Institute for Empirical Research in Economics
University of Zurich

Working Paper Series
ISSN 1424-0459

Working Paper No. 63

Testing Theories of Fairness - Intentions Matter

Armin Falk, Ernst Fehr and Urs Fischbacher

September 2000

Testing Theories of Fairness - Intentions Matter*

Armin Falk, Ernst Fehr and Urs Fischbacher

University of Zurich**

First Version: September 2000

Abstract

Recently developed models of fairness can explain a wide variety of seemingly contradictory facts. The most controversial and yet unresolved issue in the modeling of fairness preferences concerns the behavioral relevance of fairness intentions. Intuitively, fairness intentions seem to play an important role in economic relations, political struggles and legal disputes. Yet, so far there is little rigorous evidence supporting this intuition. In this paper we provide clear and unambiguous experimental evidence for the behavioral relevance of fairness intentions. Our results indicate that the attribution of fairness intentions is important both in the domain of negatively reciprocal behavior and in the domain of positively reciprocal behavior. This means that reciprocal behavior cannot be fully captured by equity models that are exclusively based on preferences over the distribution of material payoffs. Models that take into account players' fairness intentions *and* distributional preferences are consistent with our data while models that focus exclusively on intentions or on the distribution of material payoffs are not.

JEL-Classification: D63, C78, C91.

Keywords: Fairness, reciprocity, intentions, experiments, moonlighting game.

---

* Financial support by the Swiss National Science Foundation (Project 1214-05100.97) and by the MacArthur Foundation (Network on Economic Environments and the Evolution of Individual Preferences and Social Norms) is gratefully acknowledged. This paper is part of the EU-TMR Research Network ENDEAR (FMRX-CTP98-0238).

** Postal address: Institute for Empirical Research in Economics, Blümlisalpstrasse 10, CH-8006 Zurich. E-mail: falk@iew.unizh.ch, efehr@iew.unizh.ch, fiba@iew.unizh.ch

I. Introduction

A considerable body of evidence indicates that a substantial number of the people are motivated by concerns for fairness and reciprocity. Moreover, the presence of fair-minded people is likely to have important economic effects (Kahneman, Knetsch and Thaler 1986; Camerer and Thaler 1995; Bewley 1999; Fehr and Gächter 2000). This has led to the development of several fairness models (Rabin 1993; Levine 1998; Dufwenberg and Kirchsteiger 1999; Falk and Fischbacher 1999; Fehr and Schmidt 1999; Bolton and Ockenfels 2000; Charness and Rabin 2000). These models share the property that some people are assumed to have a preference for fairness – in addition to their preference for material payoffs. The impressive feature of these models is that they are capable of predicting correctly a wide variety of seemingly contradictory facts.

This paper examines the most controversial question in the modeling of fairness preferences: the role of *fairness intentions*. Do fair-minded people respond to fair or unfair *intentions* or do they respond solely to fair or unfair *outcomes*? One class of fairness models – the equity models of Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) – is based on the assumption that fairness intentions are behaviorally irrelevant. Another class of models (e.g., Rabin 1993, Falk and Fischbacher 1999, Dufwenberg and Kirchsteiger 1999) assigns fairness intentions a major behavioral role.

The answer to our question is of great practical and theoretical interest. At the theoretical level the question concerns not only the proper modeling of fairness preferences but also standard utility theory. Standard utility theory assumes that the utility of an action depends solely on the consequences of the action and not on the intention behind the action. Therefore, if the attribution of intentions turns out to be behaviorally important, the "consequentialism" inherent in standard utility models is also in doubt. At the practical level the issue is important because many relevant decisions are likely to be affected if the attribution of intentions matters. Fairness attributions are likely to influence decision-making in firms and other organizations as well as in markets and the political arena. Political decisions and business decisions, for instance, often affect the material payoffs of some parties negatively. If the response of the negatively affected parties also takes into account the decision-maker's fairness intentions, it will be much easier to prevent opposition when the decision-maker can credibly claim that he is somehow forced – by law, by international

competition or by some other external force – to take the action. It is, therefore, no coincidence that the rhetoric of politicians and business leaders often appeals to the phrase that "there is no alternative". If there is indeed no alternative it is not possible to attribute unfair intentions to the action because the decision-maker cannot be held responsible for the action. If, in contrast, there are obvious alternative actions available, it is much easier for the affected parties to attribute unfair intentions to the action and, as a consequence, their opposition will be much stronger.

The attribution of intentions is also important in law (Huang 2000). Intentions often distinguish between whether the same action is a tort or a crime and whether a tort should involve punitive damages. Other distinctions made in criminal law concern whether an action is taken purposely, knowingly, recklessly, or negligently (see Model Penal Code § 2.02(1)-(2)). Thus, the penal code (which represents a codified broad sense of justice) distinguishes quite carefully between the consequences of an action and the intentions underlying this action.

Although, at the intuitive level, the attribution of intentions seems to be important, so far it has been rather difficult to provide rigorous and unambiguous evidence for this. This is not surprising with regard to field data because outcomes and intentions are usually inextricably intertwined in the field. Yet, even in laboratory experiments the issue has been quite elusive. Charness (1996), Bolton, Brandts and Ockenfels (1998), Offerman (1999) and Cox (2000) find little or no evidence that the attribution of fairness intentions matters in the domain of positively reciprocal behavior. Blount (1995) and Offerman (1999) find evidence that it matters in the domain of negatively reciprocal behavior but, as we will argue below, these studies have some methodological problems.[1] Thus, we face the puzzle that, intuitively, the attribution of fairness intentions seems to be important while given the prevailing evidence, the issue remains highly controversial.

In this paper we provide clear and unambiguous experimental evidence for the behavioral relevance of fairness attributions. Our results indicate that the attribution of fairness intentions is important both in the domain of negatively reciprocal behavior and in the domain of positively reciprocal behavior. When the experimental design rules out the attribution of fairness intentions, reciprocal responses are substantially weaker. This result is

[1] Positive reciprocity is defined as a kind response to an action that leads to a fair outcome or is driven by fair intentions. Negative reciprocity is defined as a hostile response to an action that leads to an unfair outcome or is driven by an unfair intention.

corroborated at the individual as well as at the aggregate level. It is not only the case that some individuals show weaker reciprocal responses when it is impossible to attribute fairness intentions. A non-negligible fraction of the subjects (30 percent) exhibit *no* reciprocal behavior when fairness attributions are ruled out, i.e., they behave like selfish individuals. However, when the design allows for the attribution of fairness intentions no subject behaves in a fully selfish manner. This indicates that the recently developed equity models of Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) are incomplete because they neglect fairness intentions. The behavioral relevance of fairness intentions does, of course, not rule out that subjects also respond to unfair outcomes. Our results indicate that, on average, subjects exhibit reciprocal behavior even if they cannot attribute fairness intentions. Thus, models that are exclusively based on intention-driven reciprocal behavior (e.g., Rabin 1993; Dufwenberg and Kirchsteiger 1999) are also incomplete. Models that combine both aspects (like e.g., Falk and Fischbacher 1999) fit our data best.

Our experimental design also provides an opportunity to examine to what extent both positive and negative reciprocity is exhibited by the *same* individuals. To our knowledge there is no study that examines whether positive and negative reciprocity is correlated at the level of the individual. Previous studies can only answer the question whether a given individual exhibits a positively or a negatively reciprocal response. It turns out that – when fairness attributions are possible - 40 percent of the subjects exhibit both positively and negatively reciprocal responses. However, a surprisingly large fraction of 21 percent exhibit only positively reciprocal responses and 15 percent show only negatively reciprocal responses.

The paper is organized as follows. The next section shortly discusses potential obstacles in finding behavioral effects of fairness intentions. Section III presents the experimental design. Section IV discusses the predictions of several fairness theories. Results are presented in Section V. Section VI gives a short summary.

II. Obstacles for Finding Behavioral Effects of Fairness Intentions

Before we present our experimental design it is important to discuss the potential reasons for the lack of convincing evidence in favor of fairness intentions. In our view, there are four potential reasons. The first reason is that in some studies there is a potential confound with the efficiency motive. Andreoni and Miller (2000), Bolle and Kritikos (1998) and Charness

and Rabin (2000) report results suggesting the presence of a non-negligible fraction of subjects that is willing to increase efficiency. These subjects seem to be willing to bear some cost in order to increase the total payoff, i.e., the sum of the payoffs that accrues to all the parties. This motive could have swamped the positively reciprocal responses in the studies of Charness (1996), Bolton, Brandts and Ockenfels (1998) and Offerman (1999) because in these studies the reciprocal behavior of the second-mover was associated with large efficiency increases. It is also possible that reciprocity motives and efficiency motives interact in a yet unknown way. For this reason our design rules out that reciprocal responses increase the total payoff.

A second reason is related to the issue of repetition. In Charness (1996) subjects faced a different opponent in each of ten periods. Repetitions may create all sorts of ill-understood noise and spillovers across periods that make it difficult to isolate the attribution of fairness intentions. For this reason we conducted a one-shot experiment without any repetitions.

A third potential reason for the lack of a behavioral impact of fairness intentions could be that the treatment manipulations were not strong enough. To isolate the role of fairness intentions one ideally needs a treatment in which first-movers can signal their fairness intentions and a treatment in which such signals are ruled out completely. The signaling of fairness intentions rests on two premises: (i) The first-mover's choice set actually allows the choice between a fair and an unfair action, and (ii) the first-mover's choice is under the *full* control of the first-mover. The first premise implies that the treatment manipulation can be "too weak" because the choices available to the first-mover may not be sufficiently different, i.e., the fairness or unfairness of the available actions is not salient enough. In our design we solved this problem by giving the first-mover a choice set that allows for very different actions. In particular, the first-mover could either increase or decrease the second-movers payoff relative to a clearly defined reference point (i.e., relative to an initial endowment that was the same for both players). This distinguishes our study from the studies of Charness (1996), Bolton, Brandts and Ockenfels (1998) and Cox (2000) where the first movers could only be more or less kind to the second-movers but they could not hurt them. Perhaps, the distinction between being more or less kind was not salient enough and, as a consequence, there was little or no intention-driven reciprocal behavior in these studies.

The fourth reason is related to the second premise above. It concerns the question of how one can rule out the attribution of fairness intentions to the first mover's choice. In our

view the strongest method is to deprive the first mover of any choice at all and to make this salient to the second mover. In our experiment we achieved this by determining the first mover's "action" by a salient random device. Saliency was implemented by rolling dice in front of each second mover. Yet, if the first mover's choice is determined by a random device, the second movers might have views about what constitutes fair or unfair random devices. For example, if the random device determines with high probability a very bad outcome for the second mover, the second mover may become angry because she views this as a rather unfair device. If, in contrast, human first movers are unlikely to choose such a bad outcome the comparison of responses across the random device and the human choice condition does not isolate the impact of fairness intentions. The reason is that there is likely to be a confound due to the angry response to an unfair random device. Our solution of this problem is to implement a random device that mimics the probability distribution over the actions of human first movers.

Both in Blount (1995) and Offerman (1999) the first mover's "action" was determined by a random device. Yet, only Blount kept the probability distribution of first mover actions constant across the random device and the human choice condition. Although the study of Blount is very clean and convincing in this regard it faces other methodological problems. The results of her ultimatum game may be affected by the fact that subjects in the human choice condition had to make decisions as a proposer *and* as a responder before they were to know their actual roles. After subjects had made their decisions in both roles, the role for which they received payments was determined randomly. This means that the decision situation was not kept constant across the random device and the human choice condition because in the human choice condition the responders also had to put themselves in the shoes of the proposers while in the random device condition this was not the case.

In one of Blount's treatments deception was involved. Subjects believed that there were proposers although in fact the experimenters made the proposals. All subjects in this condition were "randomly" assigned to the responder role. It may well be the case that this kind of deception cannot be hidden from the subjects, i.e., at least a number of subjects might have noticed that they were deceived. In contrast to this setting subjects in our experiment knew their role in all conditions before they made decisions and we had real human subjects in the first-mover position and in the second-mover position.

III. Experimental Design and Procedures

Our experimental design is based on the "moonlighting game" (Abbink, Irlenbusch and Renner 2000). This game has the advantage that we can examine the impact of fairness intentions on positively as well as on negatively reciprocal responses at the individual level. In the following we first describe the moonlighting game. Then we present our two treatments, the Intention treatment and the No-Intention treatment. Finally, we report the procedures of the experiment.

The constituent game. The "moonlighting game" is a two-player sequential move game that consists of two stages. At the beginning of the game, both players are endowed with 12 points. At the *first stage* player A chooses an action $a \in \{-6, -5, \ldots, 5, 6\}$. If A chooses $a \geq 0$, he gives player B a tokens while if he chooses $a < 0$, he takes away $|a|$ tokens from B. In case of $a \geq 0$ the experimenter triples a so that B receives $3a$. If $a < 0$ A reaps $|a|$ and player B loses $|a|$. After player B observes a, she can choose an action $b \in \{-6, -5, \ldots, 17, 18\}$ at the *second stage*, where $b \geq 0$ is a reward and $b < 0$ is a sanction. A reward transfers one point from B to A. A sanction costs B exactly $|b|$ but reduces A's income by $3|b|$. After B's decision, final incomes are determined. Since As can give and take while Bs can reward or sanction, this game allows for both, positively and negatively reciprocal behavior.

In our experiment we applied the strategy method. This means that player B had to give us a response for each feasible action of player A, before B was informed about the actual choice of A. This has several advantages[2]. First, it allows us to examine the correlation between positive and negative reciprocity at the individual level. Thus, we know whether there are subjects who exhibit only negatively reciprocal responses or only positively reciprocal responses, or whether those who are negatively reciprocal are also positively reciprocal. Second, as we will see, the strategy method allows us to study the relevance of intentions for reciprocal behavior at any level of a. This is so because we have sufficiently many responses to each feasible action of A.

Treatments. As discussed above, A's action signals *fairness intentions* if (i) A's choice set allows the choice between saliently fair and saliently unfair decisions, and (ii) if A's choice is

[2] In principle, it is possible that the strategy method induces a different behavior of B relative to a situation where B has to respond to the actual move of A. However, Brandts and Charness (1998) and Cason and Mui (1998) report evidence indicating that the strategy method does not induce different behavior. Moreover, we used the strategy method in both of our treatments. Therefore, the impact of this method, if there is any, is kept constant across treatments.

under the full control of A. Condition (i) is guaranteed by our experimental game since it allows A to give or to take different amounts of money. Condition (ii) is our treatment variable. In the *Intention treatment* (I-treatment) A himself determines *a*. Thus, in the I-treatment A is responsible for the consequences of his action and, therefore, his action signals good intentions (if $a > 0$) or bad intentions (if $a < 0$). In the *No-intention treatment* (NI-treatment), on the other hand, A's move is determined by a random device. Consequently, A has no control over his action. His action therefore signals neither good nor bad intentions.

The random move of A in the NI-treatment was implemented as follows: After B had determined her strategy, the experimenter went to her place and threw two dice in front of B. Both dice were ten-sided showing numbers from 0 to 9, i.e., together they created numbers between zero and 99 with equal probability. The number that was determined was then used to determine the move of A according to Table 1. For example, if the dice showed a number between 0 and 6, A's random move was to take 6 points ($a = -6$), while if the number was, e.g., 58, player A's move was $a = 3$. After the move of A had been determined, the experimenter went to another player B and threw the dice again. This procedure was explained to the Bs in great detail in the instructions. Thus, it was completely transparent to each B that A's move was determined randomly according to Table 1. Players A also knew that their choice would be randomly determined but they did not know the probability distribution.

Notice that according to Table 1 the randomly determined moves of A are not equally likely. For example, it is more likely to randomly select $a = -6$ (7 percent chance) than $a = -5$ (2 percent chance). Table 1 is based on the *actual human decisions* of the As who participated in the moonlighting experiment by Abbink, Irlenbusch and Renner (2000). Using this table we are able to approximate a 'human choice distribution' even in the NI-treatment where choices were randomly selected. To keep everything constant except the potential for the attributions of intentions across the NI- and the I-treatment, the choice distribution given in Table 1 was also presented to the Bs in the I-condition[3]. In this condition it was pointed out to the Bs that the same experiment had already been conducted before and that the relative frequency of the decisions by the As in that experiment was identical to the frequencies in Table 1. This was done to induce players B to have the same beliefs about the choice distribution of the As in both treatment conditions. In this way we ruled out the possibility that the responses of the Bs were affected by different beliefs about the choice

[3] As in the NI-treatment, the As were not informed about the choice distribution in Table 1.

distribution of the As.[4] Thus, with regard to the probability distribution over the set of feasible actions the two treatments cannot evoke different fairness judgements.

Table 1: Probability distribution of the move of A in the NI-treatment

Realized number	**A's move *a***	**Percent**
0-6	*-6*	7
7-8	*-5*	2
9-15	*-4*	7
16-19	*-3*	4
20-21	*-2*	2
22-26	*-1*	5
27-39	*0*	13
40-46	*1*	7
47-55	*2*	9
56-62	*3*	7
63-73	*4*	11
74-75	*5*	2
76-99	*6*	24

Procedure. Before the game started, subjects were randomly assigned to their role as player A or B (in both treatments). They were seated in front of their terminal and given their instructions. To ensure the understanding of the experimental procedures all subjects had to answer several control questions and the experiment did not start until all subjects had answered all questions correctly. Procedures and payoff functions were known by all the players, i.e., they were explained in the instructions and orally summarized. We used the experimental software z-Tree (Fischbacher 1999) to run the experiments.

IV. Predictions

In the following we derive the theoretical predictions for our experimental game. First, we shortly present the economic prediction based on the assumption that it is common knowledge that all players are selfish and rational followed by the predictions of recently developed fairness theories.

[4] We also checked whether the distribution of realized choices in the I-treatment and the NI-treatment differ. Based on a Kolmogorov-Smirnov test the null hypothesis of identical distributions cannot be rejected ($p = 0.289$).

Self Interest Prediction. If it is common knowledge that all players are selfish and rational the following subgame perfect equilibrium outcome is predicted: In both treatments B will always choose $b = 0$, i.e., she will neither punish nor reward, because any other choice is costly. Therefore, in the I-treatment player A will choose $a = -6$ because he only loses if he chooses $a > 0$ and has nothing to fear if $a < 0$. In the NI-treatment, player A's move is determined by a random device.

Fairness Predictions. We now turn to the predictions of recently developed fairness theories. Our focus is the behavior of player B. The models by Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) are built on the assumption that subjects dislike inequity. In the Bolton and Ockenfels model inequity averse players have a concern for a fair relative share of the total payoffs. The fair relative share is defined as $1/n$ where n is the number of players in the game. If a player receives less than the fair relative share he tries to increase his share and vice versa. According to Fehr and Schmidt (1999) inequity averse players are concerned with the payoff differences between themselves and each other player. If player i's earnings differ from that of player j, he aims at reducing the payoff difference between himself and j. Both models predict that for sufficiently strong inequity aversion people exhibit reciprocal behavior, i.e., b is increasing in a and $b = 0$ if $a = 0$. Common to both approaches is the neglect of intentions. Only the payoff consequences are assumed to explain reciprocal responses. For our experiment this implies that for a given move of A, reciprocal responses between the I-treatment and the NI-treatment should be *exactly the same*. Since the payoff consequences of A's move are the same in both treatments, a player B who is solely concerned with payoff consequences should respond in the same way.

A different concept of reciprocity starts with the premise that reciprocal responses are triggered *exclusively* by kind or unkind *intentions* (Dufwenberg and Kirchsteiger 1999)[5]. From this premise it immediately follows that in the absence of fairness intentions there should be no reciprocal behavior at all, i.e., player B neither rewards nor punishes but pursues her material self-interest. Therefore, in the NI-treatment Dufwenberg and Kirchsteiger predict *no* reciprocal behavior at all ($b = 0, \forall a$). The prediction for the I-treatment is less clear. The reason for this is that the model exhibits multiple equilibria and some of them are compatible with b being locally decreasing in a. This is so because

[5] The model of Dufwenberg and Kirchsteiger is based on Rabin's (1993) normal form theory of fairness. Since the present game is a sequential game we restrict our analysis to the Dufwenberg and Kirchsteiger theory of sequential reciprocity.

according to the model higher values of a do not necessarily signal more friendly intentions[6]. There are, however, plausible equilibria where $a > 0$ signals good intentions and $a < 0$ signals bad intentions. In these equilibria b is increasing in a (in the I-treatment).

Finally there is the model by Falk and Fischbacher (1999) that combines a concern for a fair distribution of payoffs with the reward and punishment of fair and unfair intentions. In the I-treatment the model makes the (unique) prediction that b is increasing in a. In the NI-treatment, this reciprocal pattern is predicted to be *weaker* compared to the I-treatment. Contrary to Dufwenberg and Kirchsteiger, the model does not predict $b = 0, \forall a$ since players in this model not only have a concern for intentions but also for a fair distribution of the payoffs. However, since intentions are absent in the NI-treatment subjects react less reciprocally than in the I-treatment. The latter prediction distinguishes Falk and Fischbacher from the inequity aversion models by Bolton and Ockenfels and Fehr and Schmidt. Table 2 summarizes all predictions. Notice that all fairness theories make similar predictions in the I-treatment but differ in their predictions for the NI-treatment.

Table 2: Summary of predictions for player B

Model	I-treatment	NI-treatment
Standard prediction	$b = 0, \forall a$	$b = 0, \forall a$
Only payoff consequences matter (Fehr/Schmidt and Bolton/Ockenfels)	b increases in a	exactly the *same* behavior as in I-treatment
Only fairness intentions matter (Dufwenberg/Kirchsteiger)	b increases in a[7]	$b = 0, \forall a$
Payoff consequences and fairness intentions matter (Falk/Fischbacher)	b increases in a	b increases in a but *less* than in the I-treatment

[6] To make this point clear, consider the following example. Assume that B believes that A expects B to punish $a = -5$ with $b = -6$ while B is expected not to punish $a = -6$. In this case, the expected payoffs of B, π_B, are *higher* if $a = -6$ ($\pi_B = 6$) than if $a = -5$ ($\pi_B = 1$). This means that $a = -6$ does in fact signal more friendly intentions than $a = -5$, which in turn justifies the higher punishments. Thus, it is possible in the Dufwenberg and Kirchsteiger model that $a = -5$ is punished more than $a = -6$ in equilibrium.

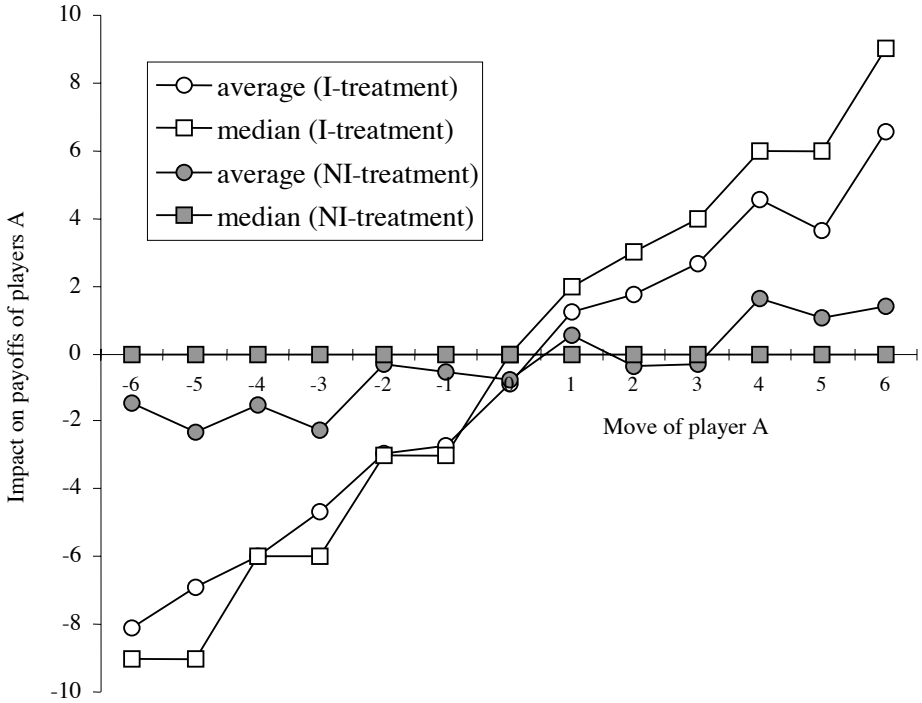
[7] See discussion in the text.

V. Results

In total 112 subjects participated in the experiment (66 in the I-treatment and 46 in the NI-treatment). All subjects were students from the University of Zurich or the Polytechnic University of Zurich, no economics students among them. The experiments were conducted in June 1998. 1 point in the experiment represented 1 Swiss Franc (CHF 1 ≈ .65 US$). Subjects received on average CHF 22.20 in the I-treatment and CHF 24.10 in the NI-treatment (including a show-up fee of CHF 10). On average, the experiment lasted 45 minutes.

Our main result is shown in Figure 1.[8] In this figure we plot the rewards and sanctions of B in both treatments, i.e., we show the impact of B's decisions on A's payoff for each possible action of A. For instance, if in the I-treatment A chooses $a = 6$ then A receives on *average* 6.55 points from B. The corresponding *median* value is 9 points.

Figure 1: Rewards and sanctions of players B dependent on decisions of players A



[8] In this section we restrict our attention to the behavior of players B. In the Appendix we also present player A decisions for the I-treatment and the random moves in the NI-treatment.

The figure reveals that in the I-treatment the Bs do reward and sanction the behavior of the As. Average and median rewards are increasing in the level of the transfer. Similarly, the more A takes away from B, the more B is willing to sanction. This behavioral pattern is in clear contradiction to the standard economic prediction ($b = 0, \forall a$). It is, however, well in line with the predictions of all fairness theories.

Figure 1 also reveals that behavior is remarkably different in the NI-treatment compared to the I-treatment. On average sanctions and rewards are much *weaker* in the NI- than in the I-treatment. Only for sufficiently high or low values of a, average sanctions and rewards differ from zero in the NI-treatment. The treatment differences between the I- and the NI-treatment are even more pronounced if we look at the median behavior. In the NI-treatment median behavior does not show any reciprocal pattern but completely coincides with the prediction of the self-interest model.

Are the differences between the I-treatment and the NI-treatment statistically significant? Table 3 provides the answer. Similarly to Figure 1, it shows the impact of B's decision on A's payoff for all moves of A. In addition to the average and median impacts, it also shows quartile values.[9] These distribution measures indicate that the reciprocal responses of the Bs are not only *on average* weaker in the NI-treatment but that the whole distribution is shifted towards zero. For example, if A chooses $a = -6$, average sanctions are lower (1.43 instead of 8.09) but so are the first quartile, the median and the third quartile values. This holds (weakly) for *all* take-decisions. Similarly if A chooses, e.g., $a = 4$ not only average rewards are lower (1.65 instead of 4.58) in the NI-treatment, but also all distribution measures. Again, this holds (weakly) for *all* give-decisions. In the last row of Table 3, we present the results of the nonparametric Mann-Whitney U-test, which was run to check whether the decisions of the Bs are significantly different across treatments. For almost all $a > 0$ and $a < 0$, behavior is indeed significantly different across treatments at the five percent level. For $a = 5$ it is significant at the ten percent level.

[9] Quartile values were determined in the following way. First all decisions of all Bs were sorted. The first quartile value represents the choice of the player B who is located at position 0.25, the median decision is that of B who is located at 0.5 and the third quartile is the decision of B at 0.75.

Table 3: Behavior of players B - Distribution measures and statistical significance

Player A's move *a*	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
I-treatment													
Average	-8.09	-6.91	-5.97	-4.70	-2.97	-2.73	-0.88	1.24	1.73	2.64	4.58	3.64	6.55
First quartile	-18	-15	-12	-9	-6	-3	0	0	0	0	3	0	1
Median	-9	-9	-6	-6	-3	-3	0	2	3	4	6	6	9
Third quartile	0	0	0	0	0	0	1	2	4	6	8	10	12
NI-treatment													
Average	-1.43	-2.35	-1.52	-2.26	-0.30	-0.57	-0.78	0.57	-0.39	-0.30	1.65	1.09	1.39
First quartile	0	-3	-3	-6	-3	-3	0	0	0	0	0	0	0
Median	0	0	0	0	0	0	0	0	0	0	0	0	0
Third quartile	0	0	0	0	0	0	0	1	2	5	5	7	8
Significance of difference between treatments*	.001	.016	.023	.025	.031	.032	.109	.032	.006	.017	.002	.069	.001

* Significance is checked by means of the non-parametric Mann-Whitney U test. Numbers are p-values. Given our hypothesis that in the NI-treatment reciprocal responses are weaker than in the I-treatment, we used a one sided test (except for the (random) move of $a = 0$, where we have no such hypothesis.

Given the results shown in Table 3 we can reject the predictions by Bolton and Ockenfels (2000) and Fehr and Schmidt (1999). Behavior in the NI-treatment is significantly different from that in the I-treatment for all give- and take-decisions. Put differently, our results indicate that on an aggregate level intentions matter both in the domain of positive as well as in the domain of negative reciprocity.

The behavioral relevance of fairness intentions can also be shown with the help of regression analysis. Table 4 shows a regression model where the impact of B's decision is regressed on A's move (a). In this model we also include a dummy variable for the I-treatment (I) and an interaction term $a \times I$. This specification allows us to estimate different linear fits for the I- and for the NI-treatment and, thus, to asses the difference between the two treatments. The result of this regression is shown in Table 4. The coefficient of the interaction term $a \times I$ is highly significant while the dummy variable for the I-treatment (I) is not significantly different from zero. This means that for $a = 0$ there is no difference between the treatments but for sufficiently high or low a, the reciprocal responses of the Bs are stronger in the I-treatment compared with the NI-treatment. This result confirms the statistical test presented in Table 3.

Table 4: Regression with impact of B's decision as dependent variable

variable	coefficient
constant	-.401 (.550)
A's decision a	.295* (.157)
dummy for I-treatment I	-.522 (1.086)
$a \times I$	.907*** (.222)

Robust standard errors are in parenthesis (subject ID as cluster variable). There are 728 observations in 56 clusters. The F-statistic equals 20.89; $p < .001$. *= significance at 10 percent level; ***= significance at 1percent level.

Furthermore, the regression allows us to test whether in the NI-treatment reciprocity is wiped out completely as predicted by Dufwenberg and Kirchsteiger (1999). Figure 1 shows that on average there are reciprocal choices for high enough give- and take-decisions, but is this reciprocal behavior significantly different from $b = 0$? The constant and the coefficient of a shown in Table 4 measure the behavior of the Bs in the NI-treatment. Notice that the constant is insignificant. If the random device determines $a = 0$, the Bs neither reward nor sanction on average. The coefficient of a is, however, positive and (weakly) significant. We thus conclude that, on average, Bs reward positive and sufficiently high a-values and sanction negative and sufficiently low a-values in the NI-treatment.[10] Thus, even though reciprocal behavior is considerably weaker in the NI-treatment, there are also significant reciprocal responses in this treatment. This suggests that reciprocal behavior is not solely intention-driven, as assumed in Dufwenberg and Kirchsteiger (1999), but that fair outcomes also play a role. Therefore, approaches that model reciprocal behavior as a response to intentions *and* material consequences, like, e.g., Falk and Fischbacher (1999), organize our results best.

So far we have restricted our analysis to aggregate behavior. However, since the Bs had to indicate a decision for each of A's possible actions, we can also study *individual patterns* of behavior.[11] The first column shows the percentage of subjects who neither reward nor sanction, i.e., whose behavior is in accordance with the standard economic prediction ($b = 0$, $\forall a$). The second column reports the percentage of subjects who reward *or* sanction. The percentage of subjects who exhibit positive as well as negative reciprocity is given in column

[10] To check the robustness of this finding, we also calculated the Spearman rank correlation between the average impact of B's decisions on A's payoff and the corresponding moves by A. The resulting coefficients are 0.8721 for the NI-treatment and 0.9945 for the I-treatment. Both coefficients are significant at any conventional level ($p < 0.001$).

[11] In the Appendix, where we show all individual decisions, we also indicate how each subject is assigned to the different behavioral categories.

3. The percentage of those who reward are listed in column 4 while the percentage of those who sanction are listed in column 5. The sixth column, finally, consists of subjects whose rewarding or sanctioning behavior is very unsystematic. Most of these subjects rewarded a particular transfer and – at the same time – sanctioned a *higher* transfer.[12]

Table 5: Individual patterns of behavior of Bs (percent)*

	Selfish	Reward or sanction	Reward and sanction	Reward	Sanction	Other patterns
I-treatment (*n=33*)	0	76	40	61	55	24
NI-treatment (*n=23*)	30	39	18	35	22	30
Significance of difference (Fisher's exact test, p-values)	.001	.005	.052	.037	.011	.607

*The classification is constructed as follows: First, all subjects who show 'other patterns' (column 5) are sorted out. This category contains (i) subjects with a negative correlation between a and b, (ii) subjects who reward a give-decision $a > 0$ while they sanctioned a *higher* give-decision $a' > a$ and (iii) subjects with an unconditional *non-zero* transfer decision. The rest of the subjects is classified into the other categories. Subjects who never reward or sanction ($b = 0$), are assigned to the first column. Subjects who reward an $a > 0$ at least once *or* sanction an $a < 0$ at least once are assigned to the second column. Subjects who reward an $a > 0$ at least once *and* sanction an $a < 0$ at least once are assigned to the third column. Subjects who reward an $a > 0$ at least once are counted in the fourth column and subjects who sanction an $a < 0$ at least once are assigned to column 5.

Table 5 shows that individual behavioral patterns between the two treatments are quite different. First notice that the percentage of choices that coincides with the prediction of the self-interest model ($b = 0, \forall a$) sharply increases in the NI-treatment (30 percent) relative to the I-treatment (zero percent). This difference suggests that a non-negligible amount of reciprocal behavior is exclusively a response to fairness intentions. Subjects who *would* reward or sanction in the I-treatment refrain from doing so (and choose $b = 0$) because the actions of the As are determined randomly and do not signal any intentions.

This interpretation is also consistent with the second result in Table 5 (see column 2): The percentage of subjects who are either positively or negatively reciprocal drops from 76 percent in the I-treatment to 39 percent in the NI-treatment. This (highly significant) difference indicates that many reciprocal players are willing to reward or sanction only if the corresponding action by A signals fair or unfair intentions. However, reciprocity in the NI-treatment is not wiped out completely. Almost 40 percent of the subjects show some reciprocal behavior. We take this evidence (which is in line with the regression results

[12] Two other subjects included in this category always indicated the exact same action (but not $b = 0$) for all possible transfers of player A. Note that (except for rounding errors) the sum of numbers in columns 1, 2 and 6 adds up to 100 percent.

presented in Table 4) as a further indication that reciprocal behavior is not solely intention-driven, but also by concerns for fair outcomes.

It is also interesting to know the fraction of subjects who exhibit both positively and negatively reciprocal behavior. Column 3 shows that 40 percent of the subjects in the I-condition and 18 percent in the NI-condition exhibit this pattern. Thus, the possibility to infer intentions also raises the percentage of subjects who exhibit both types of reciprocity significantly. Column 4 and 5 further support our previous conclusion that fairness intentions significantly increase the willingness to reciprocate and that even in the absence of fairness intentions there is a non-negligible percentage of people who reciprocate.

VI. Discussion

In his classic account of reciprocity Gouldner (1960) conjectured that the force of reciprocity depends on the *motives* imputed to the donor and the donor's *own free will*. Although this notion of reciprocity is highly suggestive it has so far been quite difficult to provide *direct* and *unambiguous* evidence for the behavioral relevance of fairness intentions. We have discussed several potential reasons for this and designed an experiment that avoids potential confounds with other sources of reciprocal behavior. Our results provide clean evidence that in judging the fairness of an action, people do not only take into account the distributive consequences of an action but also the intention that is signaled by the action. This result not only casts serious doubt on the consequentialist practice in standard economic theory that defines the utility of an action solely in terms of the consequences of this action. It also shows that the recently developed models of fairness by Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) are incomplete to the extent that they neglect "nonconsequentialist" reasons for reciprocally fair actions. Our results also show that even in an environment where actions do not signal any intentions, there is still some reciprocal behavior, i.e., the fairness of the outcome matters, too. The fact that reciprocity is not solely triggered by kind or unkind intentions implies that the pure intentions models of Rabin (1993) and Dufwenberg and Kirchsteiger (1999) are incomplete as well. Only models that combine both driving forces of reciprocal behavior (e.g., Falk and Fischbacher 1999) – are compatible with the observed behavioral patterns.

References

Abbink, K., Irlenbusch, B. and Renner, E. (2000): "The Moonlighting Game – An Experimental study on reciprocity and Retribution", Journal of Economic Behavior and Organization 42, 265-277.

Andreoni, J. and Miller, J. (2000): "Giving According to GARP: An Experimental Test of the Rationality of Altruism," Discussion paper, University of Wisconsin.

Bewley, Truman (1999): *Why Wages don't fall during a Recession*, Harvard University Press, Harvard.

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

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

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

Bolton, G. E., Brandts, J. and Ockenfels, A. (1998): "Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game", Experimental Economics 1, 207-220.

Brandts, J. and Charness, G. (1998): "Hot versus Cold: Sequential Responses and Preference Stability in Experimental Games", Discussion Paper, Universidad Autonoma de Barcelona.

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

Cason, T. and Mui, V. (1998): "Social Influence in the Sequential Dictator Game", Journal of Mathematical Psychology, forthcoming.

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

Charness, G. and Rabin M. (2000): "Social Preferences: Some Simple Tests and a New Model." *Mimeo*, University of California at Berkeley.

Cox, J. (2000): "Trust and Reciprocity: Implications of Game Triads and Social Contexts", Working paper, University of Arizona.

Dufwenberg, M. and Kirchsteiger, G. (1999): "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.

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

Fehr, E. and Gächter S. (2000): "Fairness and Retaliation – The Economics of Reciprocity", Journal of Economic Perspectives 14, 159-181.

Fischbacher, U. (1998): "z-Tree. Zurich Toolbox for Readymade Economic Experiments", Working paper No. 21, University of Zurich.

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

Huang, P. H. (2000): "Reasons within Passions: Emotions and Intentions in Property Rights Bargaining", Oregon Law Review, forthcoming.

Kahneman, Daniel and Knetsch, Jack L. and Thaler, Richard (1986): "Fairness as a Constraint on Profit Seeking: Entitlements in the Market", American Economic Review 76, 728-41.

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

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

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

Appendix

Table A1: Decisions of players A in the I-treatment

	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
Number of A-players	4	2	2	3	1	0	3	2	5	2	0	1	8
Percentage of A-players	12	6	6	9	3	0	9	6	15	6	0	3	24

Table A2: Random moves of players A in the NI-treatment

	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
Actual number of random moves	2	0	0	0	1	1	2	2	5	1	2	1	6
Percentage of random moves	9	0	0	0	4	4	9	9	22	4	9	4	26

Table A3: Individual data of players B in the I- and the NI-treatment

Treatment	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	reward	sanction	never reward nor punish	other patterns
I	4	2	0	9	7	8	2	4	4	5	7	6	10	x			
I	0	0	0	0	0	0	0	1	2	3	4	5	6	x			
I	0	0	0	0	0	0	0	1	2	4	6	7	9	x			
I	0	0	0	0	0	0	0	2	4	6	8	10	12	x			
I	0	1	2	3	4	5	6	7	8	9	10	11	12	x			
I	0	0	0	0	0	0	0	2	3	7	8	10	12	x			
I	0	0	1	1	2	2	2	3	6	6	12	12	18	x			
I	-3	-2	-2	-1	-1	-4	1	1	5	6	7	5	1	x	x		
I	-6	-2	-3	-2	0	-1	0	0	0	1	4	7	10	x	x		
I	-4	-4	-4	-4	-4	-4	-1	1	2	2	3	5	6	x	x		
I	-5	-4	-3	-3	-2	-1	0	2	3	4	4	5	5	x	x		
I	-6	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	x	x		
I	-6	-5	-5	-2	-2	-1	0	2	3	5	7	8	8	x	x		
I	-2	-2	-2	-1	-1	-1	0	2	4	5	8	9	12	x	x		
I	-5	-4	-4	-4	-1	-1	0	2	4	6	8	9	11	x	x		
I	-4	-4	-4	-3	-2	-1	0	2	4	6	8	10	12	x	x		
I	-6	-5	-5	-5	-1	-1	0	2	3	6	8	10	10	x	x		
I	-4	-4	-4	-3	-3	-2	3	4	5	7	9	11	13	x	x		
I	-6	-6	-6	-5	-4	-3	0	2	4	6	8	10	12	x	x		
I	-6	-5	-4	-3	-2	-1	0	2	4	6	8	10	12	x	x		
I	-6	-5	-1	-1	-2	-2	0	0	0	0	0	0	0		x		
I	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0		x		
I	-3	-3	-2	-2	-1	-1	0	0	0	0	0	0	0		x		
I	-5	-5	-4	-3	-2	-1	0	0	0	0	0	0	0		x		
I	-6	-6	0	0	0	0	0	0	0	0	0	0	0		x		
I	-1	17	-4	18	-3	2	-6	18	18	18	18	-6	18				x
I	3	-4	4	-5	3	-6	1	0	-6	4	6	-2	18				x
I	-5	5	1	-4	6	0	4	3	2	-3	2	14	1				x
I	3	-2	1	-1	0	5	2	0	3	-2	3	-4	3				x
I	0	3	1	0	0	2	-2	0	3	0	4	0	2				x
I	-1	-6	-5	-4	-4	-3	-6	-6	-6	-6	-6	-6	-6				x
I	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6				x
I	12	12	12	12	12	12	12	12	12	12	12	12	12				x
NI	0	0	0	0	0	0	0	0	1	5	6	7	8	x			
NI	0	0	0	0	0	0	0	1	2	3	4	6	8	x			
NI	0	0	0	0	0	0	0	2	5	6	6	7	8	x			
NI	0	0	0	0	0	0	0	1	2	4	6	8	10	x			
NI	-2	-6	-4	-3	-3	-1	0	0	2	6	4	9	0	x	x		
NI	0	-1	-1	-2	-1	-1	0	5	3	6	7	12	15	x	x		
NI	-1	-2	-3	-1	-2	-1	-1	1	2	5	7	9	11	x	x		
NI	-2	-1	-1	-1	0	0	0	1	2	4	5	7	8	x	x		
NI	-3	-3	-3	-2	-2	-1	-1	0	0	0	0	0	0		x		
NI	0	0	0	0	0	0	0	0	0	0	0	0	0			x	
NI	0	0	0	0	0	0	0	0	0	0	0	0	0			x	
NI	0	0	0	0	0	0	0	0	0	0	0	0	0			x	
NI	0	0	0	0	0	0	0	0	0	0	0	0	0			x	
NI	0	0	0	0	0	0	0	0	0	0	0	0	0			x	
NI	0	0	0	0	0	0	0	0	0	0	0	0	0			x	
NI	0	0	0	0	0	0	0	0	0	0	0	0	0			x	
NI	-4	-6	8	-4	12	4	-2	10	-4	-6	2	-4	-6				x
NI	1	8	-1	-4	6	2	0	-2	5	3	1	-5	0				x
NI	0	3	4	1	6	3	0	1	2	-2	0	15	0				x
NI	0	-3	-1	-2	1	2	3	4	-4	-5	5	-6	0				x
NI	0	0	-2	3	-1	-1	0	2	4	5	0	5	0				x
NI	2	1	1	0	-2	-3	-3	-3	-4	-5	-5	-5	-6				x
NI	0	0	0	1	1	0	0	0	-1	0	0	0	0				x