A Characterization of Intrinsic Reciprocity*

Uzi Segal† and Joel Sobel‡

June 5, 2007

Abstract

This paper studies a game-theoretic model in which players have preferences over their strategies. These preferences vary with the strategic context. The paper further assumes that each player has an ordering over an opponent's strategies that describes the niceness of these strategies. It introduces a condition that insures that the weight on an opponent's utility increases if and only if the opponent chooses a nicer strategy.

Keywords: reciprocity, extended preferences, game theory

*Segal thanks SSHRCC and Sobel thanks the Guggenheim Foundation, NSF, and the Secretaría de Estado de Universidades e Investigación del Ministerio de Educación y Ciencia (Spain) for financial support. Sobel is grateful to the Departament d'Economia i d'Història Econòmica and Institut d'Anàlisi Econòmica of the Universitat Autònoma de Barcelona for hospitality and administrative support.

†Department of Economics, Boston College, Chestnut Hill, MA 02467, U.S.A. E-mail: uzi.segal@bc.edu

‡Department of Economics, University of California, San Diego, La Jolla, CA 92093, U.S.A. and Institut d'Anàlisi Econòmica - CSIC Universitat Autònoma de Barcelona, SPAIN. E-mail: jsobel@ucsd.edu

1 Introduction

Reciprocity is an important aspect of social interaction. People act kindly to people who have helped them and unkindly to people who have harmed them. Many times when people reciprocate, they are motivated solely by their narrow self interest. This kind of *instrumental* reciprocity arises naturally in dynamic settings. In repeated games, an equilibrium strategy may specify that "bad" actions are punished and "good" actions are rewarded. Familiar arguments demonstrate how this behavior can support repeated-game equilibria that provide efficient average payoffs. There is also evidence that some reciprocity is *intrinsic*. People have an intrinsic preference for reciprocity if they are willing to sacrifice their own material payoff in order to increase the payoff of a kind opponent or decrease the payoff of an unkind opponent.[1] This paper adds to a small theoretical literature on intrinsic reciprocity by identifying a property that connects the weight one player's preferences place on the utility of another player to the "niceness" of the other player's strategy.

Geanakoplos, Pearce, and Stacchetti [4], Rabin [5], and Segal and Sobel [6] present ways to model intrinsic reciprocity in games. These papers extend standard game theory in a way that permits preferences of one player to depend on the intentions of other players. Each strategy profile in a game determines an outcome. Assume that player i has (decision-theoretic) preferences over outcomes that can be represented by a von Neumann-Morgenstern utility function u, and that preferences over strategy profiles s are represented by the expected value of the utility u obtained from the possible outcomes of s. These material utility functions are the payoffs of standard game theory. A tractable way to model intrinsic reciprocity is to assume, in addition, that players have strategic preferences over their own strategy set and that these preferences depend on the strategic context (or expected play of the game) so that, in particular, they need not agree with preferences over outcomes. Segal and Sobel [6] investigate this model and provide conditions under which player i's preferences over his set of mixed strategies (with representative element σ_i) conditional on a context on expected pattern of play (described by the mixed-strategy profile σ^*) is represented by a function $V_i(\cdot)$ that can

[1] Sobel [7] reviews the literature on reciprocity.

be written as:

$$V_{i,\sigma^*}(\sigma_i) = u_i(\sigma_i, \sigma^*_{-i}) + \sum_{j \neq i} a^j_{i,\sigma^*} u_j(\sigma_i, \sigma^*_{-i}) \tag{1}$$

When representation (1) holds, we say that player i has *preferences for reciprocity*. Equation (1) states that instead of evaluating strategies using the utility function $u_i(\cdot)$, player i's utility in a game is a weighted average of the decision theoretic utility of all of the players in the game. The weights a^j_{i,σ^*} depend on the anticipated play of the game (σ^*). One implication of the representation is that player i can positively weight the utility of his opponents in some contexts, and negatively weight it in others. That is, depending on the strategic context, an individual can exhibit altruism (utility is increasing in opponent's material payoff) or spite (utility is decreasing in the opponent's material payoff). Segal and Sobel [6] develop this theory, provide a representation theorem, and discuss equilibrium in games with preferences for reciprocity.

This paper connects the weights in equation (1) to the perceived kindness of opponents' strategies. Intuitively, one would expect that player i would increase the weight he puts on j's material payoff when i is pleased with j's behavior. If j behaves nicely (for example, by making a voluntary contribution to a public good or offering a fair division in a bargaining game), then i might be willing to sacrifice material welfare to make player j better off. That is, i may repay kindness with kindness. At the same time, i may be willing to sacrifice material utility to harm j when j is nasty. The aim of this paper is to identify the connection between nice behavior and reciprocity. We offer a condition under which it follows that nicer behavior by player j will lead player i to put a higher weight on j's utility. In Section 2 we assume that player i has preferences over strategy profiles, which describe his view of their 'niceness.' We then provide conditions on these preferences under which

$$a^j_{i,\sigma} > a^j_{i,\sigma} \text{ if and only if } \sigma \text{ is 'nicer' than } \sigma'. \tag{2}$$

This result captures the idea that a player is more likely to be kind to an opponent who treats him nicely. In order to prove the result, we introduce a Reciprocal Altruism assumption that makes precise the connection between (2) and the niceness order.

Charness and Rabin [1], Rabin [5], and Segal and Sobel [6] define pref-

erences over strategies in strategic-form models of reciprocity.[2] Section 3 discusses the niceness orders that are implicit in the papers of Rabin [5] and Segal and Sobel [6].[3] Dufwenberg and Kirchsteiger [2] and Falk and Fischbacher [3] present models of reciprocity for extensive-form games. In their model, the weight $a_{i,\cdot}$ changes with the play of the game. A generalization of our approach to extensive-form games would appear to require the definition of niceness preferences conditional on all histories and is beyond the scope of the present paper.

2 Reciprocal Altruism

Let X_i be the space of outcomes to player i, $i = 1, \ldots, I$. Each player has preferences $\succeq_i^{out}$ over $\Delta(X_i)$, the space of lotteries over X_i. A game is a collection $\mathbf{s}_i = \{s_i^1, \ldots, s_i^{n_i}\}$ of strategies for player i, $i = 1, \ldots, I$, together with the payoff function $O : \prod_{j=1}^I \mathbf{s}_j \to \prod_{j=1}^I X_j$. Let Σ_i be the space of mixed strategies of player i and extend O to be from $\prod_{j=1}^I \Sigma_j$ to $\prod_{j=1}^I \Delta(X_j)$. Throughout the paper, $\Sigma = \prod_{j=1}^I \Sigma_j$.[4]

Given a game, player i has a complete and transitive preference relation over Σ_i. These preferences depend of course on σ_{-i}, the strategies of other players, and possibly also on i's interpretation of these strategies or the "context" in which the game is being played. We assume that the context is summarized by a mixed strategy profile σ^*, which we interpret as a description of the conventional way in which the game is played.[5] It is within this context that players rank their available strategies. Formally, given $\sigma^* = (\sigma_i^*, \sigma_{-i}^*)$, player i has preferences $\succeq_{i,\sigma^*}$ over Σ_i. The statement $\sigma_i \succ_{i,\sigma^*} \sigma_i'$ says the following. Given the context σ^*, player i would prefer to play σ_i rather than σ_i'.

[2]Rabin [5] studies a special case of the general model in Segal and Sobel [6], although he does not explicitly describe his utility function as a representation of preferences over strategies.

[3]Charness and Rabin [1] uses a preference relationship that exhibits both distributional preferences and intrinsic reciprocity, but we are unable to represent these preferences in the general form studied by Segal and Sobel [6].

[4]Strategies ($\mathbf{s}_i$), strategy sets (Σ_i), outcome functions (O) and preferences over strategies (see below) can vary with the game. Our analysis always concentrates on a fixed game, however, so we suppress this dependence in our notation.

[5]Alternatively, for each i, σ_{-i}^* represents player i's beliefs about how his opponents play the game.

Segal and Sobel [6] give conditions under which $\succeq_{i,\sigma^*}$ can be represented by (1).[6] We assume that these conditions hold and, when they do, say that the players have reciprocity preferences. In this section we relate the relative size of the weight a^j_{i,σ^*} and the niceness of player j's strategy. When there are more than two players, it becomes harder to interpret preferences over opponents' strategies. For example, it is not clear how player i should evaluate player j's utility when player j is nice to k but mean to ℓ. We avoid these issues by concentrating on two-player games. Remark 1 describes one way to extend our results to games with more than two players.

We assume that $i, j \in \{1, 2\}$ and $i \neq j$. When there are only two players the weight player i gives to the utility of player j determines the representation of i's utility function over strategies. For simplicity, we drop the j superscript and denote this weight by a_{i,σ^*}.[7]

To describe player i's attitudes towards player j's behavior we assume that player i has preferences over player j's strategies. Intuitively, i will prefer one strategy profile to another if in the first player j is behaving in a way that i views as "nicer." Different possible attitudes are possible, and we discuss in detail some possibilities below. Leaving room for the most general approach, we assume that this ranking compares pairs of joint strategies. Formally, $\succeq_i^{opp}$ is defined over $\Sigma_i \times \Sigma_j$. (The superscript *opp* stands for "opponent.") The interpretation of the statement "$\sigma^1 \succeq_i^{opp} \sigma^2$" is that player i considers j to be nicer to him when she is using σ_j^1 in response to σ_i^1 than when she is using σ_j^2 in response to σ_i^2. (Of course, this definition does not preclude the possibility that the ranking by i of j's behavior is independent of i's choice of strategy.) In this section we analyze the connection between these preferences and the weight a_{i,σ^*} player i gives to j's utility. We provide conditions that enable us to say when the weight player i puts on player j's utility is an increasing function of the niceness of player j's behavior. That is, we are interested in

[6] Naturally, since the representation theorem in Segal and Sobel [6] is weaker than the standard case, the conditions are weaker that the ones used in standard game theory. Loosely, standard game theory requires that preferences over outcomes determine preferences over strategies for all contexts. Segal and Sobel assume instead that preferences over outcomes determine preferences over strategies only when in contexts where one's strategy choice does not change the material payoffs of all other players.

[7] Even though we conduct our analysis from the perspective of player i, we do not drop the subscript i in order to emphasis that different players will put different weights on the other player's utility from outcomes.

conditions on preferences that guarantee:

$$a_{i,\bar{\sigma}} \geqslant a_{i,\sigma} \text{ iff } \bar{\sigma} \succeq_i^{opp} \sigma. \tag{3}$$

Our main definition captures the intuition behind the idea that players are willing to reward nice behavior and to punish mean behavior. In what follows, we simplify notation by writing $\sigma' = (\sigma_i', \sigma_j)$; $\sigma'' = (\sigma_i'', \sigma_j)$; $\bar{\sigma}' = (\bar{\sigma}_i', \bar{\sigma}_j)$; and $\bar{\sigma}'' = (\bar{\sigma}_i'', \bar{\sigma}_j)$.

Definition 1 The preferences $\succeq_{i,\sigma}$ represent reciprocal altruism if whenever

1. $\sigma_i'' \succeq_{i,\sigma} \sigma_i'$;
2. $u_j(\bar{\sigma}'') - u_j(\bar{\sigma}') = u_j(\sigma'') - u_j(\sigma')$;
3. $u_i(\bar{\sigma}'') - u_i(\bar{\sigma}') \geqslant u_i(\sigma'') - u_i(\sigma')$; and
4. $\bar{\sigma} \succ_i^{opp} \sigma$ [resp. $\bar{\sigma} \sim_i^{opp} \sigma$];

It follows that $u_j(\bar{\sigma}'') > u_j(\bar{\sigma}')$ implies $\bar{\sigma}_i'' \succeq_{i,\bar{\sigma}} \bar{\sigma}_i'$ [resp. It follows that $\bar{\sigma}_i'' \succeq_{i,\bar{\sigma}} \bar{\sigma}_i'$].

Furthermore, if either the preference in Condition 1 or the inequality in Condition 3 is strict, then it follows that $u_j(\bar{\sigma}'') > u_j(\bar{\sigma}')$ implies $\bar{\sigma}_i'' \succ_{i,\bar{\sigma}} \bar{\sigma}_i'$ [resp. It follows that $\bar{\sigma}_i'' \succ_{i,\bar{\sigma}} \bar{\sigma}_i'$]. ■

The definition says that if "things are essentially equal," then when player j plays a nicer strategy, reciprocally altruistic player i will prefer strategies that lead to larger selfish payoffs to player j. In this way, the definition formalizes the notion that player i repays kindness with kindness. Observe however that so far we did not put any restrictions on the relation $\succeq_i^{opp}$, that is, we did not define the meaning of "player j plays a nicer strategy."

The definition appears complicated, so it deserves more discussion. Reciprocal altruism uses information about how player i ranks his own strategies in one context (σ) to draw a conclusion about his ranking in another context ($\bar{\sigma}$). In the first context (σ), player i prefers σ_i'' to σ_i'. There is no reason for player i's preferences over strategies to be preserved in the second context. When Conditions 2 and 3 hold, however, there is a sense in which the relationship between σ_i' to σ_i'' when σ is expected is comparable to the relationship between $\bar{\sigma}_i'$ and $\bar{\sigma}_i''$ when $\bar{\sigma}$ is expected: Compared to going from σ' to σ'', going from $\bar{\sigma}'$ to $\bar{\sigma}''$ leads to the same change in utility from payoffs

for player j and an increase in the change in utility from payoffs for player i. Condition 4 states that player i thinks that player j is nicer is the second context ($\bar{\sigma}$) than in the first (σ). Under these conditions the definition states that player i is reciprocally altruistic if he responds to j's nicer behavior by strengthening his preferences (preferring $\bar{\sigma}_i''$ to $\bar{\sigma}_i'$) when $\bar{\sigma}''$ leads to higher utility (over outcomes) for j than $\bar{\sigma}'$. When σ and $\bar{\sigma}$ are equally nice, then the reciprocally altruistic preferences of player i are not reversed when going from the original comparison (σ'' to σ') to the comparable comparison ($\bar{\sigma}''$ to $\bar{\sigma}'$).

One can check that in standard game theory when player i's preferences over strategies satisfy $\sigma_i'' \succeq_{i,\sigma} \sigma_i'$ iff $u_i(\sigma'') \geqslant u_i(\sigma')$, they represent reciprocal altruism provided that player i views all of j's strategies as equally nice ($\sigma \sim_i^{opp} \bar{\sigma}$ for all $\bar{\sigma}$). To see this, note that the first condition implies $u_i(\sigma'') \geqslant u_i(\sigma')$. The third condition then implies $u_i(\bar{\sigma}'') \geqslant u_i(\bar{\sigma}')$, hence, the conclusion of the definition must hold.

We argue that the condition must hold whenever equation (3) holds. That is, preferences must satisfies the Reciprocal Altruism assumption whenever weights $a_{i,\sigma}$ are increasing in the niceness of σ. Later in the section we provide conditions under which Reciprocal Altruism is in fact necessary and sufficient for equation (3).

Assume that Conditions 1, 2, and 3 of the axiom hold and one can represent player i's preferences over strategies in the form $u_i(\sigma) + a_{i,\sigma}u_j(\sigma)$. The first condition implies that

$$u_i(\sigma'') + a_{i,\sigma}u_j(\sigma'') \geqslant u_i(\sigma') + a_{i,\sigma}u_j(\sigma'),$$

which, when combined with the second and third conditions, yields

$$u_i(\bar{\sigma}'') + a_{i,\sigma}u_j(\bar{\sigma}'') \geqslant u_i(\bar{\sigma}') + a_{i,\sigma}u_j(\bar{\sigma}')$$

hence

$$u_i(\bar{\sigma}'') - u_i(\bar{\sigma}') \geqslant a_{i,\sigma}(u_j(\bar{\sigma}') - u_j(\bar{\sigma}'')). \tag{4}$$

It follows that

$$\begin{aligned}
&\bar{\sigma}_i' \succ_{i,\bar{\sigma}} \bar{\sigma}_i'' \iff \\
&u_i(\bar{\sigma}'') + a_{i,\bar{\sigma}}u_j(\bar{\sigma}'') < u_i(\bar{\sigma}') + a_{i,\bar{\sigma}}u_j(\bar{\sigma}') \iff \\
&u_i(\bar{\sigma}'') - u_i(\bar{\sigma}') < a_{i,\bar{\sigma}}(u_j(\bar{\sigma}') - u_j(\bar{\sigma}'')) \implies \\
&(a_{i,\bar{\sigma}} - a_{i,\sigma})(u_j(\bar{\sigma}'') - u_j(\bar{\sigma}')) < 0
\end{aligned} \tag{5}$$

where the last inequality follows by inequality (4). If either Condition 1 holds with strict preference or Condition 3 holds with a strict inequality, then the inequality in (4) is strict and (5) becomes

$$\bar{\sigma}_i' \succeq_{i,\bar{\sigma}} \bar{\sigma}_i'' \Longrightarrow (a_{i,\bar{\sigma}} - a_{i,\sigma})(u_j(\bar{\sigma}'') - u_j(\bar{\sigma}')) < 0 \tag{6}$$

The next result is a straightforward consequence of inequalities (5) and (6).

Theorem 1 *If players have reciprocity preferences and* $a_{i,\bar{\sigma}} \geqslant a_{i,\sigma}$ *iff* $\bar{\sigma} \succeq_i^{opp} \sigma$*, then preferences represent reciprocal altruism.*

Theorem 1 states that if the weight that player i places on player j's utility is an increasing function of the niceness of player j's behavior, then reciprocal altruism must hold. The main result of this section is a converse to this result, that is, it states conditions under which reciprocal altruism implies that the weight a_{i,σ^*} is increasing in the niceness of player j's strategy. Because of the structure of our model, this statement cannot be globally true, as there are situations where the weights a_{i,σ^*} are not uniquely defined (see Example 1) below.

There are two situations in which a_{i,σ^*} is uniquely determined by player i's preferences over strategies and choices of the utility functions u_i and u_j. When either of these situations hold, then we show in Theorem 2*(1)* below that a_{i,σ^*} is necessarily increasing in the niceness of player j's strategy. They are

1. The set of possible utility allocations generated by player i's activity given σ_j^*, $A_i(\sigma_j^*) = \{u(\sigma_i, \sigma_j^*) : \sigma_i \in \Sigma_i\}$, has a nonempty interior.

2. The set $A_i(\sigma_j^*)$ has an empty interior, but is a non-trivial indifference set of the preferences $\succeq_{i,\sigma^*}$.

There are two cases in which the first of these conditions is not satisfied. Either for *all* $\sigma_j \in \Sigma_j$, $A_i(\sigma_j)$ has an empty interior, in which case we say that the game is poor for player i, or some, but not all of sets $A_i(\sigma_j)$ have an empty interior, in which case we say that σ_j is poor. Strategies that are not poor are rich.

All games where $n_i = 2$ are poor for player i, and as we show in Appendix A, this is, to a certain extent, the only case in which games are poor. We also show there that if a game is not poor, then the set of its poor strategies is closed and of measure zero. Theorem 2*(2)* states that even if in this case

$a_{i,\sigma}$ or $a_{i,\bar{\sigma}}$ are not uniquely determined, it is still possible to choose such weights consistent with reciprocity utility functions and such that $a_{i,\bar{\sigma}} \geqslant a_{i,\sigma}$ iff $\bar{\sigma} \succeq_i^{opp} \sigma$.

Theorem 2 *Assume that the preferences represent reciprocal altruism.*

1. *If $a_{i,\sigma}$ and $a_{i,\bar{\sigma}}$ are uniquely determined, then $a_{i,\bar{\sigma}} \geqslant a_{i,\sigma}$ iff $\bar{\sigma} \succeq_i^{opp} \sigma$.*
2. *Otherwise, there exist specifications of $a_{i,\sigma}$ and $a_{i,\bar{\sigma}}$ consistent with reciprocity utility functions such that $a_{i,\bar{\sigma}} \geqslant a_{i,\sigma}$ iff $\bar{\sigma} \succeq_i^{opp} \sigma$.*

In the second case, the actual choice of $a_{i,\sigma}$ may depend on $\bar{\sigma}$ and not only on σ, and it may differ when the poor strategy σ is compared to $\bar{\sigma}$ or to $\tilde{\sigma}$. The following example shows that this problem is unavoidable.

Example 1 Consider the game

	s_j^1	s_j^2	s_j^3	s_j^4
s_i^1	$2,0$	$2,2$	$1,1$	$1,1$
s_i^2	$0,2$	$0,0$	$1,1$	$5,1$

When j plays

$$\sigma_j(\varepsilon) = \left(\tfrac{\varepsilon}{2-\varepsilon}, \varepsilon\tfrac{1-\varepsilon}{2-\varepsilon}, 1-\varepsilon-\varepsilon^2, \varepsilon^2\right)$$

the set $A_i(\sigma_j(\varepsilon))$ is the line segment connecting

$$\left(1+\varepsilon, 1-\tfrac{\varepsilon^2}{2-\varepsilon}\right) \quad \text{with} \quad \left(1-\varepsilon+4\varepsilon^2, 1+\tfrac{\varepsilon^2}{2-\varepsilon}\right).$$

Suppose further that for every $\varepsilon > 0$, player i prefers playing s_i^2 to s_i^1, [8] and that he considers $\sigma_j(\varepsilon) \succ_i^{opp} \sigma_j(\varepsilon')$ iff $\varepsilon < \varepsilon'$. [9] Now

$$\begin{aligned}
&s_i^2 \succ_{i,\sigma_j(\varepsilon)} s_i^1 \iff \\
&1-\varepsilon+4\varepsilon^2 + a_{i,\sigma_j(\varepsilon)}\left(1+\tfrac{\varepsilon^2}{2-\varepsilon}\right) > 1+\varepsilon + a_{i,\sigma_j(\varepsilon)}\left(1-\tfrac{\varepsilon^2}{2-\varepsilon}\right) \iff \\
&a_{i,\sigma_j(\varepsilon)} > \tfrac{(1-2\varepsilon)(2-\varepsilon)}{\varepsilon}
\end{aligned}$$

[8] Observe that for every $\varepsilon > 0$, the lotteries person i obtains from these two pure strategies cannot be compared by first order stochastic dominance, while the lottery player j receives from s_i^2 dominates the lottery she receives from s_i^1. Moreover, when $\varepsilon \to 0$, all lotteries converge to a sure gain of 1.

[9] For example, i considers j's behavior to be protective, and the lower the value of ε, the higher is the lowest outcome which player i may receive.

Obviously, it is impossible to assign $a_{i,\sigma_j(0)}$ a unique value such that for every $\varepsilon > 0$, $a_{i,\sigma_j(0)} > a_{i,\sigma_j(\varepsilon)}$. Note that when $\varepsilon = 0$, $A_i(\sigma_j(0))$ is the point $(1,1)$, and any value of $a_{i,\sigma_j(0)}$ will be consistent with reciprocity utility functions. (See Fig. 1 where the upper end points of each of the sets $A_i(\sigma_j(\varepsilon))$ represent the utility allocations that follow from i's optimal behavior given j plays $\varepsilon =$ 0.8, 0.5, 0.35, 0). □

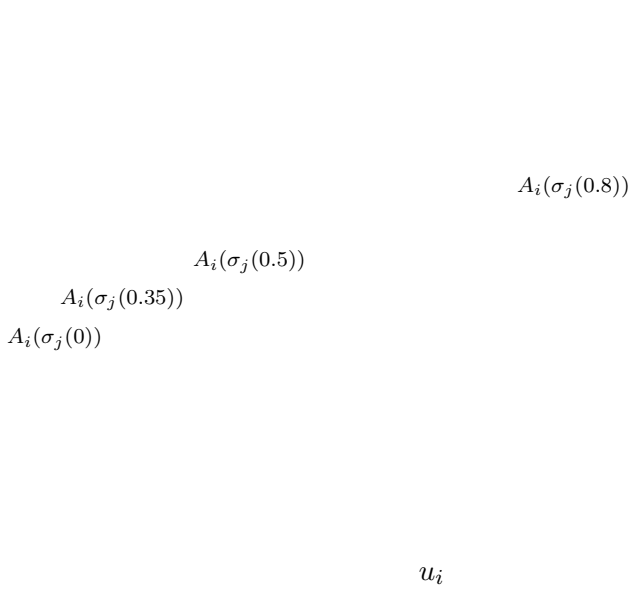

Figure 1: Example 2

Remark 1 We mentioned before that interpretation of preferences over opponents' strategies becomes harder when there are more than two players. The following is a possible extension of the definition of reciprocal altruism to the case of more than two players. In all strategy profiles, σ_k are fixed for all $k \neq i, j$.

Definition 3* The preferences $\succeq_{i,\sigma}$ represent I-person reciprocal altruism if whenever

1. $\sigma_i'' \succeq_{i,\sigma} \sigma_i'$;

2. $u_j(\bar{\sigma}'') - u_j(\bar{\sigma}') = u_j(\sigma'') - u_j(\sigma')$;

3. $u_i(\bar{\sigma}'') - u_i(\bar{\sigma}') \geqslant u_i(\sigma'') - u_i(\sigma')$;

4. For all $k \neq i, j$, $u_k(\sigma') = u_k(\sigma'')$ and $u_k(\bar{\sigma}') = u_k(\bar{\sigma}'')$; and

5. $\bar{\sigma} \succ_i^{opp} \sigma$ [resp. $\bar{\sigma} \sim_i^{opp} \sigma$];

It follows that $u_j(\bar{\sigma}'') > u_j(\bar{\sigma}')$ implies $\bar{\sigma}_i'' \succeq_{i,\bar{\sigma}} \bar{\sigma}_i'$ [resp. It follows that $\bar{\sigma}_i'' \succeq_{i,\bar{\sigma}} \bar{\sigma}_i'$].

Furthermore, if either the preference in Condition 1 or the inequality in Condition 3 is strict then

It follows that $u_j(\bar{\sigma}'') > u_j(\bar{\sigma}')$ implies $\bar{\sigma}_i'' \succ_{i,\bar{\sigma}} \bar{\sigma}_i'$ [resp. It follows that $\bar{\sigma}_i'' \succ_{i,\bar{\sigma}} \bar{\sigma}_i'$].

In words, the reciprocal altruism definition is applied to any two players provided the acts of player i do not affect the utility of all other players (except for his own utility and that of player j). While Theorem 1 and Theorem 2*(1)* extend to I-player games with this definition, we cannot prove Theorem 2*(2)*. □

3 Examples

In this section we discuss the niceness preferences that is implicit in two treatments of reciprocal preferences. Our results enable us to look at a specific functional form for preferences over strategies and determine the implicit niceness preference relationship.

The functional form Rabin uses in the body of his manuscript is equivalent to the conditional preference relationship over strategies given by equation (1).

In order to define the weight, a_{i,σ^*}, Rabin lets $u_i^h(\sigma_i^*)$ be the highest (material) payoff available to player i if player i chooses σ_i^*. That is,

$$u_i^h(\sigma_i^*) = \max_{\sigma_j \in \Sigma_j} u_i(\sigma_i^*, \sigma_j).$$

Similarly, let $u_i^{\min}(\sigma_i^*)$ be player i's lowest payoff among available payoffs; $u_i^l(\sigma_i^*)$ be player i's lowest payoff among available Pareto-efficient payoffs;

and let $u_i^e(\sigma_i^*)$ be the average of $u_i^h(\sigma_i^*)$ and $u_i^l(\sigma_i^*)$. In our notation, Rabin sets $a_{i,\sigma^*} = 0$ if $u_k^h(\sigma_k^*) - u_k^{\min}(\sigma_k^*) = 0$ for $k = i$ or j and otherwise

$$a_{i,\sigma^*} = \frac{u_i(\sigma^*) - u_i^e(\sigma_j^*)}{(u_i^h(\sigma_i^*) - u_i^{\min}(\sigma_i^*))(u_j^h(\sigma_j^*) - u_j^{\min}(\sigma_j^*))}. \tag{7}$$

We refer the reader to Rabin's article for a motivation for these preferences.[10]

The results of Section 2 link the niceness of a strategy profile σ^* to the magnitude of a_{i,σ^*}. Note that in this example, player i's niceness ranking depends on σ_i^*, the strategy that he is expected to play.

Segal and Sobel [6] argue that Rabin's representation cannot explain observed behavior in ultimatum bargaining games and propose If $u_j^h(\sigma_i^*) = \max_{s_j \in S_j} u_j(s_j, \sigma_i^*)$, $\bar{u}_j = \max_{s \in S_i \times S_j} u_j(s)$, and $\underline{u}_j = \min_{s \in S_i \times S_j} u_j(s)$, then

$$a_{i,\sigma^*} = \begin{cases} \lambda \frac{u_i^h(\sigma_j^*) - F^G}{\bar{u}_i - \underline{u}_i} & \text{if } \bar{u}_i - \underline{u}_i > 0, \\ 0 & \text{if } \bar{u}_i - \underline{u}_i = 0, \end{cases} \tag{8}$$

where F^G is a fair outcome of the game G and λ is a normalization factor. If equation (8) defines the weights, then player i views a strategy profile as nicer if it enables i to obtain a higher material payoff. In this case, a player's "niceness" preferences depend only on the strategy choice of the opponent.

Appendix A: Rich and Poor Strategies

If Player i has reciprocity preferences, then a_{i,σ^*} is uniquely determined by u_i and u_j provided that there exist $\sigma', \sigma'' \in \Sigma$ such that

$$\sigma_i' \sim_{i,\sigma} \sigma_i'' \text{ and } u_j(\sigma') \neq u_j(\sigma''). \tag{9}$$

Under these conditions, one has

$$u_i(\sigma') + a_{i,\sigma} u_j(\sigma') = u_i(\sigma'') + a_{i,\sigma} u_j(\sigma'')$$

and so

$$a_{i,\sigma} = \frac{u_i(\sigma'') - u_i(\sigma')}{u_j(\sigma') - u_j(\sigma'')}.$$

We now provide conditions under which (9) must hold.

[10] The appendix of Rabin's paper suggests alternative functional forms.

Definition 4 A game is called poor (for player i) if all the n_i points

$$\{(u(s_i^k, s_j^1), \ldots, u(s_i^k, s_j^{n_j}))\}_{k=1}^{n_i}$$

are on the same line in $\Re^{2n_j}$. A game that is not poor is rich.

In particular, all games where player i has at most two pure strategies are poor for player i. Moreover, if the game is poor for player i, then there are two pure strategies s_1^i and s_2^i such that for all k, the vector $(u(s_k^i, s_\ell^j))_{\ell=1}^{n_j}$ is a linear combination of the vectors $(u(s_1^i, s_\ell^j))_{\ell=1}^{n_j}$ and $(u(s_2^i, s_\ell^j))_{\ell=1}^{n_j}$, and the game is essentially a $2 \times n_j$ game.

A strategy σ_j is rich (for player i) if all the n_i points $(u(s_i^k, \sigma_j))_{k=1}^{n_i}$ are not on the same line in $\Re^2$. By Segal and Sobel [6, Theorem 1], if σ_j is rich, then $a_{i,\sigma}$ is uniquely determined. Strategies that are not rich are poor. The game G is poor for player i iff σ_j is poor for player i for all $\sigma_j \in \Sigma_j$.[11] A rich game will generally have poor strategies, but the next result shows that they are rare.

Lemma 1 *If the game is rich, then the set of poor strategies is a closed set of measure zero in the $n_j - 1$ dimensional simplex.*

Proof If $n_i > 2$ and the game is rich, then the strategy σ_j is poor if either *1.* all the n_i points $u(s_i^1, \sigma_j), \ldots, u(s_i^{n_i}, \sigma_j)$ are the same point, or *2.* these n_i points are not the same, but are on the same line. For $m = i, j$, denote $\sigma_j \cdot u_m(s_i^k) = \sum_{\ell=1}^{n_j} \sigma_j^\ell u_m(s_i^k, s_j^\ell)$. If all the n_i points are the same (case *1* above), then for $k = 2, \ldots, n_i$ the following linear (in $\sigma_j^1, \ldots, \sigma_j^{n_j}$) equations are satisfied:

$$\sigma_j \cdot [u_m(s_i^k) - u_m(s_i^1)] = 0, \quad m = i, j$$

If case *2* holds, then for $k = 3, \ldots, n_i$, the following $n_i - 2$ quadratic (in $\sigma_j^1, \ldots, \sigma_j^{n_j}$) equations are satisfied:

$$\frac{\sigma_j \cdot [u_j(s_i^k) - u_j(s_i^1)]}{\sigma_j \cdot [u_i(s_i^k) - u_i(s_i^1)]} = \frac{\sigma_j \cdot [u_j(s_i^2) - u_j(s_i^1)]}{\sigma_j \cdot [u_i(s_i^2) - u_i(s_i^1)]}.$$

[11] It is clear that if the game is poor then all σ_j are poor. Conversely, if σ_j is poor then there are k_1 and k_2 such that for all $k = 1, \ldots, n_i$, the utility allocation $u(s_i^k, \sigma_j)$ is on the chord connecting $u(s_i^{k_1}, \sigma_j)$ with $u(s_i^{k_2}, \sigma_j)$. Since *all* strategies σ_j are poor, one can show that k_1 and k_2 cannot depend on σ_j, hence the game G is poor.

Since G is rich for player i, the above equations do not have all of Σ_j as a solution, hence the set of solutions is of measure zero (in the (n_i-1)-dimensional simplex). The set of solutions to either set of equations is obviously closed. ■

Appendix B: Proofs

Proof of Theorem 1 Assume that the first three conditions in **RA** hold. If $\bar{\sigma} \sim_i^{opp} \sigma$, then $a_{i,\bar{\sigma}} = a_{i,\sigma}$. It follows from (5) that $\bar{\sigma}_i'' \succeq_{i,\bar{\sigma}} \bar{\sigma}_i'$. On the other hand, if $\bar{\sigma} \succ_i^{opp} \sigma$, then $a_{i,\bar{\sigma}} > a_{i,\sigma}$. Inequality (5) implies that if $u_j(\bar{\sigma}'') - u_j(\bar{\sigma}') \geqslant 0$ then $\bar{\sigma}_i'' \succeq_{i,\bar{\sigma}} \bar{\sigma}_i'$.

Now assume that the first three conditions in the definitions of reciprocal altruism hold with either $\sigma_i'' \succ_{i,\sigma} \sigma_i'$ in Condition 1 or $u_i(\bar{\sigma}'') - u_i(\bar{\sigma}') > u_i(\sigma'') - u_i(\sigma')$. If $\bar{\sigma} \sim_i^{opp} \sigma$, then $a_{i,\bar{\sigma}} = a_{i,\sigma}$. It follows from (6) that $\bar{\sigma}_i'' \succ_{i,\bar{\sigma}} \bar{\sigma}_i'$. On the other hand, if $\bar{\sigma} \succ_i^{opp} \sigma$, then $a_{i,\bar{\sigma}} > a_{i,\sigma}$. Consequently, inequality (6) implies that if $u_j(\bar{\sigma}'') - u_j(\bar{\sigma}') \geqslant 0$ then $\bar{\sigma}_i'' \succ_{i,\bar{\sigma}} \bar{\sigma}_i'$. ■

Proof of Theorem 2 Choose $\sigma = (\sigma_i, \sigma_j)$ and $\bar{\sigma} = (\bar{\sigma}_i, \bar{\sigma}_j) \in \Sigma_i \times \Sigma_j$. If $z, z' \in A_i(\hat{\sigma}_j)$ implies that $z_j = z_j'$, then $\hat{\sigma}$ is poor, $a_{i,\hat{\sigma}}$ can be any number, and the conclusion follows. We therefore assume that when $\hat{\sigma}_j = \sigma_j$ or $\bar{\sigma}_j$ there exist $z, z' \in A_i(\hat{\sigma}_j)$ such that $z_j \neq z_j'$.

Observe that $a_{i,\sigma}$ is uniquely defined either when $A_i(\sigma_j)$ has a non-empty interior or when $A_i(\sigma_j)$ is a non-trivial line segment and $\succeq_{i,\sigma}$ is trivial. Since we ruled out the case where $A_i(\sigma_j)$ is a single point, $a_{i,\sigma}$ is not uniquely defined only when $A_i(\sigma_j)$ is a non-trivial line segment and $\succeq_{i,\sigma}$ is non-trivial. Let $L = a_{i,\sigma}$ if $A_{i,\sigma}$ is well defined. Otherwise, let L be the unique value for which

$$z_i + Lz_j = z_i' + Lz_j' \text{ for all } (z_1, z_2), (z_1', z_2') \in A_i(\sigma_j). \tag{10}$$

As in Section 2, we write $\sigma' = (\sigma_i', \sigma_j)$; $\sigma'' = (\sigma_i'', \sigma_j)$; $\bar{\sigma}' = (\bar{\sigma}_i', \bar{\sigma}_j)$; and $\bar{\sigma}'' = (\bar{\sigma}_i'', \bar{\sigma}_j)$. Notice that when L is defined by eq. (10), and $\sigma', \sigma'' \in A_i(\sigma_j)$,

$$u_i(\sigma') + au_j(\sigma') > u_i(\sigma'') + au_j(\sigma'') \Longleftrightarrow (a - L)(u_j(\sigma') - u_j(\sigma'')) > 0. \tag{11}$$

It is convenient to break the proof into several cases.

Case I: $A_i(\bar{\sigma}_j)$ has a nonempty interior. For $x \in \Re^n$ and $a > 0$, let $B(x, a)$ be the ball with radius a around x. Pick $\sigma_i' \in \Sigma_i$ and $\varepsilon > 0$ such that

$B(u(\bar{\sigma}'), \varepsilon) \in \text{Int } A_i(\bar{\sigma}_j)$ and there is a strategy $\sigma_i'' \in \Sigma_i$ such that

$$u_j(\sigma'') - u_j(\sigma') > 0; \tag{12}$$

$$u_i(\sigma') + Lu_j(\sigma') = u_i(\sigma'') + Lu_j(\sigma''); \tag{13}$$

and $u(\sigma'') \in B(u(\sigma'), \varepsilon)$ (see Fig. 2). In words, when $a_{i,\sigma}$ is well defined and hence $L = a_{i,\sigma}$, given that player j is responding with σ_j to player i playing σ_i (hence the use of $a_{i,\sigma} = a_{i,(\sigma_i,\sigma_j)}$), player i is indifferent between playing σ_i' and playing σ_i''. When $a_{i,\sigma}$ is not well defined, $u(\sigma'')$ is another point in $A_i(\sigma_j)$ (satisfying (12)).

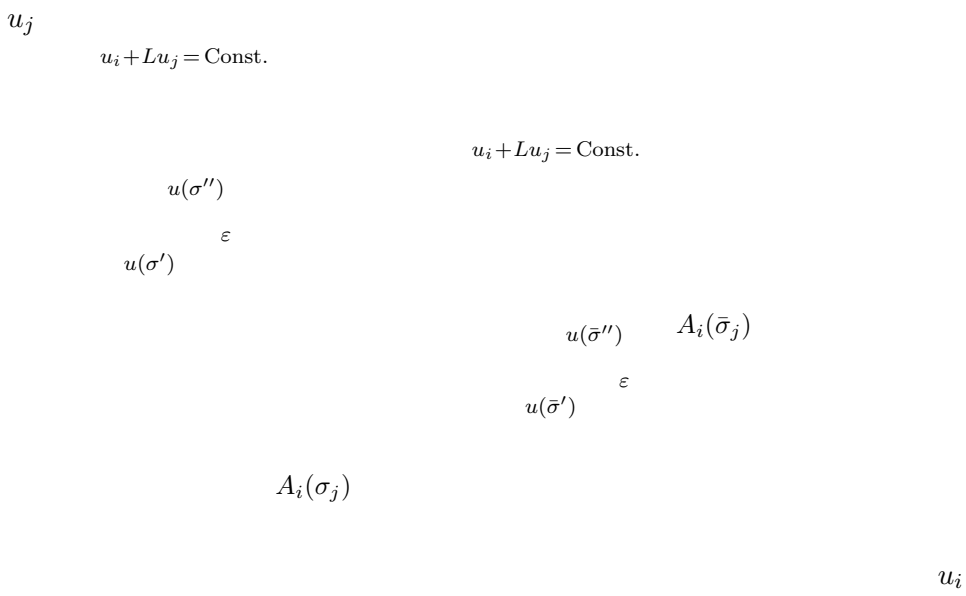

Figure 2: Case I

Since $B(u(\bar{\sigma}'), \varepsilon) \in \text{Int } A_i(\bar{\sigma}_j)$ and $u(\sigma'') \in B(u(\sigma'), \varepsilon)$, it follows that there exists $\bar{\sigma}_i'' \in \Sigma_j$ such that

$$u(\bar{\sigma}'') - u(\bar{\sigma}') = u(\sigma'') - u(\sigma'). \tag{14}$$

Eqs. (13) and (14) imply that

$$u_i(\bar{\sigma}') + Lu_j(\bar{\sigma}') = u_i(\bar{\sigma}'') + Lu_j(\bar{\sigma}'') \tag{15}$$

(geometrically, the line through $u(\sigma')$ and $u(\sigma'')$ is parallel to the line through $u(\bar{\sigma}')$ and $u(\bar{\sigma}'')$). As $A_i(\bar{\sigma}_j)$ has a nonempty interior, $a_{i,\bar{\sigma}}$ is uniquely determined. Furthermore, it follows from eq. (15) that

$$u_i(\bar{\sigma}'') + a_{i,\bar{\sigma}} u_j(\bar{\sigma}'') = u_i(\bar{\sigma}') + a_{i,\bar{\sigma}} u_j(\bar{\sigma}') + (a_{i,\bar{\sigma}} - L)(u_j(\bar{\sigma}'') - u_j(\bar{\sigma}')). \quad (16)$$

Since $u_j(\sigma'') > u_j(\sigma')$, it follows that $u_j(\bar{\sigma}'') > u_j(\bar{\sigma}')$, and we obtain from (16) that

$$u_i(\bar{\sigma}'') + a_{i,\bar{\sigma}} u_j(\bar{\sigma}'') \gtreqless u_i(\bar{\sigma}') + a_{i,\bar{\sigma}} u_j(\bar{\sigma}') \iff a_{i,\bar{\sigma}} \gtreqless L. \quad (17)$$

Case Ia: $L = a_{i,\sigma}$. We have (from eq. (13)) that $\sigma_i' \sim_{i,\sigma} \sigma_i''$ and so by eq. (14), the first three conditions in the definition of reciprocal altruism hold.

1. If $\bar{\sigma} \succ_i^{opp} \sigma$, then, by reciprocal altruism, $\bar{\sigma}_i'' \succeq_{i,\bar{\sigma}} \bar{\sigma}_i'$ iff $u_j(\bar{\sigma}'') \geqslant u_j(\bar{\sigma}')$. Since $u_j(\bar{\sigma}'') > u_j(\bar{\sigma}')$, it follows that $\bar{\sigma}_i'' \succ_{i,\bar{\sigma}} \bar{\sigma}_i'$. But

$$\bar{\sigma}_i'' \succ_{i,\bar{\sigma}} \bar{\sigma}_i' \iff u_i(\bar{\sigma}'') + a_{i,\bar{\sigma}} u_j(\bar{\sigma}'') > u_i(\bar{\sigma}') + a_{i,\bar{\sigma}} u_j(\bar{\sigma}'). \quad (18)$$

The equivalence in (17) guarantees that the inequality in the rhs of (18) holds iff $a_{i,\bar{\sigma}} > L = a_{i,\sigma}$. Therefore, $\bar{\sigma} \succ_i^{opp} \sigma$ implies $a_{i,\bar{\sigma}} > a_{i,\sigma}$.

2. If $\bar{\sigma} \sim_i^{opp} \sigma$, then by reciprocal altruism, $\bar{\sigma}_i' \sim_{i,\bar{\sigma}} \bar{\sigma}_i''$, which means $u_i(\bar{\sigma}') + a_{i,\bar{\sigma}} u_j(\bar{\sigma}') = u_i(\bar{\sigma}'') + a_{i,\bar{\sigma}} u_j(\bar{\sigma}'')$. By eq. (17), $a_{i,\sigma} = a_{i,\bar{\sigma}} = L$.

3. If $\sigma \succ_i^{opp} \bar{\sigma}$, then it follows from the definition of reciprocal altruism (when the roles of $\bar{\sigma}$ and σ are reversed in the axiom) that $\bar{\sigma}_i'' \succeq_{i,\bar{\sigma}} \bar{\sigma}_i'$ implies $\sigma_i'' \succ_{i,\sigma} \sigma_i'$. Hence, since $\sigma_i' \sim_{i,\sigma} \sigma_i''$, it must be that $\bar{\sigma}_i' \succ_{i,\bar{\sigma}} \bar{\sigma}_i''$ and so, by equivalence in (17) $a_{i,\bar{\sigma}} < L = a_{i,\sigma}$.

When $L \neq a_{i,\sigma}$, L is defined by (10). If $\sigma_i'' \succ_{i,\sigma} \sigma_i'$, then we can represent $\succ_{i,\sigma}$ by setting $a_{i,\sigma}$ equal to anything strictly greater than L. If $\sigma_i' \succ_{i,\sigma} \sigma_i''$, we can represent $\succ_{i,\sigma}$ by setting $a_{i,\sigma}$ equal to anything strictly less than L.

Case Ib: $a_{i,\sigma} > L$. We know from (11) and (12) that $\sigma_i'' \succ_{i,\sigma} \sigma_i'$. If $\bar{\sigma} \succ_i^{opp} \sigma$, then reciprocal altruism implies that if $u_j(\bar{\sigma}'') > u_j(\bar{\sigma}')$, then $\bar{\sigma}_i'' \succ_{i,\bar{\sigma}} \bar{\sigma}_i'$. Since $u_j(\bar{\sigma}'') > u_j(\bar{\sigma}')$, it follows that $\bar{\sigma}_i'' \succ_{i,\bar{\sigma}} \bar{\sigma}_i'$. But

$$\bar{\sigma}_i'' \succ_{i,\bar{\sigma}} \bar{\sigma}_i' \iff u_i(\bar{\sigma}'') + a_{i,\bar{\sigma}} u_j(\bar{\sigma}'') > u_i(\bar{\sigma}') + a_{i,\bar{\sigma}} u_j(\bar{\sigma}'). \quad (19)$$

Equivalence (17) guarantees that the inequality in the rhs of eq. (19) holds iff $a_{i,\bar{\sigma}} > L$.

1. $\bar{\sigma} \succ_i^{opp} \sigma$ implies $a_{i,\bar{\sigma}} > L$. The conclusion of the theorem holds provided that $a_{i,\sigma} \in (L, a_{i,\bar{\sigma}})$.

2. If $\bar{\sigma} \sim_i^{opp} \sigma$, then by reciprocal altruism, $\bar{\sigma}_i'' \succ_{i,\bar{\sigma}} \bar{\sigma}_i'$, which means $u_i(\bar{\sigma}'') + a_{i,\bar{\sigma}} u_j(\bar{\sigma}'') > u_i(\bar{\sigma}') + a_{i,\bar{\sigma}} u_j(\bar{\sigma}')$. By (17), $a_{i,\bar{\sigma}} > L$, so letting $a_{i,\sigma} = a_{i,\bar{\sigma}}$ satisfies the conclusion of the theorem.

3. If $\sigma \succ_i^{opp} \bar{\sigma}$, then any choice of $a_{i,\sigma}$ strictly greater than both L and $a_{i,\bar{\sigma}}$ satisfies the conclusion of the theorem.

Case Ic: $a_{i,\sigma} < L$. We know from (11) and (12) that $\sigma_i' \succ_{i,\sigma} \sigma_i''$.
1. If $\bar{\sigma} \succ_i^{opp} \sigma$, then it is clearly possible to specify $a_{i,\sigma}$ so that $a_{i,\bar{\sigma}} > a_{i,\sigma}$.

2. If $\bar{\sigma} \sim_i^{opp} \sigma$, then, by reciprocal altruism, $\sigma_i' \succ_{i,\sigma} \sigma_i''$ implies that $\bar{\sigma}_i' \succ_{i,\bar{\sigma}} \bar{\sigma}_i''$, which means $u_i(\bar{\sigma}') + a_{i,\bar{\sigma}} u_j(\bar{\sigma}') > u_i(\bar{\sigma}'') + a_{i,\bar{\sigma}} u_j(\bar{\sigma}'')$. By equivalence (17), $a_{i,\bar{\sigma}} < L$, hence setting $a_{i,\bar{\sigma}} = a_{i,\sigma}$ has the desired properties.

3. If $\sigma \succ_i^{opp} \bar{\sigma}$, then, by reciprocal altruism, $\bar{\sigma}_i'' \succeq_{i,\bar{\sigma}} \bar{\sigma}_i'$ implies $\sigma_i'' \succ_{i,\sigma} \sigma_i'$. Therefore, as $\sigma_i' \succ_{i,\sigma}$, it follows that $\bar{\sigma}_i' \succ_{i,\bar{\sigma}} \bar{\sigma}_i''$. It follows from the equivalence in (17) that $L > a_{i,\bar{\sigma}}$. Consequently we can satisfy the conclusion of the theorem provided that $a_{i,\sigma}$ is an element of the (non-empty) interval $(a_{i,\bar{\sigma}}, L)$.

Case II: Both $A_i(\sigma_j)$ and $A_i(\bar{\sigma}_j)$ have empty interiors. Let $\bar{L}$ be the unique value for which

$$z_i + \bar{L} z_j = z_i' + \bar{L} z_j' \text{ for all } (z_1, z_2), (z_1', z_2') \in A_i(\bar{\sigma}_j).$$

It is possible to find $\sigma_i', \sigma_i'', \bar{\sigma}_i'$, and $\bar{\sigma}_i'' \in \Sigma_i$ that satisfy (12), (13),

$$u_j(\bar{\sigma}'') - u_j(\bar{\sigma}') = u_j(\sigma'') - u_j(\sigma') \tag{20}$$

and

$$u_i(\bar{\sigma}') + \bar{L} u_j(\bar{\sigma}') = u_i(\bar{\sigma}'') + \bar{L} u_j(\bar{\sigma}''). \tag{21}$$

It follows from equivalence (11) and inequality (12) that

$$\sigma_i'' \succ_{i,\sigma} \sigma_i' \Longleftrightarrow a_{i,\sigma} - L > 0 \tag{22}$$

is a necessary and sufficient condition for $a_{i,\sigma}$ to represent preferences $\succ_{i,\sigma}$. Similarly, it follows from eq. (21) that

$$u_i(\bar{\sigma}'') + au_j(\bar{\sigma}'') > u_i(\bar{\sigma}') + au_j(\bar{\sigma}') \Longleftrightarrow (a - \bar{L})(u_j(\sigma'') - u_j(\sigma')) > 0. \quad (23)$$

It follows from inequality (12) and equivalence (23) that

$$\bar{\sigma}_i'' \succ_{i,\bar{\sigma}} \bar{\sigma}_i' \Longleftrightarrow a_{i,\bar{\sigma}} - \bar{L} > 0 \quad (24)$$

is a necessary and sufficient condition for $a_{i,\bar{\sigma}}$ to represent preferences $\succ_{i,\bar{\sigma}}$. Observe that (13), (20), and (21) imply that

$$u_i(\bar{\sigma}'') - u_i(\bar{\sigma}') = u_i(\sigma'') - u_i(\sigma') + (L - \bar{L})(u_j(\bar{\sigma}'') - u_j(\bar{\sigma}')). \quad (25)$$

The remainder of the proof differs depending on whether L and $\bar{L}$ are equal.

Case IIa: $\bar{L} > L$.

1. If $\bar{\sigma} \succ_i^{opp} \sigma$, then the theorem holds because it is always possible to specify $a_{i,\bar{\sigma}} > \frac{\bar{L}+L}{2} > a_{i,\sigma}$.

2. If $\bar{\sigma} \sim_i^{opp} \sigma$ and $\sigma_i'' \succ_{i,\sigma} \sigma_i'$, then it follows from (22) that any $a > L$ represents the preferences $\succeq_{i,\sigma}$ and when $\bar{L} > L$ it is possible to find an $a_{i,\bar{\sigma}}$ that represents $\succeq_{i,\bar{\sigma}}$ such that $a_{i,\bar{\sigma}} = a_{i,\sigma}$. (If $\bar{\sigma}_i' \succ_{i,\bar{\sigma}} \bar{\sigma}_i''$, $a_{i,\bar{\sigma}}$ can be chosen to be between L and $\bar{L}$).

3. If $\bar{\sigma} \sim_i^{opp} \sigma$ and $\sigma_i' \succeq_{i,\sigma} \sigma_i''$, then inequality (12), eq. (20), and eq. (25) allow us to apply the definition of reciprocal altruism to conclude that $\bar{\sigma}_i'' \succeq_{i,\bar{\sigma}} \bar{\sigma}_i' \Longrightarrow \sigma_i'' \succ_{i,\sigma} \sigma_i'$. Therefore, $\bar{\sigma}_i' \succ_{i,\bar{\sigma}} \bar{\sigma}_i''$. From equivalence (24) we can take any $a < \bar{L}$ to represent $\succeq_{i,\bar{\sigma}}$. Therefore it is possible to satisfy the conclusion of the theorem.

If $\sigma \succ_i^{opp} \bar{\sigma}$ and $\sigma_i' \succeq_{i,\sigma} \sigma_i''$, then inequality (12), eq. (20) and eq. (25) allow us to apply reciprocal altruism to conclude that $\bar{\sigma}_i' \succ_{i,\bar{\sigma}} \bar{\sigma}_i''$. It follows from equivalence (24) that we can take any $a < \bar{L}$ to represent $\succeq_{i,\bar{\sigma}}$. Therefore it is possible to satisfy the conclusion of the theorem by taking $a_{i,\bar{\sigma}} < \min\{\bar{L}, a_{i,\sigma}\}$. If $\sigma \succ_i^{opp} \bar{\sigma}$ and $\sigma_i'' \succ_{i,\sigma} \sigma_i'$, then it follows from equivalence (22) that we can take any $a > L$ to represent $\succ_{i,\bar{\sigma}}$. Therefore it is possible to satisfy the conclusion of the theorem.

Case IIb: $\bar{L} = L$. In this case it follows from (20) and (25) that Conditions 2 and 3 in the definition of reciprocal altruism hold (and Condition 3 holds as an equation). Also note that inequality (12) and eq. (20) imply that $u_j(\sigma'') > u_j(\sigma')$ and $u_j(\bar{\sigma}'') > u_j(\bar{\sigma}')$.

1. If $\bar{\sigma} \succ_i^{opp} \sigma$ and $\sigma_i'' \succeq_{i,\sigma} \sigma_i'$, then reciprocal altruism implies that $\bar{\sigma}_i'' \succ_{i,\bar{\sigma}} \bar{\sigma}_i'$. Hence, from (22) and (24) we can set $a_{i,\sigma} = L$ (if $\sigma_i'' \sim_{i,\sigma} \sigma_i'$) or $a_{i,\sigma} > L$ (if $\sigma_i'' \succ_{i,\sigma} \sigma_i'$) and $a_{i,\bar{\sigma}}$ such that $a_{i,\bar{\sigma}} > a_{i,\sigma}$ to satisfy the theorem. If $\sigma_i' \succ_{i,\sigma} \sigma_i''$, then $a_{i,\sigma}$ can be arbitrarily small by equivalence (22). So, when $\bar{\sigma} \succ_i^{opp} \sigma$, it is possible to select a value for $a_{i,\bar{\sigma}} > a_{i,\sigma}$ that represents $\succeq_{i,\bar{\sigma}}$ since $a_{i,\bar{\sigma}}$ can be close to $\bar{L}$. Similar analysis holds for the case $\sigma \succ_i^{opp} \bar{\sigma}$.

2. If $\bar{\sigma} \sim_i^{opp} \sigma$, then reciprocal altruism implies that $\sigma_i'' \succeq_{i,\sigma} \sigma_i'$ iff $\bar{\sigma}_i'' \succeq_{i,\bar{\sigma}} \bar{\sigma}_i'$, hence it is possible to satisfy the conclusion of the theorem by choosing $a_{i,\sigma} = a_{i,\bar{\sigma}}$. ■

References

[1] Gary Charness and Matthew Rabin. Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117(3):817–869, August 2002.

[2] Martin Dufwenberg and Georg Kirchsteiger. A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2):268–298, May 2004.

[3] Armin Falk and Urs Fischbacher. A theory of reciprocity. *Games and Economic Behavior*, 54(2):293–315, 2006.

[4] John Geanakoplos, David Pearce, and Ennio Stacchetti. Psychological games and sequential rationality. *Games and Ecoonomic Behavior*, 1:60–79, 1989.

[5] Matthew Rabin. Incorporating fairness into game theory. *American Economic Review*, 83(5):1281–1302, December 1993.

[6] Uzi Segal and Joel Sobel. Tit for tat: Foundations of preferences for reciprocity in strategic settings. *Journal of Economic Theory*, 2007.

[7] Joel Sobel. Interdependent preferences and reciprocity. *Journal of Economic Literature*, 43(2):396–440, June 2005.