Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Games and Economic Behavior 47 (2004) 268–298
GAMES and Economic Behavior
www.elsevier.com/locate/geb

A theory of sequential reciprocity

Martin Dufwenberg [a] and Georg Kirchsteiger [b,*]

[a] *Department of Economics, University of Arizona, Tucson, AZ 85721, USA*
[b] *Department of Economics, University of Maastricht, PO Box 616, 6200 MD Maastricht, The Netherlands*

Received 4 September 2001

Abstract

Many experimental studies indicate that people are motivated by reciprocity. Rabin [Amer. Econ. Rev. 83 (1993) 1281] develops techniques for incorporating such concerns into game theory and economics. His theory is developed for normal form games, and he abstracts from information about the sequential structure of a strategic situation. We develop a theory of reciprocity for extensive games in which the sequential structure of a strategic situation is made explicit, and propose a new solution concept—sequential reciprocity equilibrium—for which we prove an equilibrium existence result. The model is applied in several examples, and it is shown that it captures very well the intuitive meaning of reciprocity as well as certain qualitative features of experimental evidence.

JEL classification: A13; C70; D63

Keywords: Reciprocity; Extensive form games

1. Introduction

Reciprocity... is the key to every relationship.
(James Cromwell to Danny DeVito in *L.A. Confidential*)

Almost all of economic theory is built on the assumption that people act selfishly and do not care about the well-being of other human beings. Lots of recent evidence, however, contradicts pure selfishness. For example, Kahneman et al. (1986) show in a seminal paper that consumers' opinions about price increases depend crucially on the costs of the firm,

* Corresponding author.
E-mail addresses: martind@eller.arizona.edu (M. Dufwenberg), g.kirchsteiger@algec.unimaas.nl (G. Kirchsteiger).

0899-8256/$ – see front matter
doi:10.1016/j.geb.2003.06.003

but not on the market conditions—a price increase due to cost increases is regarded as justified, while a demand shock is not a valid justification. Whereas Kahneman et al. (1986) study deals with the *fairness perceptions* of consumers, experimental evidence suggest that *actual behavior* is also shaped by factors inconsistent with pure selfishness. For example, in ultimatum bargaining experiments people often reject allocations in which they receive a much smaller monetary payoff than their partners in favor of an allocation where neither player receives anything (see Roth (1995) for an overview). In gift exchange games, where two persons in turn determine how large gifts to give to one another, a large gift by the first mover is reimbursed by the second mover (see, e.g., Berg et al., 1995; Falk and Gächter, 2002; Fehr et al., 1996). If the size of the gift of the first player is determined on an auction market, these gift exchange forces are even strong enough to prevent the market from clearing (see, e.g., Fehr and Falk, 1999; Fehr et al., 1993, 1998).

These deviations from selfishness may have important economic consequences. As Fehr et al. (1997) show experimentally, the set of enforceable contracts increases considerably due to non-selfish behavior. These effects are of particular importance for understanding labor markets. In a series of theoretical papers Akerlof (1982) and Akerlof and Yellen (1988, 1990) show that fairness is a possible explanation why wages may be above the market clearing level so that involuntary unemployment occurs. Fehr and Kirchsteiger (1994) use this approach to explain why two-tier systems are rarely observed in reality. Bewley (1999) finds strong empirical evidence for the validity of these theories. When asked for the reason why wages remain above the market clearing level in recessions, managers and other labor market participants say that wage declines may destroy "working morale"—workers would decrease their working effort after a decline in wages which therefore cannot be enforced.

All this evidence suggests that people are not motivated solely by material self-interest. Also considerations of altruism, fairness, etc. play a role. Among the models designed to capture some of these phenomena two prominent classes can be distinguished: models that focus on distributional concerns, and models that focus on a concern for reciprocity.[1] The distributional approach permits decision makers to be motivated not only by their own material gain, but rather by the final distribution of the material payoff. In Fehr and Schmidt (1999), for example, it is assumed that for a given own material payoff a person's utility is decreasing in the difference between the own payoff and that of the counterpart. They show that much (selected) experimental evidence can be explained by their theory which, furthermore, has the advantage of being very close to standard models. There is, however, a certain cost. The assumption that individuals care only about final distributions implies that they must be indifferent concerning *how* distributions come

[1] Examples of the former approach are Akerlof and Yellen (1988, 1990), Bolton and Ockenfels (2000), Fehr and Kirchsteiger (1994), Fehr et al. (1998), Fehr and Schmidt (1999), Kirchsteiger (1994), and Levine (1998), where in addition to distributional concerns persons are also motivated by the degree of altruism of the partner. Rabin (1993), Segal and Sobel (1999) and our paper represent the second approach. Charness and Rabin (2002), Cox and Friedman (2002), and Falk and Fischbacher (1998) develop theories that combine elements of both approaches. The dual classification suggested here is not comprehensive; non-selfish motivation can be neither distributional nor related to reciprocity, like Andreoni's (1990) "warm glow of giving" or the emotions considered by Geanakoplos et al. (1989). See Fehr and Gächter (2000) and Sobel (2000) for further discussions of the work mentioned in this footnote.

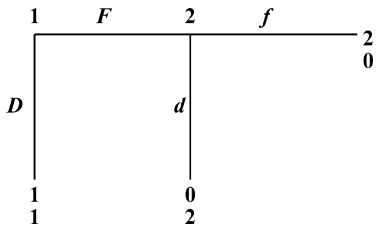

Fig. 1. Game Γ_1.

about. This is problematic if in fact individuals regard information about their co-players' specific choices or intentions as important to their decision making.[2]

Rabin (1993) convincingly argues that intentions play a crucial role when individuals are motivated by reciprocity considerations.[3] When a person wants to be kind to someone who was kind to her, and unkind to unkind persons, she has to assess the kindness (or unkindness) of her own action as well as that of others. To do this she has to consider the intentions that accompany an action. Take as an example the game Γ_1 in Fig. 1 (where the given payoffs are in monetary units).

Is F an unkind action? Clearly, this depends on what player 1 believes that player 2 will do. Suppose player 1 believes that player 2 will choose d. By choosing F player 1 then intends to give a payoff of 2 to player 2, whereas player 2 would get a payoff of only 1 if player 1 chose D. Hence, one may conclude that player 1 acts kindly if he chooses F. By an analogous argument, however, one must conclude that 1 is unkind if he chooses F while believing that 2 will choose f. This example shows not only that intentions are crucial in order to model reciprocity; it also makes clear that intentions depend on the *beliefs* of the players. Furthermore, the kindness of a player also depends on the *possibilities* he has. Change the game of Fig. 1 such that player 1's strategy set consists only of F—he has to "choose" F. In such a game a "choice" of F is of course neither kind nor unkind—it is simply the only thing that 1 can do. Hence, in order to model the impact of intentions one has to take into explicit account both the possibilities and the beliefs of the players.

This is what Rabin (1993) does. He assumes that the players in two-player normal form games experience psychological payoffs in addition to the underlying material payoffs. The former payoffs depend on the players' kindness, which in turn depends on beliefs. Given the belief of player i about the strategy choice of the other player j, i is kind to the extent that he believes he gives j a (relatively) high material payoff. In this sense, i's kindness depends on the payoff he intends to "give" to j, compared to the payoffs he believes it would be possible to give her—intentions and possibilities define the kindness of action. Similarly, how kind i believes j is depends on a belief of i about a belief of j,

[2] A related problem discussed in social choice theory concerns whether welfare assessments can be made with reference to final distributions only. See Sen (1979) for a critical discussion.

[3] A word of caution about terminology is in order, since the meaning of the term "reciprocity" varies considerably in the literature. Some papers define certain actions as reciprocal, without making explicit reference to intentions. Other authors (for example Bolton and Ockenfels, 2000) distinguish between direct and indirect reciprocity, the former being a principle like the one we describe here (and simply call "reciprocity"), whereas the latter is a pure concern for distributive justice.

since j's kindness depends on j's belief. Rabin then models the psychological payoff as a concern for reciprocity such that i wants to be kind to j if he believes j to be kind to him (as long as the material payoff does not become too important for i). Notice that since the model involves belief-dependent motivations, the utility functions have to be defined on a richer domain than in standard game theory where payoffs depend on actions only. The framework of psychological game theory, developed by Geanakoplos et al. (1989), provides appropriate tools which Rabin adopts. He shows that a reformulation of his model using standard game theory is impossible—since intentions matter, models of reciprocal behavior have to lead to different results than an approach where beliefs are not allowed to affect payoffs directly.

However, Rabin points out an important limitation of his model. As it is a normal form construct it does not take into account the dynamic structure of a strategic situation, and "[e]xtending the model to sequential games is also essential for applied research" (Rabin, 1993, p. 1296). If an equilibrium is calculated using the normal form of some extensive game, non-optimizing behavior may be prescribed at information sets that are not reached. The problem resembles that in usual game theory, where in Nash equilibrium players do not necessarily optimize off the equilibrium path. However, to handle this problem turns out to be more complicated than in usual game theory. As play unravels in a sequential game, a player who revises his beliefs may have to also revise beliefs about how kind other players are, since kindness depends on beliefs. Therefore, the way that the player is affected by reciprocity concerns may differ dramatically between different parts of the game tree. To illustrate all of this, consider the "Sequential Prisoners' Dilemma" game of Fig. 2, which is a stylized version of the experiments conducted by Fehr et al. (1993, 1998) and which has been experimentally tested also by Clark and Sefton (2001).

It can be easily shown that cooperation by player 1 (the choice C) and unconditional cooperation by player 2 (i.e., the choice c at each node controlled by 2) is one of the equilibria admitted by Rabin's theory (defined in the normal form of Γ_2), as long as the concern for material payoffs does not overcome the concern for reciprocity.[4] Unconditional cooperation of player 2, however, is very implausible. 1's choice of D guarantees that 2 gets

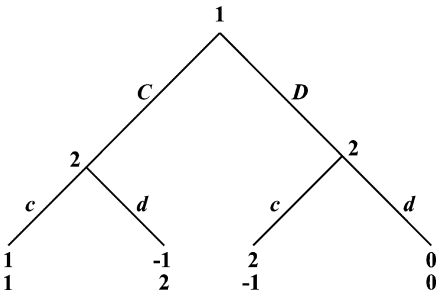

Fig. 2. Game Γ_2—the sequential prisoners' dilemma.

[4] Due to a normalization in Rabin's theory, this means that the units in Fig. 1 must be thought of as (small enough) fractions of some unit that measures material value.

a lower monetary payoff than she would get following the choice of C. If 2 is motivated by reciprocity, why should she be kind after 1 chose such an unkind action?[5]

The equilibrium is sustained because Rabin's normal form concept does not insist on optimization at 2's second node (which in the equilibrium is not reached). However, the problem cannot be solved merely by looking at the extensive form and mandating optimization at all histories of play. After all, for 2 to choose c at her rightmost node may be in her interest *if she believes 1 is kind*. The point is, though, that such a belief would make no sense, if 2 believes that 1's choice was deliberate and purposeful (as we will assume). Even if 2 initially believes that 1 is kind she should not maintain this belief after 1 chose D. Rather she should then regard 1 as unkind, so reciprocity motives (in addition to selfish motives) would motivate her to take revenge by choosing d.

The general upshot is that a sensible model of reciprocity in sequential games must handle with care how beliefs change and how this affects reciprocity considerations. Incorporating such a "sequential reciprocity principle" is important in many potential applications which have a non-trivial dynamic structure. For example, the game in Fig. 2 is a very stylized version of the fair wage effort models of Akerlof and Yellen (1988, 1990): The firm (player 1) makes a generous or greedy wage offer and the worker (player 2) decides whether to provide a high or a low working effort. Given the sequential structure of this game and other potential applications, it is crucial to derive a concept of sequential reciprocity. This is the main objective of our paper.

In our model, just as in Rabin's, kindness and perceived kindness range from negative to positive. Reciprocation entails responding to positive perceived kindness with positive kindness, and to negative perceived kindness with negative kindness; by virtue of this "sign-matching" reciprocation adds to utility. The main novelty of our approach is that we take account of how strategic choices and reciprocity motivation change as new subgames are entered, and that we impose that strategic choices prescribe best responses in all stages.

In order to highlight and isolate the consequences of sequential reciprocity, and in order to facilitate a comparison with Rabin's (1993) model, we focus exclusively on incorporating a concern for reciprocity (in addition to selfish motivation). We disregard distributional concerns. This is not to say that such concerns are unimportant. In reality both motives seem to play a role.[6] However, in this paper our objective is not to explain as many experimental findings as possible. Rather, we concentrate on reciprocal motivation and develop a model which is useful for analyzing its impact in sequential games. We

[5] In the experiments by Fehr et al. (1993, 1998) and Clark and Sefton (2001) such behavior was nearly never observed.

[6] The experimental evidence on the importance of reciprocity vis-à-vis the importance of distributional concerns is somewhat mixed. Whereas Bolton et al. (1996) find that only the final distribution matters, results of Blount (1995), Bolle and Kritikos (1999), Charness (1996), Charness and Rabin (2002), Gneezy et al. (2000) suggest that reciprocity (especially *negative* reciprocity: being unkind to someone who was unkind) as well as distributional concerns play a role. Rabin (1998) discusses empirical findings about the importance of reciprocity. Outside economics social psychologists have found strong experimental evidence of the importance of reciprocity, stressing the important role played by intentions (see e.g. Goranson and Berkowitz, 1966; Greenberg and Frisch, 1972, or Tesser et al., 1968). Also anthropologists and sociologists regard reciprocity as a main factor of human behavior, crucial for the functioning of human societies. For an overview of this literature, see Komter (1996).

regard our approach as complementary to work that aims at modeling other motivational forces.

In the next section we present our model and define the concept of a *sequential reciprocity equilibrium* (*SRE*). In Section 3 we state an existence theorem for this concept. In Section 4 we apply the SRE concept to some well-known games, and discuss how reciprocity shapes the analysis. In Section 5 our approach is compared to Rabin's (1993) approach. Section 6 contains concluding comments. The theorem of Section 3, as well as some observations of Section 4, are proved in an Appendix A.

2. The model

In the introduction we argued that whether a person's action is kind or unkind depends not only on what he does but also on what he *believes* will be the consequence of his decision, as compared to what he believes would be the consequences of other decisions. Said differently, a person's kindness depends on his intentions. When another person wants to reciprocate kindness with kindness, she must hence form beliefs about the first person's intentions. Since intentions depend on beliefs, it follows that reciprocal motivation depends on *beliefs about beliefs*.

To come to grips with such issues we work within the framework of *psychological game theory* (Geanakoplos et al., 1989). Psychological games differ from standard games in that a player's payoff depends not only on what strategy profile is played, but possibly also on what are the player's beliefs about other players' strategic choices or beliefs. The approach we use is directly inspired by Rabin (1993). We start off with a standard game, which is viewed as a description of a strategic situation which specifies only the material payoffs. We then derive a *psychological game* in which the payoff functions are redefined so as to reflect also reciprocity considerations. The main difference between our model and Rabin's is that he works with normal form representations of strategic situations, while we work with extensive forms and impose a requirement of sequential rationality.

When this is done, a subtle issue arises: If a subgame is reached, perhaps unexpectedly, this may sometimes force a player to change his beliefs about the strategy profile being played. Since kindness relates to beliefs, assessments about kindness may therefore change and affect the ways in which a player is motivated by reciprocity concerns. It becomes necessary to somehow distinguish between a player's initial and subsequent beliefs. We handle this by keeping track of how the players' beliefs change as new subgames are reached, and by assuming that whenever a player makes a choice he is motivated according to the beliefs he holds at that stage. These assumptions are central to our model. We already argued (in connection to Γ_2) in the introduction that if reciprocity is important, one may get unreasonable conclusions unless players are assumed to update their assessments of how kind their co-players are as play unravels, and then reciprocate accordingly. However, this also means that the psychological games we consider do not belong to the class of psychological games that receives most attention in Geanakoplos et al. (1989), as they confine attention to psychological games where only *initial* beliefs may influence players' valuations of different strategy profiles (although they suggest (p. 78) that other assumptions may be important).

We restrict attention to finite multi-stage games with observed actions and without nature; see Fudenberg and Tirole (1991, Chapter 3). Play proceeds in "stages" in which each player, along any path reaching that stage

(i) knows all preceding choices,
(ii) moves exactly once (a player may then have a *singleton* choice set; this trick is used to formally model games with alternating moves, although we do not depict such choices when we draw game trees), and
(iii) obtains no information about other players' choices in that stage.

Thus all instances of imperfect information arise from simultaneous moves. The restriction to multi-stage games with observed actions facilitates the description of strategies and beliefs that follows, without essentially compromising the scope of the model since most applied and experimental work is concerned with such games.

Formally, let $N = \{1, \ldots, n\}$ be the set of players where $n \geqslant 2$. Let H be the set of choice profiles, or *histories*, that lead to subgames ($\varnothing$ belongs to H and "leads to" the root). Let A_i be the set of behavior strategies of $i \in N$; each strategy assigns for each history $h \in H$ a probability distribution on the set of possible choices of i at h. Note that A_i allows for randomization. Our favored interpretation of the concept to be developed entails that players make pure choices only. Nevertheless, the concept allows for randomization, which is to be interpreted in terms of the frequencies with which pure choices may be made in a "population". More comments on this follow later in this section.

Define $A = \prod_{i \in N} A_i$. Using the assignment of payoffs to endnodes, we can derive a payoff function for each player which depends on what profile in A is played. Let $\pi_i : A \to \mathsf{R}$ denote this function. We shall refer to π_i as player i's *material payoff function*. The material payoffs represent money, or some other objectively measurable quantity.

The material payoff is not the only payoff which we shall assume motivates i in his decision making. To get i's *utility*, which is the function that i wants to maximize, we shall add a *reciprocity payoff* to i's material payoff. The reciprocity payoff depends on i's beliefs about other players' strategies and beliefs. We represent beliefs as behavior strategies. However, in order to avoid confusion, we introduce separate notation for beliefs. Let $B_{ij} = A_j$ be the set of possible beliefs of player i about the strategy of player j. Let $C_{ijk} = B_{jk} = A_k$ be the set of possible beliefs of player i about the belief of player j about the strategy of player k.

As exemplified by the discussion concerning the sequential prisoners' dilemma in the introduction, a player's kindness and perception of another player's kindness may differ after different histories. In order to capture this it is important to keep track of how each player's behavior, beliefs, kindness, and perception of others' kindness differ across histories. We do this as follows: Let Γ be a finite multi-stage games with observed actions and without nature. With $a_i \in A_i$, $h \in H$, let $a_i(h)$ be the (updated) strategy that prescribes the same choices as a_i, except for the choices that define history h which are made with probability 1. For example, if h is the node where 2 moves in Γ_1, then $D(h) = F$, $F(h) = F$, $d(h) = d$, and $f(h) = f$. Note that $a_i(h)$ is uniquely defined for any history. For beliefs $b_{ij} \in B_{ij}$ or $c_{ijk} \in C_{ijk}$, define updated beliefs $b_{ij}(h)$ and $c_{ijk}(h)$ in a fashion

completely analogous to that of updated strategies. For example, if h is the node where 2 moves in Γ_1, and if $b_{21} = D$, then $b_{21}(h) = F$.

Suppose player i plays the strategy $a_i \in A_i$, initially believing the others to play $(b_{ij})_{j\neq i}$ and to believe $(c_{ijk})_{k\neq j}$. After history h, we model i as playing strategy $a_i(h) \in A_i$, believing the others to play $(b_{ij}(h))_{j\neq i}$ and to believe $(c_{ijk}(h))_{k\neq j}$. This specification entails that players give up beliefs that involve randomization in favor of beliefs with pure choices as choices are realized. For example, consider Γ_2, let h be the node where 2 moves after 1 chooses D, and consider $b_{21} = (1-\mu)C + \mu D$ (the notation refers to the randomized choice which assigns probabilities $1-\mu$ and μ to the choices C and D, respectively). For any value of μ, we get $b_{21}(h) = D$. The interpretation is that whatever 2 believes about 1 at the root of the sequential prisoners' dilemma, when she is asked to play after observing D she judges that 1 chose D with probability one. If $\mu > 0$, this form of updating is "Bayesian," given our favored interpretation of strategies involving randomization in terms of frequencies of pure strategies rather than as consciously randomized choices (cf. our comment about the interpretation of strategies involving randomization above, and the further discussion after Definition 4 below). If $\mu = 0$ the updating is still Bayesian, but of a particular form, reflecting an assumption that players treat the choices of others as purposeful and deliberate. In a game like Γ_2 this has the upshot that when 2 is asked to play after 1 chose D, she treats 1 as if he chose D on purpose rather than as if he tried to choose C but made a mistake. This is crucial to guarantee that our theory induces a retaliatory motive in game Γ_2.

This way of updating beliefs handles the problem of how the preferences change when unexpected moves occur. Our solution is not, however, intended to tackle the general problem of what to conclude from moves that in equilibrium should not occur. Of course, in such cases the question arises why a "surprised" player should stick to his initial beliefs concerning later play, facing the *fait accompli* that the initial beliefs were incorrect for earlier play. This problem arises also in standard game theory for any concept of sequential rationality (e.g. subgame perfection), and it goes beyond the scope of this paper to address the matter. For a discussion of the problem in the context of standard games, see Reny (1992).

We wish to capture that each player i wants to some extent to be kind in return to any player j who is kind to i. What does it mean for i to be kind to j? Suppose that i chooses $a_i \in A_i$ and that he believes that all other players make choices according to the profile $(b_{ij})_{j\neq i} \in \prod_{j\neq i} B_{ij}$. Following Rabin (1993), we note that player i then believes that he chooses in such a way that j's material payoff will be $\pi_j(a_i, (b_{ij})_{j\neq i})$. He also believes that the feasible set of material payoffs for j is $\{\pi_j(a_i', (b_{ij})_{j\neq i}) | a_i' \in A_i\}$. How kind i is to j can now be measured in terms of the relative size of $\pi_j(a_i, (b_{ij})_{j\neq i})$ within this set.

While this measurement may be done in several ways, there is one particular aspect that must be handled carefully. Consider the game Γ_3, which is related to Γ_2, as shown in Fig. 3.

Suppose 1 plays the strategy D, and that he believes with probability one that player 2 is playing the strategy cd (meaning the strategy that assigns the choice c to the leftmost node and d to the rightmost node). (Any other belief will in fact also do to make our point.) One sees that 1 believes he chooses the material payoff $\pi_j(D, cd) = 0$ for player 2, from the feasible set of material payoffs for j which is $[-1000, 1]$. Within this set, 0 is a rather

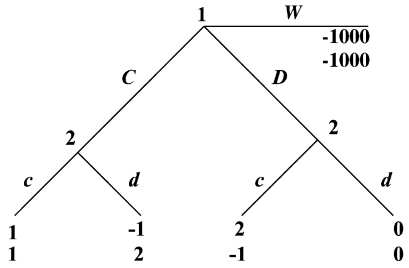

Fig. 3. Game Γ_3.

large number. Should one therefore conclude that player 1 is rather kind by choosing D? We would find this unreasonable. The fact that the internecine choice W is possible for 1 seems to be irrelevant for drawing conclusions regarding the kindness of the choices C and D. The choice of W guarantees an inefficient outcome that hurts both players. By contrast, each of the actions C and D may lead to outcomes that are efficient in terms of material payoff allocations.

We propose that 1's kindness if he chooses D in Γ_3 should be the same as if 1 chooses D in Γ_2, if 1 has the same beliefs in the two cases. That is, 1's kindness should be assessed with reference to the relative position of $\pi_2(D, cd) = 0$ for player 2 in the set $\{\pi_2(\mu \cdot C + (1 - \mu \cdot D, cd) \mid \mu \in [0, 1]\} = [\pi_2(D, cd), \pi_2(C, cd)] = [0, 1]$. Since 0 is the lowest number in this set, player 1 should be considered unkind if he chooses D.

In general, we proceed as follows. Define player i's *efficient strategies* by

$$E_i = \Bigg\{ a_i \in A_i \;\Big|\; \text{there exists no } a_i' \in A_i \text{ such that for all } h \in H,\ (a_j)_{j\neq i} \in \prod_{j\neq i} A_j,$$
$$\text{and } k \in N \text{ it holds that } \pi_k\big(a_i'(h), (a_j(h))_{j\neq i}\big) \geqslant \pi_k\big(a_i(h), (a_j(h))_{j\neq i}\big),$$
$$\text{with strict inequality for some } \big(h, (a_j)_{j\neq i}, k\big) \Bigg\}.$$

Intuitively, a strategy is inefficient if there exists another strategy which conditional on any history of play and subsequent choices by the others provides no lower material payoff for any player, and a higher material payoff for some player for some history of play and subsequent choices by the others. For example, in Γ_1 and Γ_2 all strategies are efficient for both players. In Γ_3 all strategies are efficient, except those strategies of player 1 that assign positive probability to the choice W.

If a_i is an inefficient strategy, it involves "wasteful play" following some history h. Reaching h may be inconsistent with early choices according to a_i. However, the equilibrium notion we develop (see Definition 4) requires optimal choices also after histories not played in equilibrium, with the player using strategy $a_i(h)$ instead of a_i. It seems natural to us that inefficiency concerns then relate to wasteful play from h on. Our efficiency definition picks this up.

The definition of E_i differs from the corresponding definition in Rabin (1993). In Section 5 we discuss this in detail and motivate our modeling choice further. Here we move on to our definition of kindness, which is based on an idea of Rabin's: i's kindness to j is zero if he believes that j's material payoff will be the average between the lowest and

the highest material payoff of j that is compatible with i choosing an efficient strategy.[7] It is convenient to have a special notation which describes this number as a function of i's beliefs about the profile being played. We call this function $\pi_j^{e_i}$, defined by

$$\pi_j^{e_i}\big((b_{ij})_{j\neq i}\big) = \frac{1}{2}\cdot\Big[\max\big\{\pi_j\big(a_i,(b_{ij})_{j\neq i}\big)\;\big|\;a_i\in A_i\big\} + \min\big\{\pi_j\big(a_i,(b_{ij})_{j\neq i}\big)\;\big|\;a_i\in E_i\big\}\Big].$$

Think of $\pi_j^{e_i}((b_{ij})_{j\neq i})$ as a norm for i describing the "equitable" payoffs for player j when i's beliefs about other players' behavior are summarized by $(b_{ij})_{j\neq i}$. We use $\pi_j^{e_i}((b_{ij})_{j\neq i})$ as a reference point for measuring how kind i is to j. If i chooses a strategy a_i such that $\pi_j(a_i,(b_{ij})_{j\neq i}) = \pi_j^{e_i}((b_{ij})_{j\neq i})$, then his kindness to j is zero. Otherwise i's kindness to j is proportional to how much more or less material payoff than $\pi_j^{e_i}((b_{ij})_{j\neq i})$ that i thinks will be the consequence for j. More specifically:

Definition 1. The kindness of player i to another player $j\neq i$ at history $h\in H$ is given by the function $\kappa_{ij}: A_i\times\prod_{j\neq i}B_{ij}\to\mathsf{R}$ defined by

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

Intuitively, Definition 1 reflects the idea that i's kindness to j is proportional to "the size of his gift." The definition differs from Rabin's analogous one which includes a normalization; see Section 5 for further discussion of this.

Having defined kindness, we now turn to reciprocity—the idea that if j is kind (unkind) to i, then i wants to be kind in return (take revenge). Since j's kindness depends on j's beliefs, i cannot observe j's kindness directly. However, i can consult his beliefs about j's actions and beliefs and draw inferences concerning j's kindness. We introduce a function λ_{iji} to keep track of how kind i believes that j is to i:

Definition 2. Player i's beliefs about how kind player $j\neq i$ is to i at history $h\in H$ is given by the function $\lambda_{iji}: B_{ij}\times\prod_{k\neq j}C_{ijk}\to\mathsf{R}$ defined by

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

Since $B_{ij}=A_j$ and $C_{ijk}=B_{jk}$, the function λ_{iji} is mathematically equivalent to κ_{ji}, although λ_{iji} captures a psychological component that pertains to player i, not player j.

It is now time to specify the utilities which the players are assumed to maximize:

Definition 3. Player i's utility at history $h\in H$ is a function

$$U_i: A_i\times\prod_{j\neq i}\left(B_{ij}\times\prod_{k\neq j}C_{ijk}\right)\to\mathsf{R}$$

defined by

[7] We see no deep justification for picking the average (rather than some other intermediate value), except that the choice is simple and does not affect the qualitative performance of the theory.

$$U_i\big(a_i(h), \big(b_{ij}(h), \big(c_{ijk}(h)\big)_{k\neq j}\big)_{j\neq i}\big)$$
$$= \pi_i\big(a_i(h), \big(b_{ij}(h)\big)_{j\neq i}\big)$$
$$+ \sum_{j\in N\setminus\{i\}} \big(Y_{ij}\cdot\kappa_{ij}\big(a_i(h), \big(b_{ij}(h)\big)_{j\neq i}\big)\cdot\lambda_{iji}\big(b_{ij}(h), \big(c_{ijk}(h)\big)_{k\neq j}\big)\big),$$

where Y_{ij} is an exogenously given non-negative number for each $j\neq i$.

Player i's utility is the sum of n terms. The first term is his material payoff, the remaining terms express his *reciprocity payoff with respect to each player* $j\neq i$. The constant Y_{ij} measures how sensitive i is to reciprocity concerns regarding player j. Our formulation admits that a player's reciprocity sensitivity varies depending on which other player is targeted. In application this may be useful; perhaps, for example and as assumed in Dufwenberg and Kirchsteiger (2000), a worker is reciprocally motivated towards his employer but not towards an unemployed outsider. If $Y_{ij}>0$ the following is true: If i believes that j is kind to him (i.e., $\lambda_{iji}(\cdot)>0$), then i's reciprocity payoff with respect to j is increasing in i's kindness to j. Furthermore, the higher is $\lambda_{iji}(\cdot)$, the more material payoff i is willing to give up in order to do j a favor. If i believes that j is unkind to him (i.e., $\lambda_{iji}(\cdot)<0$), then i's reciprocity payoff with respect to j is decreasing in i's kindness to j. This is the way in which U_i reflects the idea that if i thinks that j is kind (unkind) to him, then i wants to be kind in return (take revenge). Of course, when i optimizes he may have to make tradeoffs between various reciprocity payoffs with respect to different players as well as his material payoff.

We can now append any game Γ with a vector of utilities $(U_i)_{i\in N}$ defined as above and get the tuple $\Gamma^* = (\Gamma, (U_i)_{i\in N})$. We refer to any Γ^* constructed in this fashion as a *psychological game with reciprocity incentives*. Note that such a Γ^* is not a "game" in the traditional sense, since the utility functions U_i are defined on domains that include subjective beliefs, and not only strategic choices.

We propose a solution concept that is applicable to any psychological game with reciprocity incentives. We look for equilibria in which each player in each history chooses optimally given his beliefs. The players' initial first and second order beliefs are required to be correct, and following each history of play the beliefs are updated as explained above.

The definition of the equilibrium requires the following additional piece of notation: For any $a=(a_i)_{i\in N}\in A$ and history $h\in H$, let $A_i(h,a)\subseteq A_i$ denote the set of strategies that prescribe, for each player i, the same choices as the strategy $a_i(h)$ for all histories other than h. That is, $A_i(h,a)$ is the set of strategies i may use if he behaves according to $a_i(h)$ at other histories than h, but is free to make any choice at h. $A_i(h,a)$ is nonempty, since $a_i(h)\in A_i(h,a)$.

Definition 4. The profile $a^*=(a_i^*)_{i\in N}$ is a *sequential reciprocity equilibrium* (SRE) if for all $i\in N$ and for each history $h\in H$ it holds that

(1) $a_i^*(h)\in \underset{a_i\in A_i(h,a^*)}{\operatorname{argmax}}\ U_i(a_i,(b_{ij}(h),(c_{ijk}(h))_{k\neq j})_{j\neq i})$,
(2) $b_{ij}=a_j^*$ for all $j\neq i$,
(3) $c_{ijk}=a_k^*$ for all $j\neq i, k\neq j$.

By condition (1) of Definition 4, a SRE is a strategy profile such that at history h each player makes choices which maximizes his utility given his beliefs and given that he follows his equilibrium strategy at other histories. At the initial stage, conditions (2) and (3) guarantee that the initial beliefs are correct. At any subsequent history, condition (1) requires that beliefs assign probability one to the sequence of choices that define that history, but are otherwise as the initial beliefs.

Because of (1), an SRE rules out profitable "local" deviations at any particular stage. The definition does not exclude the possibility that a "joint" deviation at several successive stages might increase a player's utility *as evaluated at the first of the stages where the player deviates*. However, we do not regard this as a drawback. Note that a player's preference between two different strategies, because of the updating of strategies and beliefs with respect to histories (as stated in Definition 3), can change during the play of the game. Hence, our framework allows for dynamically inconsistent preferences.[8] A player might have a multi-stage deviation, which is profitable according to the evaluation of an early deviation stage, even if his strategy is part of an SRE. But whenever this is the case, that player will no longer have an incentive to carry out the involved deviation at some later deviation stage (as seen by the fact that condition (1) must be applied to the history where the later deviation is supposed to occur). In order to realize such a multi-stage deviation, the player must be able to bind himself at the first deviation stage to follow the deviation strategy also at the other deviation stages. If such self-commitment were feasible, it should be modeled explicitly and thus should have led to a different game tree in the first place.

With $Y_{ij} = 0$ for all $i, j \in N$ the optimality check involves material payoffs only. In this case Definition 4 requires that a^* is a subgame perfect equilibrium in Γ (this follows from the "one-shot-deviation property" of subgame perfect equilibrium; cf. Hendon et al., 1996). In the next section we prove that at least one SRE exists in every psychological game with reciprocity incentives.

Definition 4 refers to behavior strategies, which may involve randomization. However, as commented on in two places earlier in this section, our preferred interpretation of our equilibrium notion does not incorporate conscious randomization by individual players; we envisage players as choosing pure strategies only. The probabilities specified by some randomized choice rather reflect frequencies with which pure choices are made, in some society where people from time to time play a given game. Our interpretation thus parallels John Nash's (1950) "mass-action" interpretation of the Nash equilibrium concept.[9] Of course, the problem arises of what to conclude from observing a move that is

[8] Dynamically inconsistent preferences were first analyzed in Strotz (1956). Contrary to our approach, Strotz (1956) as well as more recent work inspired by Strotz (e.g. Harris and Laibson, 2001) investigate single-agent decision problems. In these papers the inconsistencies occur because decision makers are more impatient when they make short run trade-offs than when they make long run trade-offs, whereas in our strategic setting the preferences might be inconsistent due to the change of beliefs as play proceeds.

[9] Nash writes: "[T]he participants are supposed to accumulate empirical information on the relative advantages of the various pure strategies at their disposal. To be more detailed, there is a population (in the sense of statistics) of participants for each position of the game. Let us also assume that the 'average playing' of the game involves n participants selected at random from the populations, and that there is a stable frequency with which each pure strategy is employed by the 'average member' of the appropriate population. Thus the assumption we made in

inconsistent with the initial beliefs, but this issue is not idiosyncratic to our approach and it goes beyond the scope of our paper to sort it out.

3. Existence

In this section we provide an existence theorem for our concept of a *sequential reciprocity equilibrium*. The proof is in Appendix A. Here we provide intuition for the techniques used, and explain why these differ from techniques commonly used in standard game theory. We also explain why the existence theorems in Geanakoplos et al. (1989) are not applicable.

The difficulty of proving existence arises from the fact that kindness as well as believed kindness at some history h depend on the beliefs and second order beliefs about the actions following *all* histories. In particular, the belief about the behavior in histories that do not follow h is also important for the evaluation of kindness and expected kindness at h. Since in equilibrium beliefs have to be consistent with the equilibrium strategy profile, it follows that in games with reciprocity incentives it is in general impossible to determine equilibrium choices by looking at isolated subgames. Therefore, the backward inductive techniques that are usually used for proofs in standard game theory cannot be applied.

It may be illuminating to illustrate this impossibility with an example. Consider Γ_2 and the history where 2 moves after 1 has chosen C. To determine an equilibrium choice for 2 it is necessary to know her equilibrium belief of how kind is 1. Since 1's kindness depends on his belief of 2's behavior following *the other history where she moves*, it is impossible to determine 2's equilibrium behavior at the leftmost history without specifying what goes on also at the rightmost history.

The difficulty can be overcome by considering *all histories simultaneously*. Recall that Definition 4 requires robustness only against "local" deviations, at any particular history. Although a player at any history has his utility calculated with reference to strategies and beliefs that relate to choices at all histories, his actual optimization task locally refers to choices at that history only. There is sufficient structure to apply Kakutani's fixed point theorem to a best-reply correspondence that distinguish not only between players but also between the different histories at which they move. This technique allows us to show

Theorem. *There exists a SRE in every psychological game with reciprocity incentives.*

See proof in Appendix A.

This result is not covered by the existence theorems for solutions of psychological games presented in Geanakoplos et al. (1989). As mentioned in Section 2, Geanakoplos et al. focus on psychological games where only initial beliefs may influence players' valuations of different strategy profiles, while in the psychological games with reciprocity incentives subsequent beliefs also have bearing on these evaluations.

---

this 'mass-action' interpretation lead to the conclusion that the mixed strategies representing the average behavior in each of the populations form an equilibrium point." See Weibull (1994) for a discussion of this interpretation.

4. Applications

In this section we apply the concept of SRE to some well known games. This serves the purpose of showing how reciprocity motives affect the analysis. In each application we start with an extensive game that has perfect information and generic material payoffs, so the subgame perfect equilibrium for selfish players is unique and in pure strategies. We then calculate the SRE in the corresponding psychological games with reciprocity incentives. It turns out that even for generic material payoffs (and generic reciprocity parameters) the SRE need not be unique, as the first example (the Sequential Prisoners' Dilemma) shows. One might conjecture that the number of equilibria increases with the number of stages in games. However, that is not necessarily the case. Our second example (the Centipede game) shows that the SRE can be a unique in games with many stages. These examples also show that there need not exist an equilibrium in pure strategies.

The third example concerns a three-player game ("So Long, Sucker"), and indicates the usefulness of our approach for games with more than two players. Furthermore, this example shows that in some games all players who make choices are unkind to each other along the equilibrium path. This result can be contrasted with the outcome in the second example, in which for large enough reciprocity parameters there is no SRE in which the players are unkind to one another. In fact, in the second example only positive emotions are predicted, in the sense that the players are kind to one another in the unique SRE.

In all applications we restrict our attention to equilibria for non-negative reciprocity parameters that are generic in the sense that the conditions on these parameters used for the characterization of equilibria (see below) are never fulfilled with equality. We allow for the case with vanishing reciprocity parameters ($Y_{ij} = 0$ for all i, j), which coincides with the standard approach without reciprocity. For each game, we first give a brief summary of how the subgame perfect equilibrium and the sequential reciprocity equilibria look, and then provide a detailed calculation of these equilibria. The more lengthy proofs are relegated to Appendix A.

4.1. The sequential prisoners' dilemma

The first game we analyze is the sequential prisoners' dilemma Γ_2 of Fig. 2, discussed in the introduction. We will see that if player 2's sensitivity towards reciprocity[10] is strong enough, she cooperates if player 1 cooperates and defects if 1 defects. This prediction matches the experimental evidence mentioned in the Introduction. For intermediate values of Y_2 player 2 makes a randomized choice in all SRE—an SRE in pure strategies does not exist for this range of values of Y_2. For low enough values of Y_2 player 2 always defects. The possible patterns of equilibrium behavior of player 1 are more complex. In particular, when both players' reciprocity sensitivities are high enough, both cooperation and defection are compatible with equilibrium play for player 1. In the first case, the "reciprocal" behavior of player 2 induces player 1 to cooperate for material as well as

[10] Since this and the next example are two-player games, we simplify notation by using Y_1 and Y_2 instead of Y_{12} and Y_{21}.

reciprocity reasons. The second case is characterized by 'self-fulfilling expectations." Along the equilibrium path each player believes that the other one is unkind, and hence is unkind in return. In each case the equilibrium is "strict" in the sense that no player has an "unused best response" at any history.

We now move to the detailed calculations. We first analyze player 2's behavior, which is summarized by two observations:

Observation 1. If player 1 defects (chooses D), player 2 also defects in every SRE.

To see this, note that for any possible strategy of 2, player 2 gets less when 1 chooses D than when he chooses C. It follows that whatever 1 believes about 2's strategy, 1's choice of D is unkind, and hence 2 must believe that 1 is unkind. Hence, the reciprocity payoff as well as the material payoff makes player 2 choose d.

Observation 2. If player 1 cooperates, the following holds in all SRE:

(1) if $Y_2 > 1$, player 2 cooperates;
(2) if $Y_2 < 0.5$, player 2 defects;
(3) if $0.5 < Y_2 < 1$, player 2 cooperates with a probability of $p = (2 \cdot Y_2 - 1)/Y_2$.

See proof in Appendix A.

Observations 1 and 2 are intuitively very plausible. They show that as long as player 2 is sufficiently motivated by reciprocity, 2's choice depends on the behavior of 1. On the other hand, a relatively selfish player 2 defects irrespectively of 1's action. For any intermediate case (Observation 2(3)), player 2 will randomize with given probabilities. Hence, Observations 1 and 2 together imply that for a given parameter Y_2 player 2's equilibrium behavior is unique. This is, however, in general not true for 1's behavior which can be characterized by three observations:

Observation 3. If $Y_2 < 0.5$, defection is 1's unique equilibrium behavior.

To see this, notice that for $Y_2 < 0.5$ player 2 always defects (see Observations 1 and 2). Hence, only the reciprocity part of the utility function can make 1 choose C (since the material payoff alone would dictate the choice D). However, 2's strategy of always defecting is unkind. Hence, the reciprocity payoff as well as the material payoff makes player 1 choose D.

Observation 3 considers only equilibria for a player 2 who does not behave reciprocally, so player 1 has no incentive to cooperate and therefore defects.

Observation 4. If $Y_2 > 1$, 1's equilibrium behavior is characterized by one of the following three possibilities:

(1) player 1 cooperates (regardless of Y_1);
(2) $Y_1 > 1$ and player 1 defects;
(3) $Y_1 > 1$ and player 1 cooperates with probability $q = (Y_1 - 1)/2 \cdot Y_1$.

See proof in Appendix A.

Observation 4 considers only equilibria for a player 2 who behaves reciprocally. Observation 4(1) corresponds to the intuitively most plausible equilibrium—since 2 is using strategy cd, 1's material payoff as well as his reciprocity payoff leads him to cooperate. If, however, reciprocity is important enough, there also exists other equilibria that are characterized by "self-fulfilling expectations". If 1 believes that 2 initially believes that 1 chooses D, and that 2 defects in that case, then 1 believes that 2 is unkind. This in turn leads 1 to be unkind, i.e. to play D(or to randomize). Of course, this only works when 1 is sufficiently motivated by reciprocity—if this is not the case, 1's material payoff together with the reciprocal behavior of 2 would make him cooperate.

We next examine equilibrium behavior when 2 is moderately motivated by reciprocity. In this case 2 answers a cooperative choice of 1 with a randomized choice.

Observation 5. If $0.5 < Y_2 < 1$, 1's equilibrium behavior is characterized by one of the three following possibilities:

(1) $Y_2 > 2/3$ and player 1 cooperates;
(2) $Y_1 > 3 \cdot Y_2 - 2$ and player 1 defects;
(3) $Y_1 > 3 \cdot Y_2 - 2$, $Y_2 > 2/3$ and player 1 cooperates with probability $q = Y_2 \cdot (2 - 3 \cdot Y_2 + Y_1)/(2 \cdot Y_1 \cdot (2 \cdot Y_2 - 1))$.

See proof in Appendix A.

As in Observation 4, the first of these cases is the intuitively plausible one—if 2 reciprocates with a high enough probability, 1 cooperates because of his material payoff as well as because of reciprocity reasons. If, however, reciprocity is important enough, there also exist other equilibria that are characterized by self-fulfilling expectations: If 1 believes that 2 initially believes that 1 chooses D, and that 2 defects in that case, 1 expects an unkind action of 2. This in turn leads 1 to be unkind, i.e. to play D (or to randomize). Of course, this only works when 1 is sufficiently motivated by reciprocity—if this is not the case, 1's material interest together with the reciprocal behavior of 2 make him cooperate.

4.2. The centipede game

The next game we analyze is the Centipede Game, first introduced by Rosenthal (1982). This is a two-player game with $M \geqslant 2$ nodes, denoted $1, 2, \ldots, M$. At each odd node k player 1 can choose between staying in the game (choice f_k), or ending it (choice d_k). If 1 stays, the material payoff of 2 increases by two, whereas 1's payoff decreases by one, and node $k+1$ is reached (as long as the final decision node M is not yet reached). If 1 chooses d_k, the material pay-off of both players does not change, and the game ends. Hence, the strategy of player 1, s_1, determines at every odd node whether player 1 stays in or ends the game. Similarly, player 2's strategy, s_2, determines at every even node whether she stays in or ends the game. At the beginning of the game player 1 is endowed with one unit and player 2 with zero units of material payoff. The case of a Centipede with $M = 4$ is illustrated in Γ_4 in Fig. 4.

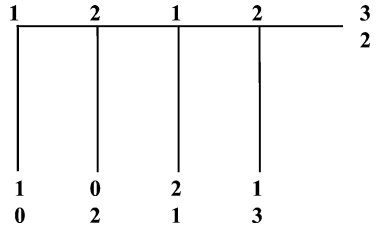

Fig. 4. Game Γ_4—a centipede game with $M = 4$.

Denote by $e(s_1, s_2)$ the first node at which one of the players ends the game, when they play according to their strategies s_1 and s_2, respectively. Then the material payoffs are given by:

$$\pi_1(s_1, s_2) = \frac{(e(s_1, s_2) + 1)}{2}, \quad \pi_2(s_1, s_2) = \frac{(e(s_1, s_2) - 1)}{2}, \quad \text{for } e(s_1, s_2) \text{ odd},$$
$$\pi_1(s_1, s_2) = \frac{e(s_1, s_2) - 1}{2}, \quad \pi_2(s_1, s_2) = \frac{e(s_1, s_2) + 1}{2}, \quad \text{for } e(s_1, s_2) \text{ even}.$$

Henceforth, we only consider the case where M is even. Qualitatively the same results as those presented below are obtained if M is odd. If both players choose to stay at all nodes, the material payoffs are

$$\pi_1(s_1, s_2) = \frac{M}{2} + 1, \qquad \pi_2(s_1, s_2) = \frac{M}{2}.$$

All Nash equilibria of this game imply that player 1 ends the game at the first node. If Rabin's model is applied to the normal form of the game, then if reciprocity is important enough there are multiple equilibria. In some of these player 1 ends the game at the first node, but there is also an equilibrium where the players stay at all nodes in the game. When a version of the Centipede game was tested experimentally, most of the subjects stayed in the game at the first nodes (see McKelvey and Palfrey, 1992).

In our model, as long as at least one of the players is sufficiently motivated by reciprocity, the unique SRE entails that both players stay in the game till the last node (where a selfish player chooses d_M). By staying a non-reciprocal player ensures that the other—reciprocal—player will regard him as kind. This finding addresses an issue raised by Rabin (1993) who asks: "Can players force emotions? That is, can a player do something which compels another player to view him favorably?" Our analysis of the centipede game shows that answer may be *yes*. This finding is consonant with that of Dufwenberg (2002), who discusses "psychological forward induction" in a psychological game where a player is motivated to avoid feeling "guilty."

In order to give a complete characterization of equilibrium behavior, we first turn to choices at the last node which is controlled by player 2.

Observation 1. In all SRE it holds at the last node M that:

(1) if $Y_2 > 2/M$, player 2 stays (chooses f_M);
(2) if $Y_2 < 2/(M + 2)$, player 2 chooses d_M;

(3) if $2/(M+2) < Y_2 < 2/M$, player 2 chooses f_M with a probability of $p = 1 + M/2 - 1/Y_2$.

See proof in Appendix A.

As to be expected, a reciprocal player will play f even at the very last node, whereas players with little reciprocity motivation will not. Note that the more stages the game possesses the less reciprocity motivation is needed to prevent the player from playing d even at the very last node. The reason is simple: The more stages the more often player 1 has already played f if the very last stage is reached. Hence 1's kindness to 2 increases with M, and therefore less reciprocity sensitivity of 2 is needed to make 2 giving up material payoff in order to do 1 a favor.

Using Observation 1 we can now analyze the conditions under which the standard solution is obtained.

Observation 2. If $Y_i < 2/(M+2)$ for $i = 1, 2$ the only SRE is given by $s_1 = (d_1, d_3, \ldots, d_{M-1})$, $s_2 = (d_2, d_2, \ldots, d_M)$.

To show the validity of Observation 2, take an arbitrary node k controlled by player i. Recall that at node M player 2 chooses d_M. We can show that the only equilibrium behavior at k is d_k, provided that both players end the game at all nodes larger k. To see this, note that the material payoff of i decreases by one unit if he chooses to stay instead of ending the game (since at $k+1$ the other player will end the game anyhow). On the other hand, the difference in kindness of player i between d_k and f_k is -2, because j's material payoff decreases by 2 if i chooses f_k instead of d_k. Hence, the difference in i's utility between choosing d_k and f_k is $1 - 2Y_i\lambda_{iji}(\cdot)$.[11] $\lambda_{iji}(\cdot)$ cannot be larger than half of the largest material payoff of the whole game, i.e. $M/4 + 1/2$. This implies that whenever $Y_i < 2/(M+2)$, i's utility of d_k is larger than i's utility of f_k—in all SRE player i chooses d_k at node k.

The next observation deals with the opposite case, where both players stay in the game at all nodes.

Observation 3. If $Y_2 > 2/M$, in the unique SRE both players stay in the game at all nodes.

See proof in Appendix A.

Observation 3 shows that both players stay in game at all nodes as long as player 2 is sufficiently motivated by reciprocity, independently of 1's reciprocity parameter. This is the "psychological forward induction" result referred to above.

The next observation deals with the question whether—similarly to Observation 3—a sufficiently reciprocal player 1 can also sustain "cooperation" independently of the type of player 2. We know from Observation 1 that if player 2 has a low sensitivity to reciprocity, she ends the game at node M (if this node is actually reached). If $M = 2$, this implies that in such a case 2's equilibrium behavior is inevitably unkind, and player 1 has never any

[11] $\lambda_{iji}(\cdot)$ denotes i's belief about j's kindness to i (cf. Section 2).

incentive to stay at the first node. If M is large enough,[12] however, the situation is different. Reaching the last but one node in that case might imply that 2's behavior is regarded as kind anyhow, i.e. independently of what 2 does at the very last node. This is because 2 has already stayed several times at the previous nodes controlled by her. This in turn might motivate a sufficiently reciprocal player 1 to stay at node $M-1$ independently of his beliefs about 2's behavior at node M. In that case, every type of player 2 has an incentive to stay at all nodes preceding node $M-1$. Therefore, player 1's sensitivity to reciprocity can be enough to prevent the game from ending before the last node. This intuition is verified by

Observation 4. If $M > 6$ and $Y_1 > 2/(M-6)$, node M is reached in all SRE.

See proof in Appendix A.

Observation 4 applies even to the extreme case where 2 is not being motivated by reciprocity at all ($Y_2 = 0$). If 2's reciprocity parameter is in a medium range such that 2 chooses to randomize at the last node, cooperation to node $M-1$ is sustained even if 1's reciprocity parameter violates the condition of Observation 4. And if player 2 stays even at the very last node, we already know from Observation 3 that player 1 stays independently of his sensitivity to reciprocity. Hence, Observations 3 and 4 show that it is enough that *one* of the players is sufficiently motivated by reciprocity for both players stay to the last or to the last but one node. Furthermore, the observations also show that the more stages there are the easier it is to establish "cooperation."

4.3. "So Long, Sucker"

We close this section by analyzing the game Γ_5 in Fig. 5, a modified version of a three-player game which Nalebuff and Shubik (1988) use to discuss certain aspects of vengefulness. It is a simple version of the game "So Long, Sucker," once invented by John Nash.

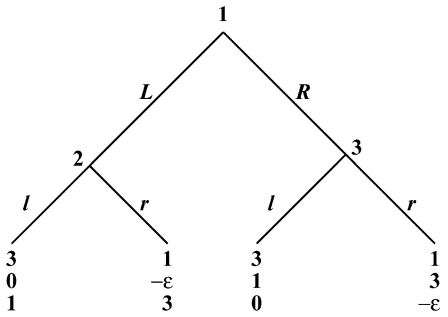

Fig. 5. Game Γ_5—"So Long, Sucker."

[12] In the present context, $M = 6$ is "high enough" (cf. Observation 4). In a more general set-up, where the changes of the material payoffs when a player stays are not one and two, the critical value of M could be different.

With $\varepsilon = 0$, Γ_5 may be thought of as a model of a strategic situation in which a \$4-pie is to be divided between three players. First player 1 has to choose which one of the other two players must get a zero payoff. Then the player who was "unfavorably" treated by 1 is called upon to decide which one of the other two will get which of the two positive monetary payoffs. Intuition may suggest that player 1 is a priori worst off of the three. Whoever he treats unfavorably will feel badly treated, and hence take revenge on 1 by awarding him the lowest possible monetary payoff. Effectively, player 1 will get a payoff of one, while players 2 and 3 look at expected payoffs of 1.5.

If each player was motivated solely by his or her own monetary income this outcome would not be guaranteed (in subgame perfect equilibrium), as players 2 and 3 would be indifferent between all their choices. In order to accommodate revenge, Nalebuff and Shubik modify the usual selfishness assumption, and assume that the players have lexicographically ordered objectives. Each player primarily maximizes his monetary rewards, but in case some choices yield *exactly the same* monetary payoff ties are broken so as to allow a player to take revenge. In Γ_5, this works to 1's disadvantage.

Our model of sequential reciprocity allows a similar conclusion, also evoking a reciprocal sensation for player 1. This is true also when 2 and 3 incur some monetary cost $\varepsilon > 0$ if they "punish" player 1. For *any* $\varepsilon \geqslant 0$, at 2's decision node 2 believes that 1 is unkind to 2 ($\lambda_{121}(\cdot) < 0$), and that 3 is neither kind nor unkind to 2 ($\lambda_{323}(\cdot) = 0$). Player 2 can get a positive reciprocity payoff only by choosing r_2, since $\kappa_{21}(r_2, \cdot) < 0 < \kappa_{21}(l_2, \cdot)$. For large enough Y_{21} player 2 will choose r_2 as her material cost is swamped by the sweetness of revenge.

Analogous remarks apply at player 3's node, so in any SRE it is true that if Y_{21} and Y_{31} are high enough, then 2 and 3 choose r_2 and r_3 respectively. Yet, there are multiple equilibria which are characterized by "self-fulfilling expectations" much like in the first example this section. Both the pure strategy profiles (L, r_2, r_3) and (R, r_2, r_3) are equilibria. The following calculations for player 1 confirm this for (L, r_2, r_3):

$$\begin{aligned}
&\kappa_{12}\big(L, (r_2, r_3)\big) = \kappa_{13}\big(R, (r_2, r_3)\big) = -1.5,\\
&\kappa_{13}\big(L, (r_2, r_3)\big) = \kappa_{12}\big(R, (r_2, r_3)\big) = 1.5,\\
&\lambda_{121}\big(r_2, (L, r_3)\big) = -1; \qquad \lambda_{131}\big(r_3, (L, r_2)\big) = 0.
\end{aligned}$$

Hence, it holds that

$$\begin{aligned}
u_1\big(L, (r_2, r_3)\big) &= 1 + Y_{12} \cdot (-1.5) \cdot (-1) + Y_{13} \cdot (1.5) \cdot 0\\
&> 1 + Y_{12} \cdot (1.5) \cdot (-1) + Y_{13} \cdot (-1.5) \cdot 0 = u_1\big(R, (r_2, r_3)\big),
\end{aligned}$$

which shows that (L, r_2, r_3) is indeed a SRE. By an analogous argument, so is (R, r_2, r_3).

5. Comparison with Rabin (1993)

Rabin develops a theory of reciprocity for normal form games with two players. If we apply the concept of SRE to any two-player (single-stage) game with simultaneous moves, we get qualitatively similar conclusions as Rabin does in most cases. This indicates that the main difference between our model and that of Rabin is the requirement of sequential

reciprocity we impose in games with an interesting dynamic structure. Yet the two models are also different in some other ways. In this section we review these differences and attempt to justify our modeling choices.

An obvious difference between Rabin's theory and ours is that we allow for more than two players (see, for example, the "So Long, Sucker" game in Section 4, or the wage-setting games analyzed in Dufwenberg and Kirchsteiger, 2000). As concerns two-player games, if we were to make the following three changes to our model, and then if we applied it to any two-player extensive game that has no proper subgames we would get *exactly* the same solutions as Rabin does in the normal form of that game:

Change 1. Substitute $(1+\kappa_{ij}(a_i,(b_{ij})_{j\neq i}))$ for $\kappa_{ij}(a_i,(b_{ij})_{j\neq i})$ in Definition 3.

Change 2. Modify Definitions 1 and 2 and redefine κ_{ij} and λ_{iji} as follows:

$$\kappa_{ij}\big(a_i,(b_{ij})_{j\neq i}\big)=\frac{\pi_j(a_i,(b_{ij})_{j\neq i})-\pi_j^{e_i}((b_{ij})_{j\neq i})}{\max\{\pi_j(a_i,(b_{ij})_{j\neq i})\mid a_i\in A_i\}-\min\{\pi_j(a_i,(b_{ij})_{j\neq i})\mid a_i\in A_i\}},$$

$$\lambda_{iji}\big(b_{ij},(c_{ijk})_{k\neq j}\big)=\frac{\pi_i(b_{ij},(c_{ijk})_{k\neq j})-\pi_i^{e_j}((c_{ijk})_{k\neq j})}{\max\{\pi_i(b_{ij},(c_{ijk})_{k\neq j}\mid b_{ij}\in B_{ij})\}-\min\{\pi_i(b_{ij},(c_{ijk})_{k\neq j}\mid b_{ij}\in B_{ij})\}}.$$

If the right-hand side denominators take the value of zero, it is furthermore assumed that $\kappa_{ij}(a_i,(b_{ij})_{j\neq i})$ and $\lambda_{ij}(b_{ij},(c_{ijk})_{j\neq i})$ take the value of zero.

Change 3. Redefine the notion of an efficient strategy (see Section 2) such that $a_i\in A_i$ is an *efficient strategy given beliefs* $(b_{ij})_{j\neq i}$ if there exists no $a_i'\in A_i$ such that for all $r\in R$, and $k\in N$ it holds that $\pi_k(a_i'(r),(b_{ij}(r))_{j\neq i})\geqslant\pi_k(a_i(r),(b_{ij}(r))_{j\neq i})$, with strict inequality for some (r,k).

Change 1 incorporates an additional motivational element which Rabin (1993, p. 1287) argues is realistic. However, for the sake of simplicity we avoid it. In principle Change 1 can be applied to our model without adverse consequences, and we will not discuss this any further here. Change 2 represents a kind of normalization of the players kindness such that the reciprocity payoff is "dimensionless," in the sense that kindness is measured in units of the material payoff *divided* by units of the material payoff. By contrast we measure kindness in the same unit as the material payoffs (for example dollars). Change 3 makes the definition of an efficient strategy dependent on players' beliefs, whereas according to our definition efficiency is a belief-independent property.[13]

Changes 2 and 3 are somewhat problematic in the context of general multi-stage games. To see this, consider the game Γ_6 in Fig. 6.

[13] There is actually another feature which creates a difference between our model and Rabin's. We use the Y_{ij} parameters to scale i's sensitivity to reciprocity concerns, whereas Rabin fixes the players' sensitivity to reciprocity and scales the importance of material payoffs (via his X). Unlike Rabin, we can look at games with $Y_{ij}=0$ where reciprocity is unimportant. However, as long as material as well as reciprocity payoffs matter, there is no real difference.

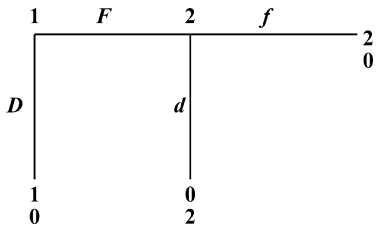

Fig. 6. Game Γ_6.

Assume first that only Change 2 is made in our theory. Suppose that in equilibrium it holds that $a_2 = b_{12} = p \cdot f + (1-p) \cdot d$ with $p < 1$. A direct calculation involving the relevant function outlined in Change 2 shows that 1's kindness is $(2 \cdot (1-p) - (1/2) \cdot (2 \cdot (1-p) + 0))/(2 \cdot (1-p) - 0) = 1/2$ when the stage where 2 moves is reached, i.e. after 1 plays F. If Y_{21} is high enough, 2 must then choose f (so that $p = 1$) which is a contradiction. Suppose instead that $a_2 = b_{12} = f$ (so that $p = 1$). Player 1's kindness of playing F is now zero, so 2 must choose d (so that $p = 0$). Again this is a contradiction. This proves that invoking Change 2 in our theory would preclude an existence theorem like that in Section 3. Note that the culprit here is the discontinuity exhibited by player 1's kindness function as $p \to 1$. In fact, for all values of $p < 1$, given Change 2, 1's kindness is *constant* $(= 1/2)$. We find this feature questionable, since the higher is p the more likely 1 believes it to be that 2 chooses f (since in equilibrium $b_{12} = p \cdot f + (1-p) \cdot d$), and the less material payoff 1 then believes that he gives to 2. By contrast, in our theory 1's kindness if he chooses F is decreasing in p (with $b_{12} = p \cdot f + (1-p) \cdot d$, 1's kindness is $1-p$).

Also Change 3 could lead to existence problems. To see this, consider again Γ_6, and assume that only Change 3 is made in our theory. Suppose that in equilibrium it holds that $a_2 = b_{12} = pf + (1-p)d$ with $p \geqslant 1/2$. Given Change 3, *only* F is an efficient strategy for 1.[14] Hence 1 is not kind when choosing F. But then 2 chooses d, which is a contradiction. Suppose instead that $a_2 = b_{12} = pf + (1-p)d$ with $p < 1/2$. Then *all* 1's strategies are efficient. Player 1 is now kind choosing F, and since $p < 1/2$ his kindness is bounded away from zero. If Y_{21} is high enough 2 must choose f, which again is a contradiction. This proves that invoking Change 3 leads to non-existence of equilibria in some games. In our theory this problem does not arise because efficiency of a strategy is a belief-independent property. According to our definition, there are no inefficient strategies in Γ_6 regardless of b_{12}.

Our efficiency definition may also shed some light on the emergence of "trust." Rabin (1993, Fig. 3) considers a "partnership game" here reproduced as G_7 in Fig. 7.

Rabin suggests that (*Trust, Share*) may be a reasonable outcome, but notes that the outcome is not an equilibrium in his model. To see why, note that with Change 3 and

[14] Rabin (1993) does not distinguish between the weak and the strong notion of Pareto efficiency. Our argument here presumes a definition corresponding to Change 3, so that a strategy is efficient if no other strategy is not worse for all players, and strictly better for some player. Alternatively a strategy may be defined as efficient if no other strategy is *strictly* better for *all* players. It is easy to verify that also then Γ_6 can be used to illustrate non-existence of the equilibrium.

	Share	*Grab*
Trust	**6, 6**	**0, 12**
Dissolve	**5, 5**	**5, 5**

Fig. 7. Game G_7—a partnership game.

beliefs corresponding to the profile (*Trust, Share*) 1's strategy *Dissolve* is inefficient, and 1 is not kind if he chooses *Trust*. Hence, even if 2 is motivated by reciprocity she would choose *Grab*. By contrast, in our theory (applied to any extensive game corresponding to G_7) no strategy is inefficient. 1 is unambiguously kind if he chooses *Trust*, and (*Trust, Share*) may be a SRE.[15]

Rabin (1993) derives many general implications of his model, and we close this section by pointing out how our theory escapes one of the more gloomy ones. Rabin's Proposition 6 assures the existence of an equilibrium where no player is kind. In many games there are also happy equilibria where (both) players are kind, but in these cases there must exist multiple equilibria and in at least one of these no player is kind. Our analysis of the Centipede game in Section 4 shows that this result has no general counterpart in our theory. If the players in Γ_4 are sufficiently motivated by reciprocity, then the SRE is unique and both players are kind.

It is, however, not true that our theory guarantees for all games the existence of a happy outcome where all players are kind. This follows from the analysis of the "So Long, Sucker" game Γ_5 in Section 4, where along the equilibrium path of any SRE player 1 and player 2 or player 3 view each other as unkind.

6. Concluding remarks

As we have seen our approach is able to capture the intuitive meaning of reciprocity as well as the stylized facts of experimental games that show the importance of reciprocal behavior. We propose that our model may serve as a useful tool for analyzing the implications of reciprocal behavior in several important economic problems. In Dufwenberg and Kirchsteiger (2000) we apply our model to investigate employer–employee relationships. We show that reciprocity can explain why employers are reluctant to hire workers who offer to work at less than the prevailing wage, a phenomenon

[15] At this point we would like to point out (and respond to) a critique which has been raised against our efficiency definition. Consider a modification of the game Γ_3 in Fig. 3, where the only change is that after 1's choice W player 2 is called upon to move. 2 must then choose between l and m, which leads to end nodes with the respective material payoffs $(-1000, -1000)$ and $(2, -2000)$. In Section 2 we argued that in Γ_3 the inclusion of the strategy W should not influence the kindness assessment of the other strategies. However, in the modified game W is efficient according to our definition. It may seem questionable that the addition of the masochistic response m for 2 changes the perception of the kindness of 1's strategies. Nevertheless, we find that 1's strategy W is not completely unreasonable in the modified game, since it is consistent with 1 attempting to get his very best material payoff. Arguably 1's strategy D is thus kind, since by choosing D player 1 foregoes any chance to get his highest material payoff. Therefore, although these are debatable matters, we feel that our efficiency definition is intuitive and works well.

frequently observed in labor markets (see Bewley, 1999). Reciprocity may moreover play an important role in contract theory, in situations that are potentially plagued by moral hazard. Our approach can then be used to analyze how reciprocity changes the set of incentive-compatible contracts, the nature of an optimal contract, and to what extent incentive problems can be overcome in such a situation.

It may also be of interest to confront our model with experimental data. In this connection we wish to point to a few considerations that may be important. First, the crucial role played by beliefs in our theory (and in Rabin's) suggests that it may be necessary to explicitly measure beliefs in the laboratory.[16] Second, some tests of our theory might restrict attention to Definitions 1–3 without connecting to the concept of SRE in Definition 4. Suppose that in an experiment first and second order beliefs are measured alongside the strategic choices, so that the experimenter can measure the subjects' kindness and perceived kindness. It is now possible to test whether *an individual subject* is reciprocally motivated and updates his or her beliefs in the way that Definitions 1–3 suggest. To also test whether *all the interacting subjects* have coordinated on a particular SRE may correspond to a further independent hypothesis. Third, it should be noted that we have chosen to work with utilities as given by Definition 3 because this is the simplest formulation we can think of that invokes a concern for reciprocity. This simplicity comes at a cost which may have bearing on experimental testing. The utility U_i, specified in Definition 3, will not represent i's preferences in a way which is invariant with respect to the choice of monetary units. To see this, note that if i's monetary payoff is measured in dollars, then the reciprocity payoff will have the dimension of dollars squared. At the cost of considerable analytic complexity, this problem can be solved by defining player i's reciprocity payoff with respect to each player j as Y_{ij} times the square root of the absolute value of $\kappa_{ij}(\cdot) \cdot \lambda_{iji}(\cdot)$, adjusted so as to maintain the right sign. If one wanted to estimate a parameter like Y_{ij} based on experimental data for different games, it might be sensible to adopt such an approach.[17]

Experimentalists should take into account, however, the limits of the scope of this paper. We deliberately focus on modeling a concern for reciprocity, and disregard other motivations like altruism, equity, envy, let-down or guilt aversion, or concern for the least well-off individual. As noted in the introduction, it is clear that this omission is not innocuous. For example, in experimental Dictator games individuals often give away lots of money (see Davis and Holt, 1993, pp. 263–269, for a discussion), something which cannot be explained by the model we propose in this paper. In reality people seem to be motivated in many different ways, and perhaps this all depends not only on the strategic nature of a situation but also on other aspects of the context where the situation occurs. For example, in the case of Dictator games the evidence reported by Hoffman et al. (1996) suggests that "social distance" is important in that context, and Kirchsteiger et al. (2001)

[16] Dufwenberg and Gneezy (2000) point to ways in which beliefs can be elicited, in an experiment related to psychological game theory. They measure beliefs of first and second order by making subjects guess one another's strategy choices and guesses, and offering monetary rewards for accuracy in the guesswork.

[17] We should note, however, that Nelson (2001) argues (with reference to empirical observations) that for moderate stakes reciprocity considerations lose importance as material payoffs increase, but that for very large stakes reciprocity considerations recover their importance. Our formulation has this property.

show that the nature of motivations may depend on "mood." We leave for future research the delicate task of determining in what context one or another motivational concern is of particular importance. It seems clear that when this issue is tackled, experimental and theoretical work should go hand in hand.

Acknowledgments

This paper was conceived while we were both at the CentER for Economic Research at Tilburg University. We appreciate helpful comments from Geir Asheim, Gary Charness, Doug DeJong, Uri Gneezy, Manfred Nermuth, Muriel Niederle, Matthew Rabin, Arno Riedl, Aldo Rustichini, three referees, the associate editor, and the participants of several seminars. Dufwenberg's research has been supported by a grant from Tom Hedelius and Jan Wallander's research foundation.

Appendix A

A.1. Proof of the Theorem

Let $X_i(h)$ be i's set of (possibly randomized) choices at history $h \in H$. If $x \in X_i(h)$, let $a_i(h)\backslash x$ be the strategy of i's which specifies the choice x at h, but which is otherwise just like $a_i(h)$. Define correspondences $\beta_{i,h} : A \to X_i(h)$ and $\beta : A \to \Pi_{(i,h)\in N\times H} X_i(h)$ by

$$\beta_{i,h}(a) = \operatorname*{argmax}_{x\in X_i(h)} U_i\big(a_i(h)\backslash x, \big(a_j(h), \big(a_k(h)\big)_{k\neq j}\big)_{j\neq i}\big),$$

$$\beta(a) = \prod_{(i,h)\in N\times H} \beta_{i,h}(a).$$

The sets $\prod_{(i,h)\in N\times H} \beta_{i,h}(a)$ and A are topologically equivalent, so β is equivalent to a correspondence $\gamma : A \to A$ which is defined in the obvious way. Fixed points under γ are SREs. To see this, note that $\beta_{i,h}$ caters to condition (1) of Definition 4, plugging in the correct beliefs as mandated by the conditions (2) and (3). Thus $\beta_{i,h}$ effectively finds the optimal *strategies* in $A_i(h, a)$, in conformance with (1), although this is here done using the optimal *choices* in $X_i(h)$. β and γ are combined best-response correspondences, and since γ is a correspondence from A to A it is amenable to fixed point analysis.

It remains to show that γ possesses a fixed point. Berge's maximum principle guarantees that $\beta_{i,h}$ is non-empty, closed-valued, and upper hemi-continuous, because $X_i(h)$ is non-empty and compact and U_i is continuous (since π_i, κ_{ij}, and λ_{iji} are all continuous). $\beta_{i,h}$ is furthermore convex-valued, since $X_i(h)$ is convex and U_i is quasi-concave (in fact linear) in i's own choice. Hence, $\beta_{i,h}$ is non-empty, closed-valued, upper hemi-continuous, and convex-valued. These properties extend to β and γ. It follows by Kakutani's fixed point theorem that γ admits a fixpoint.

A.2. Applications

A.2.1. The sequential prisoner's dilemma

Proof of Observation 2. Note first that if 1 cooperates, 2 can give 1 a material payoff of at least -1 and at most 1, so the "equitable" payoff is 0 (= the average of -1 and 1). If 2 chooses cooperation, 1 receives 1. Therefore, 2's kindness of cooperation is 1. Similarly, 2's kindness of defection is -1. In order to calculate how kind 2 believes 1 is after choosing C, we have to specify 2's belief of 1's belief about 2's choice after C.[18] Denote this by $p'' \in [0,1]$. Then 2's belief about how much payoff 1 intends to give to 2 by choosing C is $p'' \cdot 1 + (1 - p'') \cdot 2$, and since 2's payoff resulting from 1's choice of D would be zero,[19] 2's belief about 1's kindness from choosing C is $[p'' \cdot 1 + (1 - p'') \cdot 2] - [0.5(p'' \cdot 1 + (1 - p'') \cdot 2 + 0)] = 1 - 0.5 \cdot p''$, with the first term in squared brackets denoting 2's actual payoff and the second squared bracket denoting 2's "equitable payoff." This implies that when 1 cooperates and the second order belief is p'', 2's utility of cooperation is given by $1 + Y_2 \cdot 1 \cdot (1 - 0.5p'')$, whereas 2's utility of defection is $2 + Y_2(-1)(1 - 0.5 \cdot p'')$. The former is larger than the latter if $Y_2 \cdot (2 - p'') > 1$. In equilibrium, the second order belief must be correct. Hence, if in equilibrium 2 cooperates, the condition must hold for $p'' = 1$. This is the case if $Y_2 > 1$. On the other hand, if in equilibrium 2 defects, the condition must not hold for $p'' = 0$. This implies that $Y_2 < 0.5$. For intermediate values of Y_2 $(0.5 < Y_2 < 1)$ neither cooperation nor defection can be part of an equilibrium. In order to have an equilibrium that involves randomized choice, the utility of cooperation must be equal to the utility of defection. This is the case when $Y_2 \cdot (2 - p'') = 1$. Since in equilibrium the second order belief must be correct, the actual probability of cooperation, p, must be such that the condition is fulfilled. This implies that $p = (2 \cdot Y_2 - 1)/Y_2$. □

Proof of Observation 4. Note first that $Y_2 > 1$ implies that 2 cooperates when 1 cooperates and defects when 1 defects (see Observations 1 and 2). Hence, 1 can give 2 a material payoff of at least 0 and at most 1. Hence, the "equitable" payoff of 1 is 0.5. If 1 chooses cooperation, 2 receives 1. Therefore, 1's kindness of cooperation is 0.5. Similarly, 1's kindness of defection is -0.5. In order to calculate how kind 1 believes that 2 is we have to specify 1's belief about what 2 believes that 1 will do. Denote by $q'' \in [0,1]$ this second order belief of 1 choosing C. Then 1 believes that 2 believes that she gives player 1 a material payoff of $q'' \cdot 1 + (1 - q'') \cdot 0$ by choosing her equilibrium strategy. If 2 always cooperates, 1's payoff is $q'' \cdot 1 + (1 - q'') \cdot 2$, whereas if 2 always defects, 1's payoff is $q'' \cdot (-1) + (1 - q'') \cdot 0$. Hence, 1's belief about 2's kindness from choosing c after C and d after D is given by

[18] In principle we also need 2's belief about 1's behavior. However, we only care about beliefs that are in accordance with reaching the node under consideration. After 1 has already chosen C, there is only one such belief, namely 1 choosing C. To put it differently: 2 already knows what 1 has done, and 2's belief has to be in accordance with her knowledge.

[19] Recall that in any SRE player 2 defects after a defection of 1 (see Observation 1).

$$\begin{aligned}&\left[q'' \cdot 1 + (1-q'') \cdot 0\right] - \left[0.5 \cdot q'' \cdot 1 + (1-q'') \cdot 2 + q'' \cdot (-1) + (1-q'') \cdot 0\right] \\ &\quad = 2 \cdot q'' - 1,\end{aligned}$$

with the first term in squared brackets denoting 1's actual payoff and the second term in squared bracket denoting 1's "equitable payoff." This implies that when 2 plays the equilibrium strategy and the second order belief is q'', 1's utility of cooperation is given by $1 + Y_1 \cdot 0.5 \cdot (2 \cdot q'' - 1)$, whereas 1's utility of defection is $0 + Y_1 \cdot (-0.5)(2 \cdot q'' - 1)$. The former is larger than the latter if $1 + Y_1 \cdot (2 \cdot q'' - 1) > 0$. In equilibrium, the second order belief must be correct. Hence, if in equilibrium 1 cooperates, the condition must hold for $q'' = 1$, which is always the case.

On the other hand, if in equilibrium 1 defects, the condition must not hold for $q'' = 0$. This implies that $Y_1 > 1$.

In order to have a equilibrium involving randomized choices, the utility of cooperation must be equal to the utility of defection. This is the case when $1 + Y_1 \cdot (2 \cdot q'' - 1) = 0$. Since in equilibrium the second order belief must be correct, the actual probability of cooperation, q, must be such that the condition is fulfilled. This implies that $q = (Y_1 - 1)/2 \cdot Y_1$. □

Proof of Observation 5. Notice that $0.5 < Y_2 < 1$ implies that 2 cooperates with probability $p = (2 \cdot Y_2 - 1)/Y_2$ when 1 cooperates, and 2 defects when 1 defects (see Observations 1 and 2). Hence, 1 can give 2 a material payoff of at least 0 and at most $1 \cdot p \cdot + 2 \cdot (1-p)$. Hence, the "equitable" payoff of 1 is $0.5 \cdot (1 \cdot p + 2 \cdot (1-p)) = (2-p)/2$. If 1 chooses cooperation, 2 receives $1 \cdot p \cdot + 2 \cdot (1-p)$. Therefore, 1's kindness of cooperation is $(2-p)/2$. Similarly, 1's kindness of defection is $-(2-p)/2$. In order to calculate how kind 1 believes that 2 is we have to specify 1's belief about what 2 believes that 1 will do. Denote by $q'' \in [0, 1]$ this second order belief of 1 choosing C. Then 1 believes that 2 believes that she gives player 1 a material payoff of $q'' \cdot (p \cdot 1 + (1-p) \cdot (-1)) + (1-q'') \cdot 0$ by her equilibrium strategy. If 2 always cooperates, 1's payoff is $q'' \cdot 1 + (1-q'') \cdot 2$, whereas if 2 always defects, 1's payoff is $q'' \cdot (-1) + (1-q'') \cdot 0$. Hence, 1's belief about 2's kindness of her equilibrium strategy is

$$\begin{aligned}&q'' \cdot \big(p \cdot 1 + (1-p) \cdot (-1)\big) + (1-q'') \cdot 0 \\ &\quad - 0.5 \cdot \left[q'' \cdot 1 + (1-q'') \cdot 2 + q'' \cdot (-1) + (1-q'') \cdot 0\right] = 2 \cdot q'' \cdot p - 1.\end{aligned}$$

This implies that when 2 plays the equilibrium strategy and the second order belief is q'', 1's utility of cooperation is given by $p \cdot 1 + (1-p) \cdot (-1) + Y_1 \cdot (2-p) \cdot (2 \cdot q'' \cdot p - 1)/2$, whereas 1's utility of defection is $0 + Y_1 \cdot (2-p) \cdot (2 \cdot q'' \cdot p - 1)/2$. The former is larger than the latter if $2 \cdot p - 1 + Y_1 \cdot (2-p) \cdot (2 \cdot q'' \cdot p - 1) > 0$. In equilibrium, the second order belief must be correct. Hence, if in equilibrium 1 cooperates, the condition must hold for $q'' = 1$, which happens if $p > 0.5$. This in turn implies that $Y_2 > 2/3$ (using $p = (2 \cdot Y_2 - 1)/Y_2$ from Observation 2(3)).

On the other hand, if in equilibrium 1 defects, the condition must not hold for $q'' = 0$. Inserting for p and rearranging terms this leads to $Y_1 > 3 \cdot Y_2 - 2$.

In order to have a equilibrium involving randomization, the utility of cooperation must be equal to the utility of defection. This is the case when $2 \cdot p - 1 + Y_1 \cdot (2-p) \cdot (2 \cdot q'' \cdot p - 1) = 0$. Since in equilibrium the second order belief must be correct, the actual

probability of cooperation, q, must be such that the condition is fulfilled. Substituting for p this implies that $q = Y_2 \cdot (2 - 3 \cdot Y_2 + Y_1)/(2 \cdot Y_1 \cdot (2 \cdot Y_2 - 1))$. The other conditions of Observation 5(3) are necessary to guarantee that q is larger than zero and smaller than 1. □

A.2.2. The centipede game

Proof of Observation 1. Notice first that the only consistent belief of 2 about 1's strategy is that 1 stays in the game at all nodes he controls. Given this belief about 1's strategy, 2's strategy could give 1 a material payoff of at least 0 (by ending the game at node 2, i.e. the first node she controls) and at most $M/2+1$ (by staying at all nodes she controls including node M). Hence, the "equitable" payoff of 1 is $M/4 + 1/2$. If 2 chooses f_M at node M, 1 receives $M/2 + 1$. Therefore, 2's kindness of staying in the game at all nodes including M is $M/4 + 1/2$. If 2 chooses d_M at node M, 1 receives $M/2 - 1$. Hence, 2's kindness of strategy $(f_2, f_4, \ldots, f_{M-2}, d_M)$ is $M/2 - 1 - (M/4 + 1/2) = M/4 - 3/2$. In order to calculate how kind 2 believes 1 is when 1 stays in the game at all nodes he controls, we have to specify 2's belief of 1's belief about 2's choice at the node M. Consistency of beliefs about 2's strategy implies that 2 is believed to stay in the game at all nodes that precede M. Denote by p'' the second order belief that 2 chooses f_M, i.e. 2's belief of 1's belief of the probability that 2 stays in the game at node M. Then 2's belief about how much payoff 1 believes he gives to 2 by always staying in the game is $p''M/2 + (1 - p'')(M/2 + 1) = M/2 + (1 - p'')$. Clearly, this is the maximum 1 can believe he gives to 2 (with consistent beliefs about 2's strategy). On the other hand, 1 could have given 2 a material payoff of zero (by ending the game at node 1). Hence, 2's belief about 1's kindness from always staying in the game is $M/2 + (1 - p'') - 0.5(M/2 + (1 - p'') + 0) = M/4 + 1/2 - p''/2$. This implies that when the second order belief is p'', 2's utility of always staying in the game is given by $M/2 + Y_2(M/4 + 1/2 - p''/2)(M/4 + 1/2)$, whereas 2's utility of strategy $(f_2, f_4, \ldots, f_{M-2}, d_M)$ is $M/2 + 1 + Y_2(M/4 + 1/2 - p''/2)(M/4 - 3/2)$. The former is larger than the latter if $2Y_2(M/4 + 1/2 - p''/2) > 1$. In equilibrium, the second order belief must be correct. Hence, if in equilibrium 2 chooses f_M, the condition must hold for $p'' = 1$. This is the case if $Y_2 > 2/M$. On the other hand, if in equilibrium 2 chooses d_M, the condition must not hold for $p'' = 0$. This implies that $Y_2 < 2/(M+2)$. For intermediate values of $Y_2(2/(M + 2) < Y_2 < 2/M)$ neither of the pure choices f_M and d_M can be part of an equilibrium. In order to have a equilibrium involving randomization, the utility of f_M must be equal to the utility of d_M. This is the case when $2Y_2(M/4 + 1/2 - p''/2) = 1$. Since in equilibrium the second order belief must be correct, the actual probability of f_M, p, must be such that the condition is fulfilled. This implies that $p = 1 + M/2 - 1/Y_2$. □

Proof of Observation 3. We will first look at equilibrium behavior at node $M - 1$. Recall from Observation 1 that in all SRE player 2 chooses f_M at node M if $Y_2 > 2/M$. Hence, in all equilibria the only consistent belief of player 1 about 2's strategy is that 2 always stays, whenever node $M - 1$ is reached. Therefore, 1's material payoff is strictly larger if he stays at $M - 1$ than if he ends the game at $M - 1$. Furthermore, by staying player 1 gives player 2 the maximal material payoff 2 can get, given 2's strategy. Hence, staying is the kindest choice player 1 can make at node $M - 1$. On the other hand, 2's strategy of

always staying is clearly kind. Hence, as long as $Y_1 > 0$ the psychological part of 1's utility is maximized if he chooses f_{M-1}. If $Y_1 = 0$, the psychological part of the 1's utility is zero anyhow. Hence, for any Y_1 player 1's overall utility is maximized by f_{M-1}—in all SRE 1 chooses f_{M-1} at node $M-1$. Node $M-2$ is controlled by player 2. In all equilibria the only consistent belief of 2 about 1's strategy is that 1 always stays. By applying the same arguments as before we can show that in this case 2's utility is maximized by the choice of f_{M-2}—f_{M-2} is the only equilibrium choice for any Y_2. It is easy to see that by applying this line of reasoning backwards from node $M-3$ to node 1 staying is the only equilibrium choice at all nodes. □

Proof of Observation 4. We will first show that at node $M-1$ player 1, who controls this node, will choose f_{M-1} in all SRE as long as $Y_1 > 2/(M-6)$. We will then show that if player 1 chooses f_{M-1}, both players will stay at all previous nodes in all SRE. When $M-1$ is actually reached, all consistent beliefs of 1 about 2's strategy are such that 2 stays in the game at all nodes except node M. Denote by p the probability that 2 chooses f_M at node M (in equilibrium, p depends of course on Y_2—cf. Observation 1). Given the consistent beliefs about 2's strategy, 1's strategy could give 1 a material payoff of at least 0 (by ending the game at node 1, the first node he controls) and at most $M/2 + (1-p)$ (by staying at all nodes he controls, including node $M-1$). Hence, the "equitable" payoff of 2 is $M/4 + (1-p)/2$. If player 1 chooses f_{M-1} at node $M-1$, player 2 receives $M/2 + (1-p)$. Therefore, 1's kindness of staying in the game at all nodes including $M-1$ is $M/4 + (1-p)/2$. If player 1 chooses d_{M-1} at node $M-1$, 2 receives $M/2-1$. Hence, 1's kindness of strategy $(f_1, f_3, \ldots, f_{M-3}, d_{M-1})$ is $M/2 - 1 - (M/4 + (1-p)/2) = M/4 - 3/2 + p/2$. In order to calculate how kind 1 believes 2 is when 2 stays in the game at all nodes but M, note first that consistency of beliefs about 1's strategy implies that 1 is believed to stay at all nodes he controls. Then 1's belief about how much payoff player 2 believes she gives to 1 by staying in the game at all nodes but M is $p(M/2+1)+(1-p)(M/2-1) = M/2-1+2p$. Clearly, for the consistent beliefs about 1's strategy the maximum 2 can believe she gives to 1 is $M/2+1$ (by staying at all nodes including M). On the other hand, 2 could have given 1 a material payoff of zero (by ending the game at node 2). Hence, 1's belief about 2's kindness from choosing to stay in the game at all nodes but M is $M/2-1+2p-0.5(M/2+1) = M/4-3/2+2p$. This implies that 1's utility of staying at all nodes is given by $p(M/2+1) + (1-p) \times (M/2-1) + Y_1(M/4-3/2+2p)(M/4+(1-p)/2)$, whereas 1's utility of strategy $(f_1, f_3, \ldots, f_{M-3}, d_{M-1})$ is $M/2 + Y_1(M/4-3/2+2p)(M/4-3/2+p/2)$. The former is larger than the latter if $2p + Y_1(M/4-3/2+2p)(2-p) - 1 > 0$. This inequality holds for any p as long as $Y_1 > 2/(M-6)$—at the node $M-1$ player 1 chooses f_{M-1} in all SRE. Given that, the arguments of the proof of Observation 3 can be used to show that at all previous nodes the only equilibrium choice of both players is to stay. □

References

Akerlof, G., 1982. Labour contracts as a partial gift exchange. Quart. J. Econ. 97, 543–569.
Akerlof, G., Yellen, J., 1988. Fairness and unemployment. Amer. Econ. Rev. 78, 44–49.

Akerlof, G., Yellen, J., 1990. The fair-wage effort hypothesis and unemployment. Quart. J. Econ. 195, 255–284.
Andreoni, J., 1990. Impure altruism and donations to public goods: A theory of warm glow giving. Econ. J. 100, 464–477.
Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity and social history. Games Econ. Behav. 10, 122–142.
Bewley, T., 1999. Why Wages Don't Fall During a Recession. Harvard Univ. Press.
Blount, S., 1995. When social outcomes aren't fair: The effect of causal attributions on preferences. Org. Behav. Human Decis. Proc. 63, 131–144.
Bolle, F., Kritikos, A., 1999. Approaching fair behavior: self-centered inequality aversion versus reciprocity and altruism. Diskussionspapier Nr. 143. Europa-Universität Viadrina, Frankfurt/Oder.
Bolton, G., Brandts, J., Katok, E., 1996. A simple test of explanations for contributions in dilemma games. Mimeo.
Bolton, G., Ockenfels, A., 2000. ERC—A theory of equity, reciprocity and competition. Amer. Econ. Rev. 90, 166–193.
Charness, G., 1996. Attribution and reciprocity in a simulated labor market: an experimental investigation. Mimeo.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Quart. J. Econ. 117, 817–869.
Clark, K., Sefton, M., 2001. The sequential prisoner's dilemma: evidence on reciprocal altruism. Econ. J. 111, 51–68.
Cox, J., Friedman, D., 2002. A tractable model of reciprocity and fairness. Mimeo.
Davis, D., Holt, C., 1993. Experimental Economics. Princeton Univ. Press.
Dufwenberg, M., 2002. Marital investment, time consistency, and emotions. J. Econ. Behav. Organ. 48, 57–69.
Dufwenberg, M., Gneezy, U., 2000. Measuring beliefs in an experimental lost wallet game. Games Econ. Behav. 30, 163–182.
Dufwenberg, M., Kirchsteiger, G., 2000. Reciprocity and wage undercutting. Europ. Econ. Rev. 44, 1069–1078.
Falk, A., Fischbacher, U., 1998. A theory of reciprocity. Working paper No. 6. University of Zuerich.
Falk, A., Gächter, S., 2002. Reputation and reciprocity—consequences for the labour relation. Scand. J. Econ. 104, 1–26.
Fehr, E., Falk, A., 1999. Wage rigidities in a competitive, incomplete contract market. J. Polit. Econ. 107, 106–134.
Fehr, E., Gächter, S., 2000. Fairness and retaliation: The economics of reciprocity. J. Econ. Perspect. 14, 159–181.
Fehr, E., Gächter, S., Kirchsteiger, G., 1996. Reciprocal fairness and noncompensating wage differentials. J. Inst. Theoretical Econ. 152, 608–640.
Fehr, E., Gächter, S., Kirchsteiger, G., 1997. Reciprocity as a contract enforcement device: experimental evidence. Econometrica 65, 833–860.
Fehr, E., Kirchsteiger, G., 1994. Insider power, wage discrimination and fairness. Econ. J. 104, 571–583.
Fehr, E., Kirchsteiger, G., Riedl, A., 1993. Does fairness prevent market clearing? An experimental investigation. Quart. J. Econ. 108, 437–460.
Fehr, E., Kirchsteiger, G., Riedl, A., 1998. Gift exchange and reciprocity in competitive experimental markets. Europ. Econ. Rev. 42, 1–34.
Fehr, E., Schmidt, K., 1999. A theory of fairness, competition, and cooperation. Quart. J. Econ. 114, 817–868.
Fudenberg, D., Tirole, J., 1991. Game Theory. MIT Press.
Geanakoplos, J., Pearce, D., Stacchetti, E., 1989. Psychological games and sequential rationality. Games Econ. Behav. 1, 60–79.
Gneezy, U., Güth, W., Verboven, F., 2000. Presents or investments? An experimental analysis. J. Econ. Psych. 21, 481–493.
Goranson, R., Berkowitz, L., 1966. Reciprocity and responsibility reactions to prior help. J. Personality Soc. Psych. 3, 227–232.
Greenberg, M., Frisch, D., 1972. Effect of intentionality on willingness to reciprocate a favor. J. Exper. Soc. Psych. 8, 99–111.
Harris, C., Laibson, D., 2001. Dynamic choices of hyperbolic consumers. Econometrica 69, 935–957.
Hendon, E., Jacobsen, H.J., Sloth, B., 1996. The one-shot-deviation principle for sequential rationality. Games Econ. Behav. 12, 152–169.
Hoffman, E., McCabe, K., Smith, V., 1996. Social distance and other—regarding behavior in dictator games. Amer. Econ. Rev. 86, 653–660.

Kahneman, D., Knetsch, J., Thaler, R., 1986. Fairness as a constraint on profit seeking: entitlements in the market. Amer. Econ. Rev. 76, 728–741.
Kirchsteiger, G., 1994. The role of envy in ultimatum games. J. Econ. Behav. Organ. 25, 373–390.
Kirchsteiger, G., Rigotti, L., Rustichini, A., 2001. Your morals are your moods. UC Berkeley Economics Department WP E01-294.
Komter, A. (Ed.), 1996. The Gift: An Interdiciplinary Approach. Amsterdam Univ. Press.
Levine, D., 1998. Modeling altruism and spitefulness in experiments. Rev. Econ. Dynam. 1, 593–622.
McKelvey, R., Palfrey, T., 1992. An experimental study of the centipede game. Econometrica 60, 803–836.
Nalebuff, B., Shubik, M., 1988. Revenge and rational play. Discussion paper No. 138. Woodrow Wilson School.
Nash, J., 1950. Non-cooperative games. Unpublished PhD thesis. Princeton University.
Nelson, R., 2001. Incorporating fairness into game theory and economics: comment. Amer. Econ. Rev. 91, 1180–1183.
Rabin, M., 1993. Incooperating fairness into game theory and economics. Amer. Econ. Rev. 83, 1281–1302.
Rabin, M., 1998. Psychology and economics. J. Econ. Lit. 36, 11–46.
Reny, P., 1992. Backward induction, normal form perfection and explicable equilibria. Econometrica 60, 627–649.
Rosenthal, R., 1982. Games of perfect information, predatory pricing, and the chain store paradox. J. Econ. Theory 25, 92–100.
Roth, A., 1995. Bargaining experiments. In: Kagel, J., Roth, A. (Eds.), Handbook of Experimental Economics. Princeton Univ. Press, pp. 254–348.
Segal, U., Sobel, J., 1999. Tit for tat: Foundations of preferences for reciprocity in strategic settings. Discussion paper 99-10. UC at San Diego.
Sen, A., 1979. Utilitarianism and welfarism. J. Philos. 76, 463–489.
Sobel, J., 2000. Social preferences and reciprocity. Mimeo.
Strotz, R., 1956. Myopia and inconsistency in dynamic utility maximization. Rev. Econ. Stud. 23, 165–180.
Tesser, A., Gatewood, R., Driver, M., 1968. Some determinants of gratitude. J. Personality Soc. Psych. 9, 233–236.
Weibull, J., 1994. The Mass-action interpretation of Nash equilibrium. Working paper 427. The Industrial Institute for Economic and Social Research, Stockholm.