interval [0.4, 0.5]. (iii) There are almost no offers below 0.2. (iv) Low offers are frequently rejected, and the probability of rejection tends to decrease with s. Regularities (i) to (iv) continue to hold for rather high stake sizes, as indicated by the results of Cameron [1995], Hoffman, McCabe, and Smith [1996], and Slonim and Roth [1997]. The 200,000 rupiahs in the second experiment of Cameron (see Table I) are, e.g., equivalent to three months' income for the Indonesian subjects. Overall, roughly 60–80 percent of the offers in Table I fall in the interval [0.4, 0.5], while only 3 percent are below a share of 0.2.

To what extent is our model capable of accounting for the stylized facts of the ultimatum game? To answer this question, suppose that the proposer's preferences are represented by (α_1,β_1), while the responder's preferences are characterized by (α_2,β_2). The following proposition characterizes the equilibrium outcome as a function of these parameters.

PROPOSITION 1. It is a dominant strategy for the responder to accept any offer $s \geq 0.5$, to reject s if

$$s < s'(\alpha_2) \equiv \alpha_2/(1 + 2\alpha_2) < 0.5,$$

and to accept $s > s'(\alpha_2)$. If the proposer knows the preferences of the responder, he will offer

$$s^* \begin{cases} = & 0.5 & \text{if } \beta_1 > 0.5 \\ \in & [s'(\alpha_2),0.5] & \text{if } \beta_1 = 0.5 \\ = & s'(\alpha_2) & \text{if } \beta_1 < 0.5 \end{cases} \tag{3}$$

in equilibrium. If the proposer does not know the preferences of the responder but believes that α_2 is distributed according to the cumulative distribution function $F(\alpha_2)$, where $F(\alpha_2)$ has support $[\underline{\alpha}, \overline{\alpha}]$ with $0 \leq \underline{\alpha} < \overline{\alpha} < \infty$, then the probability (from the perspective of the proposer) that an offer $s < 0.5$ is going to be accepted is given by

$$p = \begin{cases} 1 & \text{if } s \geq s'(\overline{\alpha}) \\ F(s/(1 - 2s)) \in (0,1) & \text{if } s'(\underline{\alpha}) < s < s'(\overline{\alpha})) \\ 0 & \text{if } s \leq s'(\underline{\alpha}). \end{cases} \tag{4}$$

TABLE I
PERCENTAGE OF OFFERS BELOW 0.2 AND BETWEEN 0.4 AND 0.5 IN THE ULTIMATUM GAME

Study (Payment method)	Number of observations	Stake size (country)	Percentage of offers with $s < 0.2$	Percentage of offers with $0.4 \leq s \leq 0.5$
Cameron [1995] (All Ss Paid)	35	Rp 40.000 (Indonesia)	0	66
Cameron [1995] (all Ss paid)	37	Rp 200.000 (Indonesia)	5	57
FHSS [1994] (all Ss paid)	67	\$5 and \$10 (USA)	0	82
Güth et al. [1982] (all Ss paid)	79	DM 4–10 (Germany)	8	61
Hoffman, McCabe, and Smith [1996] (All Ss paid)	24	\$10 (USA)	0	83
Hoffman, McCabe, and Smith [1996] (all Ss paid)	27	\$100 (USA)	4	74
Kahneman, Knetsch, and Thaler [1986] (20% of Ss paid)	115	\$10 (USA)	?	75[a]
Roth et al. [1991] (random payment method)	116[b]	approx. \$10 (USA, Slovenia, Israel, Japan)	3	70
Slonim and Roth [1997] (random payment method)	240[c]	SK 60 (Slovakia)	0.4[d]	75
Slonim and Roth [1997] (random payment method)	250[c]	SK 1500 (Slovakia)	8[d]	69
Aggregate result of all studies[e]	875		3.8	71

a. percentage of equal splits, b. only observations of the final period, c. observations of all ten periods, d. percentage of offers below 0.25, e. without Kahneman, Knetsch, and Thaler [1986].

Hence, the optimal offer of the proposer is given by

$$(5) \qquad s^* \begin{cases} = 0.5 & \text{if } \beta_1 > 0.5 \\ \in [s'(\overline{\alpha}), 0.5] & \text{if } \beta_1 = 0.5 \\ \in (s'(\underline{\alpha}), s'(\overline{\alpha})] & \text{if } \beta_1 < 0.5. \end{cases}$$

Proof. If $s \geq 0.5$, the utility of a responder from accepting s is $U_2(s) = s - \beta_2(2s - 1)$, which is always positive for $\beta_2 < 1$ and thus better than a rejection that yields a payoff of 0. The point is that the responder can achieve equality only by destroying the entire surplus which is very costly to him if $s \geq 0.5$; i.e., if the inequality is to his advantage. For $s < 0.5$, a responder accepts the offer only if the utility from acceptance, $U_2(s) = s - \alpha_2(1 - 2s)$, is nonnegative which is the case only if s exceeds the acceptance threshold

$$s'(\alpha_2) \equiv \alpha_2/(1 + 2\alpha_2) < 0.5.$$

At stage 1 a proposer never offers $s > 0.5$. This would reduce his monetary payoff as compared with an offer of $s = 0.5$, which would also be accepted with certainty and which would yield perfect equality. If $\beta_1 > 0.5$, his utility is strictly increasing in s for all $s \leq 0.5$. This is the case where the proposer prefers to share his resources rather than to maximize his own monetary payoff, so he will offer $s = 0.5$. If $\beta_1 = 0.5$, he is just indifferent between giving one dollar to the responder and keeping it to himself; i.e., he is indifferent between all offers $s \in [s'(\alpha_2), 0.5]$. If $\beta_1 < 0.5$, the proposer would like to increase his monetary payoff at the expense of the responder. However, he is constrained by the responder's acceptance threshold. If the proposer is perfectly informed about the responder's preferences, he will simply offer $s'(\alpha_2)$. If the proposer is imperfectly informed about the responder's type, then the probability of acceptance is $F(s/(1 - 2s))$ which is equal to one if $s \geq \overline{\alpha}(1 + 2\overline{\alpha})$ and equal to zero if $s \leq \underline{\alpha}/(1 + \underline{\alpha})$. Hence, in this case there exists an optimal offer $s \in (s'(\underline{\alpha}), s'(\overline{\alpha})]$.

QED

Proposition 1 accounts for many of the above-mentioned facts. It shows that there are no offers above 0.5, that offers of 0.5 are always accepted, and that very low offers are very likely to be rejected. Furthermore, the probability of acceptance, $F(s/(1 - 2s))$, is increasing in s for $s < s'(\overline{\alpha}) < 0.5$. Note also that the acceptance threshold $s'(\alpha_2) = \alpha_2/(1 + 2\alpha_2)$ is nonlinear and has some intuitively appealing properties. It is increasing and strictly concave in α_2, and it converges to 0.5 if $\alpha_2 \to \infty$. Furthermore, relatively small values of α_2 already yield relatively large thresholds. For example, $\alpha_2 = \frac{1}{3}$ implies that $s'(\alpha_2) = 0.2$ and $\alpha_2 = 0.75$ implies that $s'(\alpha_2) = 0.3$.

In Section V we go beyond the predictions implied by Proposition 1. There we ask whether there is a distribution of preferences

that can explain not just the major facts of the ultimatum game but also the facts in market and cooperation games that will be discussed in the next sections.

B. Market Game with Proposer Competition

It is a well-established experimental fact that in a broad class of market games prices converge to the competitive equilibrium. [Smith 1982; Davis and Holt 1993]. For our purposes, the interesting fact is that convergence to the competitive equilibrium can be observed even if that equilibrium is very "unfair" by virtually any conceivable definition of fairness; i.e., if all of the gains from trade are reaped by one side of the market. This empirical feature of competition can be demonstrated in a simple market game in which many price-setting sellers (proposers) want to sell one unit of a good to a single buyer (responder) who demands only one unit of the good.[9]

Such a game has been implemented in four different countries by Roth, Prasnikar, Okuno-Fujiwara, and Zamir [1991]: suppose that there are $n - 1$ proposers who simultaneously propose a share $s_i \in [0,1]$, $i \in \{1, \ldots, n-1\}$, to the responder. The responder has the opportunity to accept or reject the *highest* offer $\overline{s} = \max_i \{s_i\}$. If there are several proposers who offered $\overline{s}$, one of them is randomly selected with equal probability. If the responder rejects $\overline{s}$, no trade takes place, and all players receive a monetary payoff of zero. If the responder accepts $\overline{s}$, her monetary payoff is $\overline{s}$, and the successful proposer earns $1 - \overline{s}$ while unsuccessful proposers earn zero. If players are only concerned about their monetary payoffs, this market game has a straightforward solution: the responder accepts any $\overline{s} > 0$. Hence, for any $s_i \leq \overline{s} < 1$, there exists an $\varepsilon > 0$ such that proposer i can strictly increase this monetary payoff by offering $\overline{s} + \varepsilon < 1$. Therefore, any equilibrium candidate must have $\overline{s} = 1$. Furthermore, in equilibrium a proposer i who offered $s_i = 1$ must not have an incentive to lower his offer. Thus, there must be at least one other player j who proposed $s_j = 1$, too. Hence, there is a unique subgame perfect

9. We deliberately restrict our attention to simple market games for two reasons: (i) the potential impact of inequity aversion can be seen most clearly in such simple games; (ii) they allow for an explicit game-theoretic analysis. In particular, it is easy to establish the identity between the competitive equilibrium and the subgame perfect equilibrium outcome in these games. Notice that some experimental market games, like, e.g., the continuous double auction as developed by Smith [1962], have such complicated strategy spaces that no complete game-theoretic analysis is yet available. For attempts in this direction see Friedman and Rust [1993] and Sadrieh [1998].

equilibrium outcome in which at least two proposers make an offer of one, and the responder reaps all gains from trade.[10]

Roth et al. [1991] have implemented a market game in which nine players simultaneously proposed s_i while one player accepted or rejected $\bar{s}$. Experimental sessions in four different countries have been conducted. The empirical results provide ample evidence in favor of the above prediction. After approximately five to six periods the subgame perfect equilibrium outcome was reached in each experiment in each of the four countries. To what extent can our model explain this observation?

PROPOSITION 2. Suppose that the utility functions of the players are given by (1). For any parameters (α_i, β_i), $i \in \{1, \ldots, n\}$, there is a *unique* subgame perfect equilibrium outcome in which at least two proposers offer $s = 1$ which is accepted by the responder.

The formal proof of the proposition is relegated to the Appendix, but the intuition is quite straightforward. Note first that, for similar reasons as in the ultimatum game, the responder must accept any $\bar{s} \geq 0.5$. Suppose that he rejects a "low" offer $\bar{s} < 0.5$. This cannot happen on the equilibrium path either since in this case proposer i can improve his payoff by offering $s_i = 0.5$ which is accepted with probability 1 and gives him a strictly higher payoff. Hence, on the equilibrium path $\bar{s}$ must be accepted. Consider now any equilibrium candidate with $\bar{s} < 1$. If there is one player i offering $s_i < \bar{s}$, then this player should have offered slightly more than $\bar{s}$. There will be inequality anyway, but by winning the competition, player i can increase his own monetary payoff, and he can turn the inequality to his advantage. A similar argument applies if all players offer $s_i = \bar{s} < 1$. By slightly increasing his offer, player i can increase the probability of winning the competition from $1/(n-1)$ to 1. Again, this increases his expected monetary payoff, and it turns the inequality toward the other proposers to his advantage. Therefore, $\bar{s} < 1$ cannot be part of a subgame perfect equilibrium. Hence, the only equilibrium candidate is that at least two sellers offer $\bar{s} = 1$. This is a subgame perfect equilibrium since all sellers receive a payoff of 0, and no player can change this outcome by changing his action. The formal proof in the Appendix extends this argument to the

10. Note that there are many subgame perfect equilibria in this game. As long as two sellers propose $s = 1$, any offer distribution of the remaining sellers is compatible with equilibrium.

possibility of mixed strategies. This extension also shows that the competitive outcome must be the unique equilibrium outcome in the game with incomplete information where proposers do not know each others' utility functions.

Proposition 2 provides an explanation for why markets in all four countries in which Roth et al. [1991] conducted this experiment quickly converged to the competitive outcome even though the results of the ultimatum game, that have also been done in these countries, are consistent with the view that the distribution of preferences differs across countries.[11]

C. Market Game with Responder Competition

In this section we apply our model of inequity aversion to a market game for which it is probably too early to speak of well-established stylized facts since only one study with a relatively small number of independent observations [Güth, Marchand, and Rulliere 1997] has been conducted so far. The game concerns a situation in which there is one proposer but many responders competing against each other. The rules of the game are as follows. The proposer, who is denoted as player 1, proposes a share $s \in [0,1]$ to the responders. There are $2, \ldots, n$ responders who observe s and decide simultaneously whether to accept or reject s. Then a random draw selects with equal probability one of the accepting responders. In case all responders reject s, all players receive a monetary payoff of zero. In case of acceptance of at least one responder, the proposer receives $1 - s$, and the randomly selected responder gets paid s. All other responders receive zero. Note that in this game there is competition in the second stage of the game whereas in subsection III.B we have competing players in the first stage.

The prediction of the standard model with purely selfish preferences for this game is again straightforward. Responders accept any positive s and are indifferent between accepting and rejecting $s = 0$. Therefore, there is a unique subgame perfect equilibrium outcome in which the proposer offers $s = 0$ which is accepted by at least one responder.[12] The results of Güth, Marchand, and Rulliere [1997] show that the standard model captures

11. Rejection rates in Slovenia and the United States were significantly higher than rejection rates in Japan and Israel.

12. In the presence of a smallest money unit, ε, there exists an additional, slightly different equilibrium outcome: the proposer offers $s = \varepsilon$ which is accepted by all the responders. To support this equilibrium, all responders have to reject $s = 0$. We assume, however, that there is no smallest money unit.

the regularities of this game rather well. The acceptance thresholds of responders quickly converged to very low levels.[13] Although the game was repeated only five times, in the final period the *average* acceptance threshold is well below 5 percent of the available surplus, with 71 percent of the responders stipulating a threshold of exactly zero and 9 percent a threshold of $s' = 0.02$. Likewise, in period 5 the average offer declined to 15 percent of the available gains from trade. In view of the fact that proposers had not been informed about responders' previous acceptance thresholds, such low offers are remarkable. In the final period *all* offers were below 25 percent, while in the ultimatum game such low offers are very rare.[14] To what extent is this apparent willingness to make and to accept extremely low offers compatible with the existence of inequity-averse subjects? As the following proposition shows, our model can account for the above regularities.

PROPOSITION 3. Suppose that $\beta_1 < (n - 1)/n$. Then there exists a subgame perfect equilibrium in which all responders accept any $s \geq 0$, and the proposer offers $s = 0$. The highest offer s that can be sustained in a subgame perfect equilibrium is given by

$$\bar{s} = \min_{i \in \{2,\ldots,n\}} \left\{ \frac{\alpha_i}{(1 - \beta_i)(n - 1) + 2\alpha_i + \beta_i} \right\} < \frac{1}{2}. \tag{8}$$

Proof. See Appendix.

The first part of Proposition 3 shows that responder competition always ensures the existence of an equilibrium in which all the gains from trade are reaped by the proposer irrespective of the prevailing amount of inequity aversion among the responders. This result is not affected if there is incomplete information about the types of players and is based on the following intuition. Given that there is at least one other responder j who is going to accept an offer of 0, there is no way for responder i to affect the outcome, and he may just as well accept this offer, too. However, note that the proposer will offer $s = 0$ only if $\beta_1 < (n - 1)/n$. If there are n

13. The gains from trade were 50 French francs. Before observing the offer s, each responder stated an acceptance threshold. If s was above the threshold, the responder accepted the offer; if it was below, she rejected s.

14. Due to the gap between acceptance thresholds and offers, we conjecture that the game had not yet reached a stable outcome after five periods. The strong and steady downward trend in all previous periods also indicates that a steady state had not yet been reached. Recall that the market game of Roth et al. [1991] was played for ten periods.

players altogether, than giving away one dollar to one of the responders reduces inequality by $1 + [1/(n-1)] = n/(n-1)$ dollars. Thus, if the nonpecuniary gain from this reduction in inequality, $\beta_1[n/(n-1)]$, exceeds the cost of 1, player 1 prefers to give money away to one of the responders. Recall that in the bilateral ultimatum game the proposer offered an equal split if $\beta_1 > 0.5$. An interesting aspect of our model is that an increase in the number of responders renders $s = 0.5$ less likely because it increases the threshold β_1 has to pass.

The second part of Proposition 3, however, shows that there may also be other equilibria. Clearly, a positive share s can be sustained in a subgame perfect equilibrium only if all responders can credibly threaten to reject any $s' < s$. When is it optimal to carry out this threat? Suppose that $s < 0.5$ has been offered and that this offer is being rejected by all other responders $j \neq i$. In this case responder i can enforce an egalitarian outcome by rejecting the offer as well. Rejecting reduces not only the inequality toward the other responders but also the disadvantageous inequality toward the proposer. Therefore, responder i is willing to reject this offer if nobody else accepts it and if the offer is sufficiently small, i.e., if the disadvantageous inequality toward the proposer is sufficiently large. More formally, given that all other responders reject, responder i prefers to reject as well if and only if the utility of acceptance obeys

$$s - \frac{\alpha_i}{n-1}(1-2s) - \frac{n-2}{n-1}\beta_i s \leq 0. \tag{9}$$

This is equivalent to

$$s \leq s_i' \equiv \frac{\alpha_i}{(1-\beta_i)(n-1) + 2\alpha_i + \beta_i}. \tag{10}$$

Thus, an offer $s > 0$ can be sustained if and only if (10) holds for *all* responders. It is interesting to note that the highest sustainable offer does not depend on all the parameters α_i and β_i but only on the inequity aversion of the responder with the lowest acceptance threshold s_i'. In particular, if there is only one responder with $\alpha_i = 0$, Proposition 3 implies that there is a unique equilibrium outcome with $s = 0$. Furthermore, the acceptance threshold is decreasing with n. Thus, the model makes the intuitively appealing prediction that for $n \to \infty$ the highest

sustainable equilibrium offer converges to zero whatever the prevailing amount of inequity aversion.[15]

D. Competition and Fairness

Propositions 2 and 3 suggest that there is a more general principle at work that is responsible for the very limited role of fairness considerations in the competitive environments considered above. Both propositions show that the introduction of inequity aversion hardly affects the subgame perfect equilibrium outcome in market games with proposer and responder competition relative to the prediction of the standard self-interest model. In particular, Proposition 2 shows that competition between proposers renders the distribution of preferences completely irrelevant. It does not matter for the outcome whether there are many or only a few subjects who exhibit strong inequity aversion. By the same token it also does not matter whether the players know or do not know the preference parameters of the other players. The crucial observation in this game is that *no single player can enforce an equitable outcome.* Given that there will be inequality anyway, each proposer has a strong incentive to outbid his competitors in order to turn part of the inequality to his advantage and to increase his own monetary payoff. A similar force is at work in the market game with responder competition. As long as there is at least one responder who accepts everything, no other responder can prevent an inequitable outcome. Therefore, even very inequity-averse responders try to turn part of the unavoidable inequality into inequality to their advantage by accepting low offers. It is, thus, the impossibility of preventing inequitable outcomes by individual players that renders inequity aversion unimportant in equilibrium.

The role of this factor can be further highlighted by the following slight modification of the market game with proposer competition: suppose that at stage 2 the responder may accept *any* of the offers made by the proposers; he is not forced to take the highest offer. Furthermore, there is an additional stage 3 at which the proposer who has been chosen by the responder at stage 2 can decide whether he wants to stick to his offer or whether he wants to withdraw—in which case all the gains from trade are lost for all

15. Note that the acceptance threshold is affected by the reference group. For example, if each responder compares his payoff only with that of the proposer but not with those of the other responders, then the acceptance threshold increases for each responder, and a higher offer may be sustained in equilibrium.

parties. This game would be an interesting test for our theory of inequity aversion. Clearly, in the standard model with selfish preferences, these modifications do not make any difference for the subgame perfect equilibrium outcome. Also, if some players have altruistic preferences in the sense that they appreciate any increase in the monetary payoff of other players, the result remains unchanged because altruistic players do not withdraw the offer at stage 3. With inequity aversion the outcome will be radically different, however. A proposer who is inequity averse may want to destroy the entire surplus at stage 3 in order to enforce an egalitarian outcome, in particular if he has a high α_i and if the split between himself and the responder is uneven. On the other hand, an even split will be withdrawn by proposer i at stage 3 only if $\beta_i > (n-1)/(n-2)$. Thus, the responder may prefer to accept an offer $s_i = 0.5$ rather than an offer $s_j > 0.5$ because the "better" offer has a higher chance of being withdrawn. This in turn reduces competition between proposers at stage 1. Thus, while competition nullifies the impact of inequity aversion in the ordinary proposer competition game, inequity aversion greatly diminishes the role of competition in the modified proposer competition game. This change in the role of competition is caused by the fact that in the modified game a single proposer can enforce an equitable outcome.

We conclude that competition renders fairness considerations irrelevant if and only if none of the competing players can punish the monopolist by destroying some of the surplus and enforcing a more equitable outcome. This suggests that fairness plays a smaller role in most markets for goods[16] than in labor markets. This follows from the fact that, in addition to the rejection of low wage offers, workers have some discretion over their work effort. By varying their effort, they can exert a direct impact on the relative material payoff of the employer. Consumers, in contrast, have no similar option available. Therefore, a firm may be reluctant to offer a low wage to workers who are competing for a job if the employed worker has the opportunity to respond to a low wage with low effort. As a consequence, fairness consider-

16. There are some markets for goods where fairness concerns play a role. For example, World Series or NBA playoff tickets are often sold far below the market-clearing price even though there is a great deal of competition among buyers. This may be explained by long-term profit-maximizing considerations of the monopolist who interacts *repeatedly* with groups of customers who care for fair ticket prices. On this see also Kahnemann, Knetsch, and Thaler [1986].

ations may well give rise to wage rigidity and involuntary unemployment.[17]

IV. Cooperation and Retaliation: Cooperation Games

In the previous section we have shown that our model can account for the relatively "fair" outcomes in the bilateral ultimatum game as well as for the rather "unfair" or "competitive" outcomes in games with proposer or responder competition. In this section we investigate the conditions under which cooperation can flourish in the presence of inequity aversion. We show that inequity aversion improves the prospects for voluntary cooperation relative to the predictions of the standard model. In particular, we show that there is an interesting class of conditions under which the selfish model predicts complete defection, while in our model there exist equilibria in which everybody cooperates fully. But, there are also other cases where the predictions of our model coincide with the predictions of the standard model.

We start with the following public good game. There are $n \geq 2$ players who decide simultaneously on their contribution levels $g_i \in [0, y]$, $i \in \{1, \ldots, n\}$, to the public good. Each player has an endowment of y. The monetary payoff of player i is given by

$$x_i(g_1, \ldots, g_n) = y - g_i + a \sum_{j=1}^{n} g_j, \qquad 1/n < a < 1, \tag{11}$$

where a denotes the constant marginal return to the public good $G \equiv \Sigma_{j=1}^{n} g_j$. Since $a < 1$, a marginal investment into G causes a monetary loss of $(1 - a)$; i.e., the dominant strategy of a completely selfish player is to choose $g_i = 0$. Thus, the standard model predicts $g_i = 0$ for all $i \in \{1, \ldots, n\}$. However, since $a > 1/n$, the aggregate monetary payoff is maximized if each player chooses $g_i = y$.

Consider now a slightly different public good game that consists of two stages. At stage 1 the game is identical to the previous game. At stage 2 each player i is informed about the contribution vector $(g_1, \ldots, g_n)$ and can simultaneously impose a punishment on the other players; i.e., player i chooses a punishment vector $p_i = (p_{i1}, \ldots, p_{in})$, where $p_{ij} \geq 0$ denotes the punishment player i imposes on player j. The cost of this

17. Experimental evidence for this is provided by Fehr, Kirchsteiger, and Riedl [1993] and Fehr and Falk [forthcoming]. We deal with these games in more detail in Section VI.

punishment to player i is given by $c \sum_{j=1}^{n} p_{ij}$, $0 < c < 1$. Player i, however, may also be punished by the other players, which generates an income loss to i of $\sum_{j=1}^{n} p_{ji}$. Thus, the monetary payoff of player i is given by

$$x_i(g_1, \ldots, g_n, p_1, \ldots, p_n) = y - g_i + a \sum_{j=1}^{n} g_j - \sum_{j=1}^{n} p_{ji} - c \sum_{j=1}^{n} p_{ij}. \tag{12}$$

What does the standard model predict for the two-stage game? Since punishments are costly, players' dominant strategy at stage 2 is to not punish. Therefore, if selfishness and rationality are common knowledge, each player knows that the second stage is completely irrelevant. As a consequence, players have exactly the same incentives at stage 1 as they have in the one-stage game without punishments, i.e., each player's optimal strategy is still given by $g_i = 0$. To what extent are these predictions of the standard model consistent with the data from public good experiments? For the one-stage game there are, fortunately, a large number of experimental studies (see Table II). They investigate the contribution behavior of subjects under a wide variety of conditions. In Table II we concentrate on the behavior of subjects in the final period only, since we want to exclude the possibility of repeated games effects. Furthermore, in the final period we have more confidence that the players fully understand the game that is being played.[18]

The striking fact revealed by Table II is that in the final period of n-person cooperation games ($n > 3$) without punishment the vast majority of subjects play the equilibrium strategy of complete free riding. If we average over all studies, 73 percent of all subjects choose $g_i = 0$ in the final period. It is also worth mentioning that in addition to those subjects who play *exactly* the equilibrium strategy there are very often a nonnegligible fraction of subjects who play "close" to the equilibrium. In view of the facts presented in Table II, it seems fair to say that the standard model "approximates" the choices of a big majority of subjects rather well. However, if we turn to the public good game with punishment, there emerges a radically different picture although the standard model predicts the same outcome as in the one-stage

18. This point is discussed in more detail in Section V. Note that in some of the studies summarized in Table II the group composition was the same for all T periods (partner condition). In others, the group composition randomly changed from period to period (stranger condition). However, in the last period subjects in the partner condition also play a true one-shot public goods game. Therefore, Table II presents the behavior from stranger as well as from partner experiments.

TABLE II
PERCENTAGE OF SUBJECTS WHO FREE RIDE COMPLETELY IN THE FINAL PERIOD OF A REPEATED PUBLIC GOOD GAME

Study	Country	Group size (n)	Marginal pecuniary return (a)	Total number of subjects	Percentage of free riders ($g_i = 0$)
Isaac and Walker [1988]	USA	4and10	0.3	42	83
Isaac and Walker [1988]	USA	4and10	0.75	42	57
Andreoni [1988]	USA	5	0.5	70	54
Andreoni [1995a]	USA	5	0.5	80	55
Andreoni [1995b]	USA	5	0.5	80	66
Croson [1995]	USA	4	0.5	48	71
Croson [1996]	USA	4	0.5	96	65
Keser and van Winden [1996]	Holland	4	0.5	160	84
Ockenfels and Weimann [1996]	Germany	5	0.33	200	89
Burlando and Hey [1997]	UK,Italy	6	0.33	120	66
Falkinger, Fehr, Gächter, and Winter-Ebmer [forthcoming]	Switzerland	8	0.2	72	75
Falkinger, Fehr, Gächter, and Winter-Ebmer [forthcoming]	Switzerland	16	0.1	32	84
Total number of subjects in all experiments and percentage of complete free riding				1042	73

game. Figure II shows the distribution of contributions in the final period of the two-stage game conducted by Fehr and Gächter [1996]. Note that the *same subjects* generated the distribution in the game without and in the game with punishment. Whereas in the game without punishment most subjects play close to complete defection, a strikingly large fraction of roughly 80 percent cooperates *fully* in the game with punishment.[19] Fehr and Gächter

19. Subjects in the Fehr and Gächter study participated in both conditions, i.e., in the game with punishment *and* in the game without punishment. The parameter values for a and n in this experiment are $a = 0.4$ and $n = 4$. It is interesting to note that contributions are significantly higher in the two-stage game already in period 1. Moreover, in the one-stage game cooperation strongly decreases over time, whereas in the two-stage game cooperation quickly converges to the high levels observed in period 10.

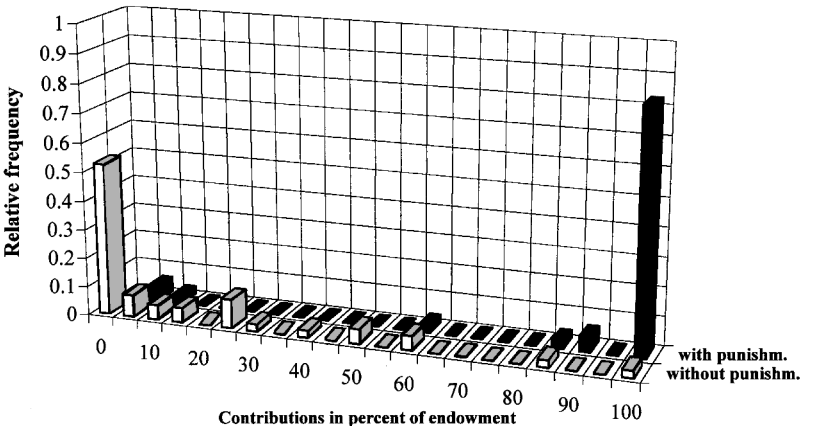

FIGURE II
Distribution of Contributions in the Final Period of the Public Good Game with Punishment (*Source:* Fehr and Gächter [1996])

report that the vast majority of punishments are imposed by cooperators on the defectors and that lower contribution levels are associated with higher received punishments. Thus, defectors do not gain from free riding because they are being punished.

The behavior in the game with punishment represents an unambiguous rejection of the standard model. This raises the question whether our model is capable of explaining both the evidence of the one-stage public good game and of the public good game with punishment. Consider the one-stage public good game first. The prediction of our model is summarized in the following proposition:

PROPOSITION 4.

(a) If $a + \beta_i < 1$ for player i, then it is a dominant strategy for that player to choose $g_i = 0$.

(b) Let k denote the number of players with $a + \beta_i < 1$, $0 \leq k \leq n$. If $k/(n-1) > a/2$, then there is a unique equilibrium with $g_i = 0$ for all $i \in [1, \ldots, n]$.

(c) If $k/(n-1) < (a + \beta_j - 1)/(\alpha_j + \beta_j)$ for all players $j \in [1, \ldots, n]$ with $a + \beta_j > 1$, then other equilibria with positive contribution levels do exist. In these equilibria all k players with $a + \beta_i < 1$ must choose $g_i = 0$, while all other players contribute $g_j = g \in [0,y]$. Note further that $(a + \beta_j - 1)/(\alpha_j + \beta_j) < a/2$.

The formal proof of Proposition 4 is relegated to the Appendix. To see the basic intuition for the above results, consider a player with $a + \beta_i < 1$. By spending one dollar on the public good, he earns a dollars in monetary terms. In addition, he may get a nonpecuinary benefit of at most β_i dollars from reducing inequality. Therefore, since $a + \beta_i < 1$ for this player, it is a dominant strategy for him to contribute nothing. Part (b) of the proposition says that if the fraction of subjects, for whom $g_i = 0$ is a dominant strategy, is sufficiently high, there is a unique equilibrium in which nobody contributes. The reason is that if there are only a few players with $a + \beta_i > 1$, they would suffer too much from the disadvantageous inequality caused by the free riders. The proof of the proposition shows that if a potential contributor knows that the number of free riders, k, is larger than $a(n - 1)/2$, then he will not contribute either. The last part of the proposition shows that if there are sufficiently many players with $a + \beta_i > 1$, they can sustain cooperation among themselves even if the other players do not contribute. However, this requires that the contributors are not too upset about the disadvantageous inequality toward the free riders. Note that the condition $k/(n - 1) < (a + \beta_j - 1)/(\alpha_j + \beta_j)$ is less likely to be met as α_j goes up. To put it differently, the greater the aversion against being the sucker, the more difficult it is to sustain cooperation in the one-stage game. We will see below that the opposite holds true in the two-stage game.

Note that in almost all experiments considered in Table II, $a \leq 1/2$. Thus, if the fraction of players with $a + \beta_i < 1$ is larger than ¼, then there is no equilibrium with positive contribution levels. This is consistent with the very low contribution levels that have been observed in these experiments. Finally, it is worthwhile mentioning that the prospects for cooperation are weakly increasing with the marginal return a.

Consider now the public good game with punishment. To what extent is our model capable of accounting for the very high cooperation in the public good game with punishment? In the context of our model the crucial point is that free riding generates a material payoff advantage relative to those who cooperate. Since $c < 1$, cooperators can reduce this payoff disadvantage by punishing the free riders. Therefore, if those who cooperate are sufficiently upset by the inequality to their disadvantage, i.e., if they have sufficiently high α's, then they are willing to punish the defectors even though this is costly to themselves. Thus, the threat to punish free riders may be credible, which may induce

potential defectors to contribute at the first stage of the game. This is made precise in the following proposition.

PROPOSITION 5. Suppose that there is a group of n' "conditionally cooperative enforcers," $1 \leq n' \leq n$, with preferences that obey $\alpha + \beta_i \geq 1$ and

$$(13) \quad c < \frac{\alpha_i}{(n-1)(1+\alpha_i) - (n'-1)(\alpha_i + \beta_i)} \quad \text{for all } i \in \{1, \ldots, n'\}.$$

whereas all other players do not care about inequality; i.e., $\alpha_i = \beta_i = 0$ for $i \in \{n'+1, \ldots, n\}$. Then the following strategies, which describe the players' behavior on and off the equilibrium path, form a subgame perfect equilibrium.

- In the first stage each player contributes $g_i = g \in [0, y]$.
- If each player does so, there are no punishments in the second stage. If one of the players $i \in \{n'+1, \ldots, n\}$ deviates and chooses $g_i < g$, then each enforcer $j \in \{1, \ldots, n'\}$ chooses $p_{ji} = (g - g_i)/(n' - c)$ while all other players do not punish. If one of the "conditionally cooperative enforcers" chooses $g_i < g$, or if any player chooses $g_i > g$, or if more than one player deviated from g, then one Nash-equilibrium of the punishment game is being played.

Proof. See Appendix.

Proposition 5 shows that full cooperation, as observed in the experiments by Fehr and Gächter [1996], can be sustained as an equilibrium outcome if there is a group of n' "conditionally cooperative enforcers." In fact, one such enforcer may be enough ($n' = 1$) if his preferences satisfy $c < \alpha_i/(n-1)(1+\alpha_i)$ and $\alpha + \beta_i \geq 1$; i.e., if there is one person who is sufficiently concerned about inequality. To see how the equilibrium works, consider such a "conditionally cooperative enforcer." For him $\alpha + \beta_i \geq 1$, so he is happy to cooperate if *all others cooperate as well* (this is why he is called "conditionally cooperative"). In addition, condition (13) makes sure that he cares sufficiently about inequality to his disadvantage. Thus, he can credibly threaten to punish a defector (this is why he is called "enforcer"). Note that condition (13) is less demanding if n' or α_i increases. The punishment is constructed such that the defector gets the same monetary payoff as the enforcers. Since this is less than what a defector would have received if he had chosen $g_i = g$, a deviation is not profitable.

If the conditions of Proposition 5 are met, then there exists a continuum of equilibrium outcomes. This continuum includes the "good equilibrium" with maximum contributions but also the "bad equilibrium" where nobody contributes to the public good. In our view, however, there is a reasonable refinement argument that rules out "bad" equilibria with low contributions. To see this, note that the equilibrium with the highest possible contribution level, $g_i = g = y$ for all $i \in \{1, \ldots, n\}$, is the unique symmetric and efficient outcome. Since it is symmetric, it yields the same payoff for all players. Hence, this equilibrium is a natural focal point that serves as a coordination device even if the subjects choose their strategies independently.

Comparing Propositions 4 and 5, it is easy to see that the prospects for cooperation are greatly improved if there is an opportunity to punish defectors. Without punishments all players with $\alpha + \beta_i < 1$ will never contribute. Players with $\alpha + \beta_i > 1$ may contribute only if they care enough about inequality to their advantage but not *too much* about disadvantageous inequality. On the other hand, with punishment *all* players will contribute if there is a (small) group of "conditionally cooperative enforcers." The more these enforcers care about disadvantageous inequality, the more they are prepared to punish defectors which makes it easier to sustain cooperation. In fact, one person with a sufficiently high α_i is already enough to enforce efficient contributions by all other players.

Before we turn to the next section, we would like to point out an implication of our model for the Prisoner's Dilemma (PD). Note that the simultaneous PD is just a special case of the public good game without punishment for $n = 2$ and $g_i \in \{0, y\}$, $i = 1,2$. Therefore, Proposition 4 applies; i.e., cooperation is an equilibrium if *both* players meet the condition $\alpha + \beta_i > 1$. Yet, if only one player meets this condition, defection of both players is the unique equilibrium. In contrast, in a sequentially played PD a purely selfish first mover has an incentive to contribute if he faces a second mover who meets $\alpha + \beta_i > 1$. This is so because the second mover will respond cooperatively to a cooperative first move while he defects if the first mover defects. Thus, due to the reciprocal behavior of inequity-averse second movers, cooperation rates among first movers in sequentially played PDs are predicted to be higher than cooperation rates in simultaneous PDs. There is fairly strong evidence in favor of this prediction. Watabe, Terai, Hayashi, and Yamagishi [1996] and Hayashi, Ostrom, Walker, and

Yamagishi [1998] show that cooperation rates among first movers in sequential PDs are indeed much higher and that reciprocal cooperation of second movers is very frequent.

V. Predictions across Games

In this section we examine whether the distribution of parameters that is consistent with experimental observations in the ultimatum game is consistent with the experimental evidence from the other games. It is not our aim here to show that our theory is consistent with 100 percent of the individual choices. The objective is rather to offer a first test for whether there is a chance that our theory is consistent with the *quantitative* evidence from different games. Admittedly, this test is rather crude. However, at the end of this section we make a number of predictions that are implied by our model, and we suggest how these predictions can be tested rigorously with some new experiments.

In many of the experiments referred to in this section, the subjects had to play the same game several times either with the same or with varying opponents. Whenever available, we take the data of the final period as the facts to be explained. There are two reasons for this choice. First, it is well-known in experimental economics that in interactive situations one cannot expect the subjects to play an equilibrium in the first period already. Yet, if subjects have the opportunity to repeat their choices and to better understand the strategic interaction, then very often rather stable behavioral patterns, that may differ substantially from first-period-play, emerge. Second, if there is repeated interaction between the same opponents, then there may be repeated games effects that come into play. These effects can be excluded if we look at the last period only.

Table III suggests a simple discrete distribution of α_i and β_i. We have chosen this distribution because it is consistent with the large experimental evidence we have on the ultimatum game (see Table I above and Roth [1995]). Recall from Proposition 1 that for any given α_i, there exists an acceptance threshold $s'(\alpha_i) = \alpha_i/(1 + 2\alpha_i)$ such that player i accepts s if and only if $s \geq s'(\alpha_i)$. In all experiments there is a fraction of subjects that rejects offers even if they are very close to an equal split. Thus, we (conservatively) assume that 10 percent of the subjects have $\alpha = 4$ which implies an acceptance threshold of $s' = 4/9 = 0.444$. Another,

TABLE III
ASSUMPTIONS ABOUT THE DISTRIBUTION OF PREFERENCES

DISTRIBUTION OF α's AND ASSOCIATED ACCEPTANCE THRESHOLDS OF BUYERS			DISTRIBUTION OF β's AND ASSOCIATED OPTIMAL OFFERS OF SELLERS		
$\alpha = 0$	30 percent	$s' = 0$	$\beta = 0$	30 percent	$s^* = 1/3$
$\alpha = 0.5$	30 percent	$s'(0.5) = 1/4$	$\beta = 0.25$	30 percent	$s^* = 4/9$
$\alpha = 1$	30 percent	$s'(1) = 1/3$	$\beta = 0.6$	40 percent	$s^* = 1/2$
$\alpha = 4$	10 percent	$s'(4) = 4/9$			

typically much larger fraction of the population insists on getting at least one-third of the surplus, which implies a value of α which is equal to one. These are at least 30 percent of the population. Note that they are prepared to give up one dollar if this reduces the payoff of their opponent by two dollars. Another, say, 30 percent of the subjects insist on getting at least one-quarter, which implies that $\alpha = 0.5$. Finally, the remaining 30 percent of the subjects do not care very much about inequality and are happy to accept any positive offer ($\alpha = 0$).

If a proposer does not know the parameter α of his opponent but believes that the probability distribution over α is given by Table III, then it is straightforward to compute his optimal offer as a function of his inequality parameter β. The optimal offer is given by

$$s^*(\beta) = \begin{cases} 0.5 & \text{if} \quad \beta_i > 0.5 \\ 0.\overline{4} & \text{if} \quad 0.235 < \beta_i < 0.5 \\ 0.\overline{3} & \text{if} \quad \beta_i < 0.235. \end{cases} \tag{14}$$

Note that it is never optimal to offer less than one-third of the surplus, even if the proposer is completely selfish. If we look at the actual offers made in the ultimatum game, there are roughly 40 percent of the subjects who suggest an equal split. Another 30 percent offer $s \in [0.4, 0.5)$, while 30 percent offer less than 0.4. There are hardly any offers below 0.25. This gives us the distribution of β in the population described in Table III.

Let us now see whether this distribution of preferences is consistent with the observed behavior in other games. Clearly, we have no problem in explaining the evidence on market games with proposer competition. Any distribution of α and β yields the competitive outcome that is observed by Roth et al. [1991] in all

their experiments. Similarly, in the market game with responder competition, we know from Proposition 3 that if there is at least one responder who does not care about disadvantageous inequality (i.e., $\alpha_i = 0$), then there is a unique equilibrium outcome with $\bar{s} = 0$. With five responders in the experiments by Güth, Marchand, and Rulliere [1997] and with the distribution of types from Table III, the probability that there is at least one such player in each group is given by $1–0.7^5 = 83$ percent. This is roughly consistent with the fact that 71 percent of the players accepted an offer of zero, and 9 percent had an acceptance threshold of $s' = 0.02$ in the final period.

Consider now the public good game. We know by Proposition 4 that cooperation can be sustained as an equilibrium outcome only if the number k of players with $a + \beta_i < 1$ obeys $k/(n - 1) < a/2$. Thus, our theory predicts that there is less cooperation the smaller a which is consistent with the empirical evidence of Isaac and Walker [1988] presented in Table II.[20] In a typical treatment $a = 0.5$, and $n = 4$. Therefore, if all players believe that there is at least one player with $a + \beta_i < 1$, then there is a unique equilibrium with $g_i = 0$ for all players. Given the distribution of preferences of Table III, the probability that there are four players with $\beta > 0.5$ is equal to $0.4^4 = 2.56$ percent. Hence, we should observe that, on average, almost all individuals fully defect. A similar result holds for most other experiments in Table II. Except for the Isaac and Walker experiments with $n = 10$ a *single* player with $a + \beta_i < 1$ is sufficient for the violation of the necessary condition for cooperation, $k/(n - 1) < a/2$. Thus, in all these experiments our theory predicts that randomly chosen groups are almost never capable of sustaining cooperation. Table II indicates that this is not quite the case, although 73 percent of individuals indeed choose $g_i = 0$. Thus, it seems fair to say that our model is consistent with the bulk of individual choices in this game.[21]

Finally, the most interesting experiment from the perspective of our theory is the public good game with punishment. While in

20. For $a = 0.3$, the rate of defection is substantially larger than for $a = 0.75$. The Isaac and Walker experiments were explicitly designed to test for the effects of variations in a.

21. When judging the accuracy of the model, one should also take into account that there is in general a significant fraction of the subjects that play close to complete free riding in the final round. A combination of our model with the view that human choice is characterized by a fundamental randomness [McKelvey and Palfrey 1995; Anderson, Goeree, and Holt 1997] may explain much of the remaining 25 percent of individual choices. This task, however, is left for future research.

the game without punishment most subjects play close to complete defection, a strikingly large fraction of roughly 80 percent cooperate *fully* in the game with punishment. To what extent can our model explain this phenomenon? We know from Proposition 5 that cooperation can be sustained if there is a group of n' "conditionally cooperative enforcers" with preferences that satisfy (13) and $a + \beta_i \geq 1$. For example, if all four players believe that there is at least one player with $\alpha_i \geq 1.5$ and $\beta_i \geq 0.6$, there is an equilibrium in which all four players contribute the maximum amount. As discussed in Section V, this equilibrium is a natural focal point. Since the computation of the probability that the conditions of Proposition 5 are met is a bit more cumbersome, we have put them in the Appendix. It turns out that for the preference distribution given in Table III the probability that a randomly drawn group of four players meets the conditions is 61.1 percent. Thus, our model is roughly consistent with the experimental evidence of Fehr and Gächter [1996].[22]

Clearly, the above computations provide only rough evidence in favor of our model. To rigorously test the model, additional experiments have to be run. We would like to suggest a few variants of the experiments discussed so far that would be particularly interesting:[23]

- Our model predicts that under proposer competition two proposers are sufficient for $s = 1$ to be the unique equilibrium outcome *irrespective of the players' preferences*. Thus, one could conduct the proposer competition game with two proposers that have proved to be very inequity averse in other games. This would constitute a particularly tough test of our model.
- Most public good games that have been conducted had symmetric payoffs. Our theory suggests that it will be more difficult to sustain cooperation if the game is asymmetric. For example, if the public good is more valuable to some of the players, there will in general be a conflict between efficiency and equality. Our prediction is that if the game is sufficiently asymmetric it is impossible to sustain cooperation even if a is very large or if players can use punishments.

22. In this context one has to take into account that the total number of available individual observations in the game with punishment is much smaller than for the game without punishment or for the ultimatum game. Future experiments will have to show whether the Fehr-Gächter results are the rule in the punishment game or whether they exhibit unusually high cooperation rates.

23. We are grateful to a referee who suggested some of these tests.

- It would be interesting to repeat the public good experiment with punishments for different values of a, c, and n. Proposition 5 suggests that we should observe less cooperation if a goes down and if c goes up. The effect of an increase in the group size n is ambiguous, however. For any given player it becomes more difficult to satisfy condition (13) as n goes up. On the other hand, the larger the group, the higher is the probability that there is at least one person with a very high α. Our conjecture is that a moderate change in the size of the group does not affect the amount of cooperation.
- One of the most interesting tests of our theory would be to do several different experiments with the same group of subjects. Our model predicts a cross-situation correlation in behavior. For example, the observations from one experiment could be used to estimate the parameters of the utility function of each individual. It would then be possible to test whether this individual's behavior in other games is consistent with his estimated utility function.
- In a similar fashion, one could screen subjects according to their behavior in one experiment before doing a public good experiment with punishments. If we group the subjects in this second experiment according to their observed inequality aversion, the prediction is that those groups with high inequality aversion will contribute while those with low inequality aversion will not.

VI. Dictator and Gift Exchange Games

The preceding sections have shown that our very simple model of linear inequality aversion is consistent, with the most important facts in ultimatum, market, and cooperation games. One problem with our approach, however, is that it yields too extreme predictions in some other games, such as the "dictator game." The dictator game is a two-person game in which only player 1, the "dictator," has to make a decision. Player 1 has to decide what share $s \in [0,1]$ of a given amount of money to pass on to player 2. For a given share s monetary payoffs are given by $x_1 = 1 - s$ and $x_2 = s$, respectively. Obviously, the standard model predicts $s = 0$. In contrast, in the experimental study of Forsythe, Horowitz, Savin, and Sefton [1994] only about 20 percent of subjects chose $s = 0$, 60 percent chose $0 < s < 0.5$, and again

roughly 20 percent chose $s = 0.5$. In the study by Andreoni and Miller [1995] the distribution of shares is again bimodal but puts more weight on the "extremes:" approximately 40 percent of the subjects gave $s = 0$, 20 percent gave $0 < s < 0.5$, and roughly 40 percent gave $s = 0.5$. Shares above $s = 0.5$ were practically never observed.

Our model predicts that player 1 offers $s = 0.5$ if $\beta_1 > 0.5$ and $s = 0$ if $\beta_1 < 0.5$. Thus, we should observe *only* very "fair" or very "unfair" outcomes, a prediction that is clearly refuted by the data. However, there is a straightforward solution to this problem. We assumed that the inequity aversion is piecewise *linear*. The linearity assumption was imposed in order to keep our model as simple as possible. If we allow for a utility function that is concave in the amount of advantageous inequality, there is no problem in generating optimal offers that are in the interior of [0,0.5].

It is important to note that nonlinear inequity aversion does not affect the qualitative results in the other games we considered. This is straightforward in market games with proposer or responder competition. Recall that in the context of proposer competition there exists a unique equilibrium outcome in which the responder receives the whole gains from trade *irrespective of the prevailing amount of inequity aversion.* Thus, it also does not matter whether linear or nonlinear inequity aversion prevails. Likewise, under responder competition there is a unique equilibrium outcome in which the proposer receives the whole surplus if there is at least one responder who does not care about disadvantageous inequality. Obviously, this proposition holds irrespective of whether the inequity aversion of the other responders is linear or not. Similar arguments hold for public good games with and without punishment. Concerning the public good game with punishment, for example, the existence of *nonlinear* inequity aversion obviously does not invalidate the existence of an equilibrium with full cooperation. It only renders the condition for the existence of such an equilibrium, i.e., condition (13), slightly more complicated.

Another interesting game is the so-called trust- or gift exchange game [Fehr, Kirchsteiger, and Riedl 1993; Berg, Dickhaut, and McCabe 1995; Fehr, Gächter, and Kirchsteiger 1997]. The common feature of trust- or gift exchange games is that they resemble a sequentially played PD with more than two actions for each player. In some experiments the gift exchange game has been embedded in a competitive experimental market. For example, a

slightly simplified version of the experiment conducted by Fehr, Gächter, and Kirchsteiger [1997] has the following structure. There is one experimental firm, which we denote as player 1, and which can make a wage offer w to the experimental workers. There are $2, \ldots, n$ workers who can simultaneously accept or reject w. Then a random draw selects with equal probability one of the accepting workers. Thereafter, the selected worker has to choose effort e from the interval $[\underline{e},\overline{e}]$, $0 < \underline{e} < \overline{e}$. In case that all workers reject w, all players receive nothing. In case of acceptance the firm receives $x_f = ve - w$, where v denotes the marginal product of effort. The worker receives $x_w = w - c(e)$, where $c(e)$ denotes the effort costs and obeys $c(\underline{e}) = c'(\underline{e}) = 0$ and $c' > 0, c'' > 0$ for $e > \underline{e}$. Moreover, $v > c'(\overline{e})$ so that $e = \overline{e}$ is the efficient effort level. This game is essentially a market game with responder competition in which an accepting responder has to make an effort choice after he is selected.

If all players are pure money maximizers, the prediction for this game is straightforward. Since the selected worker always chooses the minimum effort $\underline{e}$, the game collapses into a responder competition game with gains from trade equal to $v\underline{e}$. In equilibrium the firm earns $v\underline{e}$ and $w = 0$. Yet, since $v > c'(\overline{e})$, there exist many (w,e)-combinations that would make both the firm and the selected worker better off. In sharp contrast to this prediction, and also in sharp contrast to what is observed under responder competition *without* effort choices, firms offer substantial wages to the workers, and wages do not decrease over time. Moreover, workers provide effort above $\underline{e}$ and there is a strong positive correlation between w and e.

To what extent can our model explain this outcome? Put differently, why is it the case that under responder competition *without* effort choice the responders' income converges toward the selfish solution, whereas under responder competition *with* effort choice, wages substantially above the selfish solution can be maintained. From the viewpoint of our model the key fact is that—by varying the effort choice—the randomly selected worker has the opportunity to affect the difference $x_f - x_w$. If the firm offers "low" wages such that $x_f > x_w$ holds at any feasible effort level, the selected worker will always choose the minimum effort. However, if the firm offers a "high" wage such that at $\underline{e}$ the inequality $x_w > x_f$ holds, inequity-averse workers with a sufficiently high β_i are willing to raise e above $\underline{e}$. Moreover, in the presence of nonlinear inequity aversion, higher wages will be

associated with higher effort levels. The reason is that by raising the effort workers can move in the direction of more equitable outcomes. Thus, our model is capable of explaining the apparent wage rigidity observed in gift exchange games. Since the presence of inequity-averse workers generates a positive correlation between wages and effort, the firm does not gain by exploiting the competition among the workers. Instead, it has an incentive to pay efficiency wages above the competitive level.

VII. Extensions and Possible Objections

So far, we ruled out the existence of subjects who like to be better off than others. This is unsatisfactory because subjects with $\beta_i < 0$ clearly exist. Fortunately, however, such subjects have virtually no impact on equilibrium behavior in the games considered in this paper. To see this, suppose that a fraction of subjects with $\beta_i = 0$ exhibits $\beta_i < 0$ instead. This obviously does not change responders' behavior in the ultimatum game because for them only α_i matters. It also does not change the proposer behavior in the complete information case because both proposers with $\beta_i = 0$ and those with $\beta_i < 0$ will make an offer that exactly matches the responder's acceptance threshold.[24] In the market game with proposer competition, proposers with $\beta_i < 0$ are even more willing to overbid a going share below $s = 1$, compared with subjects with $\beta_i = 0$, because by overbidding they gain a payoff advantage relative to the other proposers. Thus, Proposition 2 remains unchanged. Similar arguments apply to the case of responder competition (without effort choices) because a responder with $\beta_i < 0$ is even more willing to underbid a positive share compared with a responder with $\beta_i = 0$. In the public good game without punishment all players with $a + \beta_i < 1$ have a dominant strategy to contribute nothing. It does not matter whether these players exhibit a positive or a negative β_i. Finally, the existence of types with $\beta_i < 0$ also leaves Proposition 5 unchanged.[25] If there are sufficiently many conditionally cooperative enforcers, it does not

24. It may affect proposer behavior in the incomplete information case although the effect of a change in β_i is ambiguous. This ambiguity stems from the fact that the proposer's marginal expected utility of s may rise or fall if β_i falls.

25. This holds true if, for those with a negative β_i, the absolute value of β_i is not too large. Otherwise, defectors would have an incentive to punish the cooperators. A defector who imposes a punishment of one on a cooperator gains $[-\beta_i/(n-1)](1-c) > 0$ in nonpecuniary terms and has material costs of c. Thus, he is willing to punish if $|\beta_i| \geq [c/(1-c)](n-1)$ holds. This means that only defectors with implausibly high absolute values of β_i are willing to punish. For

matter whether the remaining players have $\beta_i < 0$ or not. Recall that—according to Proposition 5—strategies that discipline potential defectors make the enforcers and the defectors equally well off in material terms. Hence, a defector cannot gain a payoff advantage but is even worse off relative to a cooperating nonenforcer. These punishment strategies, therefore, are sufficient to discipline potential defectors irrespective of their β_i-values.

Another set of questions concerns the choice of the reference group. As argued in Section II, for many laboratory experiments our assumption that subjects compare themselves with all other subjects in the (usually relatively small) group is a natural starting point. However, we are aware of the possibility that this may not always be an appropriate assumption.[26] There may well be interactive structures in which some agents have a salient position that makes them natural reference agents. Moreover, the social context and the institutional environment in which interactions take place is likely to be important.[27] Bewley [1998], for example, reports that in nonunionized firms workers compare themselves exclusively with their firm and with other workers in their firm. This suggests that only within-firm social comparisons but not across-firm comparisons affect the wage-setting process. This is likely to be different in unionized sectors because unions make across-firm and even across-sector comparisons. Babcock, Wang, and Loewenstein [1996], for example, provide evidence that wage bargaining between teachers' unions and school boards is strongly affected by reference wages in other school districts.

An obvious limitation of our model is that it cannot explain the evolution of play over time in the experiments discussed. Instead, our examination aims at the explanation of the stable behavioral patterns that emerge in these experiments after several periods. It is clear, that a model that solely focuses on equilibrium behavior cannot explain the time path of play. This limitation of our model also precludes a rigorous analysis of the

example, for $c = 0.5$ and $n = 4$, $|\beta_i| \geq 3$ is required. For $c = 0.2$ and $n = 4$, $|\beta_i|$ still has to exceed 0.75.

26. Bolton and Ockenfels [1997] develop a model similar to ours that differs in the choice of the reference payoff. In their model subjects compare themselves only with the average payoff of the group.

27. A related issue is the impact of social context on a person's degree of inequity aversion. It seems likely that a person has a different degree of inequity aversion when interacting with a friend in personal matters than in a business transaction with a stranger. In fact, evidence for this is provided by Loewenstein, Thompson, and Bazerman [1989]. However, note that in all experiments considered above interaction took place among anonymous strangers in a neutrally framed context.

short-run impact of equity considerations.[28] The empirical evidence suggests that equity considerations also have important short-run effects. This is obvious in ultimatum games, public good games with punishment, and gift exchange games, where equity considerations lead to substantial deviations from the selfish solution in the short and in the long run. However, they also seem to play a short-run role in market games with proposer or responder competition or public good games without punishment; that is, in games in which the selfish solution prevails in the long run. In these games the short-run deviation from equilibrium is typically in the direction of more equitable outcomes.[29]

VIII. Related Approaches in the Literature

There are several alternative approaches that try to account for persistent deviations from the predictions of the self-interest model by assuming a different motivational structure. The approach pioneered by Rabin [1993] emphasizes the role of intentions as a source of reciprocal behavior. Rabin's approach has recently been extended in interesting ways by Falk and Fischbacher [1998] and Dufwenberg and Kirchsteiger [1998]. Andreoni and Miller [1995] is based on the assumption of altruistic motives. Another interesting approach is Levine [1997] who assumes that people are *either* spiteful *or* altruistic to various degrees. Finally, there is the approach by Bolton and Ockenfels [1997] that is, like our model, based on a kind of inequity aversion.

The theory of reciprocity as developed by Rabin [1993] rests on the idea that people are willing to reward fair intentions and to punish unfair intentions. Like our approach, Rabin's model is also based on the notion of equity: player j perceives player i's intention as unfair if player i chooses an action that gives j less

28. In the short-run, minor changes in the (experimental) context can affect behavior. For example, there is evidence that subjects contribute more in a one-shot PD if it is called "community game" than if it is called "Wall Street Game." Under the plausible assumption that the community frame triggers more optimistic beliefs about other subjects inequity aversion our model is consistent with this observation.

29. Such short-run effects also are suggested by the results of Kahneman, Knetsch, and Thaler [1986] and Franciosi et al. [1995]. Franciosi et al. show that—in a competitive experimental market (without effort choices)—equity considerations significantly retard the adjustment to the (selfish) equilibrium. Ultimately, however, they do not prevent full adjustment to the equilibrium. Note that the retardation effect suggests that *temporary* demand shocks (e.g., after a natural disaster) may have no impact on prices at all because the shock vanishes before competitive forces can overcome the fairness-induced resistance to price changes.

than the equitable material payoff. The advantage of his model is that the disutility of an unfair offer can be explicitly interpreted as arising from j's judgment about i's unfair intention. As a consequence, player j's response to i's action can be explicitly interpreted as arising from j's desire to punish an unfair intention while our model does not explicitly suggest this interpretation of j's response. On the other hand, disadvantages of Rabin's model are that it is restricted to two-person normal form games and that it gives predictions if it is applied to the normal form of important sequential move games.[30]

The lack of explicit modeling of intentions in our model does, however, not imply that the model is incompatible with an intentions-based interpretation of reciprocal behavior. In our model reciprocal behavior is driven by the preference parameters α_i and β_i. The model is silent as to why α_i and β_i are positive. Whether these parameters are positive because individuals care directly for inequality or whether they infer intentions from actions that cause unequal outcomes is not modeled. Yet, this means that positive α_i's and β_i's can be interpreted as a direct concern for equality as well as a reduced-form concern for intentions. An intentions-based interpretation of our preference parameters is possible because bad or good intentions behind an action are, in general, inferred from the equity implications of the action. Therefore, people who have a desire to punish a bad intention behave as if they dislike being worse off relative to an equitable reference point and people who reward good intentions behave as if they dislike being better off relative to an equitable reference point. As a consequence, our preference parameters are compatible with the interpretation of intentions-driven reciprocity.

To illustrate this point further consider, e.g., an ultimatum game that is played under two different conditions [Blount 1995].

- In the "random" condition the first mover's offer is determined by a random device. The responder knows how the

30. In the sequentially played Prisoner's Dilemma, Rabin's model predicts that unconditional cooperation by the second mover is part of an equilibrium; i.e., the second mover cooperates even if the first mover defects. Moreover, conditional cooperation by the second mover is *not* part of an equilibrium. The data in Watabe et al. [1996] and Hayashi et al. [1998], however, show that unconditional cooperation is virtually nonexistent while conditional cooperation is the rule. Likewise, in the gift exchange game workers behave conditionally cooperative while unconditional cooperation is nonexistent. The reciprocity approaches of Falk and Fischbacher [1998] and of Dufwenberg and Kirchsteiger [1998] do not share this disadvantage of Rabin's model.

offer is generated and that the proposer cannot be held responsible for it.

- In the "intention" condition the proposer makes the offer himself and the responder knows that this is the proposer's deliberate choice.

In the intention condition the responder may not only be directly concerned about inequity. He may also react to the fairness of the perceived intentions of the proposer. In contrast, in the random condition it is only the concern for pure equity that may affect the responder's behavior. In fact, Blount [1995] reports that there are responders who reject positive but unequal offers in both conditions. However, the acceptance threshold is significantly higher in the intention condition.[31] Recall from Proposition 1 that there is a monotonic relationship between the acceptance threshold and the parameter α_i. Thus, this result suggests that the preference parameters do not remain constant across random and intention condition. Yet, for all games played in the intention condition and, hence, for all games considered in the previous sections, the preference parameters should be constant across games.

Altruism is consistent with voluntary giving in dictator and public good games. It is, however, inconsistent with the rejection of offers in the ultimatum game, and it cannot explain the huge behavioral differences between public good games with and without punishment. It also seems difficult to reconcile the extreme outcomes in market games with altruism. Levine's approach can explain extreme outcomes in market games as well as the evidence in the centipede game, but it cannot explain positive giving in the dictator game. It also *seems* that Levine's approach has difficulties in explaining that the *same* subjects behave very noncooperatively in the public good game without punishment, while they behave very cooperatively in the game with punishment.

The approach by Bolton and Ockenfels [1997] is similar to our model, although there are some differences in the details. For example, in their model people compare their material payoff with the material *average* payoff of the group. In our view the appropriate choice of the reference payoff is ultimately an empirical

31. Similar evidence is given by Charness [forthcoming] for a gift exchange game. For further evidence in favor of intentions-driven reciprocity, see Bolle and Kritikos [1998]. Surprisingly, and in contrast to these studies, Bolton, Brandts, and Katok [1997] and Bolton, Brandts, and Ockenfels [1997] find no evidence for intentions-driven reciprocity.

question that cannot be solved on the basis of the presently available evidence. There may well be situations in which the average payoff is the appropriate choice. However, in the context of the public good game with punishment, it seems to be inappropriate because it cannot explain why cooperators want to punish a defector. If there are, say, $n - 1$ fully cooperating subjects and one fully defecting subject, the payoff of each cooperator is below the group's average payoff. Cooperators can reduce this difference between own payoff and the group's average payoff by punishing one of the other players, i.e., they are indifferent between punishing other cooperators and the defector.

Bolton and Ockenfels [1997] assume that the marginal disutility of small deviations from equality is zero. Therefore, if subjects are nonsatiated in their own material payoff they will never *propose* an equal split in the dictator game. Likewise, they will—in case of nonsatiation in material payoffs—never propose an equal split in the ultimatum game unless $\alpha_2 = \infty$ for sufficiently many responders. Typically, the modal offer in most ultimatum game experiments is, however, the equal split. In addition, the assumption implies that complete free riding is the unique equilibrium in the public good game without punishment for *all* $a < 1$ and *all* $n \geq 2$. Their approach thus rules out equilibria where only a fraction of all subjects cooperate.[32]

IX. Summary

There are situations in which the standard self-interest model is unambiguously refuted. However, in other situations the predictions of this model seem to be very accurate. For example, in simple experiments like the ultimatum game, the public good game with punishments, or the gift exchange game, the vast majority of the subjects behave in a "fair" and "cooperative" manner although the self-interest model predicts very "unfair" and "noncooperative" behavior. Yet, there are also experiments like, e.g., market games or public good games without punishment, in which the vast majority of the subjects behaves in a rather "unfair" and "noncooperative" way—as predicted by the self-interest model. We show that this puzzling evidence can be explained in a coherent framework if—in addition to purely selfish people—there is a fraction of the population that cares for

32. Persistent asymmetric contributions are observed in Isaac, Walker, and Williams [1994].

equitable outcomes. Our theory is motivated by the psychological evidence on social comparison and loss aversion. It is very simple and can be applied to any game. The predictions of our model are consistent with the empirical evidence on all of the above-mentioned games. Our theory also has strong empirical implications for many other games. Therefore, it is an important task for future research to test the theory more rigorously against competing hypotheses. In addition, we believe that future research should aim at formalizing the role of intentions explicitly for the n-person case.

A main insight of our analysis is that there is an important interaction between the distribution of preferences in a given population and the strategic environment. We have shown that there are environments in which the behavior of a minority of purely selfish people forces the majority of fair-minded people to behave in a completely selfish manner, too. For example, in a market game with proposer or responder competition, it is very difficult, if not impossible, for fair players to achieve a "fair" outcome. Likewise, in a simultaneous public good game with punishment, even a small minority of selfish players can trigger the unraveling of cooperation. Yet, we have also shown that a minority of fair-minded players can force a big majority of selfish players to cooperate fully in the public good game with punishment. Similarly, our examination of the gift exchange game indicates that fairness considerations may give rise to stable wage rigidity despite the presence of strong competition among the workers. Thus, competition may or may not nullify the impact of equity considerations. If, despite the presence of competition, single individuals have opportunities to affect the relative material payoffs, equity considerations will affect market outcomes even in very competitive environments. In our view these results suggest that the interaction between the distribution of preferences and the economic environment deserves more attention in future research.

APPENDIX

Proof of Proposition 2

We first show that it is indeed a subgame perfect equilibrium if at least two proposers offer $s = 1$ which is accepted by the responder. Note first that the responder will accept any offer $\overline{s} \geq$

0.5, because

$$\text{(A1)} \qquad \bar{s} - \frac{1}{n-1}\beta_i(\bar{s} - 1 + \bar{s}) - \frac{n-2}{n-1}\beta_i(\bar{s} - 0) \geq 0.$$

To see this, note that (A1) is equivalent to

$$\text{(A2)} \qquad (n-1)\bar{s} \geq \beta_i(n\bar{s} - 1).$$

Since $\beta_i \leq 1$, this inequality clearly holds if

$$\text{(A3)} \qquad (n-1)\bar{s} \geq n\bar{s} - 1,$$

which must be the case since $s \leq 1$. Hence, the buyer will accept $s = 1$. Given that there is at least one other proposer who offers $s = 1$ and given that this offer will be accepted, each proposer gets a monetary payoff of 0 anyway, and no proposer can affect this outcome. Hence, it is indeed optimal for at least one other proposer to offer $s = 1$, too.

Next, we show that this is the unique equilibrium outcome. Suppose that there is another equilibrium in which $\bar{s} < 1$ with positive probability. This is only possible if each proposer offers $s < 1$ with positive probability. Let $\underline{s}_i$ be the lowest offer of proposer i that has positive probability. It cannot be the case that player i puts strictly positive probability on offers $s_i \in [\underline{s}_i, \underline{s}_j)$ because the probability that he wins with such an offer is zero. To see this, note that in this case player i would get

$$\text{(A4)} \qquad U_i(s_i) = -\frac{\alpha_i}{n-1}\bar{s} - \frac{\alpha_i}{n-1}(1 - \bar{s}) = -\frac{\alpha_i}{n-1}.$$

On the other hand, if proposer i chooses $s_i \in (\max_{j \neq i}[\underline{s}_j, 0.5], 1)$, then there is a positive probability that he will win—in which case he gets

$$\text{(A5)} \quad 1 - s_i - \frac{\alpha_i}{n-1}(2s_i - 1) - \frac{n-2}{n-1}\beta_i(1 - s_i)$$

$$> (1 - s_i)[1 - \frac{n-2}{n-1}\beta_i] - \frac{\alpha_i}{n-1} > -\frac{\alpha_i}{n-1}.$$

Of course, there may also be a positive probability that proposer i does not win, but in this case he again gets $-\alpha_i/(n-1)$. Thus, proposer i would deviate. It follows that it must be the case that $\underline{s}_i = \underline{s}$ for all i.

Suppose that proposer i changes his strategy and offers $\underline{s} + \varepsilon < 1$ in all states when his strategy would have required him to

choose $\underline{s}$. The cost of this change is that whenever proposer i would have won with the offer $\underline{s}$ he now receives only $1 - \underline{s} - \varepsilon$. However, by making ε arbitrarily small, this cost becomes arbitrarily small. The benefit is that there are now some states of the world which have strictly positive probability in which proposer i does win with the offer $\underline{s} + \varepsilon$ but in which he would not have won with the offer $\underline{s}$. This benefit is strictly positive and does not go to zero as ε becomes small. Hence, $\underline{s} < 1$ cannot be part of an equilibrium outcome.

QED

Proof of Proposition 3

We first show that $s = 1$, which is accepted by all responders, is indeed a subgame perfect equilibrium. Note that any offer $s \geq 0.5$ will be accepted by all responders. The argument is exactly the same as the one in the beginning of the proof of Proposition 1. The following Lemma will be useful.

LEMMA 1. For any $s < 0.5$ there exists a continuation equilibrium in which everybody accepts s.

Given that all other players accept s player i prefers to accept as well if and only if

$$\text{(A6)} \quad s - \frac{1}{n-1}\alpha_i(1 - s - s) - \frac{n-2}{n-1}\beta_i(s - 0) \geq 0 - \frac{1}{n-1}\alpha_i(1 - s) - \frac{1}{n-1}\alpha_i s,$$

which is equivalent to

$$\text{(A7)} \quad (1 - \beta_i)(n - 1) + 2\alpha_i + \beta_i \geq 0.$$

Since we assume that $\beta_i < 1$, this inequality must hold. □

Consider now the proposer. Clearly, it is never optimal to offer $s > 0.5$. Such an offer is always dominated by $s = 0.5$ which yields a higher monetary payoff and less inequality. On the other hand, we know by Lemma 1 that for any $s \leq 0.5$ there exists a continuation equilibrium in which this offer is accepted by everybody. Thus, we only have to look for the optimal s from the point of view of the proposer given that s will be accepted. His payoff function is

$$\text{(A8)} \quad U_1(s) = 1 - s - \frac{1}{n-1}\beta_1(1 - s - s) - \frac{n-2}{n-1}\beta_1(1 - s).$$

Differentiating with respect to s yields

$$\frac{dU_1}{ds} = -1 + \frac{2}{n-1}\beta_1 + \frac{n-2}{n-1}\beta_1, \tag{A9}$$

which is independent of s and is smaller than 0 if and only if

$$\beta_1 \leq (n-1)/n. \tag{A10}$$

Hence, if this condition holds, it is an equilibrium that the proposer offers $s = 0$ which is accepted by all responders. We now show that the highest offer that can be sustained in a subgame perfect equilibrium is given by (8).

LEMMA 2. Suppose that $s < 0.5$ has been offered. There exists a continuation equilibrium in which this offer is rejected by all responders if and only if

$$s \leq \frac{\alpha_i}{(1-\beta_i)(n-1) + 2\alpha_i + \beta_i} \qquad \forall i \in \{2, \ldots, n\}. \tag{A11}$$

Given that all other responders reject s, responder i will reject s as well if and only if

$$0 \geq s - \frac{\alpha_i}{n-1}(1-2s) - \frac{n-2}{n-1}\beta_i s, \tag{A12}$$

which is equivalent to (A11). Thus, (A11) is a sufficient condition for a continuation equilibrium in which s is rejected by everybody.

Suppose now that (A11) is violated for at least one $i \in \{2, \ldots, n\}$. We want to show that in this case there is no continuation equilibrium in which s is rejected by everybody. Note first that in this case responder i prefers to accept s if all other responders reject it. Suppose now that at least one other responder accepts s. In this case responder i prefers to accept s as well if and only if

(A13)

$$s - \frac{\alpha_i}{n-1}(1-2s) - \frac{n-2}{n-1}\beta_i s \geq 0 - \frac{\alpha_i}{n-1}(1-s) - \frac{\alpha_i}{n-1}s.$$

The right-hand side of this inequality is smaller than 0. We know already that the left-hand side is greater than 0 since (A11) is violated. Therefore, responder i prefers to accept s as well. We conclude that if (A11) does not hold for at least one i, then at least one responder will accept s. Hence, (A11) is also necessary. □

If $\beta_1 < (n-1)/n$, an equilibrium offer must be sustained by

the threat that any smaller offer $\tilde{s}$ will be rejected by everybody. But we know from Lemma 2 that an offer $\tilde{s}$ may be rejected only if (A11) holds for all i. Thus, the highest offer s that can be sustained in equilibrium is given by (8).

QED

Proof of Proposition 4

(a) Suppose that $1 - a > \beta_i$ for player i. Consider an arbitrary contribution vector $(g_1, \ldots, g_{i-1}, g_{i+1}, \ldots, g_n)$ of the other players. Without loss of generality we relabel the players such that $i = 1$ and $0 \leq g_2 \leq g_3 \leq \cdots \leq g_n$. If player 1 chooses $g_1 = 0$, his payoff is given by

$$U_1(g_1 = 0) = y + a \sum_{j=2}^{n} g_j - \frac{\beta}{n-1} \sum_{j=2}^{n} g_j. \tag{A14}$$

Note first that if all other players choose $g_j = 0$, too, then $g_1 = 0$ is clearly optimal. Furthermore, player 1 will never choose $g_1 > \max\{g_j\}$. Suppose that there is at least one player who chooses $g_j > 0$. If player 1 chooses $g_1 > 0, g_1 \in [g_k, g_{k+1}], k \in [2, \ldots, n]$, then his payoff is given by

$$\begin{aligned}
&U_1(g_1 > 0) \\
&= y - g_1 + ag_1 + a \sum_{j=2}^{n} g_j - \frac{\beta_1}{n-1} \sum_{j=k+1}^{n} (g_j - g_1) - \frac{\alpha_1}{n-1} \sum_{j=2}^{k} (g_1 - g_j) \\
&< y - g_1 + ag_1 + a \sum_{j=2}^{n} g_j - \frac{\beta_1}{n-1} \sum_{j=k+1}^{n} (g_j - g_1) + \frac{\beta_1}{n-1} \sum_{j=2}^{k} (g_1 - g_j) \\
&= y - g_1 + ag_1 + a \sum_{j=2}^{n} g_j - \frac{\beta_1}{n-1} \sum_{j=2}^{n} g_j \\
&\quad + \frac{\beta_1}{n-1}(n-1)g_1 \\
&= y - (1 - a - \beta_1)g_1 + a \sum_{j=2}^{n} g_j - \frac{\beta_1}{n-1} \sum_{j=2}^{n} g_j \\
&< y + a \sum_{j=2}^{n} g_j - \frac{\beta_1}{n-1} \sum_{j=2}^{n} g_j = U_1(g_1 = 0).
\end{aligned}$$

Hence, $g_i = 0$, is indeed a dominant strategy for player i.

(b) It is clearly an equilibrium if all players contribute nothing because to unilaterally contribute more than zero reduces the monetary payoff and causes disadvantageous inequality. Suppose that there exists another equilibrium with positive contribution levels. Relabel players such that $0 \le g_1 \le g_2 \le \cdots \le g_n$. By part (a) we know that all k players with $1 - \alpha > \beta_i$ must choose $g_i = 0$. Therefore, $0 = g_1 = \ldots g_k$. Consider player $l > k$ who has the smallest positive contribution level; i.e., $0 = g_{l-1} < g_l \le g_{l+1} \le \cdots \le g_n$. Player 1's utility is given by

$$\text{(A15)} \quad U_l(g_l) = y - g_l + \alpha g_l + \alpha \sum_{j=l+1}^{n} g_j - \frac{\beta_l}{n-1} \sum_{j=l+1}^{n} (g_j - g_l)$$

$$- \frac{\alpha_l}{n-1} \sum_{j=1}^{l-1} g_l = y + \alpha \sum_{j=l+1}^{n} g_j - \frac{\beta_l}{n-1} \sum_{j=l+1}^{n} g_j$$

$$- (1-\alpha) g_l + \beta_l \frac{n-l}{n-1} g_l - \alpha_l \frac{l-1}{n-1} g_l$$

$$= U_l(0) - (1-\alpha) g_l + \beta_l \frac{n-l}{n-1} g_l - \alpha_l \frac{l-1}{n-1} g_l,$$

where $U_l(0)$ is the utility player 1 gets if he deviates and chooses $g_l = 0$. Since $\alpha_l \ge \beta_l$, $l \ge k+1$, and $\beta_l < 1$, we have

$$\text{(A16)} \quad U_l(g_l) \le U_l(0) - (1-\alpha) g_l + \beta_l \frac{n-l}{n-1} g_l - \beta_l \frac{l-1}{n-1} g_l$$

$$\le U_l(0) - (1-\alpha) g_l + \beta_l \frac{n - 2(k+1) + 1}{n-1} g_l$$

$$< U_l(0) - (1-\alpha) g_l + \frac{n-2k-1}{n-1} g_l$$

$$= U_l(0) - \frac{(1-\alpha)(n-1) - (n-2k-1)}{n-1} g_l.$$

Thus if

$$\text{(A17)} \quad \frac{(1-\alpha)(n-1) - (n-2k-1)}{n-1} \ge 0,$$

player l prefers to deviate from the equilibrium candidate and to

choose $g_l = 0$. But this inequality is equivalent to

$$(A18) \qquad (1 - a)(n - 1) \geq n - 2k - 1$$

$$\Leftrightarrow a \leq 1 - \frac{n - 2k - 1}{n - 1}$$

$$\Leftrightarrow a \leq \frac{n - 1 - n + 2k + 1}{n - 1} = \frac{2k}{n - 1}$$

$$\Leftrightarrow \frac{k}{n - 1} \geq \frac{a}{2},$$

which is the condition given in the proposition.

(c) Suppose that the conditions of the proposition are satisfied. We want to construct an equilibrium in which all k players with $1 - a > \beta_i$ contribute nothing, while all other $n - k$ players contribute $g \in [0, y]$. We only have to check that contributing g is indeed optimal for the contributing players. Consider some player j with $1 - a < \beta_j$. If he contributes g, his payoff is given by

$$(A19) \qquad U_j(g) = y - g + (n - k)ag - [\alpha_j/(n - 1)]\, kg.$$

It clearly does not pay to contribute more than g. So suppose that player j reduces his contribution level by $\Delta > 0$. Then his payoff is

$$\begin{aligned} U_j(g - \Delta) &= y - g + \Delta + (n - k)ag - \Delta a \\ &\quad - \frac{\alpha_j}{n - 1} k(g - \Delta) - \frac{\beta_j}{n - 1}(n - k - 1)\Delta \\ &= y - g - (n - k)ag - \frac{\alpha_j}{n - 1} kg \\ &\quad + \Delta\left(1 - a + \frac{\alpha_j}{n - 1} k - \frac{\beta_j}{n - 1}(n - k - 1)\right) \\ &= U_j(g) + \Delta\left(1 - a + \frac{\alpha_j}{n - 1} k - \frac{\beta_j}{n - 1}(n - k - 1)\right). \end{aligned}$$

Thus, a deviation does not pay if and only if

$$1 - a + \frac{\alpha_j}{n - 1} k - \frac{\beta_j}{n - 1}(n - k - 1) \leq 0,$$

which is equivalent to

$$(A20) \qquad k/(n - 1) \leq (a + \beta_j - 1)/(\alpha_j + \beta_j).$$

Thus, if this condition holds for *all* $(n - k)$ players j with $1 - a < \beta_j$, then this is indeed an equilibrium. It remains to be shown that $(a + \beta_j - 1)/(\alpha_j + \beta_j) \leq a/2$. Note that $\alpha_j \geq \beta_j$ implies that $(a + \beta_j - 1)/(\alpha_j + \beta_j) \leq (a + \beta_j - 1)/(2\beta_j)$. Furthermore,

$$\frac{a + \beta_j - 1}{2\beta_j} \leq \frac{a}{2} \Leftrightarrow a + \beta_j - 1 \leq \beta_j a \Leftrightarrow a(1 - \beta_j) \leq 1 - \beta_j \Leftrightarrow a \leq 1,$$

which proves our claim.

QED

Proof of Proposition 5

Suppose that one of the players $i \in \{n' + 1, \ldots, n\}$ chooses $g_i < g$. If all players stick to the punishment strategies in stage 2, then deviator i gets the same monetary payoff as each enforcer $j \in \{1, \ldots, n'\}$. In this case monetary payoffs of i and j are given by

$$x_i = y - g_i + a[(n - 1)g + g_i] - n' \frac{g - g_i}{n' - c} \tag{A21}$$

$$\begin{aligned} x_j &= y - g + a[(n - 1)g + g_i] - c \frac{g - g_i}{n' - c} - \frac{n' - c}{n' - c}(g_i - g_i) \\ &= y - g_j + a[(n - 1)g + g_i] - (n' - c + c) \frac{g - g_i}{n' - c} = x_i. \end{aligned} \tag{A22}$$

Thus, given the punishment strategy of the enforcers, deviators cannot get a payoff higher than what the enforcers get. However, they get a strictly lower payoff than the nonenforcers who did not deviate. We now have to check that the punishment strategies are credible; i.e., that an enforcer cannot gain from reducing his p_{ij}. If an enforcer reduces p_{ij} by ε, he saves $c\varepsilon$ and experiences less disadvantageous inequality relative to those $(n - n' - 1)$ players who chose g but do not punish. This creates a nonpecuniary utility gain of $[\alpha_i(n - n' - 1)\ c\varepsilon]/(n - 1)$. On the other hand, the enforcer also has nonpecuniary costs because he experiences now disadvantageous inequality relative to the defector and a distributional advantage relative to the other $(n' - 1)$ enforcers who punish fully. The latter generates a utility loss of $\beta_i(n' - 1)c\varepsilon/(n - 1)$, whereas the former reduces utility by $\alpha_i(1 - c)\varepsilon/(n - 1)$. Thus, the loss from a reduction in p_{ij} is greater

than the gain if

(A23)

$$\frac{1}{n-1}\left[\alpha_i(1-c)\varepsilon + \beta_i(n'-1)c\varepsilon\right] > c\varepsilon + \alpha_i(n-n'-1)\frac{c\varepsilon}{n-1}$$

holds. Some simple algebraic manipulations show that condition (A23) is equivalent to condition (13). Hence, the punishment is credible.

Consider now the incentives of one of the enforcers to deviate in the first stage. Suppose that he reduces his contribution by $\varepsilon > 0$. Ignoring possible punishments in the second stage for a moment, player i gains $(1 - a)\varepsilon$ in monetary terms but incurs a nonpecuniary loss of $\beta_1\varepsilon$ by creating inequality to all other players. Since $1 - a < \beta_i$ by assumption, this deviation does not pay. If his defection triggers punishments in the second stage, then this reduces his monetary payoff which cannot make him better off than he would have been if he had chosen $g_i = g$. Hence, the enforcers are not going to deviate at stage 1 either. It is easy to see that choosing $g_i > g$ cannot be profitable for any player either, since it reduces the monetary payoff and increases inequality.

QED

Computation of the Probability That There Are Conditionally Cooperative Enforcers

To compute the probability that, in a randomly drawn group of four, there are subjects who obey condition (13) and $a + \beta_i \geq 1$, we have to make an assumption about the correlation between α_i and β_i. We mentioned already that the empirical evidence suggests that these parameters are positively correlated. For concreteness we assume that the correlation is perfect. Thus, in terms of Table III all players with $\alpha = 1$ or $\alpha = 4$ are assumed to have $\beta = 0.6$. This is clearly not fully realistic, but it simplifies the analysis dramatically.

In the Fehr-Gächter [1996] experiment the relevant parameters are $a = 0.4$, $n = 4$, and (roughly[33]) $c = 0.2$. The following summary states the conditions on α_i and β_i implied by Proposition 5 for a group of $n' \in \{1, \ldots, 4\}$ conditionally cooperative enforcers.

33. The cost function in Fehr and Gächter is actually convex, so that we have to slightly simplify their model. Yet, the vast majority of actual punishments occurred where $c = 0.2$.

If one of these conditions holds, cooperation can be sustained in equilibrium:

(i) $n' = 1, \alpha_i \geq 1.5$, and $\beta_i \geq 0.6$;
(ii) $n' = 2, \alpha_i \geq 1 - 0.\overline{3}\beta_i$, and $\beta_i \geq 0.6$;
(iii) $n' = 3, \alpha_i \geq 0.75 - 0.5\,\beta_i$, and $\beta_i \geq 0.6$;
(iv) $n' = 4, \alpha_i \geq 0.6 - 0.6\beta_i$, and $\beta_i \geq 0.6$.

Note that for each group n' of conditionally cooperative enforcers the conditions on α_i and β_i have to hold simultaneously. Given the discrete distribution of α and β of Table III, this can only be the case if

- there is at least one player with $\alpha_i = 4$ and $\beta_i = 0.6$, or
- there are at least two players with $\alpha_i = 1$ and $\beta_i = 0.6$, or
- both.

Given the numbers of Table III, it is not difficult to show that the probability that one of these cases applies is equal to 61.12 percent.

UNIVERSITY OF ZURICH
UNIVERSITY OF MUNICH

REFERENCES

Adams, J. Stacy, "Toward an Understanding of Inequity," *Journal of Abnormal and Social Psychology,* LXVII (1963), 422–436.

Agell, Jonas, and Per Lundberg, "Theories of Pay and Unemployment: Survey Evidence from Swedish Manufacturing Firms," *Scandinavian Journal of Economics,* XCVII (1995), 295–308.

Anderson, Simon P., Jacob K. Goeree, and Charles H. Holt, "Stochastic Game Theory—Adjustment to Equilibrium under Bounded Rationality," University of Virginia, Working Paper No. 304, 1997.

Andreoni, James, "Why Free Ride?—Strategies and Learning in Public Goods Experiments," *Journal of Public Economics,* XXXXVII (1988), 291–304.

——, "Cooperation in Public-Goods Experiments: Kindness or Confusion," *American Economic Review,* LXXXV (1995a), 891–904.

——, "Warm Glow versus Cold Prickle: The Effects of Positive and Negative Framing on Cooperation in Experiments," *Quarterly Journal of Economics,* CX (1995b), 1–21.

Andreoni, James, and John H. Miller, "Giving according to GARP: An Experimental Study of Rationality and Altruism," SSRI Working Paper, University of Wisconsin, Madison, 1996.

Babcock, Linda, Xianghong Wang, and George Loewenstein, "Choosing the Wrong Pond: Social Comparisons in Negotiations That Reflect a Self-Serving Bias," *Quarterly Journal of Economics,* CXI (1996), 1–21.

Banerjee, Abhihit V., "Envy," in: Dutta Bhaskar et al, eds., *Economic Theory and Policy, Essays in Honour of Dipak Banerjee* (Oxford; Oxford University Press, 1990).

Berg, Joyce, John Dickhaut, and Kevin McCabe, "Trust, Reciprocity, and Social History," *Games and Economic Behavior,* X (1995), 122–142.

Bewley, Truman, "A Depressed Labor Market as Explained by Participants," *American Economic Review, Papers and Proceedings,* LXXXV (1995), 250–254.

——, "Why not Cut Pay?" *European Economic Review,* XLII (1998), 459–490.

Blinder, Alan S., and Don H. Choi, "A Shred of Evidence on Theories of Wage Stickness," *Quarterly Journal of Economics,* CV (1990), 1003–1016.
Blount, Sally, "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences," *Organizational Behavior and Human Decision Processes,* LXIII (1995), 131–144.
Bolle, Friedel, and Alexander Kritikos, "Self-Centered Inequality Aversion versus Reciprocity and Altruism," Discussion Paper, Europa-Universität Viadrina, Frankfurt (Oder), 1998.
Bolton, Gary E., and Axel Ockenfels, "A Theory of Equity, Reciprocity and Competition," Discussion Paper, Pennsylvania State University, 1997.
Bolton, Gary E., Jordi Brandts, and Elena Katok, "A Simple Test of Explanations for Contributions in Dilemma Games," Discussion Paper, Pennsylvania State University, 1997.
Bolton, Gary E., Jordi Brandts, and Axel Ockenfels, "Measuring Motivations for Reciprocal Responses Observed in Simple Dilemma Games," Discussion Paper, Universität Magdeburg, 1997.
Burlando, Roberto, and John D. Hey, "Do Anglo-Saxons Free-Ride More?" *Journal of Public Economics,* LXIV (1997), 41–60.
Camerer, Colin, and Richard Thaler, "Ultimatums, Dictators, and Manners," *Journal of Economic Perspectives,* IX (1995), 209–219.
Cameron, Lisa, "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia," Discussion Paper, Princeton University, 1995.
Campbell, Carl M., and Kunal Kamlani, "The Reasons for Wage Rigidity: Evidence from a Survey of Firms," *Quarterly Journal of Economics,* CXII (1997), 759–789.
Charness, Gary, "Attribution and Reciprocity in a Labor Market: An Experimental Investigation," *Games and Economic Behavior,* forthcoming.
Clark, Andrew E., and Andrew J. Oswald, "Satisfaction and Comparison Income," *Journal of Public Economics,* LXI (1996), 359–381.
Croson, Rachel, T. A., "Expectations in Voluntary Contributions Mechanisms," Discussion Paper, Wharton School, University of Pennsylvania, 1995.
———, "Partners and Strangers Revisited," *Economics Letters,* LIII (1996), 25–32.
Davis, Douglas, and Charles Holt, *Experimental Economics* (Princeton, NJ: Princeton University Press 1993).
Davis, J. A., "A Formal Interpretation of the Theory of Relative Deprivation," *Sociometry,* XXII (1959), 280–296.
Dawes, Robyn M., and Richard Thaler, "Cooperation," *Journal of Economic Perspectives,* II (1988), 187–197.
Dufwenberg, Martin, and Georg Kirchsteiger, "A Theory of Sequential Reciprocity," Discussion Paper, CentER, Tilburg University, 1998.
Falk, Armin, and Urs Fischbacher, "A Theory of Reciprocity," Discussion Paper, University of Zürich, 1998.
Falkinger, Josef, Ernst Fehr, Simon Gächter, and Rudolf Winter-Ebmer, "A Simple Mechanism for the Efficient Provision of Public Goods—Experimental Evidence," *American Economic Review,* forthcoming.
Fehr, Ernst and Armin Falk, "Wage Rigidity in a Competitive Incomplete Contract Market," *Journal of Political Economy,* CVII (1999), 106–134.
Fehr, Ernst, and Simon Gächter, "Cooperation and Punishment—An Experimental Analysis of Norm Formation and Norm Enforcement," Discussion Paper, Institute for Empirical Research in Economics, University of Zürich, 1996.
Fehr, Ernst, Simon Gächter, and Georg Kirchsteiger, "Reciprocity as a Contract Enforcement Device," *Econometrica,* LXV, No. 4, (1996), 833–860.
Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl, "Does Fairness Prevent Market Clearing? An Experimental Investigation," *Quarterly Journal of Economics,* CVIII (1993), 437–460.
Festinger, L., "A Theory of Social Comparison Processes," *Human Relations,* VII (1954), 117–140.
Forsythe, Robert, N. E. Horowitz, L. Hoel Savin, and Martin Sefton, "Fairness in Simple Bargaining Games," *Games and Economic Behavior,* VI (1988), 347–369.
Franciosi, Robert, Praveen Kujal, Roland Michelitsch, Vernon Smith, and Gang Deng, "Fairness: Effect on Temporary and Equilibrium Prices in Posted-Offer Markets," *Economic Journal,* CV (1995), 938–950.

Frank, Robert H., *Choosing the Right Pond—Human Behavior and the Quest for Status* (Oxford: Oxford University Press, 1985).
Friedman, Daniel, and John Rust, *The Double Auction Market—Institutions, Theories and Evidence* (Reading, MA: Addison-Wesley Publishing Company, 1993).
Güth, Werner, Nadège Marchand, and Jean-Louis Rulliere, "On the Reliability of Reciprocal Fairness—An Experimental Study," Discussion Paper, Humboldt University Berlin, 1997.
Güth, Werner, Rolf Schmittberger, and Bernd Schwarze, "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization,* III (1982), 367–388.
Güth, Werner, Rolf Schmittberger, and Reinhard Tietz, "Ultimatum Bargaining Behavior—A Survey and Comparison of Experimental Results," *Journal of Economic Psychology,* XI (1990), 417–449.
Haltiwanger, John, and Michael Waldman, "Rational Expectations and the Limits of Rationality," *American Economic Review,* LXXV (1985), 326–340.
Hayashi, Nahoko, Elinor Ostrom, James Walker, and Toshio Yamagishi, "Reciprocity, Trust, and the Sense of Control: A Cross-Societal Study," Discussion Paper, Indiana University, Bloomington, 1998.
Hoffman, Elizabeth, Kevin McCabe, and Vernon Smith, "On Expectations and Monetary, Stakes in Ultimatum Games," *International Journal of Game Theory,* XXV (1996), 289–301.
Homans, G. C., *Social Behavior: Its Elementary Forms* (New York: Harcourt, Brace & World, 1961).
Isaac, Mark R., and James M. Walker, "Group Size Effects in Public Goods Provision: The Voluntary Contribution Mechanism," *Quarterly Journal of Economics,* CIII (1988), 179–199.
Isaac, Mark R., and James M. Walker, "Costly Communication: An Experiment in a Nested Public Goods Problem," in Thomas R. Palfrey, ed., *Laboratory Research in Political Economy* (Ann Arbor: University of Michigan Press, 1991).
Isaac, Mark R., James M. Walker, and Arlington M. Williams, "Group Size and the Voluntary Provision of Public Goods: Experimental Evidence Utilizing Large Groups," *Journal of Public Economics,* LIV (1994), 1–36.
Kachelmeier, Steven J., and Mohamed Shehata, "Culture and Competition: A Laboratory Market Comparison between China and the West," *Journal of Economic Organization and Behavior,* XIX (1992), 145–168.
Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler, "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review,* LXXVI (1986), 728–741.
Keser, Claudia, and Frans van Winden, "Partners Contribute More to Public Goods than Strangers: Conditional Cooperation," Discussion Paper, University of Karlsruhe, 1996.
Ledyard, John, "Public Goods: A Survey of Experimental Research," in J. Kagel and A. Roth, eds., *Handbook of Experimental Economics* (Princeton: Princeton University Press, 1995).
Levine, David K., "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics,* forthcoming (1997).
Loewenstein, George F., Leigh Thompson, and Max H. Bazerman, "Social Utility and Decision Making in Interpersonal Contexts," *Journal of Personality and Social Psychology,* LVII (1989), 426–441.
McKelvey, Richard D., and Thomas R. Palfrey, "Quantal Response Equilibria for Normal Form Games," *Games and Economic Behavior,* X (1995), 6–38.
Mueller, Denis, *Public Choice II* (Cambridge: Cambridge University Press, 1989).
Ockenfels, Axel, and Joachim Weimann, "Types and Patterns—An Experimental East-West Comparison of Cooperation and Solidarity," Discussion Paper, Department of Economics, University of Magdeburg, 1996.
Ostrom, Elinor, and James M. Walker, "Cooperation without External Enforcement," in Thomas R. Palfrey, ed., *Laboratory Research in Political Economy* (Ann Arbor: University of Michigan Press, 1991).
Pollis, N. P., "Reference Groups Re-examined," *British Journal of Sociology,* XIX (1968), 300–307.

Rabin, Matthew, "Incorporating Fairness into Game Theory and Economics," *American Economic Review,* LXXXIII (1993), 1281–1302.
Rehder, Robert, "Japanese Transplants: After the Honeymoon," *Business Horizons* (1990), 87–98.
Roth, Alvin E., "Bargaining Experiments," in J. Kagel and A. Roth, eds., *Handbook of Experimental Economics* (Princeton: Princeton University Press, 1995).
Roth, Alvin, E., and Ido Erev, "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term," *Games and Economic Behavior,* VIII (1995), 164–212.
Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir, "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review,* LXXXI (1991), 1068–1095.
Runciman, Walter G., *Relative Deprivation and Social Justice* (New York: Penguin, 1966).
Russell, Thomas, and Richard Thaler, "The Relevance of Quasi Rationality in Competitive Markets," *American Economic Review,* LXXV (1985), 1071–1082.
Sadrieh, Abdolkarim, *The Alternating Double Auction Market* (Berlin: Springer, 1998).
Skinner, Jonathan, and Joel Slemrod, "An Economic Perspective on Tax Evasion," *National Tax Journal,* XXXVIII (1985), 345–353.
Slonim, Robert, and Alvin E. Roth, "Financial Incentives and Learning in Ultimatum and Market Games: An Experiment in the Slovak Republic," *Econometrica,* LXVI (1997), 569–596.
Smith, Vernon L., "An Experimental Study of Competitive Market Behavior," *Journal of Political Economy,* LXX (1962), 111–137.
——, "Microeconomic Systems as an Experimental Science," *American Economic Review,* LXXII (1982), 923–955.
Smith, Vernon L., and Arlington W. Williams, "The Boundaries of Competitive Price Theory: Convergence Expectations and Transaction Costs," in L. Green and J. H. Kagel, eds., *Advances in Behavioural Economics,* Vol. 2 (Norwood, NJ: Ablex Publishing Corporation, 1990).
Stouffer, Samuel, A., *The American Soldier* (Princeton: Princeton University Press, 1949).
Thaler, Richard H., "The Ultimatum Game," *Journal of Economic Perspectives,* II (1988), 195–206.
Tversky, A., and D. Kahneman, "Loss Aversion in Riskless Choice: A Reference-Dependent Model," *Quarterly Journal of Economics,* CVI (1991), 1039–1062.
Watabe, M., S. Terai, N., Hayashi, and T. Yamagishi, "Cooperation in the One-Shot Prisoner's Dilemma Based on Expectations of Reciprocity," *Japanese Journal of Experimental Social Psychology,* XXXVI (1996), 183–196.
Whyte, William F., *Money and Motivation* (New York: Harper and Brothers, 1955).