universität wien

Diplomarbeit

Titel der Diplomarbeit

„Social Preferences in Bargaining Games“

Verfasserin

NADIA STEINER

angestrebter akademischer Grad

Magistra der Sozial- und Wirtschaftswissenschaften
(Mag. rer. soc. oec)

Wien, im Oktober 2008

Studienkennzahl lt. Studienblatt:	A140
Studienrichtung lt. Studienblatt:	Volkswirtschaftslehre
Betreuer:	Dr. Simon Weidenholzer

Contents

1 Introduction

"Economics can be distinguished from other social sciences by the belief that most (all?) behavior can be explained by assuming that agents have stable, well-defined preferences and make rational choices consistent with those preferences in markets that (eventually) clear. An empirical result qualifies as an anomaly if it is difficult to "rationalize" or if implausible assumptions are necessary to explain it within the paradigm." (Camerer 1995) The following pages deal with anomalies primarily observed in bargaining games. They reveal that fairness considerations - although neglected by standard game theory - affect interactions between economic agents.

There are two implications of the standard model of self regarding preferences that are in conflict with both laboratory and field experiments and the common intuition that people do care about other people. The first is the implication that agents only care about what they personally gain or lose and not other agents' gains, losses or intentions. The second implication is that agents only mind the final outcomes of economic interactions and not about the processes through which these outcomes are attained.

A person exhibits social preferences if the person does not only care about the economic resources allocated to her but also cares about the economic resources allocated to relevant reference agents (see Gintis (2005)). Research indicates that many people exhibit social preferences. Nevertheless there might as well be a substantial number of people who behave in a purely selfish

manner as assumed in the theory of self-regarding preferences. Actually simple maximization of one's own material payoff predicts behavior quite well in many contexts, for example competitive markets, one-sided auctions with independent private values, procurement contracting and search. Problems with standard game theory occur when it comes to ultimatum games, dictator games, public good games with voluntary contributions and experimental labor markets (see Cox (2004)). Fairness itself seems to be a concept that strongly depends on context. "What may be considered unfair when two people meet face to face or in a bilateral manner may be considered fair in a market context where economic survival is at stake." (Schotter 1995)

1.1 Bargaining Games

One of the most discussed games in the context of social preferences is the ultimatum game. The ultimatum game is one of the simplest strictly competitive games. It can be described as follows: One player, the proposer or allocator, is asked to divide a sum of money between himself and a second player, the recipient. If the recipient accepts the allocator's proposal, both receive the corresponding amounts of money, but if the recipient rejects the proposed division, both players receive nothing. In a subgame perfect equilibrium under standard assumptions allocators keep all for themselves and propose zero or ε to the recipient, where ε is an infinitesimally small positive number. The recipient should agree to any positive amount ε since it is better than nothing. However, experimental evidence shows unambigu-

ously that this does not correspond to the way people behave in reality. Fehr and Schmidt (1999) outline the following behavioral regularities observed in many studies:

- There are virtually no offers above one half of the stake.
- A vast majority of offers lies between 40% and 50% of the total sum of money.
- There are hardly any offers below 20% of the stake.
- Low offers are frequently rejected. The probability of rejection decreases the more generous the proposal is. Proposals of one half are very seldomly rejected.

A couple of manipulations of the standard setting have been made in order to decrease the number of equal split offers and the mean proportion allocators offered. For example in a study recipients were asked to compete in some kind of skill testing contest and the outcome determined the overall budget to be divided by the allocator. Other possibilities are the use of market terminology when describing the game or the winning of property rights to be the proposer. Although most of these manipulations of the standard setting were successful in the sense of changing players' behavior, they cannot be taken as a supportive of the standard model, because these extrinsic manipulations are not accounted for by game theory.

A similar but even simpler game is the dictator game studied by Forsythe et al. (1994), Hoffman et al. (1996) and others. In the dictator game a first mover, the so-called dictator, divides an amount of money between himself and player 2. Player 2 can do nothing but accept. The total payoff can be normalized to 1. If the dictator gives an amount x to the receiver, his own payoff is *1-x* and the receiver's payoff is x. It is obvious that standard game theory predicts that the dictator will keep all for himself. Empirically, the following behavioral pattern has been observed:

- There are practically no offers larger than half of the stake.
- Compared to the ultimatum game, offers are low. About 80% of the offers are between zero and one half. The average positive offer is 24% of the total pie.
- Roughly 20% of the offers are exactly zero.

In the following paper I will discuss and compare models of inequity aversion and reciprocity that can (at least partly) account for the observed data.

2 Inequity Aversion

Inequity aversion is a type of social preferences. Two important models of inequity aversion are those of Fehr and Schmidt (1999) and Bolton and Ockenfels (2000). The basic idea is that inequity averse people wish to achieve equitable distributions of economic resources. They want to increase other

persons' payoffs when those are below an equitable benchmark and they try to decrease other persons' payoffs as soon as they are above the equitable level. Inequity aversion is thus a form of conditional altruism.

"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 complicated social comparison processes. In social psychology and sociology the relevance of social comparison 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." (Fehr and Schmidt 1999) The relevance of relative payoffs is also supported by lots of research on labor economics, e.g. Fehr, Kirchsteiger and Riedl (1998) or Clark and Oswald (1995) amongst others.

In both the Fehr and Schmidt (1999) and the Bolton and Ockenfels (2000) models agents have a per se aversion to disparities in relative payoffs. Beliefs about the intentions of the other agents are not relevant in these models. Bolton and Ockenfels assume that people have a symmetric dislike for inequality whereas Fehr and Schmidt assume that agents - though still disliking all inequality - care more about it if it is at their own relative payoff disadvantage.

2.1 The Fehr and Schmidt model of inequity aversion

Formally, the utility function of player $i \epsilon [1, ..., n]$ is assumed to be given by

$$U_i(x) = x_i - \alpha_i \frac{1}{n-1} \sum max\{x_j - x_i, 0\} - \beta_i \frac{1}{n-1} \sum max\{x_i - x_j, 0\}$$

where $x = x_1, ...x_n$ denotes the vector of monetary payoffs. It is assumed that $\beta_i \leq \alpha_i$ and $0 \leq \beta_i < 1$. The second term of the expression gives the utility loss from disadvantageous inequality and the third term measures the loss from advantageous inequality. $\beta_i \leq \alpha_i$ captures the idea that negative deviations from the reference outcome hurt more than positive deviations. $\beta_i \geq 0$ means an abstraction of subjects who like to be better off than others. β_i has to be smaller than 1. Otherwise player i would be willing to throw away one dollar or even more in order to reduce his one dollar advantage relative to player j, which seems implausible. If there are more than two players, each of them compares their income with all other $n-1$ players. Therefore the second and third term have to be normalized by dividing by $n-1$ in order to make sure that the relative impact of inequality aversion on player i is independent of the number of players. Another implicit assumption is that player i compares herself with each of the other players, but does not care about inequalities within the group of other players. (Fehr and Schmidt 1999)

The model is applicable to the ultimatum game and can account quite well for the deviations from the predictions of standard game theory. The parameters (α_1,β_1) represent the allocator's preferences, and (α_2,β_2) repre-

sent the responder's preferences. Without loss of generality the bargaining surplus can be normalized to one where s denotes the responder's share and $1-s$ the proposer's share. The equilibrium is then characterized as follows:

From the point of the responder it is a dominant strategy to accept any offer s above 0.5 and to reject s if $s < s'(\alpha_2) \equiv \frac{\alpha_2}{1+2\alpha_2} < 0.5$. Any offer $s > s'(\alpha_2)$ should be accepted. If the allocator knows the responder's preferences she will offer $s^\star = 0.5$ if $\beta_1 > 0.5$, $s^\star \epsilon [s'(\alpha_2), 0.5]$ if $\beta_1 = 0.5$ and $s^\star = s'(\alpha_2)$ if $\beta_1 < 0.5$. If the proposer does not know exactly the responder's preferences she believes that α_2 is distributed according to a cumulative distribution function $F(\alpha_2)$. $F(\alpha_2)$ has support $[\alpha_l, \alpha_u]$ with $0 \leq \alpha_l < \alpha_u < \infty$. From the perspective of the allocator, the probability p that an offer below 0.5 will be accepted, is given by $p = 1$ if $s \geq s'(\alpha_u)$, $p = 1$ if $s \leq s'(\alpha_l)$ and $p = F(\frac{s}{1-2s}) \epsilon (0,1)$ if $s'(\alpha_l) < s < s'(\alpha_u)$. Thus the optimal offer of the proposer not knowing the responder's exact preferences is $s^\star = 0.5$ if $\beta_1 > 0.5$, $s^\star \epsilon [s'(\alpha_u), 0.5]$ if $\beta_1 = 0.5$ and $s^\star \epsilon [s'(\alpha_l), s'(\alpha_u)]$ if $\beta_1 < 0.5$.

Proof 2.1.1 *Since* s *is above 0.5, the responder's utility from accepting is* $U_2(s) = s - \beta_2(2s-1)$. *If* $\beta_2 < 1$, *this utility is always positive and preferred to a rejection which yields a payoff of zero. Equality could only be achieved by destroying the entire surplus which would be very costly for a responder who is offered more than half of the share. For any offer smaller than a half, the responder accepts, if this yields a nonnegative utility. That means* $U_2(s) = s - \alpha_2(1-2s) > 0$. *Thus* s *must exceed the acceptance threshold* $s'(\alpha_2) \equiv \frac{\alpha_2}{1+2\alpha_2} < 0.5$. *From the perspective of the proposer it does not make*

sense to offer more than half of the share. Doing so would reduce her payoff as compared with an offer of a half which would also be accepted with certainty and which would imply perfect equality. If $\beta_1 > 0.5$*, her utility is strictly increasing in* s *for all* $s \leq 0.5$*. In this case the allocator rather likes to share than to maximize her own monetary payoff. She then offers* $s = 0.5$*. In case* $\beta_1 = 0.5$*, the proposer is indifferent between keeping one dollar and giving it to the responder. She is thus indifferent between all offers the responder will accept up to a half of the share, i.e.* $s\epsilon[s'(\alpha_2), 0.5]$*. If* $\beta_1 < 0.5$*, the allocator will increase her monetary payoff even if that makes the responder worse off. However, she wants to avoid proposing less than the responder's acceptance threshold. If she has knowledge of the exact value of the acceptance threshold she will propose* $s'(\alpha_2)$*. Otherwise she has a believe about the probability of acceptance,* $F(\frac{s}{1-2s})$ *which is equal to one if* $s \geq \alpha_u(1+2\alpha_u)$ *and equal to zero if* $s \leq \frac{\alpha_l}{1+\alpha_l}$*. So in this case there is an optimal offer* $s\epsilon[s'(\alpha_l), s'(\alpha_u)]$*. (Fehr and Schmidt 1999)*

It is easy to see that there are no offers above 0.5 and that offers equal to 0.5 are always accepted, whereas very low offers are quite likely to be rejected. The probability of acceptance, $F(\frac{s}{1-2s})$*, is increasing in* s *but even relatively small values of* α_2 *imply relatively large thresholds. The acceptance threshold is nonlinearly increasing and strictly concave in* α_2*. As* α_2 *goes to infinity, it converges to 0.5. All those properties of the model go well together with intuition.*

The Fehr-Schmidt model of inequity aversion also accounts well for the

behavior observed in market and cooperation games. Nevertheless it bears some weaknesses that I will dicuss in chapters to follow.

2.2 The Bolton and Ockenfels model of inequity aversion

In many cases the Bolton and Ockenfels model (2000) makes equal or similar predictions to the Fehr and Schmidt model (1999). The fundamental difference is that Fehr-Schmidt assume that subjects dislike payoff differences to any other player, whereas Bolton and Ockenfels assume that subjects want the average payoff to be as close as possible to their own payoff.

Consider a game with n players, $i = 1, ...n$. Each player is supposed to maximize the expected value of his motivation function: $\nu_i = \nu_i(y_i, \sigma_i)$ where y_i is player i's monetary payoff and σ_i is i's relative share of the payoff, $\sigma_i = \sigma_i(y_i, c) = \frac{y_i}{c}$ if $c > 0$ and $\sigma_i = \frac{1}{n}$ if $c = 0$. c is the total payout of the game, $c = \sum_{i=1}^{n} y_i$. This motivation function captures the objectives that influence people's behavior during experimental situations. "The weights individuals give these objectives may well change over the long term, with changes in age, education, political or religious beliefs, and other characteristics. However, it is sufficient for our purposes that the trade-off be stable in the short run, for the duration of the experiment." (Bolton and Ockenfels 2000)

Several assumptions are made about the motivation function:

- The function ν_i is continuous and twice differentiable on the domain of

(y_i,σ_i). The reason for this assumption is mathematical convenience.

- Narrow self-interest: For a fixed σ and given two choices y_i^1 and y_i^2 with $\nu_i(y_i^1,\sigma) = \nu_i(y_i^2,\sigma)$ and $y_i^1 > y_i^2$, player i chooses y_i^1. That means that for a given relative payoff, players choose consistently with the standard assumption about preferences for money, i.e. "more is better".

- Comparative effect: Holding y_i fixed, the motivation function is strictly concave in σ, with a maximum at the allocation at which one's own share is equal to the average share.

The equal division is called the social reference point. Players experience a trade-off between adhering to the reference point and achieving personal gain. Different individuals react differently to this tension. Each player has two thresholds, $r_i(c)$ and $s_i(c)$ that capture at which point individual behavior diverges from "more money is preferred to less". The threshold $r_i(c)$ is defined as follows:

$$r_i(c) = argmax_{\sigma_i}\nu_i(c\sigma_i, \sigma_i)$$

where $c > 0$. Note that $y_i \equiv c\sigma_i(y_i, c, n)$. $s_i(c)$ is implicitly defined by

$$\nu_i(cs_i, s_i) = \nu_i(0, \frac{1}{n})$$

where $c > 0$ and $s_i \leq \frac{1}{n}$. Technically, both expressions are functions of n. For the two player case r_i corresponds to the division that player i would choose in a dictator game and s_i to i's rejection threshold in the ultimatum

game. The above assumptions imply that there is a unique $s_i \epsilon (0, \frac{1}{n}]$ and a $r_i \epsilon [\frac{1}{n}, 1]$ for every c.

- Heterogeneity: The full range of thresholds is represented in the player population.

An example motivation function for player i in a two-player game is:

$$\nu_i(y_i, \sigma_i) = a_i y_i - \frac{b_i}{2}\left(\sigma_i - \frac{1}{2}\right)^2$$

with $a_i \geq 0$ and $b_i > 0$. The first term catches the preferences for the monetary payoff itself. The second term shows the influence of the comparative effect. The further the allocation moves away from giving player i an equal share, the higher the loss in his or her utility. The ratio of weights that is attributed to the absolute and relative components of the motivation function, $\frac{a}{b}$, can be used to characterize a player's type. Strict relativism is represented by $\frac{a}{b} = 0$ and implies that $r = s = \frac{1}{2}$. Strict narrow self-interest is represented by $\frac{a}{b} \to \infty$ and implies that $r = 1$ and $s \to 0$.

Stable patterns of behavior can be characterized by equilibrium predictions. An equilibrium in this model is a perfect Bayesian equilibrium solved with respect to motivation functions. Players' r and s thresholds are private information but the densities f^r and f^s with which they are distributed are common knowledge.

Reconsider the ultimatum game. (P) denotes the proposer, (R) the responder and (c, σ_p) denotes the proposal to the responder. The proposer's

payoff is then $c\sigma_p$ and so the responder's payoff is given by $c - c\sigma_p$. If a responder is indifferent to both accepting and rejecting, $1 - \sigma_p = s_R(c)$, she is assumed to always accept the offer. The equilibrium then has the following properties:

- Responder behavior: For any $c > 0$, the probability that a randomly selected responder will reject the proposal, $p(c, \sigma_p)$, satisfies:

1. $p(c, \frac{1}{2}) = 0$ and $p(c, 1) = 1$. By the assumption of narrow self-interest, it is clear that $\nu_R(\frac{c}{2}, \frac{1}{2}) \geq \nu_R(0, \frac{1}{2})$. Thus an equal division is never rejected. By definition an offer is rejected, if $1 - \sigma_p < s_R(c)$ for $s_R(c) \epsilon (0, \frac{1}{2}]$. Therefore $\sigma_p = 1$offers will always be rejected.

2. p is strictly increasing in σ_p over the interval $(\frac{1}{2}, 1)$. This follows from integrating over the density f^s.

3. fixing a $\sigma_p \epsilon (\frac{1}{2}, 1)$, p is nonincreasing in c.

- Proposer behavior: For all ultimatum proposals it holds that $\frac{1}{2} \leq \sigma_p < 1$: For any c proposers prefer $\sigma_p = \frac{1}{2}$ to any $\sigma_p < \frac{1}{2}$ and they know that $\sigma_p = \frac{1}{2}$ is never rejected. It is therefore obvious that in equilibrium $\sigma_p \geq \frac{1}{2}$. $\sigma_p = 1$ is always rejected by responders so that in equilibrium $\sigma_p < 1$.

The above predictions are in line with experimental observations in ultimatum game situations. As the Fehr and Schmidt model the Bolton and Ockenfels model is also insightful when applying it to market games, dilemma

games and the gift-exchange game. However, it has its limitations. "ERC is a theory of "local behavior" in the sense that it explains *stationary patterns* for relatively *simple games*, played over a *short time span* in a *constant frame*." (Bolton and Ockenfels 2000)

2.3 Reference points and maximization of expected utility

A common concept in psychology and sociology is that of a reference level (of e.g. income), against which an individual compares herself or himself. A study of Clark and Oswald (1995) points out that workers' reported levels of well-being are only weakly correlated with absolute income. Their self-reported levels of satisfaction are inversely related to their comparison wage rates. It seems likely that judgements about fairness also rest on some kind of comparative process. Taking a reference point or reference transaction as a basis for fairness judgements is not necessarily just in itself but it seems natural from a psychological point of view.

People are often more sensitive to how their current situation differs from some reference level than to the absolute characteristics of the situation. Loss aversion seems to play an important role in people's notion of fairness. For example, most employees feel that a firm is more obliged not to hurt them relative to a reference level than it is obliged to improve terms of trade if doing so is possible. (Rabin 1998) Moreover, perceptions about what is considered fair, seem to adjust over time. Something that was perceived as unfair in the first place, might become a reference transaction once people

get used to it. "Psychological studies of adaptation suggest that any stable state of affairs tends to become accepted eventually, at least in the sense that alternatives to it no longer readily come to mind." (Kahnemann 1986)

When looking at bargaining games we find that people seem to implicitly consider equitable sharing over *changes* in total payoff (and not total payoff itself). A responder being offered 10% of the total pie in an ultimatum game does not consider the situation before the game started as his reference level. If this were the case he should gladly accept since even a small amount of money is an improvement relative to getting nothing. However, the responder's reference outcome is not zero but the other player's share. From this point of view, getting only 10% of the total payoff is disadvantageous. In two-player games it is trivial to determine the relevant reference group, i.e. the other player. In an ultimatum game it is more or less obvious that the reference point is equity. But in more complex social interactions it is often not clear who is part of the reference group and who is not and what the reference outcome looks like. "The determination of the relevant reference group and the relevant reference outcome for a given class of games 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." (Fehr and Schmidt 1999)

Another problem, apart from the determination of reference groups and outcomes in complex environments, is the motivation of the proposer in e.g an ultimatum game. Some researchers have suggested that allocators in ultima-

tum games might not truly be motivated to be fair but rather to appear fair. When anticipating that responders might reject low offers, proposers maximize their expected payoff by offering a more generous share. Experiments in which the proposer had to distribute a number of chips of different value to himself and the responder, show that the proposer's main motivation is to appear fair. A study that points into this direction is that of Kagel, Kim and Moser (1996). They analyzed behavior in ultimatum games with asymmetric information where chips with different valuation for the proposer and the responder had to be distributed. In the case where the proposer knows that a chip is worth 30 cents to himself and only 10 cents to the responder, an equal division would imply giving 75% of the chips to the responder. However, if the proposer knows that the responder is not aware of the different valuations of the chips, he might offer only 50% of the chips. This is enough to appear fair in front of the unaware responder and at the same time gives himself a higher payoff. Empirical data reveals that offers made in this treatment are actually close to 50% and rejections are rare. This suggests that proposers might act like "sophisticated profit maximizers" who only try to appear fair in order to prevent rejections. A study by Suleiman (1996), who conducted so-called δ-ultimatum games, came to similar findings.

Nevertheless, evidence from other games (e.g. the dictator game) suggests that it is not justified to conclude that positive allocations are only due to the fear of rejection. It is likely that various factors affect the importance of strategic and normative motivation and regardless of the proposer, it is un-

ambiguously true that the responder exhibits some kind of social preferences. See Dijk (2000) for further details.

2.4 Comparing models of inequality aversion

The difference between the two inequity aversion models by Fehr and Schmidt (1999), and Bolton and Ockenfels (2000) lies in the different inequality measures represented in the utility functions. Bolton and Ockenfels assume that it is the difference between the average payoff and individual payoffs that people care about. On the other hand, Fehr and Schmidt propose that people dislike payoff differences between them and any other player. According to Bolton and Ockenfels, in a population where some people are very rich and some are very poor, a player i receiving exactly the average material payoff should be as happy as if he was in a population where everyone gets the same payoff. The Fehr-Schmidt model predicts that player i prefers the second case where everyone gets the same.

Engelmann and Strobel (2002) try to compare the relative performances of the two inequality aversion models. Moreover, they aim to compare the relative importance of inequality aversion with concerns for efficiency and maximin preferences. In the case of the following distribution experiments efficiency means the sum of payoffs. Maximin preferences express the wish to maximize the minimal payoff in the group. (For the relevance of maximin preferences versus efficiency concerns see also Rabin (1998)).

Since only actions can be observed, it is often not easy or even not possible

for an observer to infer what the underlying motivation has been.

Consider the following example:

allocation	A	B
Player 1	5	4
Player 2	3	4
Player 3	1	4
Total	9	12

If it is observed that allocation B is chosen, a dislike of inequality could have been the motivation. But it could also be that B was chosen, because it is more efficient, which means that the total payoff of all subjects is maximized. A third possibility consistent with the choice of B is that the player has maximin preferences. In the experiments run by Engelmann and Strobel they tried to disentangle these different motivations to allow for a comparison of their relative importance. Simple distribution games were used in order to exclude any concern of reciprocity (which will later be discussed in detail). There were thirteen experimental treatments conducted in three sessions with in total 586 participants recruited from an introductory economics class. Each participant was asked to choose from three different allocations of money between three persons. They were told that later on they would randomly form groups of three where the three roles would also be assigned randomly. Only the choice of the participant selected as person two mattered for the distribution of real payoffs. Average payoffs of persons 1 and 3 as well

as the total payoff for each allocation were noted in the decision sheets. Note that person two could not influence his or her own payoff.

In the so-called taxation game there is a poor-, a middle- and an upper-class. The FS-model (Fehr-Schmidt) predicts the middle-class would like to tax the upper-class in order to help the poor. According to ERC (Bolton-Ockenfels) someone in the middle-class should be happy with the situation as it is. In the corresponding experiment the decision maker gets an intermediate payoff and the possibility to redistribute payoff from person 1 (who gets the highest payoff) to person 3 (who receives the lowest payoff). The crucial point in the different treatments E and F (and control treatments Ex and Fx where fixed roles were assigned in advance) is that the allocation that minimizes the difference between the payoffs of the decision makers and each of the other players at the same time maximizes the difference between the payoff of the decision maker and the average payoff and vice versa. Therefore FS and ERC predict choices of allocations that are at the opposite ends of the choice set. The decision maker's choice had different effects on the total payoffs depending on the treatment. In treatments F and Fx the allocation that is predicted by the Fehr-Schmidt model is also efficient. In treatments E and Ex the predicted allocation by Bolton and Ockenfels maximizes total payoff. The intermediate allocation should never be chosen as it is not in line with either ERC, FS, efficiency or maximin preferences. In all taxation games the predictions of the FS-model are the same as predictions of a model including only maximin preferences. This is due to the fact that FS can only

be contrary to maximin preferences if increasing the difference to the lowest payoff is compensated by reducing payoff differences that are larger or disadvantageous. The same is not true for ERC.

Treatment	F	E	Fx	Ex
Allocation	A, B, C	A, B, C	A, B, C	A, B, C
Person 1	8.2, 8.8, 9.4	9.4, 8.4, 7.4	17, 18, 19	21, 17, 13
Person 2	5.6, 5.6, 5.6	6.4, 6.4, 6.4	10, 10, 10	12, 12, 12
Person 3	4.6, 3.6, 2.6	2.6, 3.2, 3.8	9, 5, 1	3, 4, 5
Total	18.4, 18, 17.6	18.4, 18, 17.6	36, 33, 30	36, 33, 30
Efficient	A	A	A	A
Maximin	A	C	A	C
ERC pred.	C	A	C	A
FS pred.	A	C	A	C
Choices (abs.)	57, 7, 4	27, 16, 25	26, 2, 2	12, 5, 13
Choices (%)	83.8, 10.3, 5.9	39.7, 23.5, 36.7	86.7, 6.7, 6.7	40, 16.7, 43.3

Engelmann and Strobel (2002)

For treatment F the results are unambiguous. More than 83% of the decision makers chose an allocation that is consistent with the prediction of the Fehr-Schmidt model and that maximizes total utility. Almost 6% chose the allocation predicted by ERC and more than 10% stayed with the

intermediate allocation. In treatment E the results are not so clear. Almost 40% chose the allocation in line with ERC and efficiency, 36.7% chose the allocation predicted by FS and maximin preferences and 23.5% preferred the intermediate allocation. Statistical testing shows that the hypothesis that all three allocations are chosen with equal probability cannot be rejected. In total 136 choices were made in both treatments. 61.8% of them maximized total payoffs, whereas 21.3% minimized them. As opposed to the assumption by both ERC and FS that efficiency does not matter, a clear influence of efficiency is revealed by the experimental results. When subjects were asked to explain their motivation in treatment E and F, 18 stated they had been motivated by ideas of fairness. 17 of them made their choice according to the FS-model, including 8 participants who also referred to the maximal total payoff. Only one subject stated his or her concern for relative payoffs, but contrary to predictions in the sense that he or she wanted to maximize his/her own share. In the treatments Ex and Fx all of the 15 participants who affirmed they were caring for fairness chose the allocation in line with Fehr-Schmidt. Efficiency was mentioned by 16 subjects as their motivation and 6 indicated maximin preferences. As a conclusion, one can say that in the taxation game the Fehr-Schmidt model performs much better than the Bolton-Ockenfels model and that efficiency influences decisions.

As a test for the robustness of results and also as a more severe test for the inequality aversion models the so-called envy game was performed. In this game the payoff of person 2 is again intermediate. The FS-model

predicts a choice of C, which is pareto-dominated by choice B predicted by Bolton-Ockenfels. B is in turn pareto-dominated by allocation A. In different treatments of the game the payoff of the decision maker (player 2) varied in order to test whether participants were willing to give up their own payoff to reduce inequality or to increase efficiency.

Treatment	Envy Game		
Allocation	A	B	C
Person 1	16	13	10
Person 2	8	8	8
Person 3	5	3	1
Efficient	A		
Maximin	A		
ERC pred.	B		
FS pred.	C		
Choices (abs.)	21	8	1
Choices (%)	70	26.7	3.3

Engelmann and Strobel (2002)

70% of participants chose the pareto-efficient allocation. 26.7% made their decision in line with predictions of ERC and only 3.3% of the subjects chose the allocation predicted by FS. In this case the Bolton-Ockenfels model in combination with pareto-dominance clearly outperforms the Fehr-Schmidt

model. Results from control treatments with varying payoffs of the decision maker indicate that although there are minor effects on the choices made, the relative importance of the different motives does not change significantly. In face of pareto-dominance the ERC performs - although not very well - better than the FS-model. However, the predictive power of the Fehr-Schmidt model could be increased by abstracting from the linear form tolerating that the model is not neutral to scaling. Then it could also explain choices of B. Also the restriction $\beta \leq \alpha$ could be relaxed, so that even choices of A could be consistent with FS. In general, efficiency seems to be a major factor for preferences over distributions, but cannot account for all choices observed.

The last game studied is the rich and poor game in which the decision maker receives either the highest or the lowest payoff. FS and ERC predict the same choice in this game, i.e. allocation A in treatment R and allocation C in treatment P. In treatment R the decision maker receives the highest payoff. He can choose for the other subjects' payoffs to be relatively equal (C) or to be maximal in total (A). Maximin preferences predict that C is chosen. Inequality aversion models predict that the efficient allocation A is chosen. In treatment P the decision maker gets the lowest payoff. Inequality aversion predicts the least efficient allocation C. In this treatment maximin preferences cannot play a role. Therefore it allows to contrast efficiency and inequality aversion without a possible influence of maximin preferences.

Treatment	R			P		
Allocation	A	B	C	A	B	C
Person 1	11	8	5	14	11	8
Person 2	12	12	12	4	4	4
Person 3	2	3	4	5	6	7
Total	25	23	21	23	21	19
Efficient	A			A		
Maximin	C			A, B or C		
ERC pred.	A			C		
FS pred.	A			C		
Choices (abs.)	8	6	16	18	2	10
Choices (%)	26.7	20	53.3	60	6.7	33.3

Engelmann and Strobel (2002)

The results are surprising. The experiments discussed earlier indicate that both efficiency and inequality aversion are important factors. But in treatment R where both inequality aversion models predict the efficient allocation A, only 26.7% actually chose A. 53.3% chose allocation C. On the other hand, in treatment P where inequality aversion predicts C, 60% of the subjects cared for efficiency and chose A. So more subjects prefer the efficient allocation when it does not minimize inequality than when it does. The crucial difference between treatments R and P are maximin preferences.

In treatment P the minimal payoff is constant, whereas in treatment R the minimal payoff is maximized in allocation C. Consequently, it is very likely that maximin preferences play an important role. However, in treatment P where maximin preferences do not predict anything, a third of the subjects do care about inequality aversion.

In order to better evaluate the relative importance of the different motives Engelmann and Strobel pooled the data and estimated a conditional logit model. Explanatory variables are as follows (for every allocation $j \epsilon \{A, B, C\}$ that person i can choose from):

$Eff_{ij} = \sum_{k=1}^{3} x_{jk}$ efficiency

$MM_{ij} = min\{x_{jk}, k = 1, 2, 3\}$ minimax

$Self_{ij} = x_{j2}$ selfishness

$FS\alpha_{ij} = -\frac{1}{2} \sum_{k \neq 2} max\{x_{jk} - x_{j2}, 0\}$ Fehr-Schmidt

$FS\beta_{ij} = -\frac{1}{2} \sum max\{x_{j2} - x_{jk}, 0\}$ Fehr-Schmidt

$ERC_{ij} = -100 \mid \frac{1}{3} - \frac{x_{j2}}{Eff_{ij}} \mid$ Bolton-Ockenfels

These variables were used to estimate the conditional logit model:

$$V_{ij} = \gamma_1 Eff_{ij} + \gamma_2 MM_{ij} + \gamma_3 Self_{ij} + \gamma_4 FS\alpha_{ij} + \gamma_5 FS\beta_{ij} + \gamma_6 ERC_{ij}$$

x_{jk} is the payoff of person k in allocation j. The probability that person i chooses allocation j is given by

$$P_{ij} = \frac{exp(V_{ij})}{\sum_{g\epsilon\{A,B,C\}} exp(V_{ig})}$$

Neither component of the Fehr-Schmidt model has significant influence on the chosen allocations. The ERC motive has a negative and marginally significant influence. Both efficiency and even more maximin preferences have a significant impact. A combination of efficiency concerns, maximin preferences and selfishness are sufficient to explain the data. FS and ERC together can only add marginally to the explained variance. As soon as the maximin component is excluded, FS gets highly positively significant and ERC has a significant negative impact. Without controlling for maximin preferences the Fehr-Schmidt model appears to be much more accurate than the Bolton-Ockenfels model, but if maximin preferences are considered as well, the advantage of the FS-model disappears. It seems the Fehr-Schmidt model owes its success mostly to the fact that it is in line with maximin preferences per construction. Also experiments by Charness and Rabin (2001) and Charness and Grosskopf (2001) find little evidence for inequality aversion, but a significant influence of quasi-maximin preferences. On the other hand, the evident absence of strategic interaction in the discussed experiments might possibly change the importance of different distributional motives. Which motives are most influential seems to partly depend on the structure of the

game. Inequality aversion seems to play a more important role in bargaining, trust or public good games. For a deeper discussion see Engelmann (2004).

In response to the study of Engelmann and Strobel, Fehr et al. (2006) argue that other subjects than economics students value efficiency far less than equality. Therefore the results of the experiments run by Engelmann and Strobel might be biased. Indeed, a study repeating the same experiments with two subject pools - one group of economics and business administration students, one group of students of other fields - indicates that there are major differences between subject groups. In treatment Ey more than 66% of the subjects with a background of economics chose the efficient but most inegalitarian allocation A. From the other, non-economist group only 25% opted for A, whereas almost 58% chose the inefficient but most egalitarian allocation C. The difference between the two groups is statistically highly significant. This raises the question of whether there are other subject pool characteristics such as gender, age or political preferences that might influence results. In order to test for these potential effects, information was collected from the subjects after they had taken their choices in the experiments. It turned out that there are no significant differences between choices of non-economics students and employees without college education. Political attitudes, age and membership in organizations do not have significant effects either. The gender variable is weakly significant. If all data is pooled, the overall impact of an economists' and a gender dummy can be estimated. Economists are 25% less likely to choose the egalitarian but inefficient allocation C. Women

are 9% more likely to choose C. Thus Fehr concludes that indeed a majority of people have preferences for equity in simple distribution games and only a minority of subjects are more concerned about efficiency.

Another point to mention is that even in treatment P of Engelmann and Strobel, one third of the participants chose the most inefficient and most egalitarian allocation. This suggests that a third of the subjects actually are inequality averse. Fehr and Schmidt (1999) argue that in strategic interactions the heterogeneity of social preferences is crucial. Even a minority of inequality averse subjects can have a great impact in some economic environments.

A different test of equity based models is their application to three-person ultimatum games. In three-person ultimatum games, player A proposes a distribution of money between herself and players B and C. Either B or C is chosen at random and can accept or reject the offer. If the responder accepts, the distribution is implemented as proposed. If the responder rejects, he and the proposer get nothing, but the non-responder receives a positive payoff, the so-called consolation prize. As soon as the consolation prize is sufficiently large, both models of inequality aversion predict that all positive offers should be accepted. A rejection implies not only giving up one's own payoff but also generating income inequality between oneself (the responder) and the non-responding player. In an experiment conducted by Kagel and Wolfe (2001) the consolation prize took on values of \$0,\$1,\$3 and \$12 where player A had to distribute \$15. The Fehr and Schmidt model predicts that there should

be no rejections if the consolation prize is \$12. The Bolton and Ockenfels model predicts the same for all consolation prizes greater than zero.

In the experiment there were four sessions - one for each consolation prize. In each session 30 subjects played in ten rounds, where responder and proposer pairings were rotated after every round. Rejection rates varied between 15% to 22% when there were positive consolation prizes. If the consolation prize was zero, 21% of the responders rejected. Under the \$12 consolation prize treatment a rejection rate of 20.2% was observed proving not only the Bolton-Ockenfels but also the Fehr-Schmidt prediction empirically wrong. However, this test is very stringent and does not allow for any mistakes of players. For the models to be qualitatively consistent with experimental observations it would be sufficient if they were right in the prediction that the larger the consolation prize, the more likely players accept. Statistical testing reveals that there are marginally more acceptances in the positive consolation prize treatments, other things being equal. But the probability of acceptance does not increase monotonically with the amount of the consolation prize and its increase is not robust. Proposer behavior did not significantly differ among treatments. Most proposers made equal offers to players B and C and the median offer was (\$7,\$4,\$4). Even in the \$12 consolation prize treatment proposers did not take advantage of the potential protection from a responder inequality aversion. In a control treatment a negative consolation prize was introduced. All subjects received a starting capital and when an offer was rejected, the non-responder lost money. Re-

sponder behavior only changed slightly in this situation.

All in all, both models of inequality aversion do not organize the data very well in three-person ultimatum games. There are two possible explanations for the responders' ignoring of the third player. First, he might perceive the non-responder as not being part of the relevant reference group. This would be a quite dissatisfying conclusion as the need for ad hoc specifications of the relevant reference group for each setting means a serious constraint of the models' power. The other explanation is that intentionality matters and an inequality resulting from an intentional action is perceived and treated differently from an unintentional inequality. See Kagel and Wolfe (2001) for further details.

2.5 Relevance of intentions and causal attributions

"The psychology literature strongly suggests that people often consider context and interpersonal history when determining their actions in social exchange situations." (Charness 2004)

"A fundamental assumption across all attribution theory models is that individuals have a need to infer causes and to assign responsibility for why outcomes occur. [...] When assigning responsibility, people assess the degree to which the causal agent is perceived to have knowingly or unknowingly contributed to the outcome. When outcomes are unfavorable and perceptions of intention are high, there is a tendency to assign blame. [...] Within a social utility framework, the occurrence of aggression could be expected to

lead to a heightened concern for comparative payoffs and a reduced concern for absolute payoffs." (Blount 1995)

Both models of inequity aversion ignore the role of infered intentions for individuals' behavior. They proceed from the assumption that people compare only relative payoffs to judge whether they have been treated fairly. This concept is called outcome fairness in contrast to contextual fairness. In many situations outcome fairness by itself cannnot sufficiently explain why people behave the way they do. To clarify the importance of intentions many experiments have been made. Falk, Fehr and Fischbacher (2003) and Nelson (2002) have shown that identical offers in an ultimatum game lead to different rejection rates, depending on the alternatives that were available to the proposer.

In the study of Falk, Fehr and Fischbacher (2003) four so-called mini-ultimatum games were conducted. A mini-ultimatum game has the same structure as the ultimatum game with the only difference that the proposer cannot freely choose how much to offer but has the choice between only two allocations. In the experiment the allocation x stayed the same and the alternative allocation y differed from game to game. The allocation $x = (8/2)$ implies that if the responder accepts, he gets 2 points and the proposer gets 8 points. In game A the alternative offer y is $(5/5)$, in game B it is $(2/8)$, in game C there is no alternative at all and in game D the alternative is $(10/0)$. There were 90 experimental subjects that all participated in each game. Each responder was asked to decide on his reaction for both the x and the y case

without knowing in advance what the proposer's action had been. Thus responders had to specify complete strategies in the game-theoretic sense. Subjects were randomly assigned the proposer or responder role. They faced the four games in varying order and played against a different anonymous opponent in each game. Outcomes were only announced after all four games had been completed. This approach had the advantages that income effects could be avoided and that subjects' behavior was not influenced by previous decisions of their opponents.

The standard game-theoretic model with selfregarding preferences predicts that x offers are never rejected throughout all of the four settings. The Bolton-Ockenfels and Fehr-Schmidt models of inequity aversion are consistent with positive rejection rates but predict that the rejection rate of the (8/2) offer is the same for all games. Intuitively most people would expect that responder behavior does differ according to the setting they are in, and this was confirmed by the experimental data. Whether the x offer was qualified as acceptable, depended on whether the responder would have been free to chose a more equal distribution or not. In the (5/5)-game rejection rates were highest (44%) indicating that many people find it offensive not to be offered an equal share. The rejection rate was lower in the (2/8)-game (26,7%). Offering (8/2) now was probably still perceived as unfair but less so than in the (5/5)-game because the only alternative would have been another unfair distribution and from the fact that the proposer did not want to be unfair to herself one could not conclude that she meant to be unfair to the responder.

In the (8/2)-game the rejection rate was only 18%. In this game the proposer did not have any choice and so from the outcome the responder could not infer any potentially good or bad intentions. Therefore this rejection rate measures pure inequality aversion. In the last game, the (10/0)-game, the rejection rate is the lowest (8,9%). It is even lower than pure inequality aversion since offering (8/2) in this game might even be thought to be a fair or kind action.

Various statistical tests (non-parametric Cochran Q-test, pair-wise comparisons, non-parametric McNemar test) confirm that the differences in rejection rates across all games are significant and results are robust.

From the proposer's point of view the expected return from the (8/2)-offer varies across games. It was least profitable to propose (8/2) in the (5/5)-game and most profitable in the (10/0)-game. A majority of the proposers made the payoff-maximizing choice in each game, which was (5/5) in the (5/5)-game and (8/2) in the (2/8)- and (10/0)-games. It remains an open question whether proposers actually cared for fairness or just tried to maximize their payoffs anticipating their responders' behavior. For further discussion see Falk, Fehr and Fischbacher (2003).

However, the results suggest that pure inequity aversion does play some role, but the differences in rejection rates reveal that it does not tell the whole story. Similar results were provided by the experiments of Blount (1995), Brandts and Solà (2001), Güth et al. (2001) and Offerman (2002). Blount found in her experiments that people react very differently to ac-

tions, depending on whether they believe those were taken by humans or by non-human devices. When agents are human, people develop normative expectations about how these agents should behave and they experience disutility if their expectations are not met.

As a consequence, fairness models should not only take into account that many people care for the distribution of payoffs. People also try to infer and value fairness intentions behind actions. Reciprocity based models are models that try to capture this idea.

3 Reciprocity

"[...] Reciprocity is a behavioral response to perceived kindness and unkindness where kindness comprises both distributional fairness as well as fairness intentions." (Falk 2006) Unlike inequity aversion models reciprocity based models assume that people's evaluation of the kindness or fairness of an action does not only depend on its consequences but also on the actor's underlying intentions.

3.1 Rabin's model of reciprocity

Rabin (1993) developed a game-theoretic framework for incorporating the idea of reciprocity into economic models. Let us start from the following considerations: People are willing to sacrifice their own material well-being to help those who are being kind and to punish those who are being un-

kind. Both motivations have a greater effect on behavior the smaller the cost of sacrificing becomes. Outcomes that reflect these motivations are called fairness equilibria.

In the following model, payoffs depend on players' actions and on their beliefs. Whether a person's actions are kind or unkind depends on what she thinks the consequences of her actions will be. In other words, a player's kindness depends on her intentions. If the other player wants to reciprocate, he must form beliefs about the first player's intentions. Intentions also depend on beliefs. It follows that reciprocal motivation can only arise with beliefs about beliefs. Games in which payoffs also depend on players' beliefs about other player's strategic choices or beliefs are called psychological games.

Rabin's model of reciprocity is applicable to all two-person, finite-strategy normalform-games. Assume that in a two-player (i=1,2) normal form game with (mixed) strategy sets S_1 and S_2 derived from the finite pure strategy sets A_1 and A_2 material payoffs are $\pi_i : S_1 x S_2 \rightarrow \mathbb{R}$. From this "material game" the psychological game is constructed. Each player's subjective expected utility when choosing her strategy depends on (1) her strategy, (2) her beliefs concerning the strategy the other player will choose, and (3) her beliefs about the other player's beliefs concerning her own strategy. The strategies chosen by the players are represented in the following model as $a_1 \epsilon S_1$ and $a_2 \epsilon S_2$. $b_1 \epsilon S_1$ and $b_2 \epsilon S_2$ represent player 2's beliefs about the strategy player 1 is choosing and player 1's beliefs about the strategy player 2 is choosing.

$c_1 \epsilon S_1$ and $c_2 \epsilon S_2$ are player 1's beliefs about what player 2 believes player 1's strategy is, and player 2's beliefs about what player 1 believes player 2's strategy is. $f_i(a_i,b_j)$ defines a kindness function measuring how kind player i is to player j. For the sake of simplicity, it is assumed that the kindness function is the same for both players, which means that they have the same notion of kindness or fairness. Player i tries to infer the other's kindness by evaluating the payoff combination j's choice has induced. The set of all feasible payoffs is $\Pi(b_j) \equiv \{(\pi_i(a,b_j),\pi_j(b_j,a)) \mid a \epsilon S_i\}$. Player i chooses a strategy that yields payoffs $(\pi_i(a_i,b_j),\pi_j(b_j,a_i))$ given player j chooses b_j. $\pi_j^h(b_j)$ denotes the highest payoff in the set of all feasible payoffs for player j. $\pi_j^l(b_j)$ is the lowest payoff among all pareto-efficient points in $\Pi(b_j)$ for player j. From these two points one can derive an "equitable payoff"

$$\pi_j^e(b_j) = \frac{1}{2}[\pi_j^h(b_j) + \pi_j^l(b_j)]$$

This is a rough reference point which shows how fair or kind player i is towards player j. $\pi_j^{min}(b_j)$ is the worst possible payoff for player j in the set $\Pi(b_j)$. The kindness function can then be defined as

$$f_i(a_i, b_j) \equiv \frac{\pi_j(b_j, a_i) - \pi_j^e(b_j)}{\pi_j^h(b_j) - \pi_j^{min}(b_j)}$$

It measures how much more or less than her opponent's equitable payoff player i believes she is giving to player j. If player i gives player j her equitable payoff the kindness function takes a value of zero. If she gives her

less than the equitable payoff, the kindness function takes a value below zero. f_i can be negative for two reasons: either player i takes more than her equal share on the Pareto frontier of $\Pi(b_j)$ or she chooses an inefficient point in $\Pi(b_j)$.

The formally equivalent function $f_j^\circ(b_j,c_i) \equiv \frac{\pi_i(c_i,b_j)-\pi_i^e(c_j)}{\pi_i^h(c_i)-\pi_i^{min}(c_i)}$ gives player i's belief about how fair player j treats her.

Since both $f_j^\circ(\cdot)$ and $f_i(\cdot)$ are normalized, their values always lie between -1 and $\frac{1}{2}$ and they are not sensitive to positive affine transformations of the material payoffs. Players' preferences can now be specified by using the kindness functions. Every player chooses a_i in order to maximize her expected utility $U_i(a_i,b_j,c_i)$. The utility function captures material utility as well as utility gained from perceived fairness intentions:

$$U_i(a_i, b_j, c_i) \equiv \pi_i(a_i, b_j) + f_j^\circ(b_j, c_i) \cdot [1 + f_i(a_i, b_j)]$$

The idea of reciprocity is realised as follows: If player i believes that player j is unfair to her, which means that $f_j^\circ(\cdot)$ is negative, she wants to be unkind to her opponent by choosing a_i so that $f_i(\cdot)$ is low or negative. On the other side if player j acts kindly so that $f_j^\circ(\cdot)$ is positive, player i will also wish to be kind to player j. Despite of the reciprocity effect, players of course also value pure material well-being and thus have to trade off their preference for fairness against their material interests. In this context a notable property of Rabin's model should be considered: The behavior in these games is sensitive

to the scale of material payoffs. The bigger they are, the less player's behavior will reflect their concern for fairness. This is due to the fact that kindness functions are bounded. Whether this is a desirable property or a weakness of the model is not completely clear. On the one hand it is intuitive to assume that people only care for fairness when there is not much at stake and turn more selfish when high payoffs can be obtained. On the other hand experiments with varying stakes in ultimatum games have shown that even in high-stakes treatments proposers do not offer less (see Camerer and Thaler (1995)).

The solution concept used in these games is what Rabin calls fairness equilibrium. Fairness equilibria are similar to Nash equilibria for psychological games. An additional condition that must be met is that all higher-order beliefs have to match actual behavior. Fairness equilibria neither constitute a subset nor a superset of Nash equilibria. In other words, the concept of fairness equilibria can add new predictions to economic models as well as eliminate conventional predictions.

A fairness equilibrium is a pair of strategies $(a_1,a_2)\epsilon(S_1,S_2)$ for $i = 1,2$, $j \neq i$ with $a_i \epsilon argmax_{a\epsilon S_i} U_i(a_i,b_j,c_i)$ and $a_i = b_i = c_i$.

In order to simplify the analysis of properties of fairness equilibria more notation has to be introduced.

- A mutual-max outcome is a strategy pair $(a_1,a_2)\epsilon(S_1,S_2)$ for $i = 1,2, i \neq j$ and $a_i \epsilon argmax_{a\epsilon S_i} \pi_j(a,a_j)$.

- A mutual-min outcome is a strategy pair $(a_1,a_2)\epsilon(S_1,S_2)$ for $i = 1,2, i \neq j$ and $a_i \epsilon argmin_{a \epsilon S_i} \pi_j(a,a_j)$.

- An outcome is strictly positive if, for i=1,2, $f_i > 0$. If $f_i \geq 0$ an outcome is called weakly positive. Analogously an outcome is strictly negative if, for i=1,2, $f_i < 0$ and weakly negative if $f_i \leq 0$. If $f_i = 0$ an outcome is called neutral and an outcome is called mixed if, for i=1,2, $f_i, f_j < 0$.

The following propositions hold for all games, irrespective of the scale of material payoffs.

A) If (a_1,a_2) is a Nash equilibrium and either a mutual-max or mutual-min outcome, (a_1,a_2) is also a fairness equilibrium.

Proof 3.1.1 *(a_1,a_2) being a Nash equilibrium means that both players are maximizing their material payoffs. If (a_1,a_2) is a mutual-max outcome f_1 and f_2 are both nonnegative, which means that both players wish to be kind to another. Each player chooses a strategy that maximizes her own material payoff as well as her opponent's material well-being. Thus each player maximizes her overall utility. In the other case where (a_1,a_2) is a mutual-min outcome implying that f_1 and f_2 are nonpositive, both players want to decrease their opponent's material payoff. While doing so, they simultaneously maximize their own material well-being and therefore maximize their overall utility.*

B) All fairness equilibria outcomes are strictly positive or weakly negative. This means that in a fairness equilibrium it will never be the case that one person is kind while at the same time the other is unkind.

Proof 3.1.2 *Suppose there was an outcome with* $f_i > 0$ *and* $f_j \leq 0$. $f_i > 0$ *implies that player* i *could increase her payoff by maximizing her own material interest. Doing so would at the same time harm player* j. *If* $f_j \leq 0$ *utility maximization requires player* i *to do so even at the expense of player* j. *Therefore the outcome cannot be a fairness equilibrium.*

Roughly speaking, it can be asserted that in games with very small material payoffs fairness equilibria correspond to Nash equilibria in situations where each player tries to either maximize or minimize the other player's payoff. In games with very large payoffs fairness plays relatively less role. In this case Nash equilibria are fairness equilibria, too, except for some non-strict Nash equilibria.

There is an asymmetry inherent in the model. In every game there exists a weakly negative fairness equilibrium but not necessarily a positive one. The intuition behind this is the following: When a player maximizes his own material payoff, he is either mean or neutral to the other player. Being kind involves sacrificing own material well-being. There are situations where material self-interest gives people an incentive to be mean even when other players are kind, but on the other hand material self-interest will never be a

motivation to be kind when other players are being unfair. Insofar negative feelings have more influence on outcomes than positive feelings.

A weakness of Rabin's model is that it is only applicable to two-person, normal form, complete information games. Moreover, Rabin examined only qualitative predictions of his model. For further details see Rabin (1993).

3.2 Sequential Reciprocity

The Dufwenberg - Kirchsteiger approach (2004) is in a way similar to Rabin's model. Kindness depends solely on intentions and these are infered from choices given available alternatives. The advance of the Dufwenberg - Kirchsteiger model is that it can deal with the sequential structure of strategic interactions and proposes a new solution concept, the sequential reciprocity equilibrium. If the concept of sequential reciprocity equilibria is applied to any two-player normal form game, the results are qualitatively similar to Rabin's results.

To model the impact of intentions, one has to take into account the possibilities as well as the beliefs of the players. In a sequential game players will have to revise their beliefs about how kind other players are as the game proceeds. The way a player is affected by reciprocity concerns may differ depending on her position in the game tree. With each new subgame that is entered the reciprocity motivation might change. Thus a reasonable model of reciprocity in sequential games must provide a solution for the problem of changing beliefs and how they affect reciprocity considerations.

The model distinguishes between a player's initial and subsequent beliefs and keeps track of the latter as new subgames are reached. It is assumed that in each choice situation a player is motivated according to the beliefs she has at that stage. The type of games analysed are finite multi-stage games with observed actions. Games proceed in stages. Any player reaching a stage along her path knows all preceeding choices, moves exactly once and has no information about her opponents' choices at the same stage. Because decisions are made simultaneously, games are of incomplete information.

$N = \{1, ...,n\}$ is the set of players where $n \geq 2$. H is the set of histories or the set of choice profiles that lead to subgames. A_i is the set of strategies of player i. A strategy assigns for each history a probability distribution on the set of possible choices of i at h. For simplification one can imagine that players make pure strategy choices only. However, the concept also allows for randomization. This can be thought of as frequencies with which pure strategy choices are made in a population of players. A, the set of all strategy combinations, is defined as $\prod_{i \epsilon N} A_i$. Each player's payoff function depends on what profile in A is played. So $\pi_i : A \rightarrow \mathbb{R}$ where π_i is the material payoff function measuring money or some other objectively measurable quantity. The utility of player i consists of both the material payoff and the socalled reciprocity payoff. $B_{ij} = A_j$ is the set of possible beliefs of player i about the strategy of player j. $C_{ijk} = B_{jk} = A_k$ is the set of possible beliefs of player i about the belief of player j about the strategy of player k. Each player's behavior, beliefs, kindness and perception of other players' kindness might

differ across histories.

$a_i(h)$ is an updated strategy after history h. All choices that define h are given probability 1 since they are already past. For all other (future) choices the strategy corresponds to $a_i \epsilon A_i$. Assume that player i plays a_i, that she believes that the others play $(b_{ij})_{j \neq i}$ and that she believes that the others believe $(c_{ijk})_{k \neq j}$. After history h player i plays $a_i(h) \epsilon A_i$, believes that the others play $(b_{ij}(h))_{j \neq i}$ and that they believe $(c_{ijk}(h))_{k \neq j}$. The updating of beliefs follows Bayes rule. Moreover, it is assumed that players treat the choices of others as intentional and deliberate.

If i chooses a_i and believes all other players choose $(b_{ij})_{j \neq i} \epsilon \prod_{j \neq i} B_{ij}$ he expects that j's material payoff will be $\pi_j(a_i,(b_{ij})_{j \neq i})$. i believes the set of feasible material payoffs for j to be $\{\pi_j(a_i',(b_{ij})_{j \neq i} \mid a_i' \epsilon A_i\}$. The relative size of $\pi_j(a_i,(b_{ij})_{j \neq i})$ within the set of feasible payoffs determines how kind player i is to player j.

Player i's set of efficient strategies is given by

$$E_i = \{a_i \epsilon A_i \mid \text{there exists no } a_i' \epsilon A_i \text{ so that for all } h \epsilon H, (a_j)_{j \neq i} \epsilon \prod_{j \neq i} A_j, \text{ and } k \epsilon N$$

$$\text{it holds that } \pi_k(a_i'(h), (a_j(h))_{j \neq i}) \geq \pi_k(a_i(h), (a_j(h))_{j \neq i}), \text{ with strict inequality}$$

$$\text{for some } (h, (a_j)_{j \neq i}, k)\}$$

A strategy is efficient, if there exists no alternative strategy - given the history of the game and given all subsequent choices by other players - that

provides a higher material payoff for some player without making any other player worse off. Vice versa, a strategy is inefficient, if there exists another strategy - given the history of the game and all subsequent choices by the other players - that provides no lower payoff for any player and a higher material payoff for some player for some history of the game including subsequent choices of the others.

As a reference point to measure how kind player i is to player j an equitable payoff for player j given i's beliefs $(b_{ij})_{j\neq i}$ is introduced. The equitable payoff $\pi_j^{e_i}((b_{ij})_{j\neq i})$ is an average between the lowest and the highest material payoff of j, if i chooses an efficient strategy:

$$\pi_j^{e_i}((b_{ij})_{j\neq i}) = \frac{1}{2}[max\{\pi_j(a_i,(b_{ij})_{j\neq i}) \mid a_i \epsilon A_i\} + min\{\pi_j(a_i,(b_{ij})_{j\neq i}) \mid a_i \epsilon E_i\}]$$

If i chooses a strategy so that $\pi_j = \pi_j^{e_i}$ his kindness to j is zero, which means that he is neither kind nor unkind towards player j. The kindness of player i to j at history h is given by the function:

$$K_{ij} : A_{ij} \times \prod_{j\neq i} B_{ij} \to \mathbb{R}$$

$$K_{ij}(a_i(h),(b_{ij}(h))_{j\neq i}) = \pi_j(a_i(h),(b_{ij}(h))_{j\neq i}) - \pi_j^{e_i}((b_{ij}(h))_{j\neq i})$$

In words, a positive difference between the actual payoff and the equitable payoff of j arises from kind behavior of i and a negative difference from unkind behavior of i. Kindness cannot directly be observed by the players,

but they have beliefs about the other players' actions and beliefs. From these beliefs they infer how kindly their opponents are actually acting. Player i's beliefs about how kind player j is to him at history h are described by the function lll

$$\lambda_{iji}(b_{ij}(h), (c_{ijk}(h))_{k\neq j}) = \pi_i(b_{ij}(h), (c_{ijk}(h))_{k\neq j}) - \pi_j^{e_i}((c_{ijk}(h))_{k\neq j})$$

The function λ_{iji} is mathematically equivalent to K_{ij} since $B_{ij} = A_j$ and $C_{ijk} = B_{jk}$. However, λ_{iji} captures a psychological component affecting player i.

The utility of player i at h is given by a function of the form $U_i : A_i \times \prod_{j\neq i}(B_{ij} \times \prod_{k\neq j} C_{ijk}) \rightarrow R$. It is the sum of the material payoff and the reciprocity payoffs with respect to each player $j \neq i$:

$$U_i(a_i(h),\, (b_{ij}(h),\, (c_{ijk}(h))_{k\neq j}) = \pi_i(a_i(h),\, (b_{ij}(h))_{j\neq i}) +$$

$$+ \sum_{j\epsilon N\setminus i} (Y_{ij} \cdot K_{ij}(a_i(h),\, (b_{ij}(h)_{j\neq i}) \cdot \lambda_{iji}(b_{ij}(h),\, (c_{ijk}(h))_{k\neq j}))$$

where Y_{ij} is an exogenously given non-negative number for each $j \neq i$. It measures the degree to which a player is affected by reciprocity motivation. This sensitivity to reciprocity can vary, depending on which other player is concerned. If $Y_{ij} > 0$ and i believes that j is kind to him, then i's reciprocity payoff with respect to j is increasing in i's kindness to j. The higher $\lambda_{iji}(\cdot)$ is, the more material payoff player i is ready to sacrifice in order to do j a

favour. On the other hand, if i believes that j is unkind to him, which means that $\lambda_{iji}(\cdot)$ is negative, his reciprocity payoff with respect to j is decreasing the more kind i is to j. When player i is optimizing his utility, he has to take into account reciprocity payoffs with respect to all other players plus his own material payoff.

In equilibrium all players make optimal choices in each history given their beliefs. Following each history, beliefs are updated. The profile $a^\star = (a_i^\star)_{i\epsilon N}$ is a sequential reciprocity equilibrium (SRE), if for all $i\epsilon N$ and for each $h\epsilon H$ the following conditions hold:

$$(i)\, a_i^*(h)\epsilon argmax\, U_i(a_i,\, (b_{ij}(h),\, c_{ijk}(h))_{k\neq j})_{j\neq i)},$$

$$(ii)\, b_{ij} = a_j^*\, for\, all\, j \neq i,$$

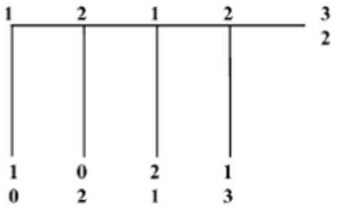
$$(iii)\, c_{ijk} = a_k^*\, for\, all\, j \neq i, k \neq j$$

Verbally, a strategy profile is a sequential reciprocity equilibrium, if at each history each player makes choices that maximize his utility given his beliefs and given that he follows his equilibrium strategy at other histories. Conditions (ii) and (iii) require that initial beliefs are correct. Condition (i) further implies that beliefs assign probability one to the sequence of choices that define the history of the game and otherwise equal the initial beliefs.

If $Y_{ij} = 0$ for all $i, j\epsilon N$, the utility functions only consist of material payoffs. In such a game, the sequential reciprocity equilibrium is a simple subgame perfect equilibrium as used in standard game theory. For $Y_{ij} > 0$,

proving existence of a sequential reciprocity equilibrium is more demanding. The problem arises from the fact that kindness and perceived kindness depend on beliefs about actions following all histories - also histories that do not follow h. Therefore in general it is not possible to determine equilibrium choices by analysing isolated subgames, which means that the usual tool of backwards induction cannot be used. The solution is to look at all histories simultaneously. Applying Kakutani's fixed point theorem to a best-reply correspondence which distinguishes between players and between different histories, one can show that there exists a sequential reciprocity equilibrium in every psychological game with reciprocity incentives.

A relatively easy application of the Dufwenberg-Kirchsteiger model is the centipede game. It is a two-player game with *1,...M* nodes where $M \geq 2$. At the beginning of the game player 1 starts with one unit of material payoff and player 2 with two units of material payoff. At each node a player can decide whether to stay in the game or whether to end it. Player 1 makes her decisions at odd nodes and player 2 at even nodes. If player 1 stays, the material payoff of player 2 increases by two units whereas her own material payoff declines by one unit and the next node is reached. The same is true for player 2 vice versa. If a player decides to finish the game, the material payoffs of both players do not change and the game ends. Each player's strategy determines at each node whether to continue or end the game.



Dufwenberg and Kirchsteiger (2004)

The figure shows a centipede game with four nodes.

$e(s_1,s_2)$ denotes the first node where one of the players ends the game given they play their strategies s_1 and s_2. If $e(s_1,s_2)$ is odd, the material payoffs are given by $\pi_1(s_1,s_2) = \frac{(e(s_1,s_2)+1)}{2}$ and $\pi_2(s_1,s_2) = \frac{(e(s_1,s_2)-1)}{2}$. If the ending node is even, material payoffs are $\pi_1(s_1,s_2) = \frac{(e(s_1,s_2)-1)}{2}$ and $\pi_2(s_1,s_2) = \frac{(e(s_1,s_2)+1)}{2}$. Because of the symmetry of the game it is sufficient to consider only the case where M is even. If both players decide to stay at all nodes, payoffs are $\pi_1(s_1,s_2) = \frac{M}{2} + 1$ and $\pi_2(s_1,s_2) = \frac{M}{2}$.

By the logic of backwards induction standard game theory predicts that in the only Nash equilibrium player 1 ends the game at the first node. However, experimental testing has shown that most people stay in the game at the first nodes. Rabin's model of reciprocity can be applied to the normal form of the game. If reciprocity motives are strong enough, the model gives multiple equilibria - in some of them player 1 ends the game at the first node and in another one players stay in the game at all nodes. In the Dufwenberg-Kirchsteiger model there is a unique sequential reciprocity equilibrium in which both players stay until the last node, provided that at least one of the players sufficiently cares for reciprocity. It is sufficient that one player is

motivated by reciprocity, because then the other non-reciprocal player can convince the reciprocal player by staying that he is being kind, no matter whether he actually cares for the opponent or not.

In all sequential reciprocity equilibria it holds at the last node M (at which player 2 makes the decision) that: (a) if $Y_2 > \frac{2}{M}$ player 2 stays, (b) if $Y_2 < \frac{2}{M+2}$ player 2 exits and (c) if $\frac{2}{M+2} < Y_2 < \frac{2}{M}$ player 2 stays with a probability of $p = 1 + \frac{M}{2} - \frac{1}{Y_2}$. It is intuitive that players with a strong concern for reciprocity will stay even at the last node whereas players who only feel little motivated by reciprocity will not. The more stages the centipede game has, the less reciprocity motivation is needed to make player 2 stay at the last node. This is because the more stages there are in total, the more often player 1 has already chosen to stay until the last node is reached. That means player 1's kindness to player 2 increases with the number of nodes and so a smaller Y_2 is needed to make player 2 give up some of his material payoff in order to reciprocate 1's kindness. A formal proof can be found in Dufwenberg and Kirchsteiger (2004).

Another observation is that if $Y_i < \frac{2}{M+2}$ for *i=1,2* the only sequential reciprocity equilibrium is such that both players want to end the game at any node that is reached.

Proof 3.2.1 *It can be shown that player 2 will exit at node* M *if* $Y_i < \frac{2}{M+2}$. *At any node* k *it is optimal to end the game given that all players exit at all nodes larger than* k. *This results from the fact that the material payoff of player* i *decreases by one unit if he stays, since at* k+1 *the other player will*

end the game anyway. The difference in player i's kindness between choosing to stay or leave is -2 because j*'s material payoff decreases by 2 if* i *chooses to stay instead of go. Therefore the difference in* i*'s utility between choosing to end the game or continue is* $1 - 2Y_i\lambda_{iji}(\cdot)$*. It is clear that* λ_{iji} *cannot be greater than* $\frac{M}{4} + \frac{1}{2}$*. So whenever* $Y_i < \frac{2}{M+2}$ *at any node* k i*'s utility from ending the game is larger than staying.*

If on the other hand $Y_2 > \frac{2}{M}$, there exists a unique sequential reciprocity equilibrium where both players stay in the game at all nodes. Whether such an SRE, where players stay until the end, exists depends on how much player 2 is motivated by reciprocity. Interestingly enough, it is not relevant whether player 1 also cares for reciprocity.

Proof 3.2.2 *It is clear that player 2 will stay at node* M*. The crucial node is thus node* M-1*. Consistency of beliefs requires that player 1 believes that 2 will always stay whenever node* M-1 *is reached. At* M-1 *player 1 wants to stay because his material payoff from staying at* M-1 *is strictly larger than from ending the game there. Moreover, staying is also the kindest choice player 1 can make at* M-1*. Also player 2's behavior is clearly kind. So if* $Y_1 > 0$*, the psychological part of 1's utility function is also maximal, if he chooses to stay. For any* Y_1 *player 1 will decide to stay at* M-1*. (If* $Y_1 = 0$ *the psychological part of 1's utility function is equal to zero.) At node* M-2 *player 2 decides whether to stay or leave the game. By the same logic as before, player 2 will choose to stay. Reasoning further backwards gives that*

at any node in the game players stay in the game.

Another question is whether the same logic can be applied if only player 1 is motivated by reciprocity and player 2 is not. It has already been shown that if 2's sensitivity to reciprocity is low, he will end the game at node *M*. If *M=2* player 2's behavior is clearly unkind and so player 1 cannot have an incentive to stay at the first node. But if the number of stages is sufficiently high and player 2 has chosen to stay in the game several times, player 1 might perceive 2's behavior as kind no matter what 2 does at the last node. A sufficiently reciprocity motivated player 1 might therefore decide to stay at *M-1,* regardless of what he thinks player 2 will chose to do at *M*. If this should be the case, player 2 - reciprocity motivated or not - should stay at all nodes preceding *M-1*. Such a sequential reciprocity equilibrium, where only player 1's reciprocity motivation guarantees that the game does not end before the last node, can only be reached if $M > 6$ and $Y_1 > \frac{2}{M-6}$. For a detailed proof see Dufwenberg and Kirchsteiger (2004).

Sequential reciprocity equilibria need not be unique and they need not exist in pure strategies. The Dufwenberg-Kirchsteiger model works well for many games, e.g. the sequential prisoners' dilemma or the centipede game as seen before. However, it fails - like other models of reciprocity - to explain behavior in dictator games.

3.3 Levine's model of reciprocity

Levine (1998) suggests a different approach to measuring kindness. In Levine's model - as opposed to other reciprocity models - players do not react to kind or unkind actions. They punish or reward spiteful or altruistic types of players. The question is not whether opponents play fairly but whether they are kind people. This has the advantage of avoiding the problem that there is no obvious notion of fairness that applies to all games. Levine uses experimental results from ultimatum games and the final round of a centipede game in order to fit his model and to deduce the distribution of spite and altruism in the population.

The basic idea of models of interdependent preferences is that utility is a linear function of both the player's own payoff and his opponent's payoff: $u_i = \pi_i + \sum_{j \neq i} \beta_{ij} u_j$. The coefficient on the own monetary payoff is normalized to one. The problem then is to determine β, the coefficient on the opponent's payoff. In fact, this specification of the model is too simplified to meet reality.

Levine (1998) develops a more adequate approach. He assumes that within a population there are different types of people, some having positive, others having negative coefficients. These are all private information and the distribution of types is fixed across different games. Players' weights on opponents' monetary payoffs depend both on their own coefficient of spite or altruism and on what they believe their opponents' coefficients are. In this perspective we are analyzing signaling games. All actions taken potentially

reveal how altruistic or spiteful players are and this is what opponents care about.

Consider n person, i=1,..., n, extensive form games. Utility functions are given by

$$U_i = \pi_i + \sum_{j \neq i} \frac{\alpha_i + \lambda \alpha_j}{1 + \lambda} \pi_j$$

π_i and π_j are material payoffs. α is the coefficient of altruism with $-1 < \alpha < 1$. It measures the relative importance of another person's payoff compared to one's own payoff. If $\alpha_i > 0$, player i is altruistic, if $\alpha_i < 0$, player i is spiteful and if $\alpha_i = 0$ he is selfish. α is bounded between -1 and 1 because no player is supposed to have a higher regard for his or her opponent than for him- or herself. λ is a universal reciprocity parameter, which means that all players are assumed to have the same λ. When $\lambda = 0$, the model is one of pure altruism. When $\lambda > 0$, the model incorporates an element of fairness. Players are more willing to be altruistic towards players who have been altruisitic towards them.

A population of players has a distribution of altruism coefficients represented by a common cumulative distribution function $F(\alpha_i)$. F has finite support for any given monetary payoffs in a game. The individual α_i is private information, whereas the distribution F is common knowledge. People have an initial prior about the type of their opponent and after observing their opponent's actions, players update their beliefs. If an action is judged as kind, the person behind the action will be judged kind, too.

The equilibrium concept used is that of a sequential equilibrium. If games are sufficiently simple, sequential equilibria coincide with perfect Bayes Nash equilibria. All equilibria that will be discussed in the following also satisfy the monotonicity requirement on beliefs in signaling games: beliefs are that the type most likely to deviate is the type for whom it is most advantageous to do so.

A drawback of the model is that it makes the analysis of games relatively complicated. For the ultimatum game the theory works fairly well. Suppose there is a total of 10 units of money to be divided between the two players.

Levine shows that regardless of F, in no sequential equilibrium any proposal will be higher than 5 and any proposal of 5 or more will be accepted. This is intuitive and consistent with actual data.

Proof 3.3.1 *Consider a general utility function with interdependent preferences:* $u_i = \pi_i + \sum_{j \neq i} \beta_{ij} u_j$. *The coefficients* β_{ij} *can be determined from players' types or other details of the game. As seen before Levine's specification of* β *corresponds to* $\sum_{j \neq i} \frac{\alpha_i + \lambda \alpha_j}{1+\lambda}$. *Responder's utility in the ultimatum game is given by* $(10 - x) + \beta x$ *where* $(10 - x)$ *is the proposal and* x *is what the allocator demands for himself. If* $\beta > -1$ *and* $x \leq 5$, *the expression is positive. So indeed any proposal of 5 or more is accepted. The proposer's utility* $x + \beta(10 - x)$ *is increasing in* x *for* $\beta > -1$. *Therefore a proposer's demand for himself below 5 can be increased without reducing the probability that the responder accepts. So it cannot be optimal to propose more than 5.*

Empirically, allocators' demands for themselves are between 5 and 7. The data we consider is the following:

Demand	Observations	Frequ. of obs.	Acc. demands	Prob. of acc.
5	37	28%	37	1
6	67	52%	55	0.8
7	26	20%	17	0.65

Roth et al. (1991)

It is assumed that the distribution of altruism coefficients is in a way that weight is placed on the points $\alpha^h > \alpha_0 > \alpha^l$, which are linked to altruistic, normal and spiteful types of players. We look for an equilibrium in which the altruistic type demands 5 (respectively proposes 5), the normal type demands 6 and the spiteful type 7. The probabilities of the types must then be 0.28, 0.52 and 0.2. A demand of 5 is accepted by all responders. A demand of 6 is accepted by the normal and altruistic types and rejected by the spiteful types of responders. A demand of 7 is accepted by 65% of the population, which corresponds to all the altruistic types and a fraction of the normal types implying that normal types must be indifferent between accepting and rejecting an offer of 3.

The parameters $1 > \alpha^h > \alpha_0 > \alpha^l > -1$ and $0 \leq \lambda \leq 1$ consistent with equilibrium must fulfil the following conditions:

(1) $$(6 + \tfrac{\alpha_0 + \lambda(0.35\alpha^h + 0.65\alpha_0)}{1+\lambda} \cdot 4)0.8 - (5 + \tfrac{\alpha_0 + \lambda(0.28\alpha^h + 0.52\alpha_0 + 0.2\alpha^l)}{1+\lambda} \cdot 5) \geq 0$$

(2) $(6+\frac{\alpha^h+\lambda(0.35\alpha^h+0.65\alpha_0)}{1+\lambda}\cdot 4)0.8-(5+\frac{\alpha^h+\lambda(0.28\alpha^h+0.52\alpha_0+0.2\alpha^l)}{1+\lambda}\cdot 5)\leq 0$

(3) $4+\frac{\alpha^l+\lambda\alpha_0}{1+\lambda}\cdot 6\leq 0$

(4) $(7+\frac{\alpha^l+\lambda(0.43\alpha^h+0.57\alpha_0)}{1+\lambda}\cdot 3)0.65-(6+\frac{\alpha^l+\lambda(0.35\alpha^h+0.65\alpha_0)}{1+\lambda}\cdot 4)0.8\geq 0$

(5) $(7+\frac{\alpha_0+\lambda(0.43\alpha^h+0.57\alpha_0)}{1+\lambda}\cdot 3)0.65-(6+\frac{\alpha_0+\lambda(0.35\alpha^h+0.65\alpha_0)}{1+\lambda}\cdot 4)0.8\leq 0$

(6) $3+\frac{\alpha_0+\lambda\alpha^l}{1+\lambda}\cdot 7=0$

(7) $(7+\frac{\alpha^l+\lambda(0.43\alpha^h+0.57\alpha_0)}{1+\lambda}\cdot 3)0.65\geq 2.8$

Proof 3.3.2 *Consider the case of a demand of 5. The proposer's adjusted utility demanding 5 for himself is thus* $(5+5\cdot\frac{\alpha+\lambda(0.28\alpha^h+0.52\alpha_0+0.2\alpha^l)}{1+\lambda})$. *The demand is made by an altruistic type. For the spiteful type to accept it must hold that* $5+\frac{\alpha^l+\lambda\alpha^h}{1+\lambda}\cdot 5\geq 0$. *For* α^l,$\alpha^h>-1$ *this inequality is always satisfied. If the spiteful type accepts, all types will accept.*

If the proposer demands 6, altruistic and normal types accept. The utility gained by the proposer is $(6+\frac{\alpha+\lambda(0.35\alpha^h+0.65\alpha_0)}{1+\lambda}\cdot 4)0.8$. *For the normal type of proposer (and therefore also for the spiteful type) this must be higher than the utility from only keeping 5. This implies constraint (1):*

$(6+\frac{\alpha_0+\lambda(0.35\alpha^h+0.65\alpha_0)}{1+\lambda}\cdot 4)0.8-(5+\frac{\alpha_0+\lambda(0.28\alpha^h+0.52\alpha_0+0.2\alpha^l)}{1+\lambda}\cdot 5)\geq 0$

On the other hand, for the altruistic type keeping 6 must lead to a lower utility than keeping 5, which leads to constraint (2):

$(6+\frac{\alpha^h+\lambda(0.35\alpha^h+0.65\alpha_0)}{1+\lambda}\cdot 4)0.8-(5+\frac{\alpha^h+\lambda(0.28\alpha^h+0.52\alpha_0+0.2\alpha^l)}{1+\lambda}\cdot 5)\leq 0$

Responder behavior must make spiteful types reject and normal types as well as altruistic types accept. The demand of 6 is by assumption made by a

normal type proposer. The spiteful type rejects if $4 + \frac{\alpha^l + \lambda\alpha_0}{1+\lambda} \cdot 6 \leq 0$ *which is constraint (3).*

The proof of conditions (4) to (7) works in a similar way and can be found in Levine (1998).

Further characteristics of the equilibrium can be examined:

There is no sequential equilibrium with $\lambda = 0$. This means that a model of pure altruism does not fit the data of the ultimatum game. This is because the high rejection rates imply that players must be relatively spiteful. But on the other hand spiteful proposers would not make as generous proposals as observed.

Proof 3.3.3 *We have* $3 + \frac{\alpha_0 + \lambda\alpha^l}{1+\lambda} \cdot 7 = 0$ *since the normal type must be indifferent to accepting or rejecting. If* $\lambda = 0$, $\alpha_0 = -\frac{3}{7}$. *This means that even the normal type is quite spiteful. If* α_0 *equals* $-\frac{3}{7}$, *the normal type proposer gets a utility of 3.43 when demanding 6 for himself, and 3.71 when demanding 7 for himself. This contradicts the incentive constraint 5.*

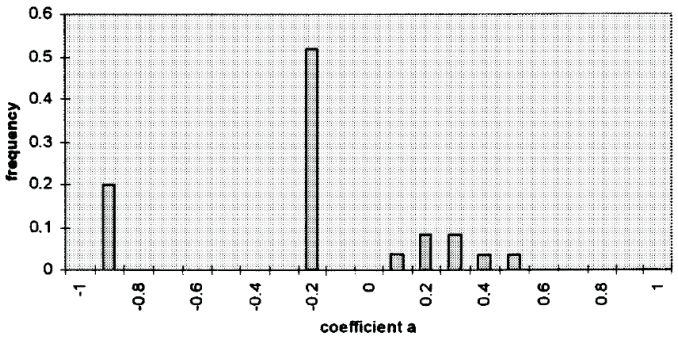
Furthermore, it can be shown that in a sequential equilibrium, $-0.301 \leq \alpha_0 \leq -0.095$, $-1 < \alpha^l < -0.73$ and $0.584 \geq \lambda \geq 0.222$. The proof relies on conditions (6) and (3) and some algebraic manipulation. For details see Levine (1998).

There are many possible sequential equilibria consistent with the data. There is most flexibility for α^h and less for α^l and λ. Probably many separating equilibria other than the one described exist and one can show that

there are two pooling equilibria - one at 7 and one at 8. The variety of possible equilibria raises the question about the predictive power of the model. Compared to the standard self-regarding theory the model proves worthwile. It is above all useful when thinking about games in which mixed strategy equilibria are observed. A selfish player can, at the Nash equilibrium, costlessly transfer money to or from his opponent because he is indifferent to doing so or not. A deviation from the Nash equilibrium depends on whether the marginal indifferent player is altruistic or spiteful. A spiteful marginal player wants to transfer money away from his opponent. Anticipating this, the opponent will adjust his strategy to keep the spiteful player indifferent. The result is that in a symmetric game equilibrium payoffs will be higher than they would be if players were purely selfish.

However, not for all games Levine's model can predict experimental outcomes. An example is the dictator game. Evidence shows that positive contributions are made though the model, because of linear utility and $\alpha_i < 1$, predicts no contributions at all.

Figure 1 shows the distribution of altruism coefficients that are consistent with the experimental data Levine (1998) used when calibrating his model. It is striking that 20% of the population have a very strong negative coefficient of - 0.9.



Levine (1998)

A study by Van Huyck et al. 1996 casts doubt on these altruism coefficients. In a public goods game experiment they found very little spitefulness and less altruism. The differences in the setting were that they implemented the game as one-shot and that they included four players. So it might be that the altruism coefficient is not independent of how many players there are. If spite is understood as a drive to increase competitiveness by harming opponents this might be plausible. It is also likely that the difference in altruism coefficients arises from the extensive form of the games Levine analyses. This would mean that retaliation for past spiteful behavior is not only due to the signaling of types.

4 Inequity aversion versus reciprocity models

4.1 Interaction of outcomes and intentions

When discussing mini-ultimatum games I came to the conclusion that models of inequity aversion fail to correctly predict different rejection rates for

identical offers. Distributive concerns are not or not solely able to deal with probelms connected with set dependence. Therefore another approach, reciprocity based models, were examined. It is left to discuss whether reciprocity models can better explain the behavior observed in mini-ultimatum games.

The purely intention based models of Rabin (1993) and Dufwenberg and Kirchsteiger (2004) can in principle account for different rejection rates across games. However, the reason for this is that they offer multiple equilibria. Rabin's model is compatible both with acceptance and rejection of the (8/2)-offer in all games. The same is true for the Dufwenberg and Kirchsteiger model in three of the four games. If the number of potential equilibria is too high, the use of these models as predictive tools is of course limited. Moreover, pure reciprocity models, where the only reason for rejecting an offer is the perceived unkindness of an intention, state that rejections cannot occur, if proposers cannot signal any intention. In the case of the (8/2)-game the proposer has no chance but to offer the responder 2. If there has been no choice for the proposer, the other player does not receive a signal about the proposer's intention. If only intentions matter, there should be no rejections in the (8/2)-game. However, empirically 18% of the responders reject the (8/2)-distribution even if proposers did not have a choice. (Falk, Fehr, and Fischbacher 2003)

The results of the mini-ultimatum game show that a fairness model that is exclusively based on either distributional preferences or on the attribution of kind or unkind intentions is incomplete. A study of Blount (1995)

also shows that both inferred intentions and outcomes determine people's behavior. Blount conducted a modified version of the ultimatum game, in which offers were determined randomly and consequently could not signal any intentions. The acceptance rate is found to be significantly higher in this non-intentional treatment, but even without signaling intentions very disadvantageous offers were sometimes rejected.

Another experiment by Falk, Fehr, and Fischbacher (2000), also points into the same direction. The game analysed is the so-called moonlighting game. This is a two-player, two-stage game with sequential moves. The initial endowment of both players is 12 units of money.

- At the first stage the first player, say player A, can choose an action $a\epsilon\{-6,-5,-4,...,4,5,6\}$.
  - If $a \geq 0$ player B is given $3a$ points and player A loses a points. If $a < 0$ player B loses $\mid a \mid$ points and player A gains $\mid a \mid$ points. If A chooses e.g. $a = -5$, she takes away 5 units of money from player B. If A chooses for example $a = 3$, she gives $3a = 9$ units of money to player B.
- After player B has observed a, she chooses an action $b\epsilon\{-6,-5,-4,...,16,17,18\}$.
  - If b takes on a positive value the action can be interpreted as a reward. Reversely a negative b can be seen as a punishment. Thus

the game allows for a distinction between positive and negative reciprocal reactions. If b is positive, player B gives away b points and A receives b points. If b is negative, B has to pay $\mid b \mid$points and A loses $3 \mid b \mid$ points. After B's decision the game ends and final payoffs are determined.

In the experiment player B had to indicate what her response would be for any possible action of player A before she was informed about the actual choice of A. The advantage of this procedure was that there was enough data to study the relevance of intentions at any level of a. Two treatments were conducted. In the intention treatment it was player A that determines her choice. She was responsible for the signal she sends to player B and for the consequences of her decision. In the no-intention treatment A was not free to choose her strategy. Her move was determined by a random device and therefore no intentions could be infered by the second player. Players were randomly assigned to the roles of player A and B.

Self-regarding preferences imply that the only subgame perfect equilibrium is that in both treatments B will always choose $b = 0$ because for any other choice she would have to give up some of her own payoff. Anticipating B's behavior, A chooses $a = -6$ in the intention treatment. In the no-intention treatment A has no influence on a. Inequity aversion models predict that sufficiently strong inequity averse players show behavioral patterns that look similar to reciprocity-induced behavior. That means b is

increasing in a and $b = 0$ if $a = 0$. Since the attribution of intentions is not assumed to have an influence on the outcome, for a given a responses in the intention treatment and the no-intention treatment must be exactly the same. The Dufwenberg and Kirchsteiger model of reciprocity assumes that reciprocal responses solely appear in reaction to infered intentions. Therefore no reciprocal behavior should appear in the no-intention treatment and b should be equal to zero. For the case of the intention treatment predictions are ambiguous. There are multiple equilibria. In some of them b is increasing in a and in others b is decreasing in a. The Falk and Fischbacher model (2006) which I will present in the following chapter combines the aspects of inequity aversion and reciprocity. For the intention treatment it predicts that there is a unique equilibrium where b is increasing in a. In the no-intention treatment b is also predicted to be increasing in a, but less so than in the intention treatment.

112 subjects participated in the experiment. Results clearly show that in the intention treatment B players exhibit strong reciprocal behavior. Average and median rewards are increasing in a and the more unkind player A, the higher is B's punishment. Comparing the no-intention treatment to the intention treatment significant differences can be observed. On average reciprocal responses, both positive and negative, are much weaker in the no-intention treatment. Only for sufficiently high or low a it is true that $b \neq 0$. However, reciprocity is statistically significant even in the no-intention treatment. For almost all $a \neq 0$ the difference between decisions of B players

between the two treatments is statistically significant at the 5%-level.

Given these results, the predictions of standard game theory with self-regarding preferences can be rejected since reciprocity is unambiguously observed. Also the predictions of the inequity aversion models do not hold because there is significant difference between behavior in the two treatments. Both positive and negative reciprocity are observed. The prediction of the Dufwenberg-Kirchsteiger model that there is no reciprocity in the no-intention treatment is also falsified.

Generally speaking, the models of Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Rabin (1993) and Dufwenberg and Kirchsteiger (2004) all lack either the intentional or the distributional aspect. Best performance can be expected from models that combine both motivations as those of Falk and Fischbacher (2006) and Charness and Rabin (2002) which I will briefly discuss in the following chapters.

4.2 Falk and Fischbacher's model of reciprocity

Falk and Fischbacher (2006) developed a model that accounts for both distributional and reciprocity motivations. The underlying assumption is that people evaluate the kindness of an action by its consequences as well as the actor's intentions. Fairness can be signaled if the actor's choice set allows the choice between a fair and an unfair action and if the actor's choice is under her full control. Falk and Fischbacher conducted a questionnaire study in order to find out how people evaluate the kindness of particular actions.

111 subjects at the University of Zurich were asked to imagine a bilateral exchange situation. Subjects had to indicate how kind or unkind they perceived different divisions of 10 Swiss Francs. They could express their judgements by assigning numbers between -100 and +100 to the proposed distributions. -100 points means the outcome is most unkind and +100 points means the outcome is kindest. The set of choices of the opponent was systematically varied and subjects were asked how they felt about different actions for each of the opponent's choice sets. Results of the study clarify that perceived kindness is monotonically increasing in the offer to the responder. It seems that an equitable share of the total payoff serves as a reference point to determine what is a fair or unfair offer. At the equitable offer of five Swiss Francs, the sign changes from minus to plus. All offers below five are judged as unkind and all offers above five as kind. When evaluating the intentions of a particular action, people take into account the actor's strategy set. If a choice set contains only one element, the perceived kindness or unkindness from this action is much weaker than for the same choice made if an alternative would have been available. However, even in case the actor cannot signal any intention because she has no alternative, the perceived kindness or unkindness of an action is not zero. This shows that also outcomes by themselves are important and reciprocity alone cannot explain the data. An action is judged as kind, if the actor could have made a less friendly offer. Likewise the perception of an unkind offer depends on whether the actor could have been more kind. In case there is an option to make a friendlier

offer, perceived intentions also depend on how much the actor would have to sacrifice in order to make the kinder choice. People often don't take an unfair offer amiss, if the alternative would be that the actor puts herself into a disadvantageous position. Generally speaking, fairness intention perception is not symmetric with respect to kindness and unkindness. For example, the kindness of giving 8 is on average judged with 62 points, whereas the unkindness of giving 2 is judged with -71.9 points. These empirical observations were used to formulate the Falk and Fischbacher model.

The model is applicable to sequential games and tries to capture reciprocity in two parts: kind or unkind behavior is captured by a kindness term φ and the behavioral reaction to that behavior is represented by a reciprocation term σ. The game that is produced is called a "reciprocity game", which belongs to the class of psychological games. The Falk-Fischbacher model requires extensive use of notation. In the following I will restrain myself to only giving a short overview of the model.

Players' utilities depend on monetary payoffs and on the so-called reciprocity utility. Formally, player i's utility function in the reciprocity game is given by

$$U_i(f, s_i'', s_i') = \pi_i(f) + \rho_i \sum_{n \to f, n \epsilon N_i} \varphi_j(n, s_i'', s_i') \cdot \sigma_i(n, f, s_i'', s_i')$$

where $\pi_i(f)$ is the monetary payoff and f is an end node that follows node n. s_i is the strategy of player i. s_i' is the first-order belief of player i, i.e.

her belief about the strategy player j will choose and s_i'' is player i's second-order belief, i.e. her belief about player j's belief about which strategy i will choose. The reciprocity utility consists of the kindness term φ_j, the reciprocation term σ_i and the reciprocity parameter ρ_i.

The kindness term φ measures how kind a person judges the action by another person. It consists of an outcome term that measures whether an outcome is advantageous or disadvantageous and of an intention factor which captures the degree of intentionality of the opponent's action. Player i compares her payoff with a reference standard. If i thinks that j wants more for himself than for i, she feels that j is acting unkindly. On the other hand, if j wants less for himself than for i, i thinks she is treated kindly. The reciprocation term σ_i measures the response to perceived kindness, i.e. how much i alters j's payoff in reaction to him at node n. The reciprocity parameter ρ_i is positive, constant and common knowledge. It measures how much player i values reciprocity utility relative to utility gained from the material payoff. If ρ_i is equal to zero, player i cares only for monetary payoff. If ρ_j is also zero and if there are only two players, the reciprocity game is reduced to a standard game. Kindness is measured in each node. When in a node n player i feels that her opponent is being kind, she can increase her utility by increasing j's payoff, given that ρ_i is positive. The overall reciprocity utility is the sum of the reciprocity utility in each node weighted with ρ.

A reciprocity equilibrium is a subgame perfect psychological Nash equilibrium. It does not always exist. In a reciprocity equilibrium utilities depend

on strategies and beliefs, where beliefs are taken as given. In an equilibrium all beliefs have to match actual behavior. Though the Falk and Fischbacher model considers sequential games, only initial believes enter the utility. The problem of the sequential structure is overcome by updating in the outcome and in the reciprocation term. Utility components are defined in each node and beliefs about actions, which are not part of the current subgame, are irrelevant for determining those components. This approach differs from the method applied in the Dufwenberg and Kirchsteiger model, i.e. maximizing of utility in each node using updated beliefs.

The Falk and Fischbacher model can be applied to various games, e.g. the ultimatum game, the gift-exchange game, reduced best-shot games, market games, the dictator game, the sequential prisoner's dilemma or the centipede game. For a detailed discussion of the model and its applications see Falk and Fischbacher (2006).

There are two major differences between the model of Falk and Fischbacher (2006) and those of Rabin (1993) and Dufwenberg and Kirchsteiger (2004). Falk and Fischbacher assume that people care for a combination of outcomes and underlying intentions, whereas the other two models adopt a purely intention driven approach to reciprocity. The other - probably less obvious - difference is the way how kindness is supposed to be evaluated by the players. The Falk and Fischbacher concept of fairness is based on payoff comparisons among players. In order to assess the kindness of a move, player j compares his resulting payoff with that of his opponents. On the

other hand, Rabin as well as Dufwenberg and Kirchsteiger do not model players' considerations about whether they have more or less than others. They assume that players compare the outcome actually chosen with the alternative actions an opponent could have chosen. Player i is supposed to be unfair if she could have taken an alternative efficient action yielding a higher payoff for player j. So kindness in the Falk and Fischbacher model is based on interpersonal comparison and in the Rabin and Dufwenberg and Kirchsteiger models it is based on comparison of alternative actions.

4.3 Charness and Rabin's model of reciprocity

Charness and Rabin (2002) examine social motivations based on 29 different games, with 467 participants making 1697 decisions. They aim to eliminate confounds within games by testing a fuller range of possible departures from self-interest. This allows to test existing theories more directly than the games commonly studied. The aim is to formulate a new model that captures patterns of behavior that previous models could not explain.

Consider the following simple model of social preferences in two-person games: π_A and π_B are the monetary payoffs of players A and B. Player B's utility function is given by

$$U_B(\pi_A, \pi_B) = (\rho \cdot r + \sigma \cdot s + \theta \cdot q) \cdot \pi_A + (1 - \rho \cdot r - \sigma \cdot s - \theta \cdot q) \cdot \pi_B$$

where $r = 1$ if $\pi_B > \pi_A$, and $r = 0$ otherwise; $s = 1$ if $\pi_B < \pi_A$, and $s = 0$

otherwise; $q = -1$ if A has been unkind, and $q = 0$ otherwise. In words, B's utility is a weighted sum of his own monetary payoff and his opponent's payoff. The weight B places on A's payoff depends on whether A has a higher or lower material payoff than B and on whether A has behaved kindly or not. Another way of writing the utility function is to distinguish two cases: when

$$\pi_B \geq \pi_A : U_B(\pi_A, \pi_B) = (1 - \rho - \theta q) \cdot \pi_B + (\rho + \theta q) \cdot \pi_A$$

and when

$$\pi_B \leq \pi_A : U_B(\pi_A, \pi_B) = (1 - \sigma - \theta q) \cdot \pi_B + (\sigma + \theta q)\pi_A$$

Different parameter ranges of the model incorporate different existing theories of social preferences. θ is a mechanism for modeling reciprocity. The parameters ρ and σ represent distributional concerns.

Distributional preferences can take the form of simple competitive preferences. If $\sigma \leq \rho \leq 0$, player B cares not only directly about his own payoff, but also strives to be as well off as possible in comparison to his opponent. A competitive player likes his or her payoff to be as high as possible relatively to the payoffs of others. The theory of inequality aversion assumes the opposite. People try to minimize differences between their own monetary payoffs and their opponents' payoffs if $\sigma < 0 < \rho < 1$. If $1 \geq \rho \geq \sigma > 0$, subjects exhibit social-welfare preferences. They always prefer more for themselves and for others, but give more weight to their own payoffs when they get less than

other players than when they get more. In the two-player case social-welfare preferences are related to the idea that players want to help all players and particularly care about the person who is worst off.

Reciprocity implies that player B's values of ρ and σ vary with B's perception of A's intentions. If players have preferences as supposed in the model of Falk and Fischbacher, the parameters should be $\sigma < 0 < \rho < 1$ if they feel positive or neutral toward another person. If A's behavior makes B think that A puts a weight $\rho \leq 0$ to his well-being, B's values for ρ and σ diminish. B retaliates against A's behavior if this is due to a small ρ but not if it is due to a small or negative σ. This means B does not want to harm A, if A is selfish when being worse off. However, B wishes to hurt A when A is selfish even when she is better off than B. Reciprocity can easily be captured by assuming that $\theta > 0$ when $q = -1$. If A has behaved unfairly, i.e. not in line with social-welfare preferences, B lowers both ρ and σ by amount θ.

The empirical data underlying the model was gained from a series of experiments. In each game, one or two participants made decisions that affected the payoffs of two or three players. Seven dictator games isolated distributional preferences. Variations of the game allowed players to sacrifice their own payoff in order to decrease inequality or give up their own payoff to increase inequality and efficiency. Also the case where players can affect inequality at no cost to themselves was examined. These empirical results allow for a characterization of ρ and σ.

Twenty response games were played. They opened a wide range of options

for both players. In some of the games entry by A hurted B and in others entry of A helped B. There were games in which this help or harm was compatible with inequality aversion or social-welfare preferences, and there were others in which it was not. These response games were used to test both for inequality aversion and reciprocity motivation. In order to facilitate inferences about reciprocity based behavior, there were many sets of games where B's choices are identical, but A's prior choice is varied.

The results of the dictator games show that the most effective type of preferences are social-welfare preferences, as long as care for reciprocity is neglected. Roughly 70% of the decisions are in line with social-welfare preferences. 20% are consistent with inequality aversion and 10% with competitiveness. Similar results were obtained by Charness and Grosskopf (2001). These proportions show how the models perform in explaining all behavior. However, it may be more insightful to test how well models match when they make unique predictions. It can again be shown that social-welfare preferences outperform both inequality aversion and competitive preferences in dictator games as well as in response games. (Charness and Rabin 2002)

An important driving force of behavior is supposed to be reciprocity. Reciprocal behavior can be observed in response games. Indeed, in these games the motivation for inequality aversion was overruled by positive reciprocity. Furthermore, the results of the ultimatum games could be much better explained by negative reciprocity than by inequality aversion. It can be concluded that inequality aversion is not a strong factor when in conflict

with other social motives. "Our results suggest that the apparent adequacy of recent difference-aversion models has likely been an artifact of powerful and decisive confounds in the games used to construct these models." (Charness and Rabin 2002)

Summarizing, it can be said that reciprocity considerations are an important component of behavior. When A hurts B, B is more likely to harm A than otherwise and less likely to sacrifice in order to help A.

In addition to the two-player games, also three-player games experiments were conducted. In multiperson models the question how players judge changes in the distribution among other players' payoffs is discussed. It is assumed that players like to improve the payoffs of everybody, but even more so when their oppponents' payoffs are lower than their own payoff. In simplified and extreme form this is equal to minimax preferences, where players like to maximize the minimum payoff among players. Empirically, people seem to care about both, efficiency and minimum payoffs. Many subjects chose to increase total payoff at the expense of the minimum payoff, while others were more willing to maximize the minimum payoff. Evidence is found against the Bolton and Ockenfels model of inequality aversion which assumes that players are concerned with the average payoff of all other players and not with the distribution of those payoffs. This could not be approved by the experimental data.

An extension of the simple model of social preferences as described before, integrates social-welfare preferences with reciprocity into a multiperson

model. It uses the framework of psychological Nash equilibria for implementing reciprocity. For a discussion of the more complex and general model of social preferences see Charness and Rabin (2002).

The model of Charness and Rabin (2002) does not incorporate the principle of sequential rationality as discussed in the models of Dufwenberg and Kirchsteiger (2004), and Falk and Fischbacher (2006). It can capture much of the experimental data but it does so at the cost of a high complexity and many parameters. The reciprocity part is hard to analyze for a particular game. Similar to the Levine and Dufwenberg and Kirchsteiger models the problem of multiple equilibria arises, which limits its use as a predictive tool.

5 Conclusion

Insights into the nature of nonself-interested behavior give an important impulse to economic theory.

This paper examined models of inequity aversion, models of reciprocity and finally models that combine both aspects. It remains to be reasoned which of the different models performs best in explaining and forecasting economic behavior. First of all, one has to be careful to be clear about what a "good performance" of a model is. On the one hand, a good model should capture psychological reality as well as possible, but on the other hand a good model should also be tractable and broadly applicable.

The concept of reciprocity is intuitive and models based on this prin-

ciple do well in explaining many games. Rabin (1993) models kindness as solely determined by intentions. His model is only applicable to two-person normal form games but found an extension in the work of Dufwenberg and Kirchsteiger (2004) who set up a model of sequential reciprocity. A different approach was taken by Levine (1998). In his model players do not punish or reward fair or unfair actions but kind or unkind types. A limitation of these models is that there are multiple equilibria possible and they are often difficult to find.

Inequality aversion models have an easier structure because they do not explicitly deal with reciprocity motives. However, this does not imply that inequity aversion models are not aware of the potential impact of perceived intentions. Inequality aversion might work as a blackbox for more complex preferences over outcomes and intentions. "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." (Fehr and Schmidt 1999) The best performance in explaining experimental data yield models that combine and explicitly model the interaction of outcomes and intentions, i.e. the Falk and Fischbacher model (2006) and the Charness and Rabin model (2002). However, these models

have a highly complicated structure and use many parameters. This makes it hard to apply them to particular games.

All models have their advantages and disadvantages. A higher predictive power and more psychological accuracy come at the price of reduced tractability. Therefore, it is a researcher's purpose that will determine which model to use.

6 References

Blount, Sally. 1995. When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences. *Organizational Behavior & Human Decision Processes* 63(2), pp. 131-144

Bolton, Gary E, and Axel Ockenfels. 2000. A Theory of Equity, Reciprocity and Competition. *American Economic Review* 100, pp. 166-193

Brandts, Jordi, and Carles Sola. 2001. Reference Points and Negative Reciprocity in Simple Sequential Games. *Games and Economic Behavior* 36, pp. 138-157

Camerer, Colin F. 2003. Behavioral Game Theory. *Princeton University Press*

Camerer, Colin, and Richard Thaler. 1995. Ultimatums, Dictators, and Manners. *Journal of Economic Perspectives* 9(2), pp. 209-219

Charness, Gary, and Matthew Rabin. 2002. Understanding Social Preferences with Simple Tests. *Quarterly Journal of Economics* 117, pp. 817-869

Charness, Gary, and Brit Grosskopf. 2001. Happiness and Relative Payoffs - An Experimental Study. *Journal of Economic Behavior and Organisation* 45(3), pp. 301-328

Clark, Andrew E., and Andrew J. Oswald. 1996. Satisfaction and Comparison Income. *Journal of Public Economics* 61(3), pp. 359-381

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. A Theory of Sequential Reciprocity. *Games and Economic Behavior* 47, pp. 268-298

Engelmann, Dirk, and Martin Strobel. 2004. Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments. *American Economic Review* 94(4), pp. 857-869

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2000. Testing Theories of Fairness - Intentions Matter. *Working Paper No.63. Institute for Empirical Research in Economics, University of Zürich*

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2003. On the Nature of Fair Behaviour. *Economic Inquiry* 41, pp. 20-26

Falk, Armin, and Urs Fischbacher. 2006. A theory of reciprocity. *Games and Economic Behavior* 54, pp. 293-315

Fehr, Ernst, and Klaus M. Schmidt. 1999. A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics* 114(3), pp. 817-868

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl. 1998. Gift exchange and reciprocity in competitive experimental markets. *European Economic Review* 42(1), pp. 1-34

Fehr, Ernst and Simon Gächter. 1998. Reciprocity and Economics, The Economic Implications of Homo Reciprocans. *European Economic Review* 42(3), pp. 845-859

Fehr, Ernst, Michael Naef, and Klaus M. Schmidt. 2006. Inequality Aversion, Efficiency and Maximin Preferences in Simple Distribution Experiments: Comment. *American Economic Review* 96(5), pp. 1912-1917

Forsythe, Robert, Joel L. Horowitz, N.E. Savin, and Martin Sefton. 1994. Fairness in simple bargaining experiments. *Games and Economic Behavior* 6(3), pp. 347-369

Gintis, Herbert, Samuel Bowles, Robert Boyd, and Ernst Fehr. 2005. Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life. *The MIT Press*

Güth, Werner, Steffen Huck, and Wieland Müller. 2001. The Relevance of Equal Splits in Ultimatum Games. *Games and Economic Behavior* 37(1), pp. 161-169

Hoffmann, Elizabeth, Kevin McCabe, and Vernon L. Smith. 1996. Social distance and other-regarding behaviour in dictator games. *American Economic Review* 86, pp. 653-660

Kagel, John, and Katherine Wolfe. 2001. Tests of Fairness Models based on Equity Considerations in a Three Person Ultimatum Game. *Experimental Economics* 4, pp. 203-220

Kagel, John, Chung Kim, and Donald Moser. 1996. Fairness in Ultimatum Games with Asymmetric Information and Asymmetric Payoffs. *Games and Economic Behavior* 13(1), pp. 100-110

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler. 1986. Fairness as a Constraint on Profit Seeking: Entitlements in the Market. *American Economic Review* 76(4), pp. 728-741

Levine, David. 1998. Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics* 1, pp. 593-622

Nelson, William Robert Jr. 2002. Equity or intention: it is the thought that counts. *Journal of Economic Behavior & Organization* 48, pp. 423-430

Offerman, Theo. 2002. Hurting hurts more than helping helps. *European Economic Review* 46(8), pp. 1423-1437

Rabin, Matthew. 1993. Incorporating Fairness into Game Theory and Economics. *American Economic Review* 83(5), pp. 1281-1302

Rabin, Matthew. 1998. Psychology and Economics. *Journal of Economic Literature* 36(1), pp. 11-46

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir. 1991. Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study. *American Economic Review* 81. pp. 1068-95

Sandbu, Martin Eiliv. 2007. Fairness and the roads not taken: An experimental test of non-reciprocal set-dependence in distributive preferences. *Games and Economic Behavior* 61, pp. 113-130

Schotter, Andrew, Avi Weiss, and Inigo Zapater. 1996. Fairness and survival in ultimatum and dictatorship games. *Journal of Economic Behavior & Organization* 31, pp. 37-56

Sen, Amartya. 1994. The Formulation of Rational Choice. *American Economic Review* 84(2), pp. 385-390

Suleiman, Ramzi. 1996. Expectations and fairness in a modified Ultimatum game. *Journal of Economic Psychology* 17, pp. 531-554

Van Dijk, Eric, and Riel Vermunt. 2000. Strategy and Fairness in Social Decision Making: Sometimes It Pays to Be Powerless. *Journal of Experimental Social Psychology* 36, pp. 1-25

Van Huyck, J., R. Battalino, and F. Rankin. 1996. On the Evolution of Convention: Evidence from Coordination Games. *Texas A&M*

7 Appendix

7.1 English Abstract

Self-regarding preferences cannot always account for empirically observed behavior. Concerns for fairness seem to affect how people behave e.g. in bargaining games. This paper focuses on the ultimatum game and examines models of reciprocity, inequality aversion and models that combine intentional and distributional aspects. The Fehr and Schmidt model of inequality aversion (1999) is more accurate than the Bolton and Ockenfels model (2000), because it is often in line with minimax preferences. However, the concept of inequality aversion does not capture the whole story since it neglects the importance of perceived intentions. Reciprocity based models explicitly model people's reactions to infered intentions. Doing so is more intuitive from a psychological point of view, but on the other hand reduces the simplicity of the models and allows for multiple equilibria. Experimental testing shows that even (pure) reciprocity models cannot explain all the behavior empirically observed. A more complete model must combine distributional concerns and reciprocity motivation, as found in the Falk and Fischbacher model (2006) or the Charness and Rabin model (2002). They are most convincing in the context of completeness and psychological plausibility, but due to their highly complex structure they are hardly applicable to concrete game-theoretic situations.

7.2 German Abstract

In einer Vielzahl von Experimenten erwies sich, dass eine signifikante Zahl von Entscheidungsträgern sich nicht wie der üblicherweise angenommene Homo Oeconomicus verhält. Diese Arbeit untersucht, vor allem anhand des Ultimatum Spiels, verschiedene Modelle sozialer Präferenzen: Ungleichheitsaversion, Reziprozität sowie Modelle, die beide Prinzipien vereinen. "Inequality Averison" Modelle haben eine simple Struktur und sind leicht anzuwenden. Das Fehr-Schmidt Modell (1999) ist dem Bolton-Ockenfels Modell (2000) meist vorzuziehen. Beide können aber nicht alle beobachteten Phänomene erklären, weil sie die Rolle von Intentionen vernachlässigen. Modelle, die auf Reziprozität basieren, modellieren explizit wie Menschen auf wahrgenommene Intentionen reagieren. Dies ist, von einem psychologischen Standpunkt betrachtet, schlüssiger, birgt aber den Nachteil größerer Komplexität der Modelle und multipler Gleichgewichte. Doch auch reine Reziprozitäts-Modelle können das in Experimenten beobachtete Verhalten nicht vollständig erklären. Die höchste Erklärungsrelevanz erzielen Modelle, die sowohl distributive Aspekte als auch aus Aktionen abgeleitete Intentionen miteinbeziehen. Beispiele dafür sind das Falk und Fischbacher Modell (2006) sowie das Charness und Rabin Modell (2002). Beide überzeugen durch größere Vollständigkeit und psychologische Schlüssigkeit, sind aber ob ihrer hohen Komplexität in konkreten Spielen schwer anwendbar.

7.3 Curriculum Vitae

ANGABEN ZUR PERSON

Name: Nadia Steiner

Adresse: Neblingergasse 6/2, 1130 Wien

E-Mail: nadia.steiner@gmx.at

Staatsbürgerschaft: Österreich

Geburtsdatum: 20.07.1986

SCHULISCHE UND UNIVERSITÄRE AUSBILDUNG

1992-1996:

Volksschule 1140 Wien, Karl Toldt Weg 12

1996-2000:

GRG 15, 1150 Wien, Auf der Schmelz 4: Gymnasium mit Französisch als 2. Fremdsprache

2000-2004:

Sir Karl Popper Schule, 1040 Wien, Wiedner Gürtel 68: neusprachliches Gymnasium mit Englisch, Französisch, Latein und Spanisch; Begabungsförderungsmodell Sir Karl Popper Schule; Matura mit ausgezeichnetem Erfolg

2004-2006:

Universität Wien: Studien der Politikwissenschaft, Französisch und Spanisch

2004-2008:

Universität Wien: Studium der Volkswirtschaftslehre, Leistungsstipendium der Universität Wien