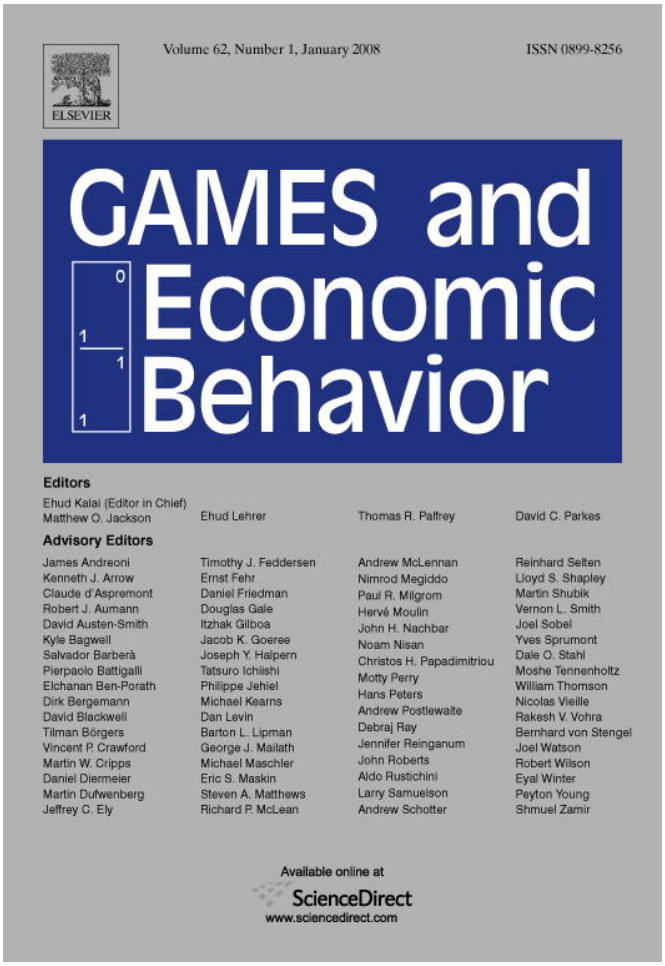

Available online at www.sciencedirect.com
ScienceDirect
Games and Economic Behavior 62 (2008) 287–303
GAMES and Economic Behavior
www.elsevier.com/locate/geb

Testing theories of fairness—Intentions matter

Armin Falk [a,b,1], Ernst Fehr [c,2], Urs Fischbacher [c,*]

[a] *IZA, Bonn, Germany*
[b] *University of Bonn, Bonn, Germany*
[c] *University of Zurich, Zurich, Switzerland*

Received 23 July 2003

Available online 16 June 2007

Abstract

Recently developed models of fairness can explain a wide variety of seemingly contradictory facts. One of the most controversial and yet unresolved issues 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 but there is surprisingly little direct evidence for its behavioral importance. We provide experimental evidence for the behavioral relevance of fairness intentions in this paper. Our main result indicates that the attribution of fairness intentions is important in both the domains of negatively and positively reciprocal behavior. This means that equity models exclusively based on preferences over the distribution of material payoffs cannot capture reciprocal behavior. Models that take players' fairness intentions and distributional preferences into account 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

* Corresponding author at: Institute for Empirical Research in Economics, Blümlisalpstrasse 10, CH-8006 Zurich, Switzerland.

E-mail addresses: falk@iza.org (A. Falk), efehr@iew.uzh.ch (E. Fehr), fiba@iew.uzh.ch (U. Fischbacher).

[1] Postal address: Institute for the Study of Labor and University of Bonn, Schaumburg-Lippe Strasse 7/9, 53113 Bonn, Germany.

[2] Postal address: Institute for Empirical Research in Economics, Blümlisalpstrasse 10, CH-8006 Zurich, Switzerland.

0899-8256/$ – see front matter
doi:10.1016/j.geb.2007.06.001

1. Introduction

A considerable body of evidence indicates that concerns for fairness and reciprocity motivate a substantial number of people. Moreover, the presence of fair-minded people is likely to have important economic effects (Kahneman et al., 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, 2004; Falk and Fischbacher, 2006; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002). 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 a wide variety of seemingly contradictory facts correctly.

This paper examines the most controversial question in the modeling of fairness preferences: the role of *fairness intentions*.[3] 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 inequity aversion 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, 2006; Dufwenberg and Kirchsteiger, 2004) 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 not only concerns the proper modeling of fairness preferences, but also standard utility theory as well. Standard utility theory assumes that the utility of an action depends solely on its consequences 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. The issue is important at the practical level 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 some parties' material payoffs negatively. If the response of the negatively affected parties also takes the decision-maker's fairness intentions into account, 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, obvious alternative actions are 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 its underlying intentions.

Gouldner (1960) points out the importance of intentions in his classic account of reciprocity by conjecturing that the force of reciprocity depends on the *motives* imputed to the donor and

[3] This paper suffered from serious editorial delays at different journals. An early version of our results can be found in Falk et al. (2000).

the donor's *own free will*. Although this notion of reciprocity is highly suggestive, providing *direct* and *unambiguous* evidence for the behavioral relevance of fairness intentions has proven very difficult up to now. This is not surprising with regard to field data because outcomes and intentions are usually inextricably intertwined in the field. Yet, the issue has been quite elusive, even in laboratory experiments. Charness (2004), Bolton et al. (1998), Offerman (2002) and Cox (2004) find little or no evidence that the attribution of fairness intentions matters in the domain of positively reciprocal behavior.[4] Blount (1995) and Offerman (2002) find evidence that it matters in the domain of negatively reciprocal behavior but, as we will argue below, these studies have some methodological problems. Thus, we face the puzzle that, intuitively, the attribution of fairness intentions seems to be important while, the issue remains controversial in light of the prevailing evidence.

We provide experimental evidence for the behavioral relevance of fairness attributions in this paper. Our main result is that the attribution of fairness intentions is important in both the domains of negatively and positively reciprocal behavior. When the experimental design rules out the attribution of fairness intentions, reciprocal responses are substantially weaker. This result is corroborated both at the individual as well as at the aggregate level. Not only do 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 inequity aversion 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 weakly 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, 2004) are also incomplete. Models that combine both aspects (like e.g. Falk and Fischbacher, 2006) fit our data best.

Our experimental design also provides an opportunity for examining the extent to which the *same* individuals exhibit both positive and negative reciprocity. To our knowledge, no study 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 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 discusses potential obstacles in finding behavioral effects of fairness intentions. Section 3 presents the experimental design. Section 4 discusses the predictions of several fairness theories. Results are presented in Section 5. Section 6 provides a short discussion.

[4] 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 unfair intentions. Note that, for convenience, we also call a response to a random event a reciprocal action.

2. Obstacles for finding behavioral effects of fairness intentions

Before presenting our experimental design, we discuss the potential reasons for the lack of convincing evidence in favor of fairness intentions; four potential reasons exist in our opinion. The first reason is that a potential confound with the efficiency motive exists in some studies. Andreoni and Miller (2002), Bolle and Kritikos (2001), Charness and Rabin (2002) and Engelmann and Strobel (2004) report results suggesting the presence of a non-negligible fraction of subjects 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 (2004), Bolton et al. (1998) and Offerman (2002) because the second mover's reciprocal behavior was associated with large efficiency increases in these studies. It is also possible that reciprocity motives and efficiency motives interact in a yet unknown way. For this reason, our design rules out an increase in the total payoff due to reciprocal responses.

A second reason is related to the issue of repetition. Subjects faced a different opponent in each of ten periods in Charness (2004). 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. Ideally, two treatments are needed to isolate the role of fairness intentions, one where first-movers can signal their fairness intentions, and one where 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 first mover's *full* control.

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. We solved this problem in our design 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-mover's 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 (2004), Bolton et al. (1998) and Cox (2004) 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. We achieved this in our experiment by determining the first mover's "action" with a salient random device. Saliency was implemented by rolling dice in front of each second mover. However, if a random device determines the first mover's choice, the second movers might have views about what constitutes fair or unfair random devices. For example, if the random device determines a very bad outcome for the second mover with high probability, 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 a confound due to the angry response to an unfair random device is likely to exist. Our solution to this problem is to implement a random device that mimics the probability distribution over the actions of human first movers.

A random device determined the first mover's "action" both in Blount (1995) and Offerman (2002).[5] However, only Blount kept the probability distribution of first mover actions constant across the random device and the human choice conditions. Although Blount's study 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 knew 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 the responders also had to put themselves in the shoes of the proposers in the human choice condition, while this was not the case in the random device condition.

Deception was involved in one of Blount's treatments. Subjects believed that there were proposers although the experimenters actually made the proposals. All subjects in this condition were "randomly" assigned to the responder role. It could well be 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 both the first-mover position and in the second-mover position.

3. Experimental design and procedures

Our experimental design is based on the "moonlighting game" (Abbink et al., 2000). This game has the advantage that we can examine the impact of fairness intentions on both positively and negatively reciprocal responses at the individual level. We first describe the moonlighting game below and then present our two treatments, the Intention treatment and the No-Intention treatment. Finally, we report the procedures of the experiment.

3.1. 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. Player A chooses an action $a \in \{-6, -5, \ldots, 5, 6\}$ in the *first stage*. If A chooses $a \geqslant 0$, he gives player B a tokens while if he chooses $a < 0$, he takes $|a|$ tokens away from B. In case of $a \geqslant 0$, 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 $b \geqslant 0$ is a reward and $b < 0$ is a sanction. A reward transfers b points from B to A. A sanction costs B exactly $|b|$ but reduces A's income by $3|b|$. Final incomes are determined after B's decision.

[5] Kagel and Wolfe (2001) suggest yet another way for studying the role of intention. They show that—in contrast to the prediction by inequity aversion models—responders reject offers even though this favors or harms a third party, resulting in inequity. However, since this third party takes no decision, this inequity is not considered as intentional and is therefore accepted.

Since As can give and take while Bs can reward or sanction, this game allows for both positively and negatively reciprocal behavior.

We applied the strategy method in our experiment. This means that player B had to give us a response for each feasible action of player A, before B was informed about A's actual choice. This has several advantages.[6] 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 only exhibit either negatively reciprocal responses or positively reciprocal responses, or whether some exhibit both types of reciprocity. Second, 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.

3.2. 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 under his full control.

Our experimental game guarantees condition (i), since it allows A to give or to take different amounts of money. Condition (ii) is our treatment variable. A himself determines a in the *Intention treatment* (I-treatment), thus making him responsible for the consequences of his action; his action therefore signals intentional kindness (if a is high) or intentional unkindness (if a is low). In contrast, a random device determines A's move in the *No-intention treatment* (NI-treatment). Consequently, A has no control over his action. His action therefore signals neither good nor bad intentions.

A's random move in the NI-treatment was implemented as follows: after B had determined her strategy, the experimenter went to her place and cast 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 cast was then used to determine A's move 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 58, for example, player A's move was $a = 3$. The experimenter entered the respective "choice". After A's move had been determined, the experimenter went to another player B and cast the dice again. This procedure was explained to Bs in great detail in the instructions.[7] 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 did not know the probability distribution.

[6] In principle, the strategy method could induce a different behavior of B relative to a situation where B has to respond to A's actual move. In fact, Güth et al. (2001) report an experiment in which they observe behavioral differences between the strategy method and the specific response method. On the other hand, Brandts and Charness (2000) and Cason and Mui (1998) report evidence indicating that the strategy method does not induce different behavior. It is important to note that we used the strategy method in both of our treatments. Therefore, our results are biased only if the impact of this method differs across treatments. A referee pointed out to us that this could be the case in our experiment: if the emotional response to an outcome is independent of the method (strategy method or specific response method) when a *person* makes the offer, but, subjects only respond emotionally to an (unfair) offer in the case of the random treatment if they actually experience it, i.e., if the specific response method is used and not the strategy method, then our method would overestimate the intention effect.

[7] The instructions are available at http://www.iew.uzh.ch/home/fischbacher/download/fafefi_test_theories_instr.pdf.

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

Notice that the randomly determined moves of A according to Table 1 are not equally likely. For example, a random selection of $a = -6$ (7 percent chance) is more likely than $a = -5$ (2 percent chance). Table 1 reflects the *actual human decisions* of the As who participated in the moonlighting experiment by Abbink et al. (2000). This table permitted us to approximate a "human choice distribution" even in the NI-treatment, where choices were random. The choice distribution given in Table 1 was also presented to Bs in the I-condition, in order to keep everything constant with the exception of the potential for the attributions of intentions across the NI- and the I-treatments.[8] The Bs informed in the I-treatment that the same experiment had already been conducted and that the relative frequency of As' decisions in that experiment was identical to those in Table 1. This was done to induce players B to have the same beliefs about As' choice distribution in both treatment conditions. This procedure ruled out the possibility that different beliefs about the choice distribution of the As affected the Bs' responses.[9] Thus, the two treatments cannot evoke different fairness judgments with regard to the probability distribution over the set of feasible actions.[10]

3.3. 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. All subjects had to answer several control questions to ensure the understanding of the experimental procedures; the experiment did not start until all subjects had answered all questions correctly. All the players knew both procedures and payoff functions, i.e., they were explained in the instructions and summarized orally. Losses were possible, and subjects had to cover them with

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

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

[10] The importance of using a human choice distribution is justified in light of the evidence of Bolton et al. (2005). They implemented a random move procedure in which they varied the probability distributions and showed that the distributions had an impact on perceived fairness.

the show-up fee in case they occurred. We used the experimental software z-Tree (Fischbacher, 2007) to run the experiments.

4. Predictions

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

4.1. Self-interest prediction

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

4.2. Fairness predictions

We now turn to the predictions of recently developed fairness theories and focus on player B's behavior. 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 those of player j, he aims at reducing the payoff difference between himself and j. Both models predict that people exhibit reciprocal behavior for sufficiently strong inequity aversion, i.e., b is increasing in a and $b = 0$ if $a = 0$. Both approaches neglect intentions; only the payoff consequences are assumed to explain reciprocal responses. This implies for our experiment that reciprocal responses between the I-treatment and the NI-treatment should be *exactly the same* for a given move of A. 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 kind or unkind *intentions exclusively* trigger reciprocal responses (Dufwenberg and Kirchsteiger, 2004).[11] It immediately follows from this premise that there should be no reciprocal behavior at all in the absence of fairness intentions, i.e., player B neither rewards nor punishes but pursues her material self-interest. Therefore, Dufwenberg and Kirchsteiger predict *no* reciprocal behavior at all in the NI-treatment ($b = 0,\ \forall a$). The prediction for the I-treatment is less clear because the model exhibits multiple equilibria, and some of them are compatible with b being locally decreasing in a. This is

[11] 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.

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[a]	$b = 0,\ \forall a$
Payoff consequences and fairness intentions matter (Falk/Fischbacher)	b increases in a	b increases in a but *less* than in the I-treatment

[a] See discussion in the text.

so because according to the model, higher values of a do not necessarily signal more friendly intentions.[12] 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, the model by Falk and Fischbacher (2006) combines a concern for a fair distribution of payoffs with the reward and punishment of fair and unfair intentions. The model makes the (unique) prediction in the I-treatment that b is increasing in a. This reciprocal pattern is predicted to be *weaker* in the NI-treatment than in 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.

5. Results

A total of 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 Swiss Federal Institute of Technology in Zurich, no economics students among them. The experiments were conducted in June 1998. 1 point in the experiment represented 1 Swiss Franc (CHF $1 \approx .65$ 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 Fig. 1,[13] where 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 of his possible ac-

[12] 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. This discussion shows that an appropriate evaluation of the Dufwenberg–Kirchsteiger model requires the elicitation of (higher order) beliefs. While we have not done this in this study, other studies have used the measurement of higher order beliefs to test theories based on psychological game theory (Dufwenberg and Gneezy, 2000; Charness and Dufwenberg, 2003).

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

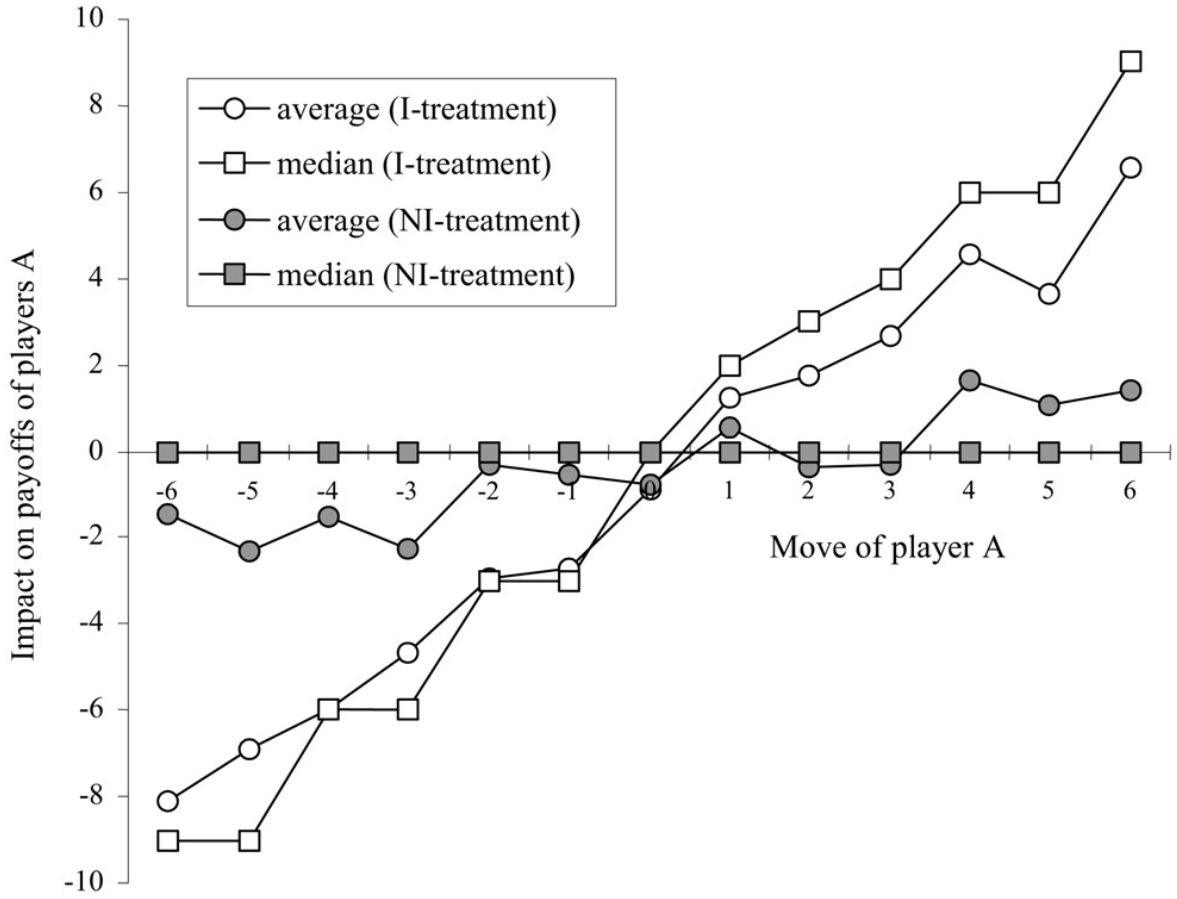

Fig. 1. Rewards and sanctions of players B dependent on decisions of players A.

tions. For instance, if A chooses $a = 6$ in the I-treatment, then he receives on *average* 6.55 points from B (see rightmost unfilled circle). The corresponding *median* value is 9 points.

The figure reveals that Bs the behavior of A's in the I-treatment. 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 differs remarkably 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. Average sanctions and rewards only differ from zero for sufficiently high or low values of a in the NI-treatment. The treatment differences between the I- and the NI-treatments are even more pronounced if we look at the median behavior. Median behavior does not show any reciprocal pattern in the NI-treatment, but completely coincides with the prediction of the self-interest model.

Are the differences between the I- and the NI-treatments statistically significant? Table 3 provides the answer. As in Fig. 1, it shows the impact of B's decision on A's payoff for all of A's moves. In addition to the average and median impacts, it also shows quartile values. These distribution measures indicate that Bs' reciprocal responses are not only weaker *on average* in the NI-treatment, but that the whole distribution is shifted towards zero. For example, if A chooses $a = -6$, both average sanctions are lower (1.43 instead of 8.09), as are the first quartile and the median values. This holds (weakly) for *all* "take" decisions. Similarly if A chooses $a = 6$, for example, not only are average rewards lower (1.39 instead of 6.55) in the NI-treatment, but all

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	−.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	−.30	−.57	−.78	.57	−.39	−.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[a]	.001	.016	.023	.025	.031	.032	.109	.032	.006	.017	.002	.069	.001

[a] Significance is checked by means of the non-parametric Mann–Whitney U-test. Numbers are p-values. Given our hypothesis that reciprocal responses are weaker in the NI-treatment 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.

distribution measures as well. Again, this holds (weakly) for *all* "give" decisions. We present the results of the nonparametric Mann–Whitney U-test, which was run to check whether the decisions of the Bs different across treatments, in the last row of Table 3. Behavior is indeed significantly different across treatments at the five percent level for almost all $a > 0$ and $a < 0$; it is significant at the ten percent level for $a = 5$.

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 intentions matter on an aggregate level 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 the results of a regression model where the impact of B's decision is regressed on A's move (a). We also include a dummy variable for the I-treatment (I) and an interaction term $a \times I$ in this model. This specification allows us to estimate different linear fits for the I- and the NI-treatments 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) does not differ significantly from zero. This means that while there is no difference between the treatments for $a = 0$, the reciprocal responses of the Bs are stronger for sufficiently high or low a in the I-treatment compared with the NI-treatment. This result confirms the statistical test presented in Table 3.

Furthermore, the regression allows us to test whether reciprocity is completely eradicated in the NI-treatment, as predicted by Dufwenberg and Kirchsteiger (2004). Figure 1 shows that there are weakly reciprocal choices for high enough "give" and "take" decisions on average, but does this reciprocal behavior differ significantly 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 $a = 0$, 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. To check the robustness of this finding, we also calculated the

Table 4
Regression with impact of B's decision as a 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 parentheses (subject ID as cluster variable). There are 728 observations in 56 clusters. The F-statistic equals 20.89; $p < .001$.

* Significance at the 10% level.
*** Significance at the 1% level.

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 ($p < 0.001$). Thus, significant reciprocal responses occur in the NI-treatment, even though reciprocal behavior is considerably weaker in this case. Although rewards and punishments are quantitatively small, the results suggest that reciprocal behavior is not solely intention-driven, but that fair outcomes also play a role.

We have restricted our analysis to aggregate behavior up to now. However, since Bs had to indicate a decision for each of A's possible actions, we can also study *individual patterns* of behavior.[14] The first column in Table 5 shows the percentage of subjects who neither reward nor sanction, i.e., whose behavior follows 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 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.[15]

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 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

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

[15] 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. Note that while the behavior classified as "other patterns" seems quite unsystematic, it is in principle possible that these subjects act according to a reciprocity model in the spirit of Dufwenberg and Kirchsteiger holding non-monotonous beliefs.

Table 5
Individual patterns of behavior of Bs (percent)[a]

	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

[a] The classification is constructed as follows: First, all subjects who show "other patterns" (column 6) are sorted out. This category contains (i) subjects with a negative correlation between a and b, (ii) subjects who reward a decision a while sanctioning 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.

that many reciprocal players are only willing to reward or sanction if the corresponding action by A signals fair or unfair intentions. However, reciprocity in the NI-treatment is not completely eradicated. Almost 40 percent of the subjects show some reciprocal behavior. We take this evidence (which is in line with the regression results presented in Table 4) as a further indication that reciprocal behavior is not solely intention-driven, but also by concerns for fair outcomes. The fraction of subjects who exhibit both positively and negatively reciprocal behavior is also of interest. 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 of inferring intentions also raises the percentage of subjects who exhibit both types of reciprocity significantly. Columns 4 and 5 further support our previous conclusion that fairness intentions significantly increase the willingness to reciprocate and that there is a non-negligible percentage of people who reciprocate, even in the absence of fairness intentions.

A final observation from Table 5 is worth noting. According to many fairness models, a person should either not respond reciprocally at all or show both positively *and* negatively reciprocal behavior. Fehr and Schmidt, for example, assume that the disutility arising from a disadvantageous inequality is at least as strong as the disutility that arises from an advantageous inequality. This implies that a player who rewards should also punish. Similarly, most reciprocity models assume a single parameter for both positive and negative reciprocity. This implies that a player who rewards also punishes and vice versa. In contrast to these predictions, Table 5 reveals that 21 percent of the subjects in the I-condition reward but do not punish. Likewise, 15 percent of the subjects punish but do not reward.

6. Discussion

Although the behavioral relevance of intention is very intuitive, it has been quite difficult to provide clean evidence for the behavioral relevance of fairness intentions up to now. 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 evidence that people not only take the distributive consequences of an action but also the intention it signals into account

when judging the fairness of an action.[16] This result casts serious doubt on the consequentialist practice in standard economic theory that defines utility of an action solely in terms of its consequences; it further shows that the 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.

Different approaches have been proposed for incorporating intentions into fairness models. Rabin (1993), Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006), and Cox et al. (2004) consider the choice set of a player and infer the intention of a particular choice from the set of possible alternatives. If, for example, there is no choice at all—as in our random treatment—no intention can be inferred from a particular move. If, however, a player actually has the choice between kind and unkind actions, the choice of a kind action allows inferring kind intentions and vice versa. Falk et al. (2003) conducted four mini-ultimatum games to directly test whether choice sets actually matter. In their experiment, one allocation x remains constant (8 points for the proposer and 2 for the responder) in all four games, while the allocation y (the "alternative" to x) differs from game to game. Although the outcome of the allocation x was constant, the rejection rate of this allocation varied depending on the available alternatives. It was highest (44%) when a fair alternative (5, 5) was available and lowest (9%) when the alternative was even more unfair (10 for the proposer, 0 for the responder). Brandts and Sola (2001) found a similar result, also showing that the choice set determines the perception of fairness of an outcome, as predicted by the models mentioned above.

The reciprocity models explain the difference between the I- and NI-treatment. However, we also observe that there is reciprocity even in an environment where actions do not signal any intention. Thus, the fairness of the outcome matters as well. This implies that the pure intention models of Rabin (1993) and Dufwenberg and Kirchsteiger (2004) are also incomplete; unlike reciprocity models that combine intentions with distributional concerns, such as Falk and Fischbacher (2006).

Levine (1998), and Charness and Rabin (2002) choose another approach for incorporating intentions. In these models, the players differ in an individual parameter—the player's type. This type measures the player's kindness. The chosen alternative allows estimating this parameter. More kind players choose more kind offers and therefore the estimate of this parameter can be interpreted as the player's intention. If player 1 chooses to take 6 points in our experiment, for example, one can infer that he is a rather unkind type, while nothing can be concluded about player 1's type in case of a random first move. Since players base their reciprocation on the assessment of the other player's type, these models predict the main difference between the I- and the NI-condition under reasonable assumptions. These models, however, fail to explain the existence of players who do both, reward and punish in the NI condition.

[16] An interesting application in the context of labor relations is whether *incompetence* is viewed and punished similarly to unfair intentions. We think that the answer to this question depends on whether an agent has the capability of delivering a good outcome but is just not trying hard, or whether he is actually incapable of providing a good outcome. In the former case, there should be reciprocal punishment since incompetent results are likely to be caused by laziness (which is a form of unkindness) and bad intentions. In the latter case, however, punishments should be less pronounced since a bad outcome can neither be attributed to laziness nor bad intentions.

Acknowledgments

This paper is part of the Research Priority Program on the "Foundations of Human Social Behavior" at the University of Zurich. We thank Simon Gächter and Chiara Gulfi for valuable comments.

Appendix A

Table A.1
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 A.2
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 A.3
Individual data of players *B* in the I- and the NI-treatment (see Table 5 for explanations)

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		

(*continued on next page*)

Table A.3 (*continued*)

Treatment	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	Reward	Sanction	Never reward nor punish	Other patterns
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

References

Abbink, K., Irlenbusch, B., Renner, E., 2000. The moonlighting game—An experimental study on reciprocity and retribution. J. Econ. Behav. Organ. 42, 265–277.

Andreoni, J., Miller, J., 2002. Giving according to GARP: An experimental test of the consistency of preferences for altruism. Econometrica 70, 737–753.

Bewley, Truman, 1999. Why Wages don't Fall during a Recession. Harvard Univ. Press, Harvard.

Blount, S., 1995. When social outcomes aren't fair: The effect of causal attributions on preferences. Organ. Behav. Human Dec. Proc. 63, 131–144.

Bolle, F., Kritikos, A., 2001. Distributional concerns: Equity- or efficiency-oriented? Econ. Lett. 73, 333–338.

Bolton, G., Ockenfels, A., 2000. ERC—A theory of equity, reciprocity and competition. Amer. Econ. Rev. 90, 166–193.

Bolton, G.E., Brandts, J., Ockenfels, A., 1998. Measuring motivations for the reciprocal responses observed in a simple dilemma game. Exper. Econ. 1, 207–220.

Bolton, G.E., Brandts, J., Ockenfels, A., 2005. Fair procedures: Evidence from games involving lotteries. Econ. J. 115, 1054–1076.

Brandts, J., Charness, G., 2000. Hot and cold decisions and reciprocity in experiments with sequential games. Exper. Econ. 2, 227–238.

Brandts, J., Solà, C., 2001. Reference points and negative reciprocity in simple sequential games. Games Econ. Behav. 36, 138–157.
Camerer, C., Thaler, R., 1995. Ultimatums, dictators, and manners. J. Econ. Perspect. 9, 209–219.
Cason, T., Mui, V., 1998. Social influence in the sequential dictator game. J. Math. Psych. 42, 248–465.
Charness, G., 2004. Attribution and reciprocity in an experimental labor market. J. Lab. Econ. 22, 665–688.
Charness, G., Dufwenberg, M., 2003. Promises & partnership. Working papers in Economics 2003-3. Stockholm University.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Quart. J. Econ. 117, 817–869.
Cox, J., 2004. How to identify trust and reciprocity. Games Econ. Behav. 46, 260–281.
Cox, J., Friedman, D., Gjerstad, S., 2004. A tractable model of reciprocity and fairness. Working paper. University of Arizona.
Dufwenberg, M., Gneezy, U., 2000. Measuring beliefs in an experimental lost wallet game. Games Econ. Behav. 30, 163–182.
Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games Econ. Behav. 47, 268–298.
Engelmann, D., Strobel, M., 2004. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. Amer. Econ. Rev. 94 (4), 857–869.
Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games Econ. Behav. 54 (2), 293–315.
Falk, A., Fehr, E., Fischbacher, U., 2000. Testing theories of fairness—Intentions matter. Working paper No. 63. Institute for Empirical Research in Economics, University of Zurich.
Falk, A., Fehr, E., Fischbacher, U., 2003. On the nature of fair behavior. Econ. Inquiry 41 (1), 20–26.
Fehr, E., Gächter, S., 2000. Fairness and retaliation—The economics of reciprocity. J. Econ. Perspect. 14, 159–181.
Fehr, E., Schmidt, K., 1999. A theory of fairness, competition, and cooperation. Quart. J. Econ. 114, 817–868.
Fischbacher, U., 2007. z-Tree: Zurich toolbox for readymade economic experiments. Exper. Econ. 10 (2), 171–178.
Gouldner, A., 1960. The norm of reciprocity. Amer. Sociological Rev. 25, 161–178.
Güth, W., Huck, S., Muller, W., 2001. The relevance of equal splits in ultimatum games. Games Econ. Behav. 37 (1), 161–169.
Huang, P.H., 2000. Reasons within passions: Emotions and intentions in property rights bargaining. Oregon Law Rev. 79, 435–478.
Kagel, J., Wolfe, K., 2001. Tests of fairness models based on equity considerations in a three-person ultimatum game. Exper. Econ. 4, 203–220.
Kahneman, D., Knetsch, J.L., Thaler, R., 1986. Fairness as a constraint on profit seeking: Entitlements in the market. Amer. Econ. Rev. 76, 728–741.
Levine, D., 1998. Modeling altruism and spitefulness in experiments. Rev. Econ. Dynam. 1, 593–622.
Offerman, T., 2002. Hurting hurts more than helping helps. Europ. Econ. Rev. 46, 1423–1437.
Rabin, M., 1993. Incorporating fairness into game theory and economics. Amer. Econ. Rev. 83, 1281–1302.