A joint Initiative of Ludwig-Maximilians-Universität and Ifo Institute for Economic Research

CESifo Working Papers

A THEORY OF RECIPROCITY

Armin Falk
Urs Fischbacher*

CESifo Working Paper No. 457

April 2001

CESifo

Center for Economic Studies & Ifo Institute for Economic Research
Poschingerstr. 5, 81679 Munich, Germany

Tel.: +49 (89) 9224-1410
Fax: +49 (89) 9224-1409
e-mail: office@CESifo.de

An electronic version of the paper may be downloaded
- from the SSRN website: www.SSRN.com
- from the CESifo website: www.CESifo.de

* Financial support by the Swiss National Science Foundation (Project 12-43590.95) and by the MacArthur Foundation (Network on Economic Environments and the Evolution of Individual Preferences and Social Norms) is gratefully acknowledged. We would like to thank Sam Bowles, Martin Brown, Simon Gächter, Herbert Gintis, Lorenz Götte, Michael Kosfeld, Matthew Rabin, Armin Schmutzler and the participants of MacArthur Foundtion meeting in Chicago 1998.

CESifo *Working Paper No. 457*
April 2001

A THEORY OF RECIPROCITY

Abstract

This paper presents a formal theory of reciprocity. Reciprocity means that people reward kind actions and punish unkind ones. The theory takes into account that people evaluate the kindness of an action not only by its consequences but also by the intention underlying this action. The theory explains the relevant stylized facts of a wide range of experimental games. Among them are the ultimatum game, the gift-exchange game, a reduced best-shot game, the dictator game, the prisoner's dilemma, and public goods games. Furthermore, the theory explains why the same consequences trigger different reciprocal responses in different environments Finally the theory explains why in bilaterial interactions outcomes tend to be 'fair' whereas in competitive markets even extremely unfair distributions may arise.

JEL Classification: C7, C91, C92, D64, H41

Keywords: Reciprocity, fairness, cooperation, competition, game theory.

Armin Falk
University of Zurich
Institute for Empirical
Economic Research
Bluemlisalpstr. 10
CH-8006 Zürich
Switzerland
Falk@iew.unizh.ch

Urs Fischbacher
University of Zurich
Institute for Empirical
Economic Research
Bluemlisalpstr. 10
CH-8006 Zürich
Switzerland
Fiba@iew.unizh.ch

Kindness is the parent of kindness.
(Adam Smith 1759)

1 Introduction

In this paper we develop a formal theory of reciprocity. According to this theory, reciprocity is a behavioral response to perceived kindness and unkindness, where kindness comprises distributional fairness as well as fairness intentions. There is by now a large body of evidence which indicates that reciprocity is a powerful determinant of human behavior: Experiments and questionnaire studies performed by psychologists and economists as well as an impressive literature in sociology, ethnology and anthropology emphasize the omnipresence of reciprocal behavior. The sociologist Gouldner (1960), for example, notes that the norm of reciprocity is "no less universal and important an element of culture than the incest taboo..." (p. 171).

The importance of reciprocity for economics has been pointed out by many scholars. Kahneman, Knetsch, and Thaler (1986), e.g., show that fairness norms prevail in a variety of business contexts. In the field of labor economics, questionnaire studies with owners and managers of firms suggest that a possible source for rigid wages is that the employers are unwilling to cut wages (Bewley (1995), Agell and Lundborg (1995), Campbell and Kamlani (1997)). According to these studies a major reason for firms' refusal to cut wages is the fear that pay cuts will adversely affect work morale. Thus concerns for reciprocity may play a key role in the explanation of downwardly rigid wages. Economic consequences of reciprocity have also been shown in many other areas, such as tax compliance (Smith (1992)), organization theory (Steers and Porter (1991), Mowday (1991)), contributions to public goods (Sugden (1984)), contract enforcement (Fehr, Gächter and Kirchsteiger (1997)), gift-giving (Ruffle (1995)) and strike breaking (Francis (1985)).[1]

The most compelling evidence for the importance of reciprocity comes from controlled laboratory experiments. In the ultimatum game, e.g., many people reject low offers in order to punish the unkindness of proposers (Güth, Schmittberger, and Schwarze (1982)).[2] The reward of kind actions is reported, e.g., in the invest-

[1] Importantly, reciprocity means a behavior that cannot be justified in terms of selfish and purely outcome oriented preferences. To avoid terminological confusion let us, therefore, clarify that reciprocity sharply distinguishes from 'reciprocal altruism' (Trivers (1971)). A reciprocal altruist is only willing to reciprocate if there are future rewards arising from reciprocal actions. In the parlance of game theory this kind of reciprocal action may be supported as an equilibrium strategy in infinitely repeated games (folk theorems) or in finitely repeated games with incomplete information (see Kreps, Milgrom, Roberts, and Wilson (1982)).

[2] All games mentioned in this section are explicitly described in the applications section.

ment game (Berg, Dickhaut and McCabe (1995)) or in the gift-exchange game (Fehr, Kirchsteiger, and Riedl (1993)). There is, however, evidence from market experiments that *seems* to be incompatible with reciprocal preferences. These experiments typically support the outcome predicted by standard economic theory which assumes completely selfish preferences. We show that our theory is capable to reconcile the seemingly contradictory evidence that in bilateral interactions outcomes seem to be 'fair' while in competitive markets 'very unfair' distributions may arise.

According to our theory, a reciprocal action is modeled as the behavioral response to an action that is perceived as either kind or unkind. The central part of the theory is therefore devoted to the question *how people evaluate whether an experienced action is kind or unkind*. In our model two aspects are essential, (i) the *consequences* of an action and (ii) the underlying *intentions*. The fact that intentions play a major role for the perception of kindness is by now well documented. In a recent experiment by Falk, Fehr and Fischbacher (2000a), e.g., second movers could reciprocate first movers' kind or unkind actions. In a treatment where first movers actually decided, we observe strong positive and negative reciprocity. In a treatment where first movers' actions are determined randomly, however, reciprocal responses to the *same* 'actions' are significantly weaker. Similarly, Falk, Fehr and Fischbacher (2000b) show that in a series of reduced ultimatum games, the exact same offer is rejected at a significantly different rate, depending on the choice set of the proposers. A given offer x is rejected at a significantly higher rate if the proposers' action signals bad intentions (because he could have chosen a more friendly offer) compared to a situation where x signals no or even fair intentions. Thus intentions clearly matter for the perception of kindness and the corresponding reciprocation[3]. Notice, however, that even in situations where intentions are absent, most people still exhibit some reciprocal behavior. In Falk, Fehr and Fischbacher (2000a) second movers punish extremely unfair offers and reward very advantageous offers, even if offers were determined randomly. This finding is corroborated by the experiments by Blount (1995) and Charness (1996) who report that in a condition where intentions play no role, reciprocity is weak but not absent. In our model we therefore incorporate *both* the concern for the consequence or outcome per se and for the underlying intentions.

Our concept of reciprocity differs from other fairness models. While in our model reciprocity is the response to kindness or unkindness, two recently developed models of inequity aversion assume that it is the dislike of inequitable distributions which triggers behavioral responses (Bolton and Ockenfels (2000) and Fehr and

[3]Experimental evidence for the importance of intentions is also found in Bolle and Kritikos (1998), Brandts and Sola (forthcoming), McCabe, Rigdon and Smith (2000), Offerman (1999), Greensberg and Frisch (1972), and Goranson and Berkowitz (1966).

Schmidt (1999)). Moreover, the inequity aversion approach takes a consequentialistic perspective: The models assume that fairness driven punishments or rewards can fully be accounted for without taking intentions into account, i.e., the distributive consequences of an action alone trigger reciprocal actions. Quite to the contrary, Dufwenberg and Kirchsteiger (1999) and Rabin (1993) assume that reciprocity is exclusively intentions driven.

Since our theory models both, intentions and outcomes, it contains variants of pure intention models and pure inequity aversion models as special cases. We readily agree that the complexity of the model is quite high and that the calculation of equilibria is more difficult compared to the inequity aversion models. We think, however, that - given its predictive power - this complexity is justified. Moreover, we would like to emphasis that the model has only two free parameters, just as many as, e.g., the Fehr and Schmidt model and surely less than the model of Charness and Rabin (2000).[4] Although our model accounts for intentions, it does not share the disadvantage of the purely intentions based theories which usually predict many equilibria. Instead, our theory makes *unique* and testable predictions. Compared to the purely intentions based theories it is therefore much better suited as a predictive tool. Notice, that all predictions in this paper are derived with a *single* utility function using the *same* parameter constellation in all games.

The remainder of this paper is organized as follows: In the following section we introduce the formal theory. Section 3 discusses some applications. The games we explain are the ultimatum game, the gift-exchange game, a reduced best-shot game, competitive market games, the dictator game, the prisoner's dilemma and public goods games. Section 4 concludes.

2 The Model

Our theory formalizes the basic structure of reciprocity which consists of a kind (or unkind) treatment by another person (represented by the *kindness term* φ) and a behavioral reaction to that treatment (represented by the *reciprocation term* σ). Our procedure is to transform a standard game into a psychological game, the so-called "reciprocity game". In this new game the players' utility depends not only on the payoffs of the original standard game but also on the kindness and the reciprocation term. In the following, we derive both terms.

Consider a two-player extensive form game with a finite number of stages and with complete and perfect information. (For notational simplicity we develop the

[4] In the Fehr and Schmidt model, econometric testing would imply the estimation of the parameters α and β. In our model, the two free parameters are the reciprocity parameter ρ and the pure outcome parameter ε (see below).

theory for the two-player case. The extension to games with more than 2 players is given in Appendix 2.) Let $i \in \{1, 2\}$ be a player in the game. N denotes the set of nodes and N_i is the set of nodes where player i has the move. Let $n \in N$ be a node of the game. Let A_n be the set of actions in node n. Let F be the set of end nodes of the game. The payoff function for player i is given by $\pi_i : F \to \mathbb{R}$.

Let $P(A_n)$ be the set of probability distributions over the set of actions in node n. Then $S_i = \prod_{n \in N_i} P(A_n)$ is player i's behavior strategy space. Thus, a player's behavior strategy puts a probability distribution on each of the player's decision nodes. Let player j be the other player[5] and let k be one of the players (either i or j). For $s_i \in S_i$ and $s_j \in S_j$ we define $\pi_k(s_i, s_j)$ as k's expected payoff, given strategies s_i and s_j.

Let $s_i \in S_i$ be a behavior strategy. We define $s_i|n$ as the strategy s_i **conditional on** node n. This strategy is simply s_i except that for all nodes n' which are ahead of n, the probability of the unique action leading to n is set to 1 and the probabilities of the other actions in nodes n' are set to 0. Furthermore, we define $\pi_k(n, s_i, s_j) := \pi_k(s_i|n, s_j|n)$ as the expected payoff conditional on node $n \in N$: It is the expected payoff of player k in the subgame starting from node n, given that the strategies s_i and s_j are played.

Let s_i' denote the **first order belief** of player i. It captures i's belief about the behavior strategy $s_j \in S_j$ which player j will choose. Similarly, the **second order belief** s_i'' of player i is defined as i's belief about j's belief about which behavior strategy i will choose. In other words, s_i'' is i's belief about s_j'. Like Rabin (1993), we assume that s_i' is an element of S_j and s_i'' is an element of S_i. A set of beliefs is said to be **consistent**, if $s_i = s_j' = s_i''$ holds for $i \neq j$.

2.1 The Kindness Term φ

The kindness term φ is the central element of our theory. It measures how kind a person perceives the action by another player. As outlined in the previous section, the perceived kindness of an action depends on the consequences or outcomes of that action and the underlying intentions. In our theory, the outcome is measured with the **outcome term** Δ where $\Delta > 0$ expresses an advantageous outcome and $\Delta < 0$ expresses a disadvantageous outcome. In order to determine the overall kindness, Δ is multiplied with the **intention factor** ϑ. This factor is a number between zero and one, where $\vartheta = 1$ captures a situation where Δ is induced fully intentionally and $\vartheta < 1$ implies a situation where an action is not fully intentional. The kindness term φ is simply the product of Δ and ϑ:

[5] Throughout the paper we will use the male form for player i (and for first movers). For player j (and second movers) we will use the female form.

Definition 1 Let strategies and beliefs be given. We define the **kindness term** $\varphi_j(n)$ in a node $n \in N_i$ as:

$$\varphi_j(n) = \vartheta_j(n)\Delta_j(n) \tag{1}$$

In the following we derive both terms (Δ and ϑ) in detail. First, we define the **outcome term**:

$$\Delta_j(n) := \pi_i(n, s_i'', s_i') - \pi_j(n, s_i'', s_i') \tag{2}$$

To interpret this expression, let us fix the intention factor $\vartheta_j(n)$. For a given $\vartheta_j(n)$, the outcome term $\Delta_j(n)$ captures the kindness of player j: The kindness of player j in node n is, ceteris paribus, higher, the more she offers to player i. This is expressed in the term $\pi_i(n, s_i'', s_i')$. From j's perspective, j is offering $\pi_i(n, s_j', s_j)$ to i. This is the payoff i is expected to get, if j chooses s_j and expects i to choose s_j'. Player i's belief about this offer is $\pi_i(n, s_i'', s_i')$.

The sign of $\Delta_j(n)$ determines whether an action is considered as kind or unkind. In order to determine the sign of $\Delta_j(n)$, i needs to compare the offer $\pi_i(n, s_i'', s_i')$ with a *reference standard*. From many experiments it is known, that an *equitable* share of payoffs is a salient and commonly held standard.[6] The expression $\pi_j(n, s_i'', s_i')$ serves that purpose. It is i's belief about what payoff j wants to keep for herself. Taken together, if $\pi_i(n, s_i'', s_i') > \pi_j(n, s_i'', s_i')$ holds, player i thinks that j wants him to get more than j wants for herself, i.e., i believes that j is acting kindly. If, on the other hand, $\pi_i(n, s_i'', s_i') < \pi_j(n, s_i'', s_i')$ holds, i believes that j claims more for herself than she is willing to leave for i. In this case, i perceives j as being unkind.[7]

Let us now derive the **intention factor** ϑ which measures the importance of fairness intentions. The signaling of fairness intentions rests on two premises: (i) Player j's choice set actually allows the choice between a fair and an unfair action, and (ii) j's choice is under her full control. From these two premises it immediately follows that in order to evaluate the intentions of a particular action of j, player i has to look at the alternatives j had, i.e., he takes into account j's strategy set. In order to get a better understanding about *how* j's strategy set shapes i's perception of j's kindness, we conducted a questionnaire study with 111 subjects. In this study each subject i was in a hypothetical bilateral exchange situation with another subject j. Subjects i were asked to indicate how *kind* or *unkind* they perceive different divisions

[6] The idea that equity is a salient reference standard was first developed in the so-called *equity theory*. Beginning in the late sixties social psychologists developed *equity theory* as a special form of *social exchange theory*. Compare, e.g., Adams (1965) and Walster and walster (1978). See also Loewenstein, Thompson and Bazerman (1989).

[7] Note that we talk a bit loosely about the kindness of an *action*. The way we model kindness comprises both the kindness of actually occurred actions as well as anticipated future actions. The actions which already occurred are represented in node n because being in node n results from player j's actually chosen actions. The anticipated actions of j are simply captured by s_i'.

of an endowment, where it was always j who divides the pie between herself and i. The study reveals *five* interesting observations which we use to model the intention factor:[8]

First, if j's strategy set contains only one element, i.e., if j has no alternative to choose, the kindness or unkindness of an offer is much weaker, compared to a situation where j can choose between fair and unfair offers. *Second*, even if j has no alternative and therefore cannot signal any intention, perceived kindness or unkindness is not zero. People have a dislike for an unfair outcome even if this outcome was caused unintentionally. *Third*, even if j's strategy space is *limited*, a friendly offer x is viewed as similarly kind compared to an unlimited strategy space as long as j could have made a less friendly offer to i. This means that x signals fair intentions if j could have been less friendly. By the same token, the kindness of the x-offer is lower if player j does not have the chance to make a less friendly offer. The intuition for the latter result is straightforward. If j has no chance to behave more 'opportunistically', how should i infer from a friendly action, that j really wanted to be kind? After all she took the least friendly action. *Fourth*, the perception of an unfriendly offer y depends on j's possibility to make a more friendly offer: If j has no chance to make a more friendly offer, y is not viewed as very unkind. The intuition is that you cannot blame a person for being mean if - after all - he did the best she could. *Fifth*, if j does have the option to make a more friendly offer z, the perception of y depends on how much j has to sacrifice in order to make the more friendly offer z. If z implies that i earns more than j, y is still perceived as unkind but not that much. The intuition is that one cannot unambiguously infer from j's unwillingness to propose an *unfair offer to herself* that she wants to be unfair to i. If, however, there is an offer z which is more friendly and does *not* imply that j earns less than i, player i thinks that it is reasonable to get offer z while getting y is considered as quite unkind.

These five observations will be used to model the interaction of alternatives and perceived kindness. To do so let us first state precisely what we mean by alternative payoff combinations. Let S_j^p be the set of pure strategies of j. For given strategies and beliefs we define in a node n:

$$\Pi_i(n) := \left\{ (\pi_i(s_i''|n, s_j^p), \pi_j(s_i''|n, s_j^p)) \mid s_j^p \in S_j^p \right\} \tag{3}$$

$\Pi_i(n)$ is a set of payoff combinations. This set contains the payoff combinations, player j can induce by choosing a pure strategy s_j^p given her beliefs about player i's strategy . Since $\Pi_i(n)$ is determined from player i's perspective, player i takes into

[8] The study was performed under anonymous conditions with students from the University of Zurich and the Polytechnical University of Zurich. It was conducted in May and June 1998. A full description of the study and its results is given in Falk and Fischbacher (2001).

account his belief about which strategy player j believes he will choose, namely $s_i''|n$. In short, $\Pi_i(n)$ is player i's belief about all payoff combinations player j considers as her payoff opportunity set.

As we have discussed, player i's perception of j's intentions depends on the options available to j. Our five observations are approximated with the help of Ω.[9] In this function, the value of Ω indicates how intentional player j's choice of a given payoff combination (π_i^0, π_j^0) is perceived, given j's alternatives $(\tilde{\pi}_i, \tilde{\pi}_j)$. If the choice was fully intentional Ω equals 1, if the choice is considered as not fully intentional, however, Ω is smaller than one. The Ω-function is defined as follows:

$$\Omega(\tilde{\pi}_i, \tilde{\pi}_j, \pi_i^0, \pi_j^0) := \begin{cases} 1 & \text{if } \pi_i^0 \geq \pi_j^0 \text{ and } \tilde{\pi}_i < \pi_i^0 \\ \varepsilon_i & \text{if } \pi_i^0 \geq \pi_j^0 \text{ and } \tilde{\pi}_i \geq \pi_i^0 \\ 1 & \text{if } \pi_i^0 < \pi_j^0,\ \tilde{\pi}_i > \pi_i^0 \text{ and } \tilde{\pi}_i \leq \tilde{\pi}_j \\ \max\left(1 - \frac{\tilde{\pi}_i - \tilde{\pi}_j}{\pi_j^0 - \pi_i^0}, \varepsilon_i\right) & \text{if } \pi_i^0 < \pi_j^0,\ \tilde{\pi}_i > \pi_i^0 \text{ and } \tilde{\pi}_i > \tilde{\pi}_j \\ \varepsilon_i & \text{if } \pi_i^0 < \pi_j^0 \text{ and } \tilde{\pi}_i \leq \pi_i^0 \end{cases} \tag{4}$$

where ε_i is an individual parameter with $0 \leq \varepsilon_i \leq 1$.[10]

The first two rows capture situations where j has treated i in a kind way ($\pi_i^0 \geq \pi_j^0$). In these situations the value of Ω depends on whether j could have reduced i's payoff ($\tilde{\pi}_i$ compared to π_i^0) or not. Assume, as shown in the first row, player j *has* the alternative to lower player i's payoff (i.e., $\tilde{\pi}_i < \pi_i^0$). Then player i considers the kind action as fully intentional. This corresponds to our third observation. If, however, $\tilde{\pi}_i \geq \pi_i^0$ holds, j has no alternative to be less kind. Our third observation suggests that in this case j's kindness is considered as less intentional. After all, it is not j's merit that she was 'kind'. In this situation Ω equals ε_i.

The other three rows represent instances where j puts i in a disadvantageous situation, i.e., where $\pi_i^0 < \pi_j^0$ holds. If j *has* the alternative to improve i's payoff without putting herself in a disadvantageous situation ($\tilde{\pi}_i > \pi_i^0$ and $\tilde{\pi}_i \leq \tilde{\pi}_j$), her unkindness is fully intentional. Therefore, Ω is equal to 1 (see our fifth observation). Now suppose that there *is* an alternative to improve i's payoff, but this alternative leads to a disadvantageous situation for j. The more this alternative is disadvantageous for player j, the less reasonable it is considered. As a consequence, the choice of π_i^0 is not considered as fully intentionally unkind and Ω is equal to $\max\left(1 - \frac{\tilde{\pi}_i - \tilde{\pi}_j}{\pi_j^0 - \pi_i^0}, \varepsilon_i\right) \leq 1$. The expression $1 - \frac{\tilde{\pi}_i - \tilde{\pi}_j}{\pi_j^0 - \pi_i^0}$ measures 'how much j must put herself into a disadvantageous situation' if she wants to improve i's payoff - related to the reference situation

[9] We do not claim that Ω completely captures the richness of the relationship between alternatives and intentions. For reasons of simplicity we think that it is justified to restrict Ω to some key aspects.

[10] This parameter is called the pure outcome concern parameter. It is interpreted below.

(π_i^0, π_j^0).[11]. Finally, if j's only alternative is to choose an even lower payoff for player i, i.e., $\tilde{\pi}_i \leq \pi_i^0$, i cannot infer that j wanted to treat him in an unkind fashion. Consequently, the action was unintentionally 'unkind' yielding $\Omega = \varepsilon_i$ (see our fourth observation).

As the reference distribution (π_i^0, π_j^0) we use the payoffs that determine the outcome term $\Delta_j(n)$, namely $\pi_i(n, s_i'', s_i')$ and $\pi_j(n, s_i'', s_i')$. Thus, we define the **intention factor**:

$$\vartheta_j(n) = \max\left\{\Omega(\tilde{\pi}_i, \tilde{\pi}_j, \pi_i(n, s_i'', s_i'), \pi_j(n, s_i'', s_i')) \mid (\tilde{\pi}_i, \tilde{\pi}_j) \in \Pi_i(n)\right\} \tag{5}$$

The maximum-operator guarantees that a particular action is considered as intentional if there is *any* 'true' alternative. Note that in the special case where player j has *no* alternative at all (i.e., $|\Pi_i(n)| = 1$), $\vartheta_j(n)$ equals ε_i (this corresponds to our first observation).

The individual parameter ε_i is called the **pure outcome concern parameter**. It measures a player's *pure* concern for an equitable *outcome*: If, e.g., ε_i is equal to zero, player i considers a particular outcome only as kind or unkind if it was caused intentionally, i.e., if the other player had an alternative to act differently. The assumption of $\varepsilon_i = 0$ would coincide with a purely intentions driven notion of fairness. However, according to our second observation (see above), we expect ε_i to be strictly larger than zero. Assuming $\varepsilon_i = 1$ implies that i cares *only* about the consequences of j's action, i.e., intentions play no role. Thus, saying that a person's ε_i is equal to 1 is equivalent to saying that this person is purely outcome oriented as suggested by the models of Bolton and Ockenfels (2000) and Fehr and Schmidt (1999). Insofar, their models can be viewed as a special case of our theory.

2.2 The Reciprocation Term σ

The second ingredient of our theory concerns the formalization of reciprocation. Let us fix an end node f that follows (directly or indirectly) node n. Then we denote by $\nu(n, f)$ the unique node that directly follows node n on the path that leads from n to f.

Notation: Let n_1 and n_2 be nodes. If node n_2 follows node n_1 (directly or indirectly), we denote this by $n_1 \rightarrow n_2$.

Definition 2 Let strategies and beliefs be given as above. Let i and j be the two players and n and f be defined as above. Then we define

$$\sigma_i(n, f) := \pi_j(\nu(n, f), s_i'', s_i') - \pi_j(n, s_i'', s_i') \tag{6}$$

[11] If, e.g., j must put herself only a little bit to the disadvantageous side (the numerator is small) the alternative action will *ceteris paribus* be considered as rather reasonable. If, however, the numerator is large (in particular, if the numerator is larger than the denominator) Ω is equal to ε_i.

as the **reciprocation term** of player i in node n.

The *reciprocation term* expresses the response to the experienced kindness, i.e., it measures how much i alters the payoff of j with his move in node n. Given i's belief about j's expectations about her payoff in node n (i.e., given $\pi_j(n, s_i'', s_i')$), i can - in node n - choose an action. The reciprocal impact of this action is represented as the *alteration* of j's payoff from $\pi_j(n, s_i'', s_i')$ to $\pi_j(\nu(n, f), s_i'', s_i')$ (always from i's perspective). For a given $\pi_j(n, s_i'', s_i')$, i can, thus, choose to either reward or to punish j. A rewarding action implies a positive, whereas a punishment implies a negative *reciprocation term*.

2.3 The Utility Function

Having defined the kindness and reciprocation term we can now derive the players' utility of the transformed "reciprocity game":

Definition 3 Let i and j be the two players of the game. Let f be an end node of the game. We define the utility in the transformed reciprocity game as:

$$U_i(f) = \pi_i(f) + \rho_i \sum_{\substack{n \to f \\ n \in N_i}} \varphi_j(n)\sigma_i(n, f) \tag{7}$$

According to Definition 3 player i's utility in the reciprocity game is the sum of the following two terms: The first summand is simply player i's **material payoff** $\pi_i(f)$. The second summand - which we call **reciprocity utility** - is composed of the reciprocity parameter ρ_i, the kindness term $\varphi_j(n)$ an the reciprocation term $\sigma_i(n, f)$.

The **reciprocity parameter** ρ_i is a positive constant. It is an individual parameter which captures the strength of player i's reciprocal preferences. The higher ρ_i, the more important is the reciprocity utility as compared to the utility arising from the material payoff. Note that if ρ_i is zero, i's utility is equal to his material payoff. If, in addition, ρ_j is also zero, the reciprocity game collapses into the standard game.

The product of the *kindness* and the *reciprocation term* measures the reciprocity utility in a particular node. If the kindness term in a particular node n is greater than zero, player i can *ceteris paribus* increase his utility if he chooses an action in that node which increases j's payoff. The opposite holds if the kindness term is negative. In this case, i has an incentive to reduce j's payoff. Since kindness is measured in each node where i has the move, the overall reciprocity utility is the sum of the reciprocity utility in all nodes (before the considered end node), weighted with the reciprocity parameter.

2.4 The Reciprocity Equilibrium

The introduced preferences form a psychological game (Geanakoplos, Pearce and Stacchetti (1989)). In psychological games, the utility of a player i does not only depend on the selected strategies of the players but also on the i's beliefs (compare Definition 3). Note, however, that beliefs are not part of the action space. Put differently, beliefs cannot be formed strategically, i.e., they are taken as given. Given the beliefs, player i chooses his optimal strategy. The additional requirement in a psychological Nash equilibrium as compared to a Nash equilibrium is that all beliefs match actual behavior. This means, that an optimal strategy is only part of an equilibrium if the beliefs are also consistent with actual behavior.

Geanakoplos, Pearce and Stacchetti (1989) show that the refinement concept of subgame perfectness can also be applied to psychological Nash equilibria. In our reciprocity game we will call a subgame perfect psychological Nash equilibrium a **reciprocity equilibrium**. If $\rho_i = \rho_j = 0$, the definition of a reciprocity equilibrium is equivalent to the definition of a subgame perfect Nash equilibrium.[12]

3 Applications

In this section we discuss the predictions of our theory in different experimental games. The games under study are the ultimatum game, the gift-exchange game, a reduced best-shot game, market games with proposer or responder competition, the dictator game, the prisoner's dilemma and public goods games.[13] Appendix 3 contains the propositions that describe the reciprocity equilibria as well as the corresponding proofs.[14]

3.1 Negative Reciprocity: The Ultimatum Game

The most known game in which negative reciprocity applies is the ultimatum game. In this two person sequential move game, the first mover ("proposer") is allocated an amount of money (which we normalize to 1). The proposer has to divide this amount between himself and a second mover ("responder"). He may offer any feasible amount c to the responder, i.e., $0 \leq c \leq 1$. After the offer is revealed to the responder, the latter can either accept or reject it. If she accepts, the resulting payoffs are $1 - c$ for

[12] A remark on the existence of reciprocity equilibria: In the present form, a reciprocity equilibrium does not always exist because the function Ω is discontinuous at $\tilde{\pi}_i = \pi_i^0$. A minor technical modification of Ω, however, guarantees the existence of a reciprocity equilibrium. For the ease of exposition we delegate the existence proof to Appendix 1.

[13] Notice that in all figures which illustrate our propositions we use the *same* set of parameters ($\rho_i = 2$, $\varepsilon_i = 0.2$).

[14] This appendix is available as a pdf-document at: http://www.iew.unizh.ch/home/falk/fafiA3.pdf

the proposer and c for the responder. If the responder rejects the offer, payoffs are zero for both parties. Given the standard assumptions, the outcome according to the subgame perfect Nash equilibrium is ($c = 0$; accept).

The ultimatum game has been studied intensively. Overviews of experimental results are presented, e.g., in Güth, Schmittberger and Schwarze (1982), Thaler (1988), Güth (1995), Camerer and Thaler (1995) and Roth (1995). The reported behavioral regularities are quite robust and can be summarized as follows: There are (i) practically no offers that exceed 0.5, (ii) the modal offers lie in a range between 0.4 and 0.5, (iii) offers below 0.2 are extremely rare, and (iv) whereas offers close to 0.5 are practically never rejected, the rejection rate for offers below 0.2 is rather high. These stylized facts are in strong contrast to the standard prediction.

We now state our predictions. Upon acceptance, material payoffs are $1-c$ for the proposer and c for the responder, respectively. Let p denote the probability that the responder accepts the offer.

Proposition 1 *If ρ_1 and ρ_2 are positive there is a unique reciprocity equilibrium (c^*, p^*) in the ultimatum game as follows:*

$$p^* = \begin{cases} \min\left(1, \frac{c}{\rho_2 \cdot (1-2c)(1-c)}\right) & if\ c < \frac{1}{2} \\ 1 & if\ c \geq \frac{1}{2} \end{cases} \tag{8}$$

$$c^* = \max\left[\frac{1+3\rho_2 - \sqrt{1+6\rho_2+\rho_2^2}}{4\rho_2}, \frac{1}{2} \cdot (1 - \frac{1}{\rho_1})\right] \tag{9}$$

If either ρ_1 or ρ_2 is zero p^* and c^* are the limits of the above formulas where ρ_1 and ρ_2 approach zero from above.

If ρ_1 and ρ_2 are both zero, $p^* = 1$ and $c^* = 0$.

Discussion: Let us first look at the responder's behavior. Equation (8) reveals the conditions that determine the acceptance probability p^* of an offer c in the reciprocity equilibrium: First, if the proposer's offer is equal or higher than half of the pie, i.e., $c \geq \frac{1}{2}$, the responder will *always* accept the offer. This holds irrespective of the responder's concern for reciprocity (compare the second row of Equation (8)). Second, for offers smaller than half of the pie the willingness to accept an offer is increasing in the level of the offer and decreasing in the responder's concern for reciprocity, ρ_2 (see the first row of Equation (8)).

We now turn to the proposer. Equation (9) shows that his equilibrium choice of c depends on two expressions. While the first expression depends on the *responder's* reciprocal inclination, i.e., on ρ_2, the second expression depends on the *proposer's* concern for reciprocity, i.e., on ρ_1. The second expression represents the proposer's

intrinsic concern for a fair outcome. If he is an egoistic player the expression is zero. If ρ_1 is large, however, he will offer a positive c. The first expression can be interpreted as an *extrinsic* constraint to offer a positive c: This expression corresponds to the offer that maximizes the material payoff of the proposer, given the rejection behavior of the responder. It is equal to the smallest offer that just guarantees an acceptance probability of 1. (We call this offer c_0, i.e., $c_0 := \frac{1+3\rho_2-\sqrt{1+6\rho_2+\rho_2^2}}{4\rho_2}$). Obviously, the responder's concern for reciprocity is crucial for the value of c_0. For example, if the responder is a selfish player (with $\rho_2 = 0$) the value of c_0 aproaches zero. The higher player 2's concern for reciprocity, the higher the value of c_0. As ρ_2 gets very large, c_0 approaches $\frac{1}{2}$.

The equilibrium offer c^* is the maximum of the first and the second expression of Equation (9): This means, e.g., that in case a selfish proposer plays against a reciprocal responder he will offer a higher share, the higher ρ_2. In this case the extrinsic constraint is binding. If, however, the responder has a very low ρ_2, i.e., he accepts practically any offer, the equilibrium offer is determined by the proposer's concern for an equitable outcome: If he is rather selfish too, his offer will be (close to) zero. If ρ_1 is rather high, however, the offer will be high as well.[15]

Before we turn to the next game we will briefly discuss the predictions of our theory for the non-intentional treatment reported by Blount (1995)[16]. In her treatment the proposers' offers were not chosen by human subjects but instead randomly selected. Consequently, a low offer did not signal any (bad) intentions. As the data of Blount's experiment reveals, the acceptance rate for a given offer is *much higher* than in the 'regular' treatment. However, even in the absence of intentions some subjects reject extremely disadvantageous offers. Our theory predicts exactly these two stylized facts. In the non-intentional treatment the equilibrium acceptance rate for p^* is given by:

$$p^* = \begin{cases} \min\left(1, \frac{c}{\varepsilon_2 \rho_2 \cdot (1-2c)(1-c)}\right) & \text{if } c < \frac{1}{2} \\ 1 & \text{if } c \geq \frac{1}{2} \end{cases}$$

Figure 1 depicts the predicted acceptance probabilities in the 'regular' ultimatum game and in Blount's treatment for a given ρ_2. The upper graph corresponds to the Blount-experiment whereas the lower graph shows the acceptance behavior in the 'regular' treatment. As can be seen in the figure, a responder's acceptance probability for low offers is higher if intentions are absent.

The lower the outcome concern parameter ε_2, the more the upper graph shifts to the left. Put differently, the more a responder is sensitive with respect to intentions

[15] The range of ρ_1- and ρ_2- combinations where the equilibrium offer c^* equals c_0 is given by $\rho_2 \geq \frac{\rho_1(\rho_1-1)}{\rho_1+1}$. This holds in particular if $\rho_2 \geq \rho_1$.

[16] For similar results and a discussion on intentions, see Falk, Fehr and Fischbacher (2000a).

relative to outcomes, the more likely she will accept very low offers. On the other hand, if a responder is purely outcome oriented, i.e., if $\varepsilon_2 = 1$ holds, she exhibits the *same* behavioral pattern as in the 'regular' treatment. As Blount's data reveals, however, people care for both, outcomes and intentions (i.e., $0 < \varepsilon_i < 1$).

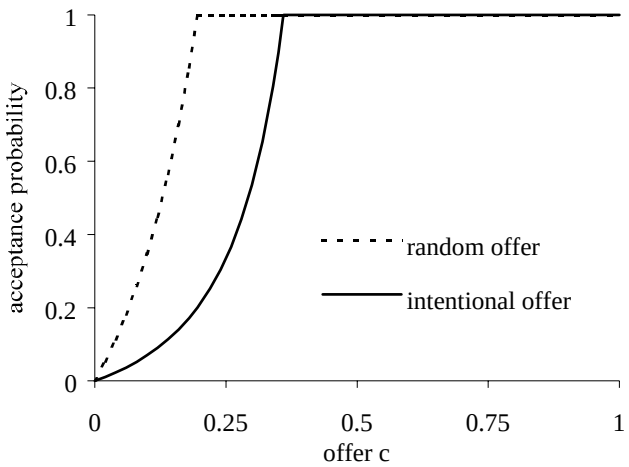

Figure 1: Acceptance probabilities in the ultimatum game with intentions (lower graph) and without intentions (upper graph) depending on the offer. The parameters $\rho_2 = 2$ and $\varepsilon_2 = 0.2$ are given.

3.2 Positive Reciprocity: The Gift-Exchange Game

One well known experiment which shows the importance of positive reciprocity is the gift-exchange game. In this two-person sequential game, the first mover (called an employer) offers a wage w to a second mover (called a worker). After receiving the wage, a worker has to make an effort decision e. Providing effort above the minimum effort level is costly with $c(e)$ being a convex effort cost function. Payoff functions are given by $\pi_1 = ve - w$ for employers and $\pi_2 = w - c(e)$ for workers, respectively. A rational and selfish worker will - irrespective of the wage - choose the minimum effort level. With backward induction, the employer will offer only the lowest possible wage. Contrary to this prediction the main experimental findings are (i) that wages clearly exceed the lowest possible wage and (ii) that there is a positive wage-effort relation. Both findings are remarkably robust. They hold in bilateral institutions as well as in competitive market institutions (see, e.g., Fehr, Kirchsteiger and Riedl (1993), Fehr and Falk (1999) and Gächter and Falk (1997)).

Our model predicts the main stylized facts. In equilibrium a worker's effort choice

equals $e^* = 0$ (if $\rho_2 = 0$) and $e^* = \min\left(1, \frac{-2\alpha-\rho_2+\sqrt{(2\alpha+\rho_2)^2+8\alpha\rho_2^2 w}}{2\alpha\rho_2}\right)$ if $\rho_2 > 0$. This prediction is illustrated in Figure 2 where we depict a worker's effort choice depending on the wage given her reciprocal motivation ρ_2. This figure shows the essence of *positive reciprocity* as reported in the gift-exchange experiments. From the worker's perspective, the higher the wage paid by the firm, the higher is the kindness term φ. A reciprocal worker (with a sufficiently high reciprocity parameter ρ_2) improves her utility if she responds in kind, i.e., if she provides more than the minimum effort. As a result, reciprocal workers provide higher effort levels the higher the wage paid by 'their' firm. Moreover, workers will - for a given wage - choose higher effort levels, the stronger their reciprocal inclination. Given that reciprocal workers reward high wages with high efforts, firms' best response is to pay wages strictly above zero.

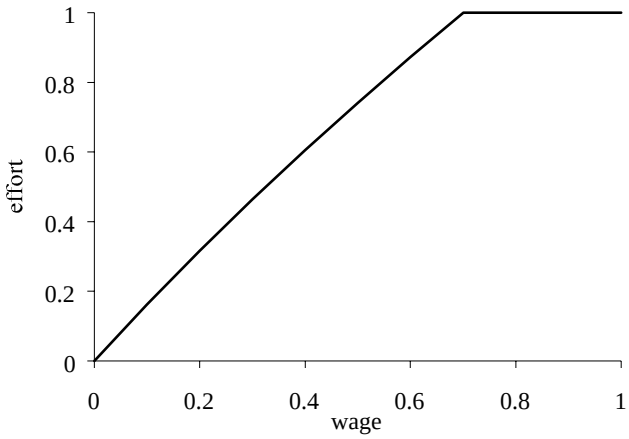

Figure 2: Effort choice depending on the wage paid for a given ρ_2 and a given α ($\rho_2 = 2$ and $\alpha = 0.2$).

There is also a non-intentional treatment of the gift-exchange game where wages are determined by a third party or a random mechanism (Charness 1996). Compared to the 'regular' treatment, this leads to a *weaker* correlation between wages and effort levels. The reason is that in the random treatments a high wage does not signal any (kind) intentions. Our model explains this pattern. Since the kindness term captures both, the concern for outcomes *and* intentions, the kindness term is smaller in the random treatment. For a given wage, the reciprocal action is therefore weaker as well.

3.3 Further Games

In this section we briefly discuss the intuition of our theory's predictions in some further games.

3.3.1 Identical Outcomes Yield Different Responses: A Comparison Between Best-Shot and Ultimatum Games.

The best-shot game was introduced by Harrison and Hirshleifer (1989) and Prasnikar and Roth (1992). The interesting feature of this experiment is that second movers are willing to accept a higher degree of inequity than in ultimatum games. This difference cannot be explained by the inequity aversion models by Bolton and Ockenfels (2000) and Fehr and Schmidt (1999). Our explanation for the difference between behavior in the best-shot game and the ultimatum game rests on the importance of intentions. This will become clear with the help of two stylized games, a reduced ultimatum and a reduced best-shot game (Falk, Fehr and Fischbacher 2000b).

In the reduced best-shot game a first mover can offer two different payoff distributions (2/8) and (8/2) which can be accepted or rejected by the second mover. If the second mover accepts, the offered payoff distribution is implemented. Otherwise both players receive nothing. The crucial feature of this game (and the richer original best-shot game) is that the first mover can only offer a payoff share that is either very advantageous or very disadvantageous to himself (8/2 or 2/8). This is different in the reduced ultimatum game where the first mover can either choose the offer (8/2) or the fair offer (5/5).

According to our theory, the predicted acceptance probability q^* for the unkind offer (8/2) in the reduced best-shot game is given by $q^* = \min\left(1, \frac{5}{12}\frac{1}{\rho_2}\right)$ while it is $q^* = \min\left(1, \frac{5}{12}\frac{1}{\varepsilon_2\rho_2}\right)$ in the reduced ultimatum game. In both games, the acceptance probability of the unkind offer decreases in the second mover's concern for reciprocity, i.e., in ρ_2. We also see, however, that for a given ρ_2 the acceptance probability is lower in the ultimatum game compared to the best-shot game. Thus, in the *best-shot game a reciprocal second mover is willing to accept a higher degree of inequity.*

The reason is that in the best-shot game the only alternative to offering (8/2) is to offer a very disadvantageous offer, namely (2/8). Since such an alternative implies that the first mover switches from an advantageous to a disadvantageous situation, player 2 will not find it very unkind if he gets the 'unkind' offer, because she cannot infer very bad intentions from this offer. Things are quite different in the reduced ultimatum game. Here, the first mover has a 'reasonable alternative', namely to choose the fair (5/5)-offer. Consequently, it *does* signal bad intentions

and is considered as quite unkind if this opportunity is not chosen. Thus, dependent on the first mover's alternatives the *same* offer will be accepted with a different probability. This difference is confirmed in the experimental study by Falk, Fehr, and Fischbacher (2000b): The rejection rate of the (8/2)-offer is 27 percent in the reduced best-shot game and 44 percent in the reduced ultimatum game.

3.3.2 Competition

In the preceding games we have analyzed only bilateral interactions. In particular, we restricted our analysis to games without any competition at all. In this section we, therefore, apply our theory to games with more than 2 players and show how competitive pressure interacts with reciprocal preferences.[17]

It is a well established fact in experimental economics that in competitive institutions market outcomes converge very well towards the outcome predicted by standard economic theory (Smith (1982), Davis and Holt (1993)). This holds even in markets where the equilibrium outcome is very "unfair" in the sense that almost the whole surplus is reaped by one side of the market. Put differently, it seems that in a competitive environment subjects behave as if they were completely selfish. In this section we show that our theory explains why in bilateral institutions outcomes tend to be "fair" while in markets reciprocal subjects' behavior gives rise to equilibria that imply an extremely uneven distribution of the gains from trade.

As an illustration we analyze a market game with proposer competition which has been studied by Roth, Prasnikar, Okuno-Fujiwara, and Zamir (1991). In this game there are $n-1$ proposers who simultaneously propose an offer $c_i \in [0,1]$ to the responder with $i \in [1, n-1]$. These offers are revealed to the responder who has to decide whether to accept or reject the highest offer c_{max}. If more than one proposer offers c_{max} a random mechanism determines whose offer will be selected. Payoffs are exactly as in the ultimatum game, i.e., the proposer whose offer is accepted receives $1-c_{max}$ and the responder gets c_{max}. A proposer whose offer is not accepted receives a payoff of zero. If the responder rejects c_{max}, all receive nothing. Assuming rational and purely selfish players the subgame perfect outcome is straightforward: At least two proposers offer $c_{max} = 1$ which is accepted by the responder. In fact, this prediction is supported by the experimental data. Roth, Prasnikar, Okuno-Fujiwara, and Zamir (1991) report that in their experimental sessions (with 9 proposers) this equilibrium outcome was reached after a few periods. Moreover, these results were very robust and showed up in four different countries. Given the experimental results of the preceding bilateral games this is remarkable since

[17] This requires an extension of our theory to games with more than 2 players. This extension is simply notational and will be discussed in Appendix 2.

in this equilibrium the proposers earn practically nothing while the responder gets the whole surplus. As it turns out, the prediction of our theory coincides with the standard prediction. In a reciprocity equilibrium in the market game with proposer competition, at least 2 proposers offer $c_{max} = 1$ which is accepted by the responder.

The striking feature of this prediction is that - independent of their reciprocal inclination - proposers will accept a very uneven distribution of the pie. The intuition of that result is that in a competitive market a proposer has *no chance* to achieve a "fair" outcome: From the *ultimatum game* we know that a responder will accept any offer larger or equal to 0.5 with certainty. Since offering more decreases a proposer's material and reciprocity utility a proposer will - in the ultimatum game - never offer more than that. Assume that in a market game with two proposers a reciprocal proposer i refuses to offer more than 0.5. What will the other proposer do? By infinitesimally overbidding player i's offer he can on the one hand increase his material payoff by a positive amount (because he can increase the winning probability from 0.5 to 1). On the other hand, the reciprocity disutility resulting from the unfair relation to the responder does only change infinitesimally. This means that player i's refusal to propose more than 0.5 is not an effective tool to achieve a "fair" outcome. As a consequence, he tries to overbid the other proposer to get at least a minimal share of the pie. This "overbidding" inevitably leads to the equilibrium prediction.

3.3.3 The Dictator Game

The dictator game is a very simple two person game. The task of the first mover (the so-called "dictator") is to divide an amount of money between himself and a counterpart (the "receiver"). Let 1 be the amount of money and c the share for the receiver. The dictator is free to choose any feasible division he wants to ($0 \leq c \leq 1$). The payoff for the receiver is c while the dictator's payoff is given by the residual amount $1 - c$.

The dictator game has been studied, e.g., by Forsythe, Horowitz, Savin and Sefton (1994) and by Hoffman, McCabe, Shachat and Smith (1994) and Eckel and Grossman (1998). The stylized facts can be summarized as follows. (i) Dictator offers larger than half of the pie, i.e., $c > \frac{1}{2}$ are practically never observed. (ii) Roughly 80 percent of the offers are between zero and half of the pie, i.e., $0 < c \leq \frac{1}{2}$. However, *compared to the ultimatum game, the distribution of offers is shifted towards zero.* (iii) About 20 percent of the dictators offer the amount predicted by standard game theory, i.e., they offer exactly zero.[18]

[18] It should be noted that the results of the dictator game are not very robust with respect to treatment variations. Increasing, e.g., the social distance among participants of an experiment and the experimenter (double blind treatment) increases the percentage of zero proposals (compare Hoffman,

Compatible with these stylized facts our theory predicts a unique reciprocity equilibrium. In this equilibrium the dictator offers $c^* = \max\left[0, \frac{1}{2} \cdot \left(1 - \frac{1}{\varepsilon_1 \rho_1}\right)\right]$. Thus, a dictator's proposal depends (i) on ρ_1, i.e., on how strong his other-regarding preferences enter his utility and (ii) on his pure outcome concern parameter ε_1. If $\varepsilon_1 \rho_1 > 1$, the dictator offers a positive amount of money. Even for very high values of $\varepsilon_1 \rho_1$, however, the dictator's offer will never exceed $\frac{1}{2}$. If $\varepsilon_1 \rho_1 \leq 1$, the dictator chooses $c = 0$. Comparing the equilibrium offers in the dictator game with those of the ultimatum game, we see that the equation that determines the equilibrium offer in the dictator game equals the second expression in equation (9) of Proposition 1 - if we replace ρ_1 by $\varepsilon_1 \rho_1$. Since $\varepsilon_1 \leq 1$, the same person will always offer at least as much in the ultimatum game as in the dictator game. Thus, consistent with stylized facts (i) to (iii), our theory predicts that dictators offer between zero and half of the pie and that the distribution of offers in the dictator game is shifted downwards compared to the corresponding distribution in the ultimatum game.

3.3.4 The Prisoner's Dilemma and Public Goods Games

In a sequential prisoner's dilemma player 1 can either cooperate or defect. After observing player 1's choice player 2 has the same choice. Assuming rational and selfish actors, the subgame perfect solution is that both players defect. Contrary to this prediction, our theory predicts the following: First, if player 2 is sufficiently reciprocally motivated, there is a positive probability that player 2 rewards player 1's cooperation with cooperation. Second, player 2 always defects if player 1 has defected beforehand. Taken together our theory predicts conditional cooperation and excludes the possibility of altruistic behaviour, i.e., non-conditional cooperation.

Experimental studies of sequential versions of the prisoner' s dilemma are reported in Bolle and Ockenfels (1990) and Clark and Sefton (1998). The results of their studies are in line with our predictions. In particular, unconditional cooperation is practically inexistent. In their conclusion Clark and Sefton, e.g., note that "cooperation is better regarded as reciprocation rather than unconditional altruism: second-movers cooperate quite frequently in response to cooperation by first-movers, but rarely cooperate after the first-mover defects" (p. 17).

We also analyze the simultaneous prisoner's dilemma. According to our theory, we get *less* cooperation if players choose simultaneously compared to the sequential move structure. There is strong evidence in favor of this prediction. Experimental studies by Watabe, Terai, Hayashi, and Yamagishi (1996) and Hayashi, Ostrom, Walker, and Yamagishi (1998) show that cooperation rates among first movers in the sequential prisoner's dilemma are higher, compared to simultaneous move games.

McCabe and Smith (1996)).

The strategic structure of the prisoner's dilemma is very similar to that of *public goods games*. The major difference is that in a public goods game players have more strategies than just to cooperate or to defect. Most public goods experiments have been conducted as simultaneous move games. However, there is a recent experiment by Fischbacher, Gächter and Fehr (Forthcoming) where subjects could *conditionally* indicate how many tokens they wanted to contribute to the public good. Despite the fact that the best reply is to provide zero tokens irrespective of the other group members' contributions, subjects on average contributed more the higher the contributions of the other group members. This 'conditional cooperation strategy' was in most cases specified such that subjects provided less than the group average. This is exactly what our theory predicts: In a public goods game subjects with a sufficiently high reciprocal inclination will conditionally cooperate but always cooperate slightly less than the other player(s). Moreover, our theory replicates another stylized fact of public goods games, namely that the propensity to cooperate increases in the marginal per capita return of an investment into the public good (see Ledyard (1995)).

4 Concluding Remarks

In this paper we have presented a formal theory of reciprocity. According to the theory people reward kind and punish unkind actions. Kindness comprises both the consequences as well as the intentions of an action. In this sense, the theory distinguishes from the consequentialistic inequity aversion-models as well as from pure intentions driven models of fairness. Our theory captures the empirical finding that the same consequences of an action are perceived and reciprocated differently, depending on the underlying intentions. The theory is also capable to reconcile the puzzling evidence that in competitive experimental markets very unfair outcomes emerge, while in bilateral bargaining situations outcomes tend to be fair.

Additionally to the different treatment of intentions there is another distinguishing feature between the concept of inequity aversion and our reciprocity approach which is worth to be emphasized: According to the inequity aversion approach, a person will exhibit reciprocal behavior only if this reduces the inequity between the person and his opponent. In our theory a person's reciprocal action is driven by the desire to reward or to punish. Put differently, the reciprocal action aims at reducing or raising the other person's payoff, regardless of whether this reduces the inequity between the players or not. This difference is important not only because it concerns the proper understanding and modelling of the nature of fair behavior. It also implies different predictions: In Falk, Fehr and Fischbacher (2000b) this question is addressed in great detail. One experiment in this paper is a three person prisoner's dilemma

with a subsequent punishment stage. In this experiment punishments imply that the inequity between cooperators and defectors is not reduced but even *increased*. As a consequence, inequity aversion models predict *no* punishments in this situation. Since defection is regarded as an unkind act, however, our theory predicts that reciprocally motivated cooperators do punish defectors. The results are unambiguous: 46.8 percent of the cooperators punish defectors. These results suggest that reciprocal behavior is primarily driven as a response to kindness - not as a desire to reduce inequity.

References

Adams, J. S. (1965): "Inequity in Social Exchange", in: Leonhard Berkowitz (ed.), Advances in Experimental Psychology 2, New York: Academic Press, 267-299.

Agell, J. and Lundborg, P. (1995): "Theories of Pay and Unemployment: Survey Evidence from Swedish Manufacturing Firms", Scandinavian Journal of Economics 97, 295 - 307.

Berg, J., Dickhaut J. and McCabe K. (1995): "Trust, Reciprocity, and Social History", Games and Economic Behavior 10, 122-142.

Bewley, T. (1995): "A Depressed Labor Market as Explained by Participants", American Economic Review 85, Papers and Proceedings, 250 - 254.

Blount, S. (1995): "When Social Outcomes aren't Fair: The Effect of Causal Attributions on Preferences", Organizational Behavior & Human Decision Processes 63, 131-144.

Bolle, F. and Ockenfels, P. (1990): "Prisoner's Dilemma as a Game with Incomplete Information", Journal of Economic Psychology, 11, 69-84.

Bolle, F. and Kritikos, A. (1998): "Self-Centered Inequality Aversion versus Reciprocity and Altruism" mimeo, 14/95, Europe-University Viadrina, Frankfurt/Oder.

Bolton, G. and Ockenfels, A. (2000): "ERC - A Theory of Equity, Reciprocity and Competition", American Economic Review 90, 166-193.

Brandts, J. and Sola, C. (1999): "Reference Points and Negative Reciprocity in Simple Sequential Games", forthcoming in: Games and Economic Behavior.

Camerer, C. and Thaler, R. (1995): "Ultimatums, Dictators, and Manners", Journal of Economic Perspectives 9, 209-219.

Campbell, C. M. and Kamlani, K. (1997): "The Reasons for Wage Rigidity: Evidence from a Survey of Firms", Quarterly Journal of Economics 112, 759-789.

Charness, G. (1996): "Attribution and Reciprocity in a Simulated Labor Market: An Experimental Investigation", mimeo, University of Berkeley.

Charness, Gary and Matthew Rabin (2000), "Some Simple Tests of Social Preferences and a New Model", Discussion Paper, University of California, Berkeley.

Clark, K. and Sefton, M. (1998): "The Sequential Prisoner's Dilemma: Evidence on Reciprocal Altruism", mimeo, University of Manchaster.

Davis, D. and Holt, C. (1993): Experimental Economics, Princeton University Press, Princeton.

Dufwenberg, M. and Kirchsteiger, G. (1998): "A Theory of Sequential Reciprocity", mimeo, CentER for Economic Research, Tilburg.

Eckel, C. and Grossman, P. (1998): "Are Women Less Selfish Than Men? Evidence from Dictator Experiments", Economic Journal 108, 726-35.

Falk, A, Fehr E., and Fischbacher U. (2000a): "Testing Theories of Fairness – Intentions Matter", Working paper No. 63, Institute for Empirical Research in Economics, University of Zurich.

Falk, A, Fehr E., and Fischbacher U. (2000b): "Informal Sactions", Working paper No. 59, Institute for Empirical Research in Economics, University of Zurich.

Falk, A, and Fischbacher U. (2001): "Modeling Fairness and Reciprocity", written for: Strong Reciprocity: Modeling Cooperative Behavior, ed. by S. Bowles, R. Boyd, E. Fehr and H. Gintis.

Fehr, E. and Falk, A. (1999): "Wage Rigidities in a Competitive, Incomplete Contract Market", Journal of Political Economy 107, 106-134.

Fehr, E., Gächter, S., and Kirchsteiger G. (1997): "Reciprocity as a Contract Enforcement Device: Experimental Evidence", Econometrica 65, 833-860.

Fehr, E., Kirchsteiger, G., and Riedl, A. (1993): "Does Fairness Prevent Market Clearing? An Experimental Investigation", Quarterly Journal of Economics 108, 437-460.

Fehr, E. and Schmidt, K. (1999): "A Theory of Fairness, Competition, and Cooperation", Quarterly Journal of Economics 114, August 1999, 817-868.

Fischbacher, U., Gächter, S. and Fehr, E. (1999): "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment", forthcoming in: Economics Letters.

Forsythe, R., Horowitz, J., Savin, N. and Sefton, M. (1994): "Fairness in Simple Bargaining Experiments", Games and Economic Behavior 6, 347-369.

Francis, H. (1985): "The Law, Oral Tradition and the Mining Community", Journal of Law and Society 12, 2267-2271.

Fudenberg, D. and Tirole, J. (1991): Game Theory. The MIT Press, Cambridge, Massachussets.

Gächter, S. and Falk, A. (1997): "Reputation or Reciprocity?", mimeo, University of Zurich.

Geanakoplos, J., Pearce, D. and Stacchetti, E. (1989): "Psychological Games and Sequential Rationality", Games and Economic Behavior 1, 60 - 79.

Gouldner, A. (1960): "The Norm of Reciprocity", American Sociological Review 25, 161 - 178.

Goranson, R. E. and Berkowitz, L. (1966): "Reciprocity and Responsibility Reactions to Prior Help", Journal of Personality and Social Psychology 3, 227-232.

Greenberg, M. S. and Frisch, D. M. (1972): "Effect of Intentionality on Willingness to Reciprocate a Favord", Journal of Experimental Social Psychology 8, 99-111.

Güth, W., Schmittberger, R. and Schwarze, B. (1982): "An Experimental Analysis of Ultimatium Bargaining", Journal of Economic Behavior and Organization 3, 367-88.

Güth, W. (1995): "On Ultimatum Bargaining Experiments - A Personal Review", Journal of Economic Behavior and Organization 27, 329 - 344.

Harrison, G. W. and Hirshleifer J. (1989): "An Experimental Evaluation of Weakest Link/Best Shot Models of Public Goods", Journal of Political Economy 97, 201-225.

Hayashi, N., Ostrom, E., Walker, J., and Yamagishi, T. (1998): "Reciprocity, Trust, and the Sense of Control: A Cross-Societal Study", Discussion paper, Indiana University, Bloomington.

Hoffman, E., McCabe, K. and Smith, V. L. (1996): "Social Distance and Other-Regarding Behavior in Dictator Games", American Economic Review 86, 653-660.

Hoffman, E., McCabe, K., Shachat, K., and Smith, V. (1994): "Preferences, Property Rights, and Anonymity in Bargaining Games", Games and Economic Behavior 7, 346-380.

Kahneman, D., Knetsch, J. and Thaler, R. (1986): "Fairness as a Constraint on Profit-Seeking: Entitlements in the Market", American Economic Review 76, 4, 728-741.

Kreps, D., Milgrom, P., Roberts, J. and Wilson, R. (1982): "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma", Journal of Economic Theory 27, 245-252.

Ledyard, J. (1995): "Public Goods: A Survey of Experimental Research", in: Handbook of Experimental Economics, ed. by J. Kagel and A. Roth, Princeton: Princeton University Press.

Loewenstein, G. F., Thompson, L., and Bazerman, M. H. (1989): "Social Utility and Decision Making in Interpersonal Contexts", Journal of Personality and Social Psychology

57, 426-441.

McCabe, K., Rigdon, M. and Smith, V. (2000): "Positive Reciprocity and Intentions in Trust Games", mimeo, University of Arizona at Tucson.

Mowday, R. T. (1991): "Equity Theory Predictions of Behavior in Organizations", in: R. M. Steers and L. W. Porter (eds.), Motivation and Work Behavior, New York: McGraw-Hill, 111-130.

Offerman, T. (1999): "Hurting Hurts More that Helping Helps: The Role of the Self-Serving Bias", Working paper, University of Amsterdam.

Prasnikar, V. and Roth, A. E. (1992), "Considerations of Fairness and Strategy: Experimental Data from Sequential Games", Quarterly Journal of Economics, 865-888.

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

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

Roth, A., Prasnikar, V., Okuno-Fujiwara, M. and Zamir, S. (1991): "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study", American Economic Review 81, 1068-1095.

Ruffle, B. (1995): "Gift Giving with Emotions", mimeo, Dept. of Economics, Princeton University.

Smith, A. [1759]: The Theory of Moral Sentiments. Edited by D.D. Raphael and A.C. Macfie, Oxford, Clarendon Press (1976).

Smith, K. W. (1992): "Reciprocity and Fairness: Positive Incentives for Tax Compliance", in: Why People Pay Taxes - Tax Compliance and Enforcement, ed. by Joel Slemrod, The University of Michigan Press, Ann Arbor, 223-250.

Smith, V. L. (1982); "Microeconomic Systems as an Experimental Science", American Economic Review 72, 923-955.

Steers, R. M. and Porter, L. W. (1991): Motivation and Work Behavior, Fifth Edition, New York: McGraw-Hill.

Sugden R. (1984): "Reciprocity: The Supply of Public Goods Through Voluntary Contributions", The Economic Journal 94, 772-787

Thaler, R. H. (1988): "Anomalies: The Ultimatum Game", The Journal of Economic Perspectives 2, 195-206.

Trivers, R. (1971): "The Evolution of Reciprocal Altruism", Quaterly Review of Biology 46, 35-57.

Walster, E. and Walster, G. W. (1978): Equity - Theory and Research, Boston, Allyn and Bacon.

Watabe, M., Terai, S., Hayashi, N., and Yamagishi, T. (1996): "Cooperation in the One-Shot Prisoner's Dilemma Based on Expectations of Reciprocity", Japanese Journal of Experimental Social Psychology, XXXVI, 183-196.

Appendix 1: Existence of Reciprocity Equilibria

In the presented form, the existence of a reciprocity equilibrium is not always guaranteed. A game where a reciprocity equilibrium may not exist is shown in Proposition 10 in the Appendix 3.

The reason why the existence of an equilibrium is not guaranteed has to do with the discontinuity of function Ω. To show this, we define for a (small) positive number

λ a continuous approximation Ω^λ for Ω.[19] We set

$$\Omega^\lambda(\tilde{\pi}_i, \tilde{\pi}_j, \pi_i^0, \pi_j^0) := \begin{cases} \min(1, \varepsilon_i + \frac{1}{\lambda}(\pi_i^0 - \tilde{\pi}_i)) & \text{if } \pi_i^0 \geq \pi_j^0 \text{ and } \tilde{\pi}_i < \pi_i^0 \\ \varepsilon_i & \text{if } \pi_i^0 \geq \pi_j^0 \text{ and } \tilde{\pi}_i \geq \pi_i^0 \\ \min(1, \varepsilon_i + \frac{1}{\lambda}(\tilde{\pi}_i - \pi_i^0)) & \text{if } \pi_i^0 < \pi_j^0,\ \tilde{\pi}_i > \pi_i^0 \text{ and } \tilde{\pi}_i \leq \tilde{\pi}_j \\ \begin{array}{l} \max(\varepsilon_i, \min(1 - \frac{\tilde{\pi}_i - \tilde{\pi}_j}{\pi_i^0 - \pi_j^0}, \\ \varepsilon_i + \frac{1}{\lambda}(\tilde{\pi}_i - \pi_i^0)) \end{array} & \text{if } \pi_i^0 < \pi_j^0,\ \tilde{\pi}_i > \pi_i^0 \text{ and } \tilde{\pi}_i > \tilde{\pi}_j \\ \varepsilon_i & \text{if } \pi_i^0 < \pi_j^0 \text{ and } \tilde{\pi}_i \leq \pi_i^0 \end{cases}$$

Given the continuous variant Ω^λ of Ω, we define a modified kindness function φ^λ and a utility function U^λ. We call a λ-reciprocity equilibrium a subgame perfect equilibrium of the psychological game with utility U^λ. This modification now guarantees the existence of an equilibrium as the following theorem shows:

Theorem 2 (Existence Theorem) *Let Γ be a finite two person extensive form game with complete information. Let $\lambda > 0$. Then Γ has a λ-reciprocity equilibrium.*

Proof of the Existence Theorem: Let $n \in N_i$ be a node of the game. Then $S^n = \{(p_a)_{a \in A_n} \mid p_a \in [0,1], \sum_a p_a = 1\}$ is the set of mixed strategies in this node. Let $S = \prod_{n \in N} S^n$ be the set of behavior strategy combinations. It includes the strategies of both players. Let $S^{-n} = \prod_{m \neq n} S^m$ be the strategies at all other nodes. Let $s = (s^n, s^{-n})$ be a behavior strategy combination with $s^n \in S^n$ and $s^{-n} \in S^{-n}$. Let s' and s'' the beliefs of first and second order. We define $V(n, (s^n, s^{-n}), s', s'')$ as the utility U_i^λ conditional on node n, i.e. it is the expected utility of the player who is at move in node n given this plaer knows he is in node n. We now define the best reply correspondence

$$B : S \rightrightarrows S : s \mapsto B(s) \subset S$$

The component $B^n : S \rightrightarrows S^n$ is defined as

$$B^n(s) := \arg\max_{\tau^n \in S^n} V(n, (\tau^n, s^{-n}), s, s)$$

We get

$$B(s) := \{(b^n)_{n \in N} \mid b^n \in B^n(s)\}$$

This definition is the best reply correspondence of the agent-strategic form (see Fudenberg and Tirole (1991)): The player behaves as if there was an agent in every decision node. A behavior strategy that optimizes in the agent-strategic form corresponds to a subgame perfect Nash equilibrium.

As S is the product of convex and compact sets, it is convex and compact. Since S is compact and V is continuous in s^n, $B(s)$ is not empty. Because U^λ is linear in the strategies, $B(s)$ is convex.

We now show that B is upper-hemi continuous: First, we show that V depends continuously on the strategies and beliefs: $\Delta_j(n)$ is continuous and $\Delta_j(n) = 0$ holds for $\pi_i(n, s_i'', s_i') = \pi_j(n, s_i'', s_i')$. Because Ω^λ is bounded (by 1), we get the desired continuity of $\vartheta_j^\lambda(n)\Delta_j(n)$ at $\pi_i(n, s_i'', s_i') = \pi_j(n, s_i'', s_i')$. By construction of Ω^λ, $\vartheta_j^\lambda(n)$ is continuous in strategies and beliefs if $\pi_i(n, s_i'', s_i') \neq \pi_j(n, s_i'', s_i')$. Therefore, this

[19] We have $\Omega(\tilde{\pi}_i, \tilde{\pi}_j, \pi_i^0, \pi_j^0) = \lim_{\lambda \to 0} \Omega^\lambda(\tilde{\pi}_i, \tilde{\pi}_j, \pi_i^0, \pi_j^0)$ for any choice of $\tilde{\pi}_i, \tilde{\pi}_j, \pi_i^0, \pi_j^0$.

also holds for $\vartheta_j^\lambda(n)\Delta_j(n)$. Hence, $\varphi_j^\lambda(n)$ is continuous. The reciprocation term and the material profit are obviously continuous and therefore V is continuous. If any function $f : X \times Y \to \mathbb{R}$ is continuous, then the best reply correspondence $R : X \rightrightarrows Y : x \mapsto \arg\max_y f(x,y)$ is upper hemi continuous. Hence, because V is continuous, the best reply correspondence B is upper hemi-continuous.

Therefore, we can apply the fixed point theorem of Kakutani and get an $s^* \in S$ with $s^* \in B(s^*)$. This strategy s^* with first order belief s^* and second order belief s^* now forms a reciprocity equilibrium: The triple (s^*, s^*, s^*) trivially satisfies the consistency of the beliefs. By construction of B each player optimizes his utility in each node - given the beliefs and given the strategies of the other players. This is exactly the definition of the subgame perfect psychological equilibrium. ■

Appendix 2: Extension To Games With More Than 2 Players

The idea behind the generalization to games with more than 2 players persons consists of independently considering the reciprocity relation of player i towards each of the other players $j \neq i$.

Let $s_i \in S_i$ be player i's behavior strategy and let $s_i^{(j)} \in S_j$ be player i's first order belief about player j's strategy. Further let $s_i^{(jk)} \in S_k$ be player i's (second order) belief about what he thinks is player j's belief about k's strategy. In the 2 player case, we get $s_i' = s_i^{(j)}$ and $s_i'' = s_i^{(ji)}$. Furthermore, we use the notation $(-ij)$ to express the set of players other than players i and j. So, $s_i^{(j(-ij))}$ is player i's belief about what j believes what the other players will do. In analogy to the definitions (2) to (7) we define in a node n - for a given set of first and second order beliefs - the following expressions:

$$\Delta_{j\to i}(n) := \pi_i(n, s_i^{(ji)}, s_i^{(j)}, s_i^{(j(-ij))}) - \pi_j(n, s_i^{(ji)}, s_i^{(j)}, s_i^{(j(-ij))}) \tag{10}$$

(In the two player case $\Delta_{j\to i}(n) = \Delta_j(n)$ holds.)

$$\Pi_i^j(n) := \left\{ \left(\pi_i(s_i^{(ji)}|n, s_j^p, s_i^{(j(-ij))}|n), \pi_j(s_i^{(ji)}|n, s_j^p, s_i^{(j(-ij))}|n) \right) \mid s_j^p \in S_j^p \right\} \tag{11}$$

$$\vartheta_{j\to i}(n) := \max \left\{ \Omega(\tilde{\pi}_i, \tilde{\pi}_j, \pi_i(n, s_i^{(ji)}, s_i^{(j)}, s_i^{(j(-ij))}), \pi_j(n, s_i^{(ji)}, s_i^{(j)}, s_i^{(j(-ij))})) \mid (\tilde{\pi}_i, \tilde{\pi}_j) \in \Pi_i^j(n) \right\} \tag{12}$$

$$\varphi_{j\to i}(n) = \vartheta_{j\to i}(n)\Delta_{j\to i}(n) \tag{13}$$

$$\sigma_{i\to j}(n, f) := \pi_j(\nu(n, f), s_i^{(ji)}, s_i^{(j)}, s_i^{(j(-ij))}) - \pi_j(n, s_i^{(ji)}, s_i^{(j)}, s_i^{(j(-ij))}) \tag{14}$$

(In the two player case $\sigma_{i\to j}(n, f) = \sigma_i(n, f)$ holds.)

$$U_i(f) = \pi_i(f) + \rho_i \sum_{j\neq i} \sum_{\substack{n\to e \\ n\in N_i}} \varphi_{j\to i}(n)\sigma_{i\to j}(n, f) \tag{15}$$

A reciprocity equilibrium is again a set of strategies and first and second order believes such that the strategies maximize the utility in Equation (15) and such that strategies and beliefs are consistent.