Distributional Consequences and Intentions in a Model of Reciprocity

Armin Falk and Urs Fischbacher*
UNIVERSITY OF ZURICH
Institute for Empirical Economic Research
Bluemlisalpstrasse 10, CH-8006 Zürich
falk@iew.unizh.ch, fiba@iew.unizh.ch

December 18, 2000

Abstract

Two types of theoretic approaches have been developed to model fair behavior. According to the 'equity approach' fairness is solely driven by distributional concerns. Opposed to this consequentialistic perspective, the 'reciprocity approach' stresses the importance of intentions. We discuss experimental evidence which demonstrates that in fact both intentions and distributional consequences matter. Neither the models that are based solely on outcomes nor those exclusively based on intentions can explain this evidence. In this paper, we present a formal model of reciprocity which captures both, outcomes and intentions. This model is compatible with many of the observed experimental findings.

*Financial support by the MacArthur Foundation (Network on Economic Environments and the Evolution of Individual Preferences and Social Norms) is gratefully acknowledged. We would like to thank Sam Bowles, Simon Gächter, Herbert Gintis and the participants of the conference on Social Interactions and Economic Behavior" held in Paris 1999.

1 Outcomes and Intentions

Economics usually takes a consequentialistic perspective. Utility functions, for example, are defined solely on outcomes and consequences. It is implicitly assumed that subjects' utility depends only on the outcome per se and not on how a particular outcome came about. This perspective, though appropriate in many contexts, is misleading when it comes to the judgement of kindness, fairness or justice. As a prime example take the criminal law. It distinguishes carefully between criminal activities that are committed negligently and those committed with criminal intent. In order to find an appropriate judgement and punishment it advises a judge to take both into consideration: the consequences of a (criminal) activity and the action's underlying motivation.

This holds more generally: According to a commonly held sense of justice, people evaluate the kindness or unkindness of a treatment not only by the consequences of a particular treatment but also by the intention that guided the treatment. Put differently, "people determine their dispositions toward others according to motives attributed to these others, not solely according to actions taken" (Rabin (1998), p.22). Of course, attributional issues may have behavioral consequences. People's willingness to reciprocate, e.g., may very well vary with the motives attributed to others. Since reciprocity is the behavioral response to a perceived kindness or unkindness, intentions may play a key role on the strength of the reciprocal action.

A large body of evidence indicates that reciprocity is a powerful description of human behavior. The results of numerous ultimatum games, for example, have provided evidence that, contrary to their immediate self-interest, subjects are willing to sacrifice substantial amounts of money in order to punish people who have treated them in an unkind fashion (negative reciprocity). Similarly, many people show an inclination to return a favor or a gift, i.e., to reward kindness (positive reciprocity). It is not the main purpose of the present paper, however, to highlight the fact that people exhibit reciprocal behavior. Instead, we focus on whether the intensity of people's reciprocal behavior is shaped by the motives attributed to their opponents.

In the recently developed game theoretic models that account for reciprocal behavior observed in the laboratory, two quite distinct points of view have evolved: According to the 'equity-approach' reciprocity is solely driven by distributional inequality (compare, e.g., Bolton and Ockenfels (1998), Fehr and Schmidt (1999), Bolton (1991)). According to these models, reciprocal actions are triggered solely by the fact that people are inequity averse. In an ultimatum game, e.g., these models predict that a responder who is offered let's say 20 percent of the pie punishes the proposer (who keeps 80 percent for himself) only because the responder is exposed to a very disadvantageous share of the pie. Importantly, the rejection behavior is completely independent of the proposer's motives. Even if the proposer is *forced* to make the 20 percent offer (maybe because this offer was the only feasible offer) the predicted rejection rate remains unchanged.

Contrary to the 'inequity approach', the 'reciprocity approach' models reciprocity

as a response to a perceived kindness or unkindness (compare, e.g., Rabin (1993), Falk and Fischbacher (1999), Bowles and Gintis (1998), Dufwenberg and Kirchsteiger (1999) and Charness and Rabin (1999)). The main difference of these models to the inequity aversion models stems from the fact that perceived kindness depends crucially on the underlying intention. In the models of Rabin (1993) and Dufwenberg and Kirchsteiger (1999) reciprocity is *solely* driven by intentions. This implies that on should observe reciprocal actions only in situations where intentions are involved. For example, in a situation where a player has no alternatives to choose from, these models predict behavior that coincides with that derived with the standard selfishness assumptions. According to the models by Falk and Fischbacher (1999), Charness and Rabin (1999) and Levine (1998) reciprocity is driven by intentions *and* distributional consequences.

In this paper, we will discuss the Falk and Fischbacher model in some detail. According to their model, reciprocal subjects may perceive the same outcome very differently. Offering a 20 percent share of the pie in the 'regular' ultimatum game clearly signals bad or greedy intentions. After all, the proposer could have chosen a less unkind offer. However, if the proposer does not have the opportunity to propose a less unkind offer (maybe because he has no alternative actions available) choosing the offer does not signal any (bad) intentions. The perceived unkindness will be much lower compared to the regular ultimatum game and the reciprocal response will be weaker, too. As a consequence, subjects will respond differently to the same outcome.

The rest of the paper is organized as follows. In the next section we review some experimental findings on reciprocity and intentions. We show that in these experiments, intentions clearly matter. However, even in the absence of intentions some subjects exhibit reciprocal actions which indicates that it is not intentions alone that are responsible for the observed reciprocity. Section 3 outlines a formal model developed in Falk and Fischbacher (1999). This model takes up the idea that both, intentions and distributional aspects are relevant in the fairness context. In Section 4 we shortly present some applications of our model. As we will show, the model has more explanatory power than models which exclusively focus on outcomes *or* on intentions. Section 5 concludes.

2 Experimental evidence

2.1 Four mini ultimatum games

The first experiment we will discuss is a study reported by Falk, Fehr and Fischbacher (1999). In this experiment subjects participated in four different games. All of these mini ultimatum games are extremely simple and share the same structure (see Figures 1a -1d). In all games the proposer P is asked to divide 10 points between himself and the second mover. The responder R can either accept a or reject r the offer. Accepting the offer leads to a payoff distribution according to the proposer's offer. A rejection implies zero profits for both players.

The proposer can choose between two allocations, x and y (see Figures 1a - 1d). The choice of x leads to the same allocation in each of the games. If the proposer

chooses x and the responder accepts this offer the proposer gets 8 points while the responder gets 2 points. In addition to this (8,2)-offer, there is always an alternative offer at the disposal of the proposer (denoted by y in Figures 1a - 1d). These alternative offers distinguish the games from each other: In the first game, the alternative offer is the offer (5,5). The game is therefore called the (5,5)-game. The second game is called the (2,8)-game for in this game the alternative move is to keep 2 points and to offer 8 points to the responder. In the third game the proposer has no alternative at all. He has only a formal choice between two (8,2)-offers. We call it the (8,2)-game. Finally, in our fourth game the alternative offer is (10,0). This game is called the (10,0)-game.

Figure 1

Thus, in this experimental set-up, one alternative is kept constant (8,2) while the other is varied. This allows to study directly the impact of different alternatives on the rejection rate of the (8,2) offer. Why is this a test for the role of intentions? The point is that in order to infer intentions, people take into consideration the alternatives that are at their opponent's disposal. Thus, depending on the alternatives, people possibly infer different intentions from the same offer.

In the following we will concentrate on the rejection behavior that follows an offer of (8,2). Applying the standard assumptions (selfish preferences and rationality) this offer should always be accepted in a subgame perfect equilibrium. Notice, however, that this offer is a very 'unkind' offer in the sense that it implies an extremely uneven share of the pie. From the experimental literature we know that in a 'regular' ultimatum game such an offer is - contrary to material self-interest - very often rejected (compare, e.g., Roth (1995) and Slonim and Roth (1998)). Therefore, we expect that in our games there is also a non-negligible fraction of subjects that will in fact reject this offer.

As outlined in the introduction there are two competing hypotheses with respect to the influence of intentions on the incidence of reciprocity. The first hypothesis claims that reciprocity is purely driven by outcomes. According to this perspective, people have preferences over distributions and will reject a particular offer if it gives rise to a very unequal share of the pie. This view is advocated in the 'equity-models' of Bolton and Ockenfels (1998) and Fehr and Schmidt (1999). Since in their models the rejection behavior of a person is triggered solely by distributional considerations they predict that in all four games the rejection rate of the (8,2)-offer is exactly the same.

The competing view claims that reciprocity is the behavioral response to a kind or unkind treatment. Evidently, the outcome of the (8,2)-offer is the same in all four games. Intentions differ though. In the (5,5)-game, offering 20 percent of the pie signals bad intentions since the proposer could have achieved an equal split by choosing the (5,5)-offer. In the (2,8)-game, choosing (8,2) leads to the same disadvantageous distribution for the responder. However, if the proposer chooses (2,8) he brings *himself* in a very disadvantageous situation. Therefore, the responder cannot expect the proposer to choose this offer. Given the proposer has only the unreasonable alternative (2,8), the responder will consider the (8,2)-offer as less unfair compared to

the game where there is the (5,5)-alternative. Accordingly, the rejection rate of the (8,2)-offer is lower. In the (8,2)-game the proposer has no choice at all. Obviously, a responder cannot infer any (bad) intentions from the 'choice' of (8,2). As a consequence, the responder will find it even less unkind if he gets the offer (8,2) compared to the (2,8)-game. Thus, he will reject this offer at a lower rate. Finally, we turn to the (10,0)-game. In this game, offering (8,2) actually signals 'good' intentions since the proposer chooses the less unkind of two unkind alternatives.

In summary, the 'reciprocity-approach' claims that offering (8,2) signals very bad intentions in the (5,5)-game, less bad intentions in the (2,8)-game, no intentions in the (8,2)-game and possibly good intentions in the (10,0)-game. Given that intentions matter, rejection rates should look accordingly. Thus, the 'reciprocity-approach' predicts that the rejection rate of the offer (8,2) is different in all games. It is larger in the (5,5)-game than in the (2,8)-game, larger in the (2,8)-game than in the (8,2)-game and larger in the (8,2)-game than in the (10,0)-game.

Figure 2 shows the observed rejection rates of the (8,2)-offer across all four games. The rejection rate in the (5,5)-game is highest (44 percent) and declines across the other games. 27 percent rejected the (8,2)-offer in the (2,8)-game, 18 percent in the (8,2)-game and 9 percent in the (10,0)-game. Thus, the 'reciprocity-approach' is supported by the data whereas the pure outcome oriented 'equity-approach' is clearly rejected by the data. In their decision to reject an offer, i.e., in their reciprocal behavior, subjects clearly differentiate in a systematic way whether the offer (8,2) signals bad intentions or not. We interpret these findings as support for the claim that intentions play an important role in the psychology of reciprocity. Notice, however, that even in the (8,2)- and the (10,0)-game where intentions are absent or even good some people nevertheless reject distributionally unfair offers. Thus: Outcomes *and* intentions matter.

Figure 2

2.2 Further examples

In this section we briefly report additional evidence in favor of the importance of intentions. The first example is Blount (1995) who compared a 'regular' *ultimatum game*[1] with an ultimatum game in which the proposer's choice was determined by a random device. In this 'random treatment' a low 'offer' does not signal any (bad) intention. According to the 'equity-approach' this should not matter for the rejection behavior since the payoff consequences of a random offer are exactly the same as the consequences of an offer made by a human being. Blount finds, however, that

[1]The ultimatum game is a two person sequential move game. The first mover ("proposer") is allocated an amount of money M, say. He has to divide this amount between himself and a second mover (called "responder"). The proposer may offer any feasible amount c to the responder, i.e., $0 \leq c \leq M$. After the offer is revealed to the responder, the latter decides to either accept or reject the offer. If she accepts, the resulting payoffs amount to $M - c$ for the proposer and c for the responder. If the responder rejects the offer, payoffs are zero for both parties. In the subgame perfect Nash equilibrium the proposer offers the lowest share c, the responder is just willing to accept. If M can continuously be devided, the only subgame perfect equilibrium outcome is ($c = 0$; accept).

rejection rates - for a given offer - are significantly lower compared to the 'regular' ultimatum game.

A similar experiment that was conducted in the domain of positive reciprocity is reported by Charness (1996). He conducted two experimental versions of the *gift-exchange game* (Fehr, Kirchsteiger and Riedl (1993)). In this game a firm offers a worker a wage. Upon acceptance the worker has to choose a costly effort level. Since effort cannot be specified in advance, the gift-exchange game captures an incomplete labor relation. Assuming selfish preferences, the worker will always choose the lowest, i.e., the cost minimal, effort level. With backward induction the firm will find it optimal to choose the lowest possible wage. This standard prediction is refuted by the experimental data. Effort levels turn out to be highly correlated with wages and both wage and effort levels are well above the level predicted by standard assumptions (see also Fehr and Falk (1999)). Charness replicated this finding in a treatment where human beings submitted the wage offers. In a second treatment, he introduced randomness similar to Blount's experiment. This 'random treatment' is characterized by the fact that firms can no longer make a wage offer themselves. Instead a random device or a third party determines the wage offer. In the latter treatments, a high wage does not signal (good) intentions. Therefore, the 'reciprocity-approach' predicts a weaker correlation between wages and efforts, i.e., less reciprocal behavior. This prediction is supported by the data.

Since in the non-intentional treatments of Blount and Charness first movers cannot choose anything, these treatments are similar to our (8,2)-game reported in the previous section. In all of these treatments, intentions are absent. However, we still observe reciprocity-like behavior: In our (8,2)-game rejection rates are very low, but not zero. In Blount's random treatment extremely low offers were rejected and also in Charness' experiment the positive correlation between wages and efforts is not wiped out. Thus, it seems that both intentions *and* outcomes do affect reciprocal choices. This conclusion is supported also supported by Falk, Fehr and Fischbacher (2000) and Offerman (1999).

Additional evidence for the shortcomings of consequentialistic models of human behavior are found in the social psychology literature. Goranson and Berkowitz (1966), for example, analyze the intensity of positive reciprocity as a response to received prior help. Prior help was specified in three forms, 'voluntary', 'compulsory' and 'refused'. Reciprocation was significantly higher when prior help was provided voluntarily compared to the compulsory condition. While voluntary help signals good intentions and is strongly reciprocated, instructed help does not signal good intentions and is therefore reciprocated less. This finding has been confirmed in many other psychological experiments, e.g., in Frisch and Greenberg (1968), Lerner and Lichtman (1968) and Hornstein, Fisch and Holmes (1968).

Other papers are closely connected to the discussion on intentions. In their mini-ultimatum games Güth, Huck and Müller (1998) find that removing the opportunity to offer an equal split produces a remarkable 'behavioral discontinuity' in the rejection behavior. Even though the authors do not directly address the issue of intentions one can interpret their findings in the light of the above discussion. Similar to our results they find that responders reject a given unfair offer less often when all al-

locations imply a payoff advantage for the proposer. Brandts and Sola (1998) lend support to the idea that in evaluating the fairness of an offer, alternatives matter. In their reduced ultimatum games, one possible offer was kept constant across the games while the alternative offer varied. They report substantial differences in rejection of the same offer depending on the available alternative. Another interesting study by McCabe and Smith (1997) investigates why subjects, contrary to the game theoretic assumption, act differently in games presented in normal form compared to (strategically equivalent) games that are presented in extensive form. One difference concerns reciprocity which seems to be stronger in games presented in extensive form. As an explanation for this finding, McCabe and Smith point to the importance of intentions: In normal form games reciprocity is less likely to occur because intentionality detection is more difficult in normal form games compared to extensive form games.

The only empirical evidence that sheds doubt on the 'reciprocity-approach' is reported in Bolton, Brandts and Ockenfels (1998). In their study, (positive) reciprocation of a given kind action was not found to vary significantly across treatments where players had different alternatives. Their result may be influenced, however, by the fact that the behavior of the second mover was associated by large efficiency gains and that reciprocity and efficiency motives interact in a yet unknown way.

3 Modeling reciprocity

Summing up the preceding section, perceived kindness contains both (i) the concern for the distributional consequences per se and (ii) the underlying intention. The more a particular action signals good or bad intentions, the stronger the reciprocal response will be. Models which are based exclusively on distributional consequences or on intention cannot account for this evidence. Therefore, we present our model of reciprocity which tries to incorporate both aspects. The purpose of the presentation in this section is only expositional. We restrict our presentation to the key aspects of our model and omit all technical details. The technically interested reader is relegated to the appendix where we formally develop our model of reciprocal preferences.[2]

According to our model, reciprocity consists of a kind (or unkind) treatment by another person (represented by the *kindness term* φ) and a behavioral reaction to that treatment (represented by the *reciprocation term* σ). Our procedure is to transform the standard game (e.g., the ultimatum game) into a so-called "reciprocity game". In this new game the players' utility depends not only on the material payoffs of the original standard game (denoted by π_i) but also on the kindness and the reciprocation term. The utility function for player i is given by the following definition:

Definition Let player i and j be the two players of the game. We define player i's utility U_i in the transformed reciprocity game as:

$$U_i = \pi_i + \rho_i \varphi \sigma \tag{1}$$

According to the definition player i's utility in the reciprocity game is the sum of the following two terms: The first summand is simply player i's **material payoff**

[2]The full model is in detail explained and discussed in Falk and Fischbacher (1999).

π_i. This material payoff corresponds to the payoffs which are induced by the experimenter. The second summand - which we call **reciprocity utility** - is composed of the following terms:

- The positive constant ρ_i is called the **reciprocity parameter**. This constant is an individual parameter which captures the strength of player i's reciprocal preferences. The higher ρ_i, the more important is the reciprocity utility as compared to the utility arising from the material payoff. Note that if ρ_i equals zero, player i's utility is equal to his material payoff. Put differently, if $\rho_i = 0$, the player has *homo economicus* preferences just as assumed in standard game theory. If, in addition, ρ_j also equals zero, the reciprocity game collapses into the standard game.

- The **kindness term** φ measures the kindness player i experiences from j's actions. As outlined in the previous section, the perceived kindness depends on the *distributional consequences* of an action and its underlying *intention*.

  - The consequences determine the sign and (partly) the size of the kindness term. In order to determine the sign and the size of φ, player i needs to compare the payoff consequences of player j's action with a given reference standard. In principle, there are many possible reference standards against which subjects can compare a particular outcome. From many experiments it is known, however, that an *equitable* share of payoffs is a salient and commonly held standard.[3] We therefore implemented an equitable outcome share as reference standard. Given this equitable reference standard, player i can either be favored in an exchange (positive kindness term) or player i can be in a disadvantageous position (in which case the kindness term is negative). If φ is positive, player j is considered as kind. This holds, e.g., in the gift-exchange game if the firm offers the worker a generous wage. Of course, φ can also be negative, meaning that the action of player j is considered as unkind. This holds, e.g., in the ultimatum game if the offer is low. The lower the offer, the more negative is φ.
  - Notice that the kindness term comprises consequences *and* intentions. Our model takes intentions into account by looking at player j's alternatives. An action by player j is induced intentional if player j has reasonable alternatives to choose from. The more an action is caused intentionally the stronger it is weighted in the kindness term. For example, if player j has been *forced* to perform a particular action the size of the kindness term is smaller compared to a situation where player j has *chosen* the action.

- The **reciprocation term** σ measures the effect of the reciprocal action of player i on the payoff of player j. In a game with only two moves σ is simply player j's payoff.

[3] On the concept of equity as a reference standard compare, e.g., Adams (1965) and Loewenstein, Thompson and Bazerman (1989).

- The product of the *kindness* and the *reciprocation term* measures the reciprocity utility. If the kindness term is greater than zero, player i can *ceteris paribus* increase his utility if he chooses an action which increases player j's payoff. This is the case in the gift-exchange game if a high wage is answered by a high effort choice. The opposite holds if the kindness term is negative. In this case, player i has an incentive to reduce player j's payoff. As an example take the ultimatum game. If the offer was very low, a reciprocal player i can increase his utility by rejecting the offer which reduces player j's payoff.

4 Applications of the model

In the previous section we have outlined the basic structure of our model. In this section we discuss some of our model's predictions. We will concentrate on the difference between the 'equity approach' and the 'reciprocity approach'. We refer to three experimental games: two of the mini ultimatum games (Section 2.1), the gift-exchange game, and the ultimatum game. We will restrict the analysis to the second mover behavior. In Falk and Fischbacher (1999) we present a full characterization of our predictions. There you find all formal propositions and proofs as well as further applications.[4]

4.1 Reduced ultimatum games

In Section 2.1 we have discussed the experimental study by Falk, Fehr, and Fischbacher (1999). In this section we show that our model is capable to explain the difference between the (5,5)-game and the (8,2)-game.[5]

As Figure 3 shows, our model predicts a higher acceptance probability for the unkind offer (8,2) in the reduced (8,2)-game, compared to the (5,5)-game. The upper graph shows the acceptance probability of the unkind offer in the (8,2)-game, the lower shows that of the (5,5)-game. In both games, we see that the acceptance probability of the unkind offer decreases in the responder's concern for reciprocity, i.e., in ρ_R. We also see, however, that for a given ρ_R the acceptance probability is lower in the (5,5)-game compared to the (8,2)-game.

Figure 3

Thus, in the (8,2)-game a reciprocal responder is willing to accept a higher degree of inequity. The reason is that in the (8,2)-game the responder does not find it very unkind if he gets the 'unkind' offer, because she cannot infer any intentions from this offer. Things are quite different in the (5,5)-game. Here, the first mover

[4] In all applications, we first normalize the game in order to get payoffs between 0 and 1. Without such a normalization of the games, the reciprocity term would dominate the material payoff as stake size increases. Since most fairness experiment results have shown to be robust with respect to stake size (see, e.g., Fehr and Tougareva 1995), we think this normalization is appropriate. Furthermore, in all figures, we use the same range of the reciprocity parameter.

[5] The model explains the higher rejection rate of the (5,5)-game compared to the other three games, but does not explain the differences between the (2,8)-, the (8,2)- and the (10,0)-game.

has the alternative to choose an equitable outcome. Consequently, it *does* signal bad intentions and it is considered as quite unkind if this opportunity is not chosen. Thus, dependent on the first mover's alternatives the *same* offer will be perceived differently, which leads to a different kindness term. As a consequence, the acceptance rates differ just as we have observed in the experimental data (see Figure 2). Of course, the models following the 'equity approach' cannot explain this data.

Notice further that even in the (8,2)-game, rejections of the (8,2)-offer are not zero if the second mover is sufficiently reciprocal. This prediction confirms the experimental data (compare Figure 2). Models which claim that reciprocal responses are solely driven by intentions, however, predict a rejection rate of zero in this experiment. In a situation where a player has no alternatives to choose from, these models predict behavior that coincides with that derived with the standard selfishness assumptions.

4.2 Gift-exchange game

The gift-exchange game has been shortly introduced above. Contrary to the standard prediction it has been found that (i) wages and effort levels clearly exceed their lowest possible levels and (ii) that there is a positive wage-effort relation.

These two stylized facts are replicated by our model and are illustrated with the help of Figures 4 and 5. In Figure 4 we depict how - in equilibrium - a worker's effort choice depends on the wage paid by the firm given his reciprocal motivation, ρ_W. This figure captures the essence of *positive reciprocity* as reported in the gift-exchange experiments. From the worker's perspective, the higher the wage paid by the firm, the higher is the kindness term φ (compare Section 3). A reciprocal worker (with a sufficiently high reciprocity parameter ρ_W) improves his utility if he responds in kind, i.e., if he provides more than the minimum effort. As a result, reciprocal workers provide higher effort levels the higher the wage paid by 'their' firm. Moreover, workers will, for a given wage, choose higher effort levels, the stronger their reciprocal inclination is. Note that the effort level is zero if and only if $\rho_W = 0$ or if the wage is equal to zero.

Figure 4

Let us turn to the gift-exchange treatment reported by Charness (1996). Recall that in his random treatment a random mechanism determines the wage and *not* the firm. Compared to the 'regular' treatment, this leads to a *weaker* correlation between wages and effort levels. The reason is that in Charness' treatment a high wage does not signal any (kind) intentions. Our model captures this difference. Since the kindness term captures the concern for intentions, the kindness term is smaller in this situation. As a consequence the reciprocal action is weaker as well. For a given reciprocal motivation of the worker ρ_W, the model predicts a weaker correlation between wages and effort levels. Put differently, the *same* worker will supply a *lower* effort for a particular wage in the random-treatment compared to the 'regular' treatment. Figure 5 shows that result.

Figure 5

Figure 5 also shows that even in the non-intentional treatment the wage-effort relation is not absent. Put differently, a subject's reciprocal inclination is not totally wiped out - neither according to the data nor according to our model. This again shows the importance of modeling both, the concern for intentions and distributional concerns.

4.3 Ultimatum game

Given standard selfishness assumptions, the outcome of the subgame perfect Nash equilibrium in the ultimatum game is to offer the smallest possible offer which is accepted by the responder. The experimental results, however, show that low offers are frequently rejected and - as a consequence - most offers lie in a range between 40 and 50 percent of the total pie. Thus, the standard subgame perfect prediction is strongly refuted by the data. In the following we show that our model does a much better job to organize the data.

The most important finding in the experiments is that responders are more likely to accept higher offers than lower offers. With the help of our model we can calculate such an acceptance probability p. Figure 6 depicts this probability as a function of the level of the offer and for a given reciprocal inclination of the responder. The figure neatly captures the essence of *negative reciprocity*: A low offer is an unkind act which results in a negative kindness term. A utility-maximizing responder who is sufficiently reciprocally motivated can improve his utility by rejecting the offer. As a result, the lower the offer, the higher is the willingness of a reciprocal responder to punish the proposer by rejecting the offer. Note that as the responder's reciprocity concern ρ_R gets higher (lower) the acceptance-curve shifts to the right (left). This means that a person who is more motivated by reciprocity than another person will accept a low offer with a lower probability than the person who is less reciprocal. However, no matter how strong a responder is reciprocally motivated he will always accept an offer with probability one if the offer is equal or higher than half of the pie, a prediction that is supported by the experimental evidence.

Figure 6

Let us now turn to the predictions of our model if applied to the 'non-intentional' experiment by Blount (1995). Remember that in her treatment the proposers' offers were randomly selected, i.e., a low offer did not signal any (bad) intentions. As the data of Blount's experiment reveals, the acceptance rate for a given offer is *higher* than in the 'regular' treatment. However, even in the absence of intentions, many subjects dislike very unfair distributions and will, for that reason, reject very disadvantageous offers.

Our model predicts exactly these stylized facts. Figure 7 depicts the predicted acceptance probabilities in the 'regular' ultimatum game and in Blount's treatment for a given ρ_R. The upper graph corresponds to the Blount-experiment whereas the lower graph shows the acceptance behavior in the 'regular' treatment. As can be seen in the figure, a responder's acceptance probability for low offers is higher if intentions are absent. The reason for this is that the experienced unkindness and thus the

(negative) kindness term is smaller. As a result the reciprocal response is smaller as well. Put differently, the same person may accept a particular offer if generated by a random device and may reject it when offered by a human being.

Figure 7

5 Concluding remarks

In his classic account of reciprocity, Gouldner (1960) points out that the force of reciprocity is - among other factors - variable with respect to the motives imputed to the donor and the donor's own free will. These two factors also characterize the 'reciprocity-approach' which received strong support in many fairness experiments. In practically all reported experiments that were devoted to test for the relevance of intentions, the latter have a significant impact on reciprocal behavior. We take this as a serious drawback for the inequity aversion models. These consequentialistic models provide an oversimplified view of the psychological determinants of fair behavior and are therefore incompatible with important experimental findings.

However, the purely intentions-based models overestimate the role of intentions and underestimate the importance of distributional concerns. As we have emphasized in our discussion, people exhibit reciprocal responses even in situations where intentions are absent. People not only care about intentions but also experience the *outcome per se* as either advantageous or disadvantageous. We interpret this concern for the outcome per se as envy or the desire to be generous: If the (non-intentional) consequences of an action are very disadvantageous, people experience feelings of envy. If, on the other hand, one is favored in an exchange people may feel a desire to be generous or to "share-the-wealth" (Rabin 1998, p. 24). Both feelings, envy as well as the desire to be generous may trigger reciprocity-like reactions.

We think that a model should take into account both, distributional concerns and intentions. The model suggested in this paper provides a first attempt in this direction.

6 Appendix

In this appendix we present the technical details of our model of reciprocity. Consider a two-player extensive form game with a finite number of stages and with complete and perfect information. Let i be a player in the game. N_i denotes the set of nodes where player i has the move with n being a node of this player. Let A_n be the set of actions in node n. Let F be the set of end nodes of the game. The payoff function for player i is given by $\pi_i : F \to \mathbb{R}$.

Let $P(A_n)$ be the set of probability distributions over the set of actions in node n. Then $S_i = \prod_{n \in N_i} P(A_n)$ is player i's behavior strategy space. For $s_i \in S_i$ and $s_j \in S_j$ we define $\pi_i(s_i, s_j)$ and $\pi_j(s_i, s_j)$ as the players' expected payoffs, given strategies s_i and s_j. Furthermore, we define $\pi_i(n, s_i, s_j)$ as the expected payoff conditional on node n: It is the expected payoff of player i in the subgame starting from node n, given that the strategies s_i and s_j are played and given it is known that player i is at node n.[6]

The kindness term φ is the central element of our model. It measures how kind a person perceives the action by another player. It depends on the outcome and the intention. The outcome is measured with the **outcome term** Δ where $\Delta > 0$ expresses an advantageous outcome and $\Delta < 0$ expresses a disadvantageous outcome. In order to determine the overall kindness, Δ is multiplied with the **intention factor** $\vartheta \geq 0$. This factor is a number between zero and one, where $\vartheta = 1$ captures a situation where Δ is induced fully intentionally and $\vartheta < 1$ expresses a situation where less or no intentions are involved. The kindness term φ is simply the product of Δ and ϑ. All terms are derived in the following.

First, we define the **outcome term**:

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

For a given ϑ, the outcome term Δ measures the kindness of player j. It captures the knowledge of player i in node n about the two players' expected payoffs. Since ϑ is positive, the sign of the kindness term, i.e., whether an action is considered as kind or unkind, is determined by the sign of Δ. The term Δ is positive if player i thinks he gets more than the other player. It is negative if player i thinks to get less than the other player.

In judging player j's kindness player i is well aware of the fact that player j might not have caused a particular outcome intentionally, i.e., we have to consider the alternative payoff combinations of player j. Let us state precisely what we mean by alternative payoff combinations. Let S_j^p be the set of pure strategies of player j. For given strategies we define:

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

$\Pi_i(s_i)$ is a set of payoff combinations π_i and π_j. These are payoffs player j can induce by choosing a pure strategy s_j^p given she expects player i to play s_i. Whether player i

[6] Because we are working with behavior strategies, the strategies s_i and s_j induce a strategy in the subgame.

experiences a particular outcome as chosen intentionally by player j depends on the options available to player j. The question to ask is: How intentional is a particular payoff distribution (π_i^0, π_j^0) induced by player j - given that player j had an alternative payoff distribution (π_i, π_j) he could have chosen? The answer to this question is given in function Ω. In this function, the value of Ω expresses how intentional player j's choice of (π_i^0, π_j^0) was, given his alternatives (π_i, π_j). If the choice was fully intentional Ω equals 1, if the choice is considered as not fully intentional, however, Ω is smaller than one. The Ω-function is defined as follows:

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

where ε_i is an individual parameter with $0 \leq \varepsilon_i \leq 1$.[7]

The first two rows capture situations where player j has treated player i in a kind way ($\pi_i^0 \geq \pi_j^0$). In these situations the value of Ω depends on whether player j could reduce player i's payoff (π_i compared to π_i^0) or not. Player i considers the kind action as fully intentional if and only if player j has an alternative to decrease player i's payoff, i.e., player j has *an alternative to be less kind* ($\pi_i < \pi_i^0$).

The other three rows represent instances where player j puts player i in a disadvantageous situation, i.e., where $\pi_i^0 < \pi_j^0$ holds. In this case player j's behavior is considered to be intentional if and only if player j has an alternative to be less kind ($\pi_i > \pi_i^0$) and if this alternative is not considered as an *unreasonable sacrifice* (line 4 of the above definition).

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

$$\vartheta(n, s_i, s_j) = \max\{\Omega(\pi_i, \pi_j, \pi_i(n, s_i, s_j), \pi_j(n, s_i, s_j)) \mid (\pi_i, \pi_j) \in \Pi_i(s_i)\} \tag{5}$$

The maximum-operator guarantees that a particular action is considered as intentional if there is *any* 'true' alternative. We now come to the definition of the kindness term.

Definition Let strategies be given. We define the **kindness term** $\varphi(n, s_i, s_j)$ in a node $n \in N_i$ as:

$$\varphi(n, s_i, s_j) = \vartheta(n, s_i, s_j)\Delta(n, s_i, s_j) \tag{6}$$

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

[7] This parameter is called the 'pure outcome concern parameter'. It is an individual parameter which measures a player's concern for the outcome per se, i.e., it captures the experienced kindness in the absence of intentions.

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

$$\sigma(n, f, s_i, s_j) := \pi_j(\nu(n, f), s_i, s_j) - \pi_j(n, s_i, s_j) \tag{7}$$

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

The *reciprocation term* expresses the response to the experienced kindness, i.e., it measures how much player i alters the payoff of player j with his move in node n. The reciprocal impact of this action is represented as the *alteration* of player j's payoff from $\pi_j(n, s_i, s_j)$ to $\pi_j(\nu(n, f), s_i, s_j)$.

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

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

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

$$U_i(f, s_i, s_j) = \pi_i(f) + \rho_i \sum_{\substack{n \to f \\ n \in N_i}} \varphi(n, s_i, s_j)\sigma(n, f, s_i, s_j) \tag{8}$$

For fixed (s_i, s_j), this utility function defines a new game $\Gamma(s_i, s_j)$. If (s_i, s_j) is a subgame perfect Nash equilibrium in $\Gamma(s_i, s_j)$, we call (s_i, s_j) a **reciprocity equilibrium**.

The strategies s_i and s_j in the game $\Gamma(s_i, s_j)$ can be interpreted as beliefs of the players. For instance, player i believes player j will use strategy s_j and he thinks player j expects him to use strategy s_i. Given this belief, player i chooses an optimal strategy. A reciprocity equilibrium can then be considered as a combination of strategies and beliefs in which the strategies are optimal and consistent with the beliefs. The presentation of our theory in this form (without beliefs) follows an idea of Gintis (2000).

7 References

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

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

Bolton, G. (1991): 'A Comparative Model of Bargaining: Theory and Evidence', *American Economic Review* 81, 1096-1136.

Bolton, G. E., Brandts, J. and Ockenfels, A. (1998): 'Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game', *Experimental Economics* 1, 207-219.

Bolton, G. and Ockenfels, A. 'ERC - A Theory of Equity, Reciprocity and Competition', forthcoming in: *American Economic Review*.

Bowles, S., and Gintis, H. (1998) 'The Evolution of Strong Reciprocity', *mimeo*, University of Massachusetts.

Brandts, J., and Sola, C. (1998) 'Reference Points and Negative Reciprocity in Simple Sequential Games', *mimeo*, University of Barcelona.

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

Charness, G. and Rabin, M. (1999) 'Social Preferences: Some Simple Tests and a New Model', *mimeo*, University of Berkeley.

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

Falk, A. and Fischbacher, U. (1999) 'A Theory of Reciprocity', *Working paper* no. 6, University of Zurich.

Falk, A, Fehr E., and Fischbacher U. (1999) 'On the Nature of Fair Behavior', *Working paper* no. 17, University of Zurich.

Falk, A, Fehr E., and Fischbacher U. (2000) 'Testing Theories of Fairness - Intentions Matter', *Working paper* no. 63, University of Zurich.

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

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

Fehr, E. and Schmidt, K. 'A Theory of Fairness, Competition, and Cooperation', forthcoming in: *Quarterly Journal of Economics*.

Fehr, E. and Tougareva, E. (1995) 'Do High Stakes Remove Reciprocal Fairness - Evidence from Russia', *Discussion paper*, University of Zurich.

Frisch, D. M. and Greenberg, M. S. (1968): 'Reciprocity and Intentionality in the giving of help', Proceedings of the 76th Annual Convention of the American Psychology Today 2, 31-34.

Gintis, H. (2000) 'Game Theory Evolving', (Princeton: Princeton University Press).

Goranson, R. E. and Berkowitz, L. (1966): 'Reciprocity and Responsibility Reactions to Prior Help', *Journal of Personality and Social Psychology* 8, 99-111.

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

Güth, W., Huck, S. and Müller, W. (1998): 'The Relevance of Equal Splits - On a Behavioral Discontinuity in Ultimatum Games', *Discussion paper*, Humboldt-University Berlin.

Hornstein, H. A., Fisch, E. and Holmes, E. (1968): 'The influence of a model's feeling about his behavior and his relevance as a comparison other on observers' behavior', *Journal of Personality and Social Psychology* 10, 222-226.

Lerner, M. J. and Lichtman, R. R. (1968): 'Effects of Perceived Norms on Attitudes and Altruistic Behavior toward a Dependent Other', *Journal of Personality and Social Psychology* 5, 319-325.

Levine, D. (1998) 'Modeling Altruism and Spitefulness in Experiments', *Review of Economic Dynamics* 1, 593-622.

Loewenstein, G. F., Thompson, L., and Bazerman, M. H. (1989) 'Social Utility and Decision Making in Interpersonal Contexts', *Journal of Personality and Social Psychology* 57, 426-441.

McCabe, K. A. and Smith, V. L. (1997): 'Intentionality Detection and "Mindreading": Why does Game Form Matter?', *Discussion paper*, University of Arizona.

Offerman, T. (1999): 'Hurting hurts more than helping helps: the role of the self-serving bias', *Discussion paper*, University of Amsterdam.

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

Rabin, M. (1998) 'Psychology and Economics', *Journal of Economic Literature* 36, 11 - 46.

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

Slonim, R. and Roth, A. (1997) 'Financial Incentives and Learning in Ultimatum and Market Games: An Experiment in the Slovak Republic', *Econometrica* 66, 569-596.

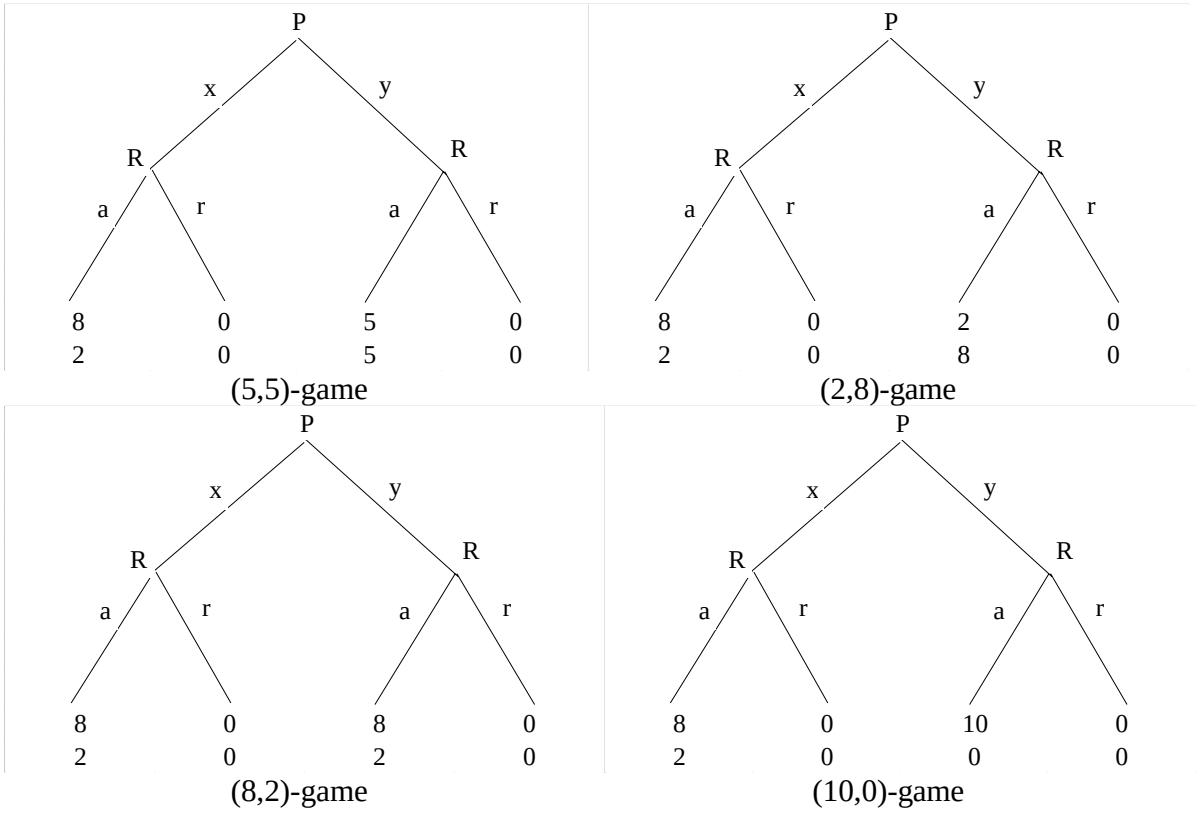

Figure 1: Game trees of the mini ultimatum games

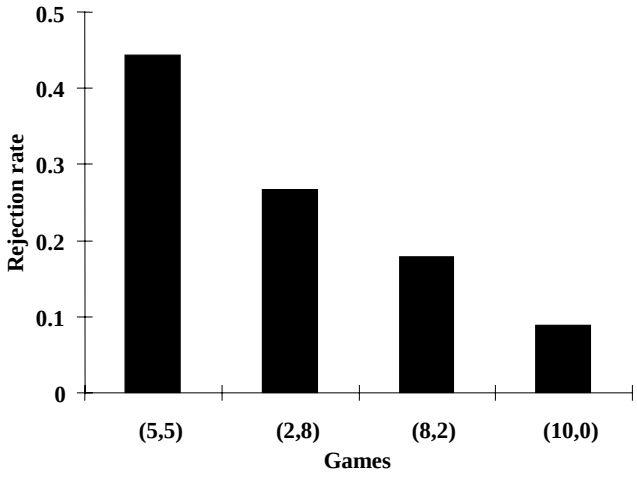

Figure 2: Rejection rate of the (8,2) offer in the four mini-ultimatum games (n=45). Source: Falk, Fehr and Fischbacher (1999).

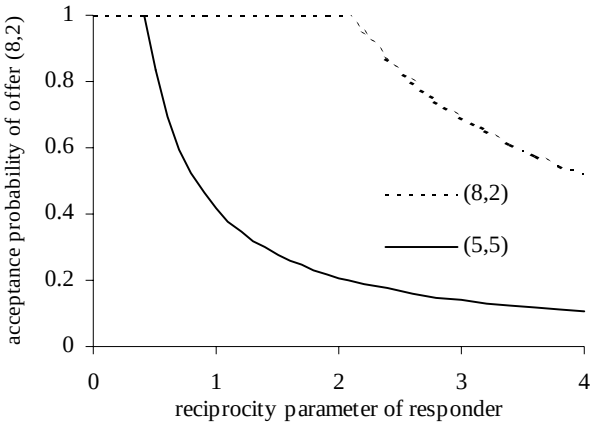

Figure 3: Predicted rejection rate of the (8,2) offer for the (normalized) (8,2)- and the (5,5)-game

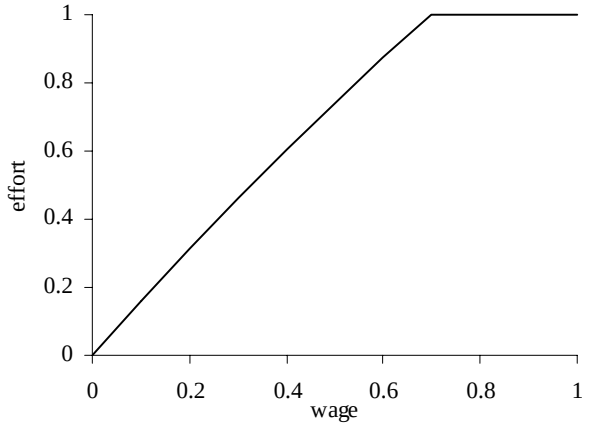

Figure 4: Predicted effort choice for a given reciprocity parameter of the worker (ρ_2= 2). The payoff functions are: *effort – wage* for the firm and *wage - 0.2*effort*^{2} for the worker.

Figure 5: Predicted effort choice for the treatment where the firm chooses the wage (intentional wage) and for the treatment where the wage is determined randomly (random wage). Parameters are as in Figure 4.

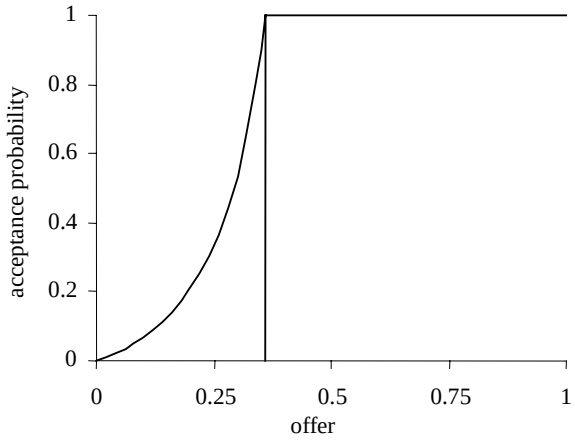

Figure 6: Predicted acceptance probability in a normalized ultimatum game for a given reciprocity parameter of the responder ($\rho_2 = 2$)

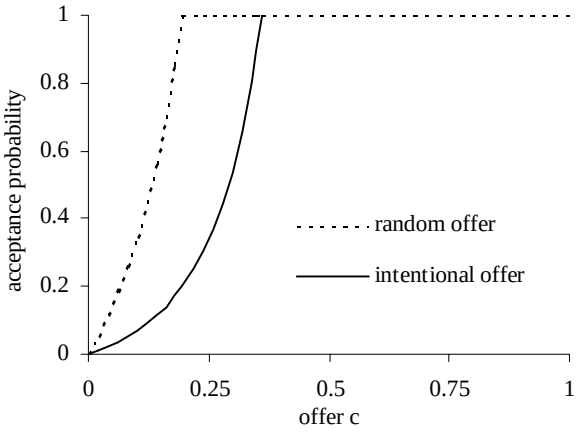

Figure 7: Predicted acceptance probability in a normalized ultimatum game where the proposer makes the offer (intentional offer) and where the offer is determined by a random device (random offer)