Journal of Economic Behavior & Organization 84 (2012) 892–905

Contents lists available at SciVerse ScienceDirect

Journal of Economic Behavior & Organization

journal homepage: www.elsevier.com/locate/jebo

# On the nature of reciprocity: Evidence from the ultimatum reciprocity measure[☆]

Andreas Nicklisch[a,*], Irenaeus Wolff[b,c]

[a] University of Hamburg & Max Planck Institute for Research on Collective Goods, Bonn, Germany
[b] University of Konstanz, Germany
[c] Thurgau Institute of Economics (TWI), Kreuzlingen, Switzerland

ARTICLE INFO

*Article history:*
Received 17 February 2009
Received in revised form 19 October 2012
Accepted 22 October 2012
Available online 1 November 2012

*JEL classification:*
C91
D03
D63

*Keywords:*
Distributional fairness
Experiments
Intention-based fairness
Reciprocity
Ultimatum bargaining

ABSTRACT

We experimentally show that current models of reciprocity are incomplete in a systematic way using a new variant of the ultimatum game that provides second-movers with a marginal-cost-free punishment option. For a substantial proportion of the population, the degree of first-mover unkindness determines the severity of punishment actions even when marginal costs are absent. The proportion of these participants strongly depends on a treatment variation: higher fixed costs of punishment more frequently lead to extreme responses. The fractions of purely selfish and inequity-averse participants are small and stable. Among the variety of reciprocity models, only one accommodates (rather than predicts) parts of our findings. We discuss ways of incorporating our findings into the existing models.

## 1. Introduction

Despite a tradition of research on reciprocal behavior that spans almost three decades, the development of theories of reciprocal behavior still is far from complete. One indication is that there has been a proliferation of reciprocity models (e.g., Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Sobel, 2005; Falk and Fischbacher, 2006; Cox et al., 2007) that all seem to fit specific situations better than others, and yet there is no clear indication of which model to choose in what situation. In his 2005 review article, Sobel criticizes the existing models of reciprocal behavior for presenting a utility function of others' and own income without providing an explanation for how much weight players are likely to put on others' income relative to their own. More specifically, all of the models posit that the harshness of a reaction to an unkind action is determined by the trade-off between a reduction in the other player's payoff and the costs of punishment. For costs of punishment that are sufficiently low, these models therefore predict the harshest-possible reaction to even the slightest degree of unkindness. We

[☆] We gratefully acknowledge the many useful comments provided by Christoph Engel, Bernd Irlenbusch, Hans-Theo Normann, the TWI research group, an associate editor, two anonymous referees, and the participants of the Economic Science Association meeting 2011 in Luxemburg, the 4th Maastricht Behavioral and Experimental Economics Symposium 2011, and the European Economic Association conference 2011 in Oslo. We are deeply indebted to Jörg Oechssler for pushing us in the right direction, and the Max Planck Society for financial support.

[*] Corresponding author at: School of Business, Economics and Social Science, University of Hamburg, von-Melle-Park 5, 20146 Hamburg, Germany. Tel.: +49 40 428388976; fax: +49 40 428386713.
*E-mail addresses:* andreas.nicklisch@wiso.uni-hamburg.de, nicklisch@coll.mpg.de (A. Nicklisch), wolff@twi-kreuzlingen.ch (I. Wolff).

0167-2681/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jebo.2012.10.009

argue – and show empirically – that this is wrong. However, as long as the marginal costs of punishment are strictly positive, it is impossible to falsify the above-mentioned models along these lines: it is always possible to adjust the reciprocation parameters such as to accommodate the data, given the reciprocation-parameter distribution is left unspecified in the model expositions. This substantiates a second criticism Sobel (2005, p. 407) expresses, namely that the ability of intention-based models of reciprocity to account for experimental results is "a tribute to their flexibility rather than actual support for the formulation." To corroborate the argument, we introduce the *ultimatum reciprocity measure* which eliminates the marginal costs of punishment altogether. Our experimental data show that a substantial proportion of the population deviates from the models' extreme predition in a systematic way, providing valuable insights into how existing models need to be amended.

In a recent contribution, Cox et al. (2008) abandon the domain of explicit functional forms and make a first step to address Sobel's (2005) first criticism. Our experiment suggests that their model may be an important step forward, being able to accommodate 27–47% of our observations in addition to what can be explained using the more conventional models. Nevertheless, the model still is prone to Sobel's second criticism of a lack of specificity: as we discuss in Section 3, the model accommodates rather than predicts our observations. The ways in which it fails on the specificity domain will provide guidance with respect to the direction in which to refine the model.

Another question that has attracted increased attention in the recent scholarly discussion is that of preference heterogeneity. In the context of our game, this particularly concerns the relative importance of intention-based reciprocal motives and inequity aversion (notably proposed by Bolton and Ockenfels (2000) and Fehr and Schmidt (1999)). Depending on the situation, one or the other seems to dominate. In fact, there is some indication that both play a role: the results of the mini-ultimatum game experiments by Falk et al. (2003) and Cox and Deck (2005) demonstrate the importance of both approaches. When the proposer has the option to offer an equal distribution of earnings and an unequal one favoring herself, the responder rejects the latter significantly more often than when the proposer has to choose between the unequal and an even more unequal distribution of earnings (in Falk et al., 44.4% versus 8.9%). Obviously, this result points to the importance of reciprocity. However, when the proposer has no option but to choose the unequal offer, still a substantial number of responders (18%) reject. As there is no intention to favor herself on the part of the proposer, this observation suggests that inequity aversion is a second empirically relevant trigger for rejections. Other experiments have shown similar patterns (e.g., on the convex ultimatum game, Andreoni et al. (2003), on three-person ultimatum games, Bereby-Meyer and Niederle (2005), and on a three-person gift exchange game, Thöni and Gächter (2007)).

The *ultimatum reciprocity measure* (URM game) has the following structure: a proposer makes a proposal of how to divide an endowment *E*.[1] The responder can either accept or reject. In the first case, the proposal is implemented, in the second, the responder obtains a fixed fraction $\kappa$, $\kappa < 1$, of the offer $x$ and freely chooses the proposer payoff from the interval $[0, E - \kappa x]$. The important feature of the URM game is that (in contrast to most other games with punishment in the literature) punishment is free of marginal costs, only coming at a cost that is fixed once the offer is made.[2] This fixed cost is either equal to half the offer or to three quarters of the offer, depending on the treatment. As we will show below, models of inequity-aversion and reciprocity lead to very different predictions for behavior in the URM game: the first class of models predicts that responders – if they reject an offer – leave the proposers with a payoff which equals their earnings. In contrast, the majority of reciprocity models predicts that responders leave the proposers with zero earnings.

The results we obtain are striking. Less than 10% of the observations can be characterized as stemming from payoff-maximizers, models of inequity aversion account for 16–18%, conventional models of reciprocity for 17–38%.[3] At the same time, we find a substantial fraction of a fourth type that deviates from these predictions in a systematic way, which we call *gradual reciprocators*. These players are characterized by punishment patterns that leave their proposers with payoffs that are increasing in the offer made but generally lead to unequal payoffs. Moreover, the fraction of these players is determined by the treatment parameter. In the treatment with a high fixed cost of punishment, 20% of the population seem to switch from being gradually reciprocal to conforming to conventional reciprocity models. These observations call for an extension of existing models of reciprocity in the spirit of Sobel's first criticism: a characterization of the situation that leads to the prediction of the type distribution induced by the situation.

In Section 5, we discuss a number of approaches of how to modify the existing models in light of our observations. In particular, we characterize the gradual-reciprocator type within the framework of Cox et al. (2008), having dismissed the idea of matching the other's degree of kindness due to a lack of observations of the corresponding response-pattern predictions. With respect to our treatment effect, we note that what appears as an auxiliary assumption that is "sometimes (. . .) useful" (Cox et al., 2008, p. 34) seems to be an *essential* ingredient of a theory of reciprocal behavior. As an alternative, we propose the situation's coerciveness as a promising explanation, defined in terms of the gap between the highest payoff the player can obtain in the given situation and the next-lower obtainable payoff. An evaluation of the idea's predictive power, however, is beyond the scope of this article and is left for future research.

The remainder of the paper is organized as follows: Section 2 introduces the URM game and presents the experimental design and procedure. Section 3 analyzes the game according to the payoff-maximization model, inequity aversion, and

[1] A symbols table can be found in Appendix A.

[2] For games that allow for a change in the other player's payoff free of marginal costs, cf., e.g., Engelmann and Strobel (2004), or Fisman et al. (2007), who examine this question in the dictator game.

[3] Note that we do not consider the proposers in our game; cf. Section 3.

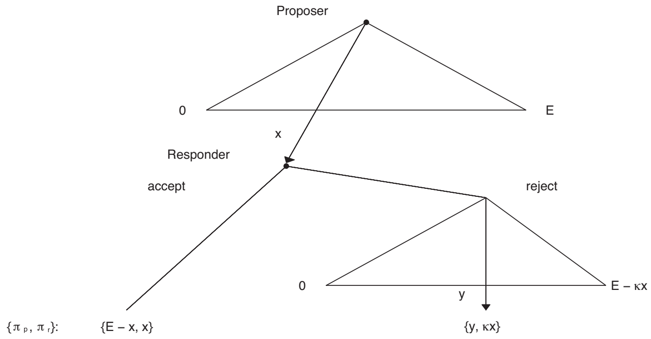

**Fig. 1.** Game tree of the ultimatum reciprocity measure.

several types of reciprocity models, always focusing on responder behavior. Subsequently, we analyze the experimental data with respect to these predictions and point to the existence of a player type that has received little attention in the literature so far in Section 4. In Section 5, we explore possible directions in which to extend existing models of reciprocal behavior to enable them to predict the kind of behavior observed. Finally, we summarize our findings and conclude in Section 6.

## 2. The game, experimental design, and procedure

### 2.1. The ultimatum reciprocity measure (URM game)

Like the classic ultimatum game, the URM game has two players, a proposer and a responder. The proposer is given an endowment of $E$ and offers $x$, $0 \le x \le E$, to the responder. If the responder accepts the offer, the proposer keeps $E - x$, while the responder earns $x$. If the responder rejects the offer, the responder earns $\kappa x$ (the *conflict payoff* $\pi_r^c$) with a commonly known parameter $\kappa \in [0, 1)$, while the proposer's conflict payoff $\pi_p^c$ is any amount $y$, $y \in [0, E - \kappa x]$, where $y$ is freely chosen by the responder. Therefore, the payoff functions for the proposer, $\pi_p$, and the responder, $\pi_r$, respectively, are

$$\pi_p = \begin{cases} E - x, & \text{in case of acceptance} \\ y, & \text{otherwise,} \end{cases} \quad \text{and}$$
$$\pi_r = \begin{cases} x, & \text{in case of acceptance} \\ \kappa x, & \text{otherwise.} \end{cases}$$

Fig. 1 illustrates the game tree of the URM game. Note that restricting the response set to $y = 0$ and setting $\kappa = 0$ yields the standard two-person ultimatum game (Güth et al., 1982). Due to these restrictions, the standard ultimatum game provides little information about negative reciprocity as a driver for rejection (since it reduces the responder's decision to a choice between only two alternatives). In contrast, by imposing no marginal costs on responders to alter proposers' payoffs after a rejection, the 'unrestricted' URM game is able to provide a very detailed picture of participants' motivations for rejections (as will become clear from the discussion of theoretical predictions in the next section). In particular, the lack of a trade-off between own monetary income and proposer payoff provides new insights into the nature of other-regarding preferences.

### 2.2. Experimental design and procedure

Each participant played one anonymous URM game either in the role of the proposer or in the role of the responder. In the instructions, we referred to proposers as *person A* and to responders as *person B*. The pie size was set to $E = 12$ Euros. Offers could only be made in integers. In order to analyze individual heterogeneity of responses corresponding to different offers in greater detail, we applied the strategy-vector method to elicit responders' choices (Selten, 1967). This means that responders had to make a decision for each possible (integer) offer before they were informed about the actual offer. Then, the offer and the corresponding responder decision determined the payoffs. This procedure implied that responders had to make a total of 13 acceptance/rejection decisions. Additionally, they had to determine the payoff of proposers for any offers rejected.

In contrast to the standard procedure of the strategy-vector method, responders were not provided with a choice menu, that is, a decision sheet that presents all potential offers in an ascending or descending order. Rather, potential offers were presented sequentially without a possibility of reviewing earlier decisions, and the order of possible offers differed randomly for all responders. We introduced this procedure for several reasons. The one-by-one procedure was chosen to make each decision as salient as possible. Further, eliciting decisions one by one in combination with a random order was intended to keep any potential experimenter-demand effect small by isolating decisions as much as possible: to 'smoothen' a response-pattern over all decisions out of a taste for consistency would inflict high cognitive costs on participants. Consequently, a smooth response-pattern should only be observed if participants exhibited underlying preferences giving rise to it. Finally, the order was randomly determined for each participant individually, in order to control for possible order effects.

The experiment started such that copies of the instructions were handed out to participants and read aloud. Subsequently, participants' questions concerning the experiments were answered privately by the instructors. Finally, all participants had to answer an electronic questionnaire testing their understanding of the game and the payoff structure.[4] Before participants answered the questionnaire, it was made clear that the only purpose of the questionnaire was to improve the understanding of the rules of the game. Wrong answers were privately explained and corrected before the experiment started.

After they had made all payoff-relevant decisions, responders were asked to state which offer they considered as fair, and which offer they expected to receive. Subsequently, we randomly matched each responder to a proposer and payoffs were realized according to the decisions made. Participants were informed about their payoffs and asked to answer a short socio-demographic questionnaire, before privately being paid.

In order to learn more about the nature of reciprocal preferences, we played the game under two treatment conditions. In the HIGH-$\kappa$ condition, the commonly known parameter $\kappa$ was set to $\kappa = 0.5$, while in the LOW-$\kappa$ condition, we set $\kappa = 0.25$. As we will show below, this (rather innocent) variation has little implication for the predictions of the considered social-preference models, while there are important differences in actual behavior. In total, 76 pairs of proposers and responders participated in the HIGH-$\kappa$ treatment, while we had 77 pairs in the LOW-$\kappa$ condition.

The laboratory experiments were conducted at the EconLab at the University of Bonn, Germany, in October and November 2006.[5] In total, 306 participants participated; 50% of the participants were female, the median age was 23 years. Participants were mostly undergraduate students from various fields of studies. Approximately one third of the students were economists or mathematicians. Further information concerning the socio-demographic background of the participants is summarized in the online supplementary material (available on the journal's website). Average payment was 5.15 Euros (no show-up fee) for an average duration of 30 min, including the instruction time and the time for paying participants.

## 3. Theoretical predictions

Our central research interest lies in the empirical analysis of reciprocal behavior. For this reason, we will focus on the behavior of responders throughout the paper. Proposer behavior is unsuitable for our purposes: proposals reflect both proposers' other-regarding preferences as well as proposers' strategic considerations concerning the other-regarding preferences of responders.

We will analyze responders' best-response functions according to all major models that are potential candidates for the explanation of reciprocal behavior. For brevity and ease of exposition, we refrain from presenting the complete sets of equilibria as they do not shed further light on our research question. In the following, we discuss three (groups of) models, the 'standard' game-theoretic prediction, models of inequity-aversion, and intention-based models of reciprocal behavior. Before we do so, let us clarify some notation. If a model predicts rejection of an offer, it will have to specify a value for the response $y$ that may be different depending on the offer. To reflect this, we will write $y = y(x)$ to denote the *(offer) response function*. Yet, there is a second way to think about responses, which will prove useful particularly in the context of treatment comparisons. For this purpose, we introduce the *conflict-payoff response function* (defining the response $y$ in terms of the conflict payoff $\pi_r^c = \kappa x$), which we will denote by $y = \gamma(\pi_r^c)$.

### 3.1. Pure payoff-maximizing preferences

The best reply of a responder exclusively driven by material self-interest is obvious: given $0 < \kappa < 1$, we have $x > \kappa x$ for any $x > 0$, and $x = \kappa x$ for $x = 0$. Consequently, payoff-maximizing responders' best reply is to always accept any positive offer $x$, and arbitrarily accept or reject a proposal of $x = 0$. Given this feature, we will not observe values of $y$ for these players. If at all, we observe a value for $y$ in response to $x = 0$; however, the theory does not give any prediction for this value. Therefore, payoff-maximizing responders' best-response function is given by $br_{pm} : x \to (\delta, y)$, where $\delta \in \{0, 1\}$ represents rejection, $\delta = 0$, or acceptance, $\delta = 1$:

$$br_{pm}(x) = \begin{cases} (1, .) & \text{if}\, x > 0, \\ (\delta', y') & \text{if}\, x = 0, \end{cases} \tag{1}$$

where $(\delta', y') \in \left\{(\delta, y) | \delta \in \{0, 1\}, y \in [0, E]\right\}$. Of course, no treatment differences are expected.

### 3.2. Inequity-averse preferences

In a first step, note that inequity-averse responders will always choose to equalize payoffs after a rejection, since it is costless to alter the proposer's payoff once the costs of rejecting are sunk. In other words, their response $y$ will be $y(x) = \kappa x$ for all rejected offers $x$. Which offers will be rejected? Both the model by Fehr and Schmidt (1999) and by Bolton and Ockenfels (2000) predict accepted offers $x$ to come from a convex set $[\underline{x}; \overline{x}]$, where $0 \leq \underline{x} < E/2 < \overline{x} \leq E$. The specific values of $\underline{x}$ and $\overline{x}$

[4] Translations of the German instructions and the questionnaire are provided in the online supplementary material (available on the journal's website).

[5] Experiments were computerized using zTree (Fischbacher, 2007). For the recruitment of participants, we used ORSEE (Greiner, 2004).

depend on the parameters of the model, notably on $\kappa$ (since it determines the monetary earnings in case of a conflict) and the importance the individual responder places on equity concerns. To indicate the dependence between $\underline{x}$ and $\kappa$, and $\overline{x}$ and $\kappa$, we will write $\underline{x}_\kappa$ and $\overline{x}_\kappa$. Both models would suggest there to be heterogeneity in the cut-off values for rejections, while all models of inequity aversion make the unique prediction $y(x) = \kappa x$. In summary, we obtain the following best-reply function $br_{ia} : x \rightarrow (\delta, y)$:

$$br_{ia}(x) = \begin{cases} (0, \kappa x) & \text{if} \, x > \overline{x}_\kappa \text{ or } x < \underline{x}_\kappa, \\ (1, .) & \text{if} \, \underline{x}_\kappa \leq x \leq \overline{x}_\kappa. \end{cases} \tag{2}$$

The predicted treatment effects are evident: an increase in $\kappa$ shifts both acceptance thresholds 'inwards' towards the egalitarian payoff distribution $(E/2, E/2)$. With respect to responses as a function $\gamma(\pi_r^c)$ of conflict payoffs, no treatment differences are expected.

### 3.3. Intention-based preferences

For our discussion of these models, we sub-divide this class into four sub-classes: (i) one in which utility functions consist of a linear combination of own income and a reciprocity term (which itself is a product of several terms as described below, cf. Rabin, 1993; Levine, 1998[6]; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006), (ii) the non-linear model of Cox et al. (2007), (iii) models mixing reciprocity concerns and inequality aversion, and finally, (iv) the model presented by Cox et al. (2008).

*'Linear' reciprocity models.* In the models subsumed under this class, utility is a linear combination of own income and reciprocity. Here, reciprocity is a product of three terms. The first weights the importance of reciprocal behavior to the person. The second term captures the kindness of the other player's past behavior. Offers are ranked from unkind (i.e., small) to kind (i.e, large) ones, such that the term for an offer which is neither kind nor unkind is zero and increases (decreases) monotonically with each rank above (below) that offer. Consequently, unkind offers have negative values and kind offers have positive values. The third term measures the degree of kindness in the person's reaction. Again, responses are ranked such that a response that is neither kind nor unkind corresponds to a value of zero, and responses above (below) that lead to values increasing (decreasing) monotonically with each rank. Due to the monotonicity of kindness, accepted offers form an interval: if rejecting a (un)kind offer yields less utility than acceptance, rejecting a less (un)kind offer yields also less utility than acceptance.

At the time of the responder's decision in the ultimatum reciprocity measure, the reciprocity weight and the kindness term in the acting player's utility function are fixed. Consequently, maximization of utility in combination with the possibility of choosing the proposer's payoff free of marginal cost implies the following for rejected offers: the best reply to any unkind offer $x$ must be the most unkind response possible, that is, $y(x) = 0, \ \forall x < \underline{x}_\kappa$. Conversely, any rejected kind offer must be answered with $y(x) = E - \kappa x, \ \forall x > \overline{x}_\kappa$, the kindest response possible. In other words, there cannot be a rejection followed by a response $y'(x_r)$, so that $0 < y'(x_r) < E - \kappa x_r$. As for the inequity-aversion models above, the switching point between acceptance and rejection is player-specific and generally cannot be predicted. The predicted best-reply function $br_{rl} : x \rightarrow (\delta, y)$ is given by:

$$br_{rl}(x) = \begin{cases} (0, E - \kappa x) & \text{if} \, x > \overline{x}_\kappa, \\ (1, .) & \text{if} \, \underline{x}_\kappa \leq x \leq \overline{x}_\kappa \\ (0, 0) & \text{if} \, x < \underline{x}_\kappa. \end{cases} \tag{3}$$

No treatment variations are predicted with respect to $y(x)$ or $\gamma(\pi_r^c)$. The lower acceptance threshold $\underline{x}_\kappa$ rises with $\kappa$, as a higher $\kappa$ makes rejection less costly. At the same time, there is no clear prediction with respect to $\overline{x}_\kappa$: while a higher $\kappa$ implies a higher 'conflict' payoff $\kappa x$, it also leads to a lower potential for rewarding actions: $E - \kappa' x < E - \kappa'' x$ for $\kappa' > \kappa''$. Therefore, the sign of the change in the upper acceptance threshold depends on the weight the responder places on reciprocity.

*Non-linear models of reciprocity*. Even though Cox et al. (2007) propose a remarkable model that generalizes the above reciprocity-models in an important way, it yields the same predictions for responder behavior in the ultimatum reciprocity measure as the 'linear' reciprocity models. In fact, utility is again a linear combination of own income and a reciprocity term, where the latter multiplies the proposer-payoff with an "emotional-state" function $\theta$. $\theta$ is a function of the proposer's previous behavior, and therefore, a fixed factor at the time of the responder's decision. Consequently, the arguments from our discussion of the 'linear' models carry over and hence, the predicted best-reply function has the same form as Eq. (3) above.

[6] Strictly speaking, the model of Levine (1998) is different from the other models listed in a number of important aspects. However, the best replies are very similar, given Levine (1998) defines a player's utility function as a linear combination of all players' monetary payoffs.

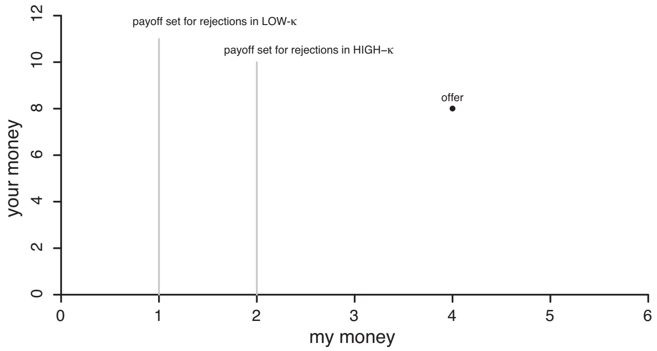

**Fig. 2.** Payoff space for responders.

*Mixed approach*. A special variation of reciprocity models is the approach by Charness and Rabin (2002) which mixes reciprocity concerns and inequality aversion.[7] In this model, a responder's utility function adds own payoff and the proposer's payoff, weighted by a term that integrates inequality as well as reciprocity concerns. In particular, this weight is lowered if the proposer receives more money than the responder and if the proposer misbehaves. However, even if the weight for reciprocity depends on the degree of misbehavior (as in the extended model in the appendix of Charness and Rabin, 2002), the sum of weights is either positive or negative.[8] Once again, the same arguments as for the 'linear' reciprocity models apply, leading to the same predictions.

*General approach to reciprocity*. Cox et al. (2008) present their novel approach to reciprocal behavior within the framework of the proposer-payoff–responder-payoff space. In this space, the choice set $\mathcal{S}$ of the responder consists of one point and a ray parallel to the proposer-payoff axis. The point describes the offer, while the ray characterizes possible payoff combinations in case of a rejection, as depicted in Fig. 2. Notice that our treatment variation does not change the location of the point, but shifts the ray in LOW-$\kappa$ closer to the proposer-payoff axis compared to the situation in HIGH-$\kappa$.

Responder preferences are represented by indifference curves $\lambda \in \Lambda$, where $\Lambda$ is a player's indifference-curve set for a given situation. To illustrate, indifference curves of payoff-maximizing players are lines parallel to the proposer-payoff axis, those of inequity-averse players are either convex (Bolton and Ockenfels, 2000) or piece-wise linear with a kink at the 45-degree line (Fehr and Schmidt, 1999), indifference curves in the 'linear' reciprocity models (as well as in the 'mixed' model by Charness and Rabin, 2002) are straight lines that are either upward-sloping (negative reciprocity) or downward-sloping (positive reciprocity), while the model of Cox et al. (2007) generalizes the 'linear' reciprocity models by allowing the indifference curves to be non-linear; however, their slope cannot change signs. Irrespective of their shape, indifference curves always can be ranked such that $\lambda'$ is said to be 'higher' than $\lambda''$ if points associated with $\lambda'$ are preferred to points associated with $\lambda''$. Finally, whenever we talk about the $\Lambda$-*defined point in* $\mathcal{S}$, we mean the point associated with the highest indifference curve in $\Lambda$ which still is in the choice set $\mathcal{S}$.[9]

At the center of the approach by Cox et al. (2008) are two basic definitions, one concerns perceived kindness and the other kindness in (re-)actions. First, let us define perceived kindness, or "generosity". In this model, the notion of *generosity* is attached to responders' opportunity sets, or, more precisely, to opportunity sets after they have been altered by the action of the proposer.[10] Particularly, consider the set $\mathcal{S}_x$ of possible payoff combinations $(\pi_p, \pi_r)$ which proposer and responder can gain after the proposer has chosen $x$. Let us define $\hat{\pi}_i(x) = \sup_{\pi_i} \mathcal{S}_x$ for $i = p, r$. A set $\mathcal{S}_{x'}$ is called "more generous than" a set $\mathcal{S}_{x''}$ if $(i)$ $\hat{\pi}_r(x') - \hat{\pi}_r(x'') \geq 0$ and $(ii)$ $\hat{\pi}_r(x') - \hat{\pi}_r(x'') \geq \hat{\pi}_p(x') - \hat{\pi}_p(x'')$. In other words, the proposer is *more generous* by choosing $x'$ than by choosing $x''$ if $(i)$ the proposer's choice of $x'$ over $x''$ does not lead to a decrease in the maximum payoff

[7] In fact, the approach by Falk and Fischbacher (2006) also represents a mixture of reciprocity and inequality considerations, as reciprocation by the responder is triggered by proposer choices that lead to unequal payoffs.

[8] This assertion is not completely correct. Under a specific parameter combination, the extended model in the appendix of Charness and Rabin (2002) allows for rejections in conjunction with responder utility increasing in proposer income if the latter is close to 0, and decreasing if it is above a threshold of $\kappa x - b$, where $b$ measures how strongly an undeserving poorest society member is disregarded. In that case, responses are predicted to be $y(x) = \max\{0, \kappa x - b\}$, i.e., the response function is parallel to the response function of an inequity-averse player (neglecting the corner solutions for low $x$). As we find only two out of 153 participants in our data whose response pattern is in line with this prediction (apart from the inequity-aversion equivalent $b = 0$), we hold that this special case can be neglected for ease of exposition. The specific parameter constellation requires that the combined weight placed on a Rawlsian social optimum, $\delta\lambda$, is close to one (but $\delta < 1$), the spite parameter with respect to undeserving players, $f$, is sufficiently small, $f < \delta/(1-\delta)$, and the undeserving are not disregarded in the total-surplus-maximising part, i.e., $k \approx 0$ for the parameter $k$ measuring the discounting of undeserving proposers' payoffs in this part, nor in the Rawlsian social-welfare part, i.e., $b < \kappa x$ for some rejection-worthy offers.

[9] We do not refer to this point as the tangential point, as in case of acceptance as well as for some of the models, it would be inadequate to speak of tangents: there cannot be a tangent to a point, and in some cases, the (highest) indifference curve will have a kink at the $\Lambda$-defined point.

[10] Strictly speaking, the notion of an opportunity set as used by Cox et al. (2008) would rule out application of their model to our game, as they require opportunity sets to be convex. However, we do not see why non-convexity of opportunity sets would lead to problems in the analysis. Hence, we drop the convexity assumption, as we are convinced that their model is an important tool to understand behavior in our game.

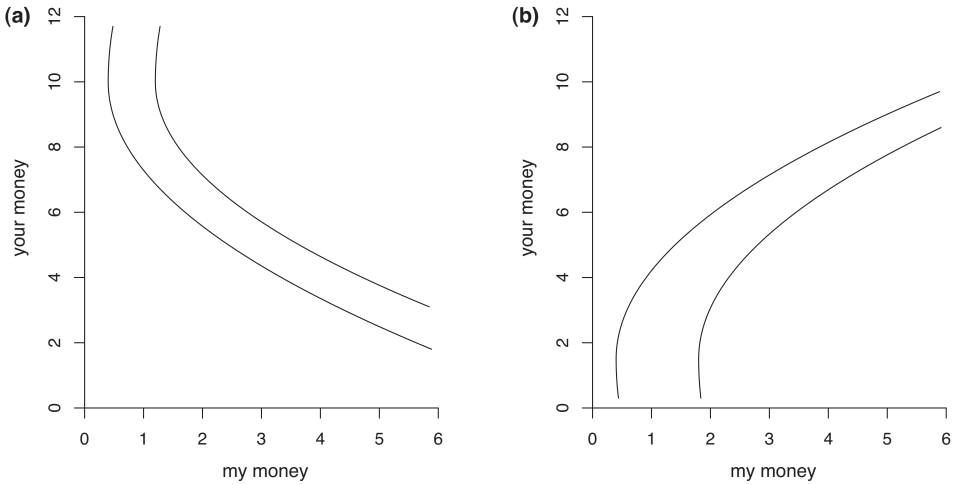

**Fig. 3.** Indifference curves (a) $\Lambda'$ and (b) $\Lambda''$ ($\Lambda'$ being more altruistic than $\Lambda''$).

the responder can earn, and (*ii*) the increase of responder's payoff as a result of decision $x'$ compared to $x''$ is not less than the corresponding increase in the proposer's payoff. According to this definition, an offer $x'$ in the URM game is *more generous than* $x''$ if and only if $x' \geq x''$.

Second, let us define kindness in action, which is termed "altruism". *Altruism* refers to to the responder's utility function $u_r(\pi_r, \pi_p)$ and the corresponding curvature of the responder's convex indifference curves $\Lambda$ in the payoff-space $\{(\pi_p, \pi_r)\}$. For convenience, it is defined in terms of a player $i$'s willingness-to-pay for a marginal increase in the payoff of player $j$, $WTP_i = [\partial u_i(\pi_i, \pi_j)/\partial \pi_j]/[\partial u_i(\pi_i, \pi_j)/\partial \pi_i]$, rather than $i$'s marginal rate of substitution $MRS_i = 1/WTP_i$. The responder's utility function $u_r'(\pi_r, \pi_p)$ is said to be "more altruistic than" $u_r''(\pi_r, \pi_p)$ if $WTP_r' \geq WTP_r''$, $\forall(\pi_r, \pi_p)$. Equivalently, the utility function associated with indifference curves $\Lambda'$ is more altruistic than the function associated with $\Lambda''$ if, compared to $\Lambda''$, the curves in $\Lambda'$ are rotated counter-clockwise (compare Fig. 3). As a consequence, the proposer payoff $\pi_p'$ in the $\Lambda'$-defined point in $\mathcal{S}_x$ must not be smaller than $\pi_p''$ in the $\Lambda''$-defined point in the same set.

Within this framework, *reciprocity* is defined as follows: a proposer's decision leading to $\mathcal{S}_{x'}$ rather than $\mathcal{S}_{x''}$ ($\mathcal{S}_{x'}$ being *more generous than* $\mathcal{S}_{x''}$) induces indifference curves $\Lambda'$ rather than $\Lambda''$ on the part of the responder (with $u_r'(\pi_r, \pi_p)$ *more altruistic than* $u_r''(\pi_r, \pi_p)$). Loosely speaking, more generous offers lead to more altruistic preferences.

Having outlined the model, we now apply it to the URM game. Recall that the convex indifference curves are rotated clockwise for less generous offers. That is, the smaller the offer, the steeper – or flatter, in case of upward-sloping curves – the indifference curves. As a consequence, the intersection between the highest indifference curve and the choice set decreases or remains constant, but never increases in the proposer-payoff dimension for offers of decreasing generosity.[11] Since altering the proposer's payoff is costless and $\inf_{\pi_p} \mathcal{S}_x = 0, \forall x$ (i.e., the lower bound of the choice set does not change for different offers) we can conclude that for two rejected offers $x'$ and $x''$ such that $x' > x''$, $y(x') \geq y(x'')$ must hold.

Which offers will be rejected? Like in any of the other models presented, the model proposed by Cox et al. (2008) assumes that utility from own income is traded off against a second utility component that is influenced by others' income. If the responder rejects an offer, the utility gains from this second component must outweigh the decrease in one's own income. Hence, the responder must have a positive *WTP* in response to very *generous* offers (e.g., rejecting $x \geq \overline{x}$ and responding by $y > E - x$) – although this scenario appears hardly intuitive at the first glance – or have a negative *WTP* for $\pi_p$ in response to very *ungenerous* offers (e.g., rejecting $x \leq \underline{x}$ and responding by $y < E - x$), so that accepted offers come from a convex set $[\underline{x}; \overline{x}]$. The specific values of $\underline{x}$ and $\overline{x}$ again depend on the importance the individual responder places on reciprocity. Thus, we obtain the following best-reply function $br_{re} : x \to (\delta, y)$:

$$br_{re}(x) = \begin{cases} (0, y'(x)) & \text{if}\, x > \overline{x}_\kappa \,\text{or}\, x < \underline{x}_\kappa, \\ (1, .) & \text{if}\, \underline{x}_\kappa \leq x \leq \overline{x}_\kappa. \end{cases} \tag{4}$$

where $y'(x)$ must satisfy $\partial y'/\partial x \geq 0$.

With respect to treatment effects, we first turn to changes in the offer-response function $y(x)$. In response to an increase in $\kappa$, the model allows for both monotonic increases and invariance at any given level, merely ruling out reductions in the

[11] Strictly speaking, this argument requires preferences to have the *increasing benevolence* property, which Cox et al. (2008) define as a willingness to pay for the other player's income that does not decrease in own income.

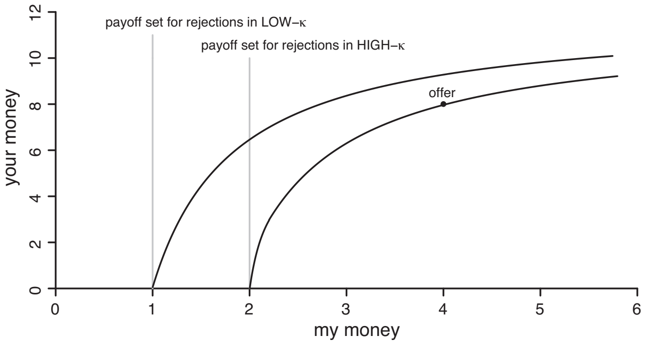

**Fig. 4.** Example for the influence of a $\kappa$ variation on offer acceptance.

response.[12] Turning to treatment effects on the conflict-payoff response function, recall that our treatment variation does not alter the supremum of $\pi_r$ in the set $\mathcal{S}_x$ for a given offer $x$, since the value of $\kappa$ changes the ray, but not the point (i.e., the offer, which defines the supremum). Hence, the treatment variation does not change the generosity of offers.[13] On the other hand, the same conflict payoff $\pi_r^c$ is associated with different offers in the different treatments: under a low $\kappa$, a higher offer is associated with the same conflict payoff than under a high $\kappa$. At the same time, higher offers are attached to higher levels of *generosity* and therefore, met with higher degrees of *altruism*. As a consequence, the conflict-payoff response function $\gamma(\pi_r^c)$ may differ between the treatments: for $\kappa' < \kappa''$, the same conflict payoff $\pi_r^c = \kappa' x' = \kappa'' x''$ implies $\gamma(\kappa' x') \geq \gamma(\kappa' x'')$.

With respect to acceptance thresholds, the model predicts a decrease of the acceptance threshold $\underline{x}$ for decreasing $\kappa$. To see this, take the offer $x'$ that makes the responder indifferent between accepting and rejecting under $\kappa'$. Let us now decrease $\kappa$. Recall that a change in $\kappa$ leaves the responder-payoff supremum unaffected and hence, indifference curves do not change as the responder's degree of *altruism* remains the same. But with the indifference curves remaining the same and the ray of $\mathcal{S}'_x$ shifting left, the responder must now prefer to accept the offer, as shown in Fig. 4. By the same token, changing $\kappa$ may change the upper threshold $\bar{x}$. By the convexity of preferences and the linearity with respect to variations of $\kappa$ of the maximum-possible reward $(E - \kappa x)$, it is immediately obvious that an increase in $\kappa$ cannot be associated with an increase in the upper acceptance threshold (and more often than not, it will lead to a decrease in $\bar{x}$).

In the preceding paragraphs, we have presented qualitative predictions that can be derived from the general model by Cox et al. (2008). Virtually all of these predictions have been weak, in the sense exemplified by the statement that "for offers of decreasing generosity, the response cannot increase." By employing *weak* inequalities in all of their definitions, Cox et al. (2008) encompass all of the existing model predictions in one framework. The model would be able to account even for offer-response functions equal to a strictly positive constant, in contrast to any of the other models. It does place a number of restrictions on behavior that can be expected, most notably perhaps the requirement of a certain degree of consistency. However, it does not make clear predictions like the remaining models presented. In other words, what the model gains in generality, it looses in terms of specifity. We consider this an important shortcoming and briefly review potential directions of model refinement to eschew this problem in Section 5 of this article. To prepare the floor for the results, we summarize the predictions from the different models in Table 1.

## 4. Results

We structure the presentation of our results as follows: first, we characterize rejection behavior of responders. Second, we analyze response patterns. Third, we systematically relate response patterns to rejection behavior. We relegate presentation of data on offers, expected offers, as well as of role-contingent average payoffs to the online supplementary material (available on the journal's website), as these are not in the focus of this study.

### *4.1. Rejections*

84% of actual offers in the HIGH-$\kappa$ condition (81% in the LOW-$\kappa$ condition) were accepted. Following our theoretical discussion from the previous section, we define an upper and a lower acceptance threshold for each responder $i$, $\bar{x}_i$ and $\underline{x}_i$, as follows:

[12] Once again, this requires invoking the *increasing benevolence* property, cf. footnote 11.

[13] Strictly speaking, this statement is not correct. Please, refer to the discussion in Section 5 for why we hold the above assertion to be in the spirit of the model.

**Table 1**
Predictions of the models discussed.

| | | |
|---|---|---|
| **Response $y(x)$ to offers** | $\mathbf{0 < x < \underline{x}_\kappa}$ | $\mathbf{x > \overline{x}_\kappa}$ |
| Payoff-maximization | Not applicable: all offers are accepted | |
| Inequity aversion | $\kappa x$ | $\kappa x$ |
| Reciprocity (linear/non-linear/mixed) | 0 | $E - \kappa x$ |
| Cox et al. (2008) | $\partial y/\partial x \geq 0$ | $y > E - \overline{x}_\kappa$, $\partial y/\partial x \geq 0$ |
| **Treatment effect on** $y(x)$ | $\mathbf{0 < x < \underline{x}_\kappa}$ | $\mathbf{x > \overline{x}_\kappa}$ |
| Payoff-maximization | Not applicable: all offers are accepted | |
| Inequity aversion | + | + |
| Reciprocity (linear/non-linear/mixed) | 0 | – |
| Cox et al. (2008) | + or 0 | – |
| **Acceptance thresholds** | $\boldsymbol{\partial \underline{x}_\kappa / \partial \kappa}$ | $\boldsymbol{\partial \overline{x}_\kappa / \partial \kappa}$ |
| Payoff-maximization | Not applicable: all offers are accepted | |
| Inequity aversion | + | – |
| Reciprocity (linear/non-linear/mixed) | + | (0/– or 0/0) |
| Cox et al. (2008) | + or 0 | – or 0 |

*Note*: Derivations for predictions under excessively nice offers are omitted.

**Table 2**
Numbers of responders according to acceptance thresholds.

| | $\underline{x}_i$ | | | | | | | | $\overline{x}_i$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | −1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 9 | 10 | 11 | 12 |
| HIGH-$\kappa$ | 0 | 3 | 5 | 5 | 9 | 42 | 11 | 1 | 1 | 1 | 10 | 64 |
| LOW-$\kappa$ | 1 | 8 | 6 | 9 | 19 | 27 | 6 | 1 | 0 | 1 | 5 | 71 |

$$\begin{aligned} \overline{x}_i &= max\{x|\delta_i(x)=1\} \text{ and} \\ \underline{x}_i &= \begin{cases} max\{x|x \leq 6, \delta_i(x)=0\}, & \text{if } \{x|x \leq 6, \delta_i(x)=0\} \neq \emptyset, \text{ and} \\ -1, & \text{otherwise,} \end{cases} \end{aligned} \quad (5)$$

where $\delta_i(x)$ denotes the acceptance decision of responder $i$ for a certain offer $x$. Note that inequity aversion and reciprocity models predict regularity with respect to rejections, that is, $\delta_i(x) = 1$, $\forall x \in (\underline{x}_i, \overline{x}_i]$, and $\delta_i(x) = 0$, otherwise. In total, 32 out of 153 responders exhibit rejection decisions that violate regularity. While this is more than typically observed (see Camerer, 2003), only 9 of them (6% of all responders) make more than one decision that would contradict regularity. We attribute the remaining 23 violations to the difficulty arising from the random-order one-by-one presentation of possible offers. To account for this fact and use as much information as possible, we chose the above definition of $\underline{x}_i$.[14] The further analysis includes the data of all responders. Responders are classified according to their acceptance thresholds; Table 2 reports the number of responders in each lower and upper acceptance class, $||\underline{x}_i||$ and $||\overline{x}_i||$, respectively.

The lower thresholds $\underline{x}_i$ are significantly higher in HIGH-$\kappa$ (with an average of 3.57 vs. 2.97 in LOW-$\kappa$, $p = 0.003$; also, as can be easily seen from Table 2, $\underline{x}_i$ from the HIGH-$\kappa$ treatment first-order statistically dominates $\underline{x}_i$ from the LOW-$\kappa$ condition).[15] At the same time, the treatment difference between upper acceptance thresholds $\overline{x}_i$ fails to reach significance (11.80 vs. 11.91, $p = 0.114$). However, this is not a strong indication that there is no effect: while most responders never reject an offer above the equal split, the number of those who do in HIGH-$\kappa$ (12 out of 76) is double the corresponding number from the LOW-$\kappa$ treatment (6 out of 77; again, statistical dominance holds).

## *4.2. Responses*

Given our main research interest lies in the study of reciprocal behavior, responses following a rejection are the central element in our analysis. In the following, we will identify rejected offers as $x_r$, so that the response to a rejected offer is $y(x_r)$. In light of the exploratory nature of our study, we want to get as close as possible to the raw data. Therefore, we classify the response functions according to mutually exclusive type categories. We define the types based on the theoretical predictions summarized in Table 1, focussing on the negative-reciprocity part, given there is little variance in the domain of positive

[14] Our qualitative results and statistical inferences do not change if we define $\underline{x}_i$ using the more straightforward definition $\underline{x}_i = min\{x|\delta_i(x) = 1\}$, indicating the robustness of our findings.

[15] Unless otherwise indicated, all comparisons are based on two-sided Wilcoxon rank-sum tests. Behavior of economists/mathematicians and other participants does not differ significantly with respect to any variable measured.

**Table 3**
Frequency of responder types.

| | HIGH-$\kappa$ | (in %) | LOW-$\kappa$ | (in %) |
|---|---|---|---|---|
| Null | 3 | 3.9 | 6 | 7.8 |
| Accepters | 0 | 0.0 | 1 | 1.3 |
| Symbolic | 2 | 2.6 | 1 | 1.3 |
| Gentle punishers | 3 | 3.9 | 3 | 3.9 |
| Inequity-averse | 14 | 18.4 | 12 | 15.6 |
| Reciprocal: linear | 13 | 17.1 | 29 | 37.7 |
| Reciprocal: gradual | 33 | 43.4 | 20 | 26.0 |
| Between | 3 | 3.9 | 3 | 3.9 |
| Unclassified | 5 | 6.6 | 2 | 2.6 |
| Total | 76 | | 77 | |

reciprocity. Participants whose responses could not be fitted into the categories defined by the presented models were grouped according to the broad characteristics of their responses.

*Null.* Subjects falling into this category accept all offers except for offers $x = 0$, in which case they respond by $y(0) = 0$. This behavior can be classified as either selfish, inequity-averse or reciprocal, and therefore does not provide much information about the *nulls*' motivations.

*Accepters.* Subjects falling into this category accept all offers.

*Symbolic.* Subjects falling into this category accept most offers but reject at least one. However, their rejection is only *symbolic*–they administer the proposer the amount the latter asked for: $y(x_r) = E - x_r$.

*Gentle punishers.* Subjects falling into this category reject some offers but leave the proposer better off than what they were offered for $x_r < 6$, i.e., $y(x_r) > x_r$.

*Inequity-averse. Inequity-averse* players conform to the predictions of the corresponding models: whenever they reject an offer $x_r$ which they do for at least two offers $x$, they respond by choosing $y(x_r) = \kappa x_r$.

*Reciprocal: linear.* Subjects falling into this category conform to the predictions of all major reciprocity models: whenever they reject an offer $x_r$ which they do for at least two offers $x$, they respond by punishing the other player as harshly as possible, $y(x_r) = 0$.

*Reciprocal: gradual.* Subjects falling into this category exhibit two characteristics: (i) they are not inequity-averse players and (ii) their offer response function fulfills $y(x_j) \geq y(x_i)$ for any pair of rejected offers $x_i$ and $x_j$ such that $x_i < x_j$ and $y(x_j) > y(x_i)$ for at least one such pair.

*Between.* Subjects are categorized to fall *in between* the other categories if they reject various offers and choose $y(x_r)$ such that they would belong to different categories for different $x_r$.

Table 3 summarizes the results of our classification analysis.[16] We make the following observations: First, taking into account all response patterns that could potentially result from preferences of a payoff-maximizing player – *null, accepters*, one *symbolic* type in each treatment plus one *gentle-punisher* – we count only 13 participants (4 in HIGH-$\kappa$, 9 in LOW-$\kappa$) out of 153 (8%). Hence, compared to typical results from other variations of the ultimatum game (e.g., see Andreoni et al., 2003), the URM game yields much less 'selfish' behavior by responders.[17]

Second, the sum of all players whose behavior can be described by one of the theoretic models outlined in Section 3 excluding the model by Cox et al. (2008) makes up for only 31 out of 76 in HIGH-$\kappa$ and 50 out of 77 in LOW-$\kappa$. In other words, conventional models account for only about 40% (65%) of the observed response patterns in HIGH-$\kappa$ (LOW-$\kappa$). More specifically, we observe a stable 16-18% inequity-averse players, while the number belonging to different subclasses of reciprocity differs substantially across treatment conditions. Most participants not exhibiting behavior as predicted by the above models can be categorized as gradually reciprocal. The model of Cox et al. (2008) can account for these observations. However, it *accommodates* rather than *predicts* them. Below, we explore a number of ways in which our understanding of gradually reciprocal behavior may be characterized on the basis of their model.

Third, in HIGH-$\kappa$, the fraction of players categorized as *gradual reciprocators* is higher than in LOW-$\kappa$ by almost 20%. At the same time, the HIGH-$\kappa$ fraction of *linear-reciprocal* players is lower by the same 20%. In other words, the data look as if a change in $\kappa$ from 0.5 to 0.25 changed the response function of about 20% of the population such that they no longer differentiate the severity of punishment with respect to an offer's unkindness.[18] Whether this is an actual type shift or whether it is merely the slope of the response function being shifted downward very strongly is something we cannot answer. What we do know is that the fraction of gradual reciprocators with a response–function slope larger than 0.4 changes from 17 out of 33 in

[16] Note that we allowed for a single deviation from the respective predictions; this could be an acceptance below $\underline{x}_i$ or a slight non-monotonicity, e.g., for gradual reciprocators. Not doing so would leave us with 24 (12) gradual reciprocators in HIGH-$\kappa$ (LOW-$\kappa$), and with 20 (13) unclassified responders.

[17] A plausible reason for this observation is that punishment costs are lower in our setup. We are thankful to an anonymous referee for pointing this out.

[18] A $\chi^2$-test suggests that the type distributions of participants classifiable as payoff-maximising, inequity-averse, linear-reciprocal, gradually reciprocal, and others, differ between treatments, ($p = 0.015$).

**Table 4**
Mean lower acceptance thresholds by treatment and type.

| Responder type | HIGH | LOW |
|---|---|---|
| Inequity-averse | 4.08 | 3.50 |
| Reciprocal: linear | 3.71 | 3.31 |
| Reciprocal: gradual | 4.00 | 3.15 |

HIGH-$\kappa$ to 5 out of 20 in LOW-$\kappa$. In other words, responses in general do get harsher with increasing fixed costs of punishment also within the group of gradual reciprocators.

In our view, these observations are critical: further development of any reciprocity model should account for both gradually reciprocal behavior and what looks like a parameter-induced shift in the type distribution, if it is to be seen as a step forward in our understanding of reciprocal behavior. For this reason, we devote Section 5 to some ideas on possible directions in which to extend existing models of reciprocity, discussing them in light of our observations. Before we do so, we shed some light on the interaction between response patterns and rejections in the following part.

### *4.3. The interaction between response patterns and rejections*

In the following paragraph, we briefly report the results of a comparison of lower acceptance thresholds between the three main types, reported in Table 4. While we do not find any significant differences of lower acceptance thresholds between inequity-averse, linear-reciprocal, and gradually reciprocal players within each treatment (all pair-wise comparisons yield $p > 0.15$), we observe a very differentiated picture across treatments. Both for inequity-averse and linear-reciprocal players, the treatment difference in lower acceptance thresholds is in the predicted direction but clearly fails to be significant ($p = 0.249$ and $p = 0.301$, respectively). However, for participants classified as gradual reciprocators, there is a treatment effect: in LOW-$\kappa$, they accept significantly lower offers than in HIGH-$\kappa$ ($p < 0.001$). Thus for a substantial fraction of these players, fairness considerations are substantially influenced by a situational variation. One possible reading of this is that players with high acceptance thresholds exhibit particular sensitivity to the fixed costs of punishment: if their response function shifts enough so that in LOW-$\kappa$, they are classified as linear reciprocators, this would explain the (non-)significance of the treatment comparisons of both linear and gradual reciprocators. Once again, further research is needed to assess the plausibility of this interpretation.

## 5. Generalized-reciprocal behavior

In this section, we set out to explore possible directions in which existing models may be changed to account for our findings. Our data call for two things. The presence of a substantial fraction of participants who can be categorized as gradual reciprocators calls for a theoretic characterization of such players. And the shift in the type distribution in response to our treatment variation calls for a theory that is able to predict that shift. Our discussion of ways to meet these challenges will be divided into two parts. First, we review and discard a simple extension of linear models of reciprocity. Subsequently, we provide a more detailed discussion within the model of Cox et al. (2008), paying tribute to the fact that it is the only available model able to accommodate our findings.

### *5.1. Linear reciprocity models*

In the reciprocity models reviewed in this paper, there are two components of reciprocal behavior: an assessment of the other player's kindness, or *generosity*, and the degree of reciprocation, or *altruism* in a player's response. Linear models like Charness and Rabin (2002), Dufwenberg and Kirchsteiger (2004), or Falk and Fischbacher (2006) aggregate an action's degree of *generosity* into a (relative) weight that is put on the other player's payoff; the degree of kindness of – or *altruism* in – a response then *follows from* the maximization of the weighted payoff sum. A very simple idea that would be able to meet both challenges posed by our data is to modify the models by specifying players' utility function such that it is maximized if the degree of *altruism* meets a certain target, namely the degree of *generosity* of the other player's action. This is akin to what most legal systems do: matching punishment to the severity of an offence, rather than assigning the maximum penalty to all infringements alike.

In principle, there are three ways to apply this idea to our game; however, only one of them can address both challenges posed by our data. As in the earlier reciprocity models, a responder would evaluate kindness against the 'fairness' or neutrality benchmark (in our game, presumably corresponding to the equal split), so that the degree of generosity to be matched is given by $x/6$. After a rejection, the responder would assign the proposer the above fraction (i) of the proposer's fair share of 6, yielding $y(x) = 6 \cdot x/6 = x$, (ii) of the proposer's claim, so that $y(x) = (12 - x) \cdot x/6$, or (iii) of the proposer's claim after shrinking it in the same way as the offer is shrunk, yielding $y(x) = \kappa(12 - x) \cdot x/6$.[19] Only the third possibility would predict a treatment difference in the offer response function $y(x)$ as we see in our data. However, we do not want to propagate this possibility due to our empirical findings: the above conjecture (iii) makes precise point-predictions for the values of the response $y$ that do not conform with the results for the vast majority of players categorized as gradual reciprocators even if we allow for rounding. More specifically, no universal point-prediction would be adequate, as the response–function variance within this group is rather large. In light of this fact, it seems unsatisfactory to choose conjecture (iii) as a suitable step forward in modeling reciprocal behavior.

### 5.2. *Gradual reciprocity in the model of Cox et al. (2008)*

Before we dwell on potential ways to account for our findings, we need to discuss an earlier imprecision in our exposition related to the reciprocity model by Cox et al. (2008).[20] We claimed that the model does not discern in terms of generosity between the same offer made in both treatments. However, blindly applying the definition of *generosity*, we would conclude that a given offer $x$ is more generous when made in HIGH-$\kappa$ than when made in LOW-$\kappa$. To see this, note that the first part of the definition, $\hat{\pi}_r^{high}(x) - \hat{\pi}_r^{low}(x) \geq 0$, trivially holds – the maximum the responder can obtain under offer $x$ is identical in both treatments (namely, the offer itself), and thus, $\hat{\pi}_r^{high}(x) - \hat{\pi}_r^{low}(x) = 0$. At the same time, the second part also holds, as $\hat{\pi}_r^{high}(x) - \hat{\pi}_r^{low}(x) > \hat{\pi}_p^{high}(x) - \hat{\pi}_p^{low}(x)$, since the right-hand side of this equation equals $(12 - (x/2)) - (12 - (x/4)) = -(x/4)$. In other words, the opportunity set defined by an offer $x$ in HIGH-$\kappa$ is *more generous than* the same offer in LOW-$\kappa$. However, this is not because the responder's conflict payoff is lower but because it leads to the same payoff maximum for the responder and the proposer cannot be rewarded as much in HIGH-$\kappa$. In our view, this does not make sense conceptually: in the domain of negative reciprocity, any option of *rewarding* the proposer after an offer that is too low to be accepted should be irrelevant (at the very least if there is a punishment option as in the URM game).[21] Hence, we contend that the model does not provide sufficient reason to predict the shift in response harshness we observe in the experiment.[22]

Having discussed Cox et al.'s definition of *generosity* at length, we are ready to review possible modifications. First of all, we suggest characterizing our gradual reciprocators by requiring the monotonicity of responses to opportunity sets – as ordered corresponding to their generosity – to be strict. In terms of the model, these players are characterized as follows: if $\mathcal{S}_{x'}$ is strictly more generous than $\mathcal{S}_{x''}$ (in the sense that the first inequality in the definition is strict), then $WTP_r{}' > WTP_r{}'' \quad \forall(\pi_r, \pi_p)$.[23] The resulting class of participants differentiates their altruism according to the severity of an offence. On the other hand, participants with flat response patterns, most notably, linear-reciprocal participants, no longer fall into the category.

### 5.3. *The treatment effect on the type classification*

Our discussion from the preceding paragraphs may suggest that the critical aspect of the model on which refinement may be necessary is its definition of *generosity*. While we think that adjustments of the definition are necessary, they are not needed to account for the apparent difference in the type distribution. As long as responder preferences after rejection-worthy offers exhibit a strong degree of *increasing benevolence*, the model can accommodate what appears to be a shift in the type distribution.[24] In this view, what appears to be a *linearly reciprocal* response pattern in LOW-$\kappa$ simply is a sequence of corner-solution choices by a *gradual reciprocator*. According to Cox et al. (2008, p.34), the *increasing-benevolence* property is "sometimes (...) useful". Our experiment gives a hint as to the types of situations in which the property is *essential*. As in the examples given in Cox et al., *increasing benevolence* seems to be particularly articulated when reciprocal behavior is concerned.

An alternative explanation would be that the degree of coerciveness of the situation plays an additional role in determining the slope of the responder's indifference curves. To exemplify this, we propose the following argument: a responder's position in LOW-$\kappa$ seems much less comfortable than in HIGH-$\kappa$, given – holding the offer the same – the responder in LOW-$\kappa$ has to renounce a larger amount of money than the responder in HIGH-$\kappa$. So, we may expect the responder in LOW-$\kappa$ to be more reluctant to reject a given offer than the responder in HIGH-$\kappa$, and therefore, that the proposer's position is more powerful

[19] Another option would be to evaluate the kindness of the offer after rejection, that is, of the conflict payoff against the equal split; however, this would mean that even the equal split itself would be unkind, which is counterintuitive.

[20] A second imprecision is that the model is not applicable to our game if we stick to Cox et al.'s exposition, given the opportunity sets in our game are not convex, and therefore, not *opportunity sets* in the sense of their definition. However, we think it would not be conducive not to consider the model on these grounds, as it is a powerful tool to think about reciprocal behavior, and one that deserves further development.

[21] Note that a simple variation of our game such as, e.g., restricting responses to $y(x) \leq E - x + \kappa$ would give rise to the opposite prediction while presumably leading to similar behavior as in the experiments conducted.

[22] As stressed above, this does not mean the model cannot *accommodate* the observations.

[23] Allowing for rounding (but not allowing for errors), this leaves us with 24 (15) gradual reciprocators in HIGH-$\kappa$ (LOW-$\kappa$); not allowing for rounding, the corresponding numbers are 9 (8).

[24] For a thorough discussion of the *increasing-benevolence* property and its relationship with the convexity of preferences, see Cox et al. (2008); applied to our setting, the definition states that the responder's willingness-to-pay for proposer income (weakly) increases in the responder's income. We are grateful to an anonymous referee for pointing us to the possible explanation.

in LOW-$\kappa$ as compared to HIGH-$\kappa$. In other words, the situation in LOW-$\kappa$ can be interpreted as being more coercive; if the coerciveness of the situation determines the intensity of responders' reaction to a given offer, then we should observe the type shift we observe: responders in LOW-$\kappa$ are less prone to reject a given offer, but if they do, they respond more harshly.[25] A possible mechanism that may give rise to the postulated effect would be that responders display an aversion to proposers abusing their power.

Is there a way in which to characterize the coerciveness of an offer? An intuitive way would be to compare the highest and the second-highest possible responder payoffs, potentially normalized using the highest-possible responder payoff. In the games examined by Cox et al. (2008), coerciveness would always be zero as they require opportunity sets to be convex; however, when applying the model to discrete opportunity sets as in our game, the hypothesis can, indeed, distinguish between different situations.[26] To develop this idea fully and incorporate it into a modified version of the model goes beyond the scope of this article.

## 6. Summary and discussion

In this article, we present the ultimatum reciprocity measure (URM game) as an analytical tool for the inquiry into the nature of reciprocal behavior. In contrast to many other games (e.g., the ultimatum game or the trust game), it gives rise to very clear and distinct predictions of models of inequity aversion on the one hand, and 'conventional' models of reciprocity, on the other. The model of Cox et al. (2008) accommodates both predictions as well as data that fall in between the two extremes. An important empirical aspect of our study is to provide data on the relative frequency of these types (and possibly others, if they were to be observed). Using second-movers' response patterns, we classify our participants. Our findings are remarkable. Less than 10% of the responders in our study exhibit behavior that can be explained by payoff-maximization; little more than 15% can be classified as inequity-averse; 'conventional' models of reciprocity account for another 17–38%, depending on the treatment. This means that the main models discussed in the literature account for only 41–65% of the observations. Adding the model of Cox et al. (2008), this number increases to 84–90%. We count this as evidence that the latter model is an important step forward in the quest for understanding reciprocal behavior.

At the same time, we observe what looks like a systematic shift of behavior between treatments that is unaccounted for by any of the 'conventional' models discussed in the literature. Decreasing the responder's conflict payoff by one half leads to a strong decrease in the frequency of players characterized as gradual reciprocators; at the same time, the frequency of linear-reciprocity types increases by the same amount. It seems as if the parameter difference induced one fifth of the population to respond in a *qualitatively* different way. In Section 5, we discuss a number of possible ways to account for this shift. The two explanations that seem to be most convincing to us are (i) that situations invoking reciprocity lead to a strong degree of *increasing benevolence*, that is, offended players' willingness to pay for offenders' income increases very strongly in the formers' own income and (ii) that the *coerciveness* of the situation influences responders' reactions: the higher the fraction of their potential earnings they have to give up in order to be able to punish, the harsher will be their response. This could be explained if we assume that people display an aversion to the abuse of power by others.

This paper contributes to the literature in a number of important ways. It introduces the ultimatum reciprocity measure as a powerful tool that provides new insights into both the nature of reciprocal behavior and the heterogeneity of preferences. We thereby extend the results of previous experiments that estimate interdependent preferences by using decisions in dictator games (e.g., Andreoni and Miller, 2002; Fisman et al., 2007) and other modified ultimatum games[27] which have focused predominantly on the robustness of the prediction based on inequity aversion (Kagel and Wolfe, 2001; Andreoni et al., 2003; Garrod, 2008). Moreover, the ultimatum reciprocity measure conveys valuable insights into how the heterogeneous type distribution changes as a consequence of differences in the situation, as exemplified by a simple parameter change within our game.

Our data provide evidence of a player type that has received little attention in the literature so far. This player type aims to level the punishment of unkind behavior according to the degree of unkindness, rather than merely restricting their punishment in response to increasing punishment costs, as in the more conventional models of reciprocity. To the best of our knowledge, only the model proposed by Cox et al. (2008) can accommodate this behavior. However, one might argue that the model can accommodate the behavior because of its flexibility rather than its accuracy. We discuss one potential weakness of the model and point to a number of possible ways of how to improve on prediction accuracy, evaluating them in light of our findings. Our treatment variation seems to suggest that the *increasing-benevolence* assumption is not just a helpful auxiliary assumption to arrive at clearer predictions for certain games, but an essential ingredient of any theory of reciprocal behavior–unless we draw on further aspects of a situation such as its coerciveness.

[25] One may argue that proposers who make the same offer to the responder in LOW-$\kappa$ than in HIGH-$\kappa$ do so *despite* their more powerful position. Therefore, if proposer do not exploit their powerful position, one could conjecture that this non-exploitation is a kind act that should be rewarded under a reciprocity hypothesis. However, this is not consistent with our data: the same offer is punished as harshly as possible by a larger fraction of participants in LOW-$\kappa$, rather than the other way around.

[26] A thought experiment suggested by an anonymous referee suggests, however, that our tentative definition of coerciveness is too simplistic: a variant of our game in which the responder can choose between accepting, rejecting, and a third option in which payoffs are $(E-x, x-\varepsilon)$, $\varepsilon \approx 0$, would presumably lead to similar results as our experiment despite the huge difference in terms of coerciveness suggested by the tentative definition.

[27] For instance, see the experiments on the convex ultimatum game (e.g., see Suleiman, 1996; Charness and Rabin, 2002).

## Appendix A.

**Table A.1**
Symbols table.

| Variable | Explanation |
|---|---|
| $\gamma(\pi_r^c)$ | The responder's conflict-payoff response function: choice of $y$ as a function of $\pi_r^c = \kappa x$ |
| $\delta$ | The responder's choice variable of accepting ($\delta = 1$) or not |
| $\lambda$ | An indifference curve |
| $\Lambda$ | The set of (the responder's) indifference curves |
| $\kappa$ | Fraction of the proposer-offer the responder keeps after rejection |
| $\pi_p$ | The proposer's payoff |
| $\pi_r$ | The responder's payoff |
| $\pi_r^c$ | The responder's (conflict) payoff in case of rejection |
| $\hat{\pi}_i^{(trmt)}(x)$ | $i$'s supremum payoff in the responder's choice set $\mathcal{S}_x$ (in treatment $trmt$) |
| $br_j$ | a $j$-type's best-response function |
| $E$ | A pair's endowment |
| $u_r(\pi_r, \pi_p)$ | The responder's utility function |
| $WTP_i$ | $i$'s willingness to pay for a marginal increase in $j$'s payoff |
| $x$ | the proposer's offer |
| $\underline{x}_{(\kappa)}$ | The lower acceptance threshold |
| $\overline{x}_{(\kappa)}$ | The upper acceptance threshold |
| $y$ | Proposer payoff after rejection as determined by the responder |
| $y(x)$ | The responder's offer response function |
| $\mathcal{S}_{(x)}$ | The responder's choice set in the proposer-payoff–responder-payoff space (as determined by $x$) |

## Appendix B. Supplementary material

Supplementary material associated with this article can be found in the online version, at http://dx.doi.org/10.1016/j.jebo.2012.10.009.

## References

Andreoni, J., Castillo, M., Petrie, R., 2003. What do bargainers' preferences look like? Experiments with a convex ultimatum game. American Economic Review 93, 672–685.

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

Bereby-Meyer, Y., Niederle, M., 2005. Fairness in bargaining. Journal of Economic Behavior and Organization 56, 173–186.

Bolton, G.E., Ockenfels, A., 2000. ERC: a theory of equity, reciprocity, and competition. American Economic Review 90, 166–193.

Camerer, C.F., 2003. Behavioral Game Theory. Princeton University Press, Princeton.

Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Quarterly Journal of Economics 117, 817–869.

Cox, J.C., Deck, C.A., 2005. On the nature of reciprocal motives. Economic Inquiry 43, 623–635.

Cox, J.C., Friedman, D., Gjerstad, S., 2007. A tractable model of reciprocity and fairness. Games and Economic Behavior 59, 17–45.

Cox, J.C., Friedman, D., Sadiraj, V.V., 2008. Revealed Altruism. Econometrica 76, 31–69.

Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games and Economic Behavior 47, 268–298.

Engelmann, D., Strobel, M., 2004. Inequity aversion, efficiency, and maximin preferences in simple distribution experiments. American Economic Review 94, 857–869.

Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior 54, 293–315.

Falk, A., Fehr, E., Fischbacher, U., 2003. On the nature of fair behaviour. Economic Inquiry 41, 20–26.

Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Quarterly Journal of Economics 114, 817–868.

Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10, 171–178.

Fisman, R., Kariv, S., Markovits, D., 2007. Individual preferences for giving. American Economic Review 97, 1858–1876.

Garrod., L., 2008. Do people really prefer equitable distributions? Working paper.

Greiner, B., 2004. An online recruitment system for economic experiments. In: Kremer, K., Macho, V. (Eds.), Forschung und wissenschaftliches Rechnen 2003, GWDG Bericht 63, Göttingen. Gesellschaft für Wissenschaftliche Datenverarbeitung, pp. 79–93.

Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization 3, 367–388.

Kagel, J., Wolfe, K., 2001. Tests of fairness models based on equity considerations in a three-person ultimatum game. Experimental Economics 4, 203–219.

Levine, D., 1998. Modeling altruism and spitefulness in experiments. Review of Economic Dynamics 1, 593–622.

Rabin, M., 1993. Incorporating fairness into game theory and economics. American Economic Review 83, 1281–1302.

Selten, R., 1967. Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In: Sauermann, H. (Ed.), Beiträge zur experimentellen Wirtschaftsforschung. J.C.B. Mohr, Tübingen, pp. 136–168.

Sobel, J., 2005. Interdependent preferences and reciprocity. Journal of Economic Literature 43, 392–436.

Suleiman, R., 1996. Expectations and fairness in a modified ultimatum game. Journal of Economic Psychology 17, 531–554.

Thöni, C., Gächter, S., 2007. Understanding social interaction effects in the workplace, Working paper, University of St. Gallen.