The Canonical Type Space for Interdependent Preferences[†]

Faruk Gul

and

Wolfgang Pesendorfer

Princeton University

February 2005

Abstract

We develop a model of interdependent preference types to capture situations where an agent's preferences depend on the characteristics and personalities of others. We define a canonical type space and provide conditions under which an abstract type space is a component of the canonical type space. As an application, we develop a model of reciprocity in which agents reward kindness of other agents. We show that this model of reciprocity can address key experimental findings in the literature.

[†] This research was supported by grants SES9911177, SES0236882, SES9905178 and SES0214050 from the National Science Foundation.

1. Introduction

In many situations a person's preferences depend on the characteristics and personalities of those around him. This dependence may be due to the social influence that others exert on the decision-maker or may stem from the fact that the decision-maker cares about the consequences of his choices on those around him. For example, a person's preference over consumption goods may depend on how others in his peer group value these goods. Alternatively, the individuals inclination to charitable giving may depend on the characteristics of the recipients of his charity. Hence, the preferences of an individual are a function of his social environment.

We model the social environment in a simple way and assume that there is one individual other than the decision-maker. Hence, the description of the social environment consists of the attributes of the decision-maker and of this other individual, where these attributes (i.e., type) define how an individual responds to the attributes (type) of others. The extension of our model to more than two agents is straightforward.

The difficulty in modelling situations of interdependent preferences comes from the potential circularity of the formulation of types. Person 1's type tells us how person 1 responds to the various types of person 2. Hence, to define person 1's type we need person 2's types to be well-defined. Conversely, defining person 2's type requires a well defined type space for person 1 and so on. The central contribution of this paper is to find a formulation of interdependent types that is not circular and allows us to interpret those types in a straightforward manner.

In our model, a type has two components, (f_0, f). The parameter f_0 captures all of the relevant attributes of the individual that can be described without explicit reference to his behavior. We call those attributes the agent's "characteristics." The personality of the agent, i.e., how the individual responds to other types, is captured by f. For example, if the agent's inclination to charitable giving depends on the ethnicity of the other agent, then the characteristic would be a description of a person's ethnicity while the personality would capture how an individual reacts to the other agent's ethnicity and personality.

Our notion of a type avoids the circularity mentioned above by requiring that the personality be identifiable by a hierarchy of preference statements that gradually reveal

the type. In round 1 the preference statement depends only on the characteristics of the opponent. More precisely, round 1 specifies a set of possible preferences as a function of the characteristic of the other player. In round $n > 1$ the preference statement depends on the other player's statements in the previous rounds. More precisely, round n specifies a set of preferences for each possible statement of the other player in the previous rounds.

For example, consider a situation where the agents must choose between a generous action G and a selfish action S and the characteristic identifies an agent's ethnicity. Round 1 specifies the possible optimal actions for each ethnicity of the other player. For example, the round 1 statement of player 1 could be:

For every ethnicity of player 2 both choices (G and S) are possibly optimal.

In that case, the characteristic does not determine player 1's preference. Player 1 cannot narrow down his preferences if all he knows about player 2 is his ethnicity. Suppose round 1 of player 2's personality specifies:

I will choose G if player 1 has my ethnicity and S otherwise.

In this case, round 1 determines a unique action (preference) for each contingency. Player 2's preference is fully determined once he knows the characteristic of player 1. Next, consider the following round 2 personality statement for player 1:

I will choose G if player 2's round 1 choice is independent of (my) ethnicity (i.e., is a constant function) and S otherwise.

Hence, player 1 chooses the selfish action if player 2's round 1 statement discriminates based on the ethnicity of his opponent; otherwise, player 1 chooses the generous action. Therefore, round 2 specifies a unique preference for player 1 for every possible round 1 statement of player 2. Substituting the personality for player 2 described above, we find that player 1 will choose the selfish action because player 2's round 1 statement depends on the ethnicity of player 1. In this particular example, the personality of player 2 is identified in the first round while the personality of player 2 is identified after round 2.

We require that the hierarchy of preference statements eventually lead to a unique preference for each contingency. Hence, for each type of player 1 and each type of player 2

there must exist a finite number of rounds after which player 1 can determine his preference. The collection of interdependent preference models that can be described in this way is our canonical type space.

Our main result relates the canonical type space to the following "reduced form" description of interdependent preferences. There is a compact type space T and a continuous function γ such that $\gamma(t, t')$ is the preference of player 1 if he is type t and player 2 is type t'. In addition, there is a continuous function ω that describes the characteristic of each type. We call $M = (T, \gamma, \omega)$ an interdependent preference model (IPM). Our main result (Theorem 2) identifies a simple condition (*validity*) that is necessary and sufficient for M to be a component of the canonical type space. Validity can be interpreted as a consistency requirement that ensures that a player's type can be inferred from observed behavior without assuming that the agents have prior knowledge of other players' types. Validity is formally defined and explained in section 2.

Our canonical type space provides a foundation for valid IPMs that is analogous to the Mertens and Zamir (1985)/Brandenburger and Dekel (1994) foundations for informational (Harsanyi) types. Mertens and Zamir/Brandenburger and Dekel define a type as an infinite hierarchy of beliefs over a set of possible parameters. Those parameters – or *payoff types* (Battigalli and Siniscalchi (2003)) – are by assumption exogenous and therefore require no further explanation. The interdependence of Harsanyi types arises from the interaction of the agents' beliefs. Agent 1's type influences agent 2's payoff because 1 has information about a payoff relevant parameter. Interdependence in our setting is not related to a player's information. A player's personality specifies how the player reacts to the characteristics and personalities of other players and is independent of what the player knows. As a result, the standard definition of a type cannot capture interdependent preference types.

Our construction of preference types and the Mertens-Zamir/Brandenburger and Dekel construction of epistemic types are complementary. In a more general model, epistemic types (i.e., the infinite hierarchies of beliefs) can be defined over interdependent preference types (i.e., the set of parameters).[1]

[1] Alternatively, one could develop a model in which interdependent preference types and epistemic types are constructed simultaneously. Such a model would allow for an interaction between preference interdependence and belief interdependence. We leave it for future research to analyze such a model.

In section 4, we present an application of our model. We present a formal notion of reciprocity and identify a class of interdependent preference models based on this notion. Experimental results suggest that subjects reciprocate generous behavior even if reciprocating is not in the their material self-interest. Camerer and Thaler (1995) survey results related to the ultimatum bargaining game. In that game, player 1 proposes a division of surplus and player 2 accepts or rejects. If player 2 rejects both players receive nothing. In experiments it is routinely observed that subjects reject the proposed division even though rejection makes them strictly worse off. One explanation of this and related experimental findings is that subjects care about the payoff of both players. Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) provide such models.

Experiments by Blount (1995) and Falk, Fehr, and Fishbacher (2000) (FFF) demonstrate that subjects care not only about outcomes but also about the opponent's intentions, i.e., types. FFF consider a simple sequential game. In the first stage, player 1 can make a (positive or negative) transfer to player 2. Increasing player 2's payoff by two units costs player 1 a unit of his own payoff. In the second stage, player 2 can reward or punish player 1. To examine whether intentions affect behavior, FFF examine two treatments of this experiment. In the first treatment, each subject is free to choose any strategy. In the second treatment, a randomization device, calibrated to match the distribution of aggregate play in the first treatment, makes player 1's choice for him. The key finding is that in the second treatment subjects (in the role of player 2) are less inclined to punish or reward their opponents than in the first treatment. The interpretation is that play in the second treatment does not reveal the opponent's type and hence removes a motive for punishment or reward.

Section 4, we develop a model of reciprocity. Let $A = [0,1] \times [0,1]$ denote the possible outcomes. For $(a_1, a_2) \in A$, the quantity a_1 is the individual's own reward and a_2 is the opponent's reward. Assume that preferences are described by a single parameter $\beta \in [\beta_*, \beta^*]$ where β captures the weight the agent puts on the opponent's payoff.[2] The preference β maximizes

$$\beta u(a_2) + (1-\beta)u(a_1)$$

[2] Note that β_* may be negative. A preference $\beta < 0$ is spiteful, i.e., the individual is willing to trade off a reduction in his own reward for a reduction in the opponent's reward.

for some fixed u. The parameter β depends on the player's type and on his opponent's type. Suppose a type is a pair of numbers $t = (t_1, t_2)$ with $t_n \in [0, 1]$ for $n = 1, 2$. The function γ specifies a player's β given his own type t and the his opponent's type t'. Hence, $\beta = \gamma(t, t')$ where

$$\gamma(t, t') = b_0 + b_1 \cdot t_1 + b_2 \cdot t_2 \cdot t'_1 \tag{1}$$

The constants b_1 and b_2 are positive numbers and $b_0 \geq \beta_*$. The parameters of a type $t = (t_1, t_2)$ have a simple interpretation: the parameter t_1 measures the minimal weight β over all opponent types. To see this note that for type (t_1, t_2) the preference β satisfies $\beta \geq t_1 + b_0$ irrespective of the type of the opponent. Hence, t_1 measures the *kindness* of the type (t_1, t_2). The parameter t_2 measures how this type responds to opponents' kindness. Hence, t_2 measures the type's *reciprocity*.

In experiments, when players are matched to play a game they do not know their opponent's personality. Players must form beliefs about their opponent's type and choose their actions accordingly. Hence, a player's action will reflect both his own type and his beliefs regarding his opponent's type. The reciprocity model above can be used to analyze the FFF experiment. In treatment 1, player 1 chooses the transfer and therefore reveals information about his type to player 2. A higher transfer will typically mean a higher parameter t_1 for player 1. A reciprocating player 2 – that is, if player 2's type is (t'_1, t'_2) with $t'_2 > 0$ – will put more weight on his opponent's payoff if he believes that the opponent's t_1 is higher. Hence, player 2 will put a higher weight on player 1's utility if player 1 chooses a higher transfer. In contrast, the transfer in treatment 2 reveals no information about player 1's type and therefore player 2's weight on his opponent's utility is independent of the chosen transfer.

Levine (1998) provides a model that is similar to our model of reciprocity above. His formula is a special case of equation (1) where types have the form (t_1, λ) for some fixed λ. Hence, in Levine's model all types have the same parameter of reciprocity but are differentiated according to their kindness. In section 4, we provides a model of reciprocity that generalizes equation (1). Our Theorem 3 gives conditions on the type space that identify this generalization. We use the class of preferences identified by Theorem 3 to address the key experimental findings in the literature.

1.1 Related Literature: Psychological Games

Our model maintains the separation of preferences and beliefs. This is in contrast to the model proposed by Geanakoplos, Pearce and Stacchetti (1989) (GPS) where players preferences may depend directly on the beliefs of players. In section 5, we provide a detailed comparison of GPS and our model.

Based on GPS, several authors[3] have proposed models of reciprocity. Rabin (1993) develops a theory of fairness and reciprocity in normal form games. Dufwenberg and Kirchsteiger (2005) propose such a theory for extensive form games. Segal and Sobel (2003) assume that a player's utility function is parameterized by the [player's belief about the] strategy profile in the game. They provide axioms yielding a separable utility function that incorporates the opponent's welfare.

The models based on GPS require a strategic context to define reciprocity or other personality traits of individuals. Each agent formulates beliefs about the opponents' behavior and the opponent's beliefs. Those beliefs in turn trigger a desire to reward or punish opponents. In our approach the desire to punish or reward is triggered not by beliefs but by the *personality of the opponent*. An agent's personality is defined independently of the strategic context and is characterized by how this agent's preferences change as he meets opponents of different types. For example, a generous person is someone who puts a large weight on the other agents' payoff irrespective of their type. An reciprocating agent is someone who puts more weight on the opponent's payoff if the opponent is more generous, etc. The advantage of our model is that types can be identified independently of the particular strategic context.

[3] See Sobel (2004) for a survey. See also Charness and Rabin (2002), Cox and Friedman (2002), Falk and Fischbacher (1998) for other models related to GPS.

2. Valid Interdependent Preference Models

Let A denote the compact metric space of alternatives. A binary relation R on A is *transitive* if xRy, yRz implies xRz for all $x, y, z \in Z$. The binary relation R is complete if either xRy or yRx holds for all $x, y \in Z$. If R is both transitive and complete, we say that R is a preference relation. The binary relation R is continuous if for all $x \in Z$, the sets $\{y \in Z \,|\, yRx\}, \{y \in Z \,|\, xRy\}$ are closed subsets of Z. Let $\mathcal{R} \subset A \times A$ be a nonempty and compact set of continuous preference relations on A.

When X_j is a metric spaces for all j in some countable or finite index set J, we endow $\times_{j \in J} X_j$ with the sup metric. For any compact metric space X, we endow $\mathcal{H}_X$, the set of all nonempty, closed subsets of X with the Hausdorff metric. Note that $\mathcal{R} \subset \mathcal{H}_{A \times A}$ is a compact metric space.

Let Ω denote a compact metric space of characteristics and let T denote the compact metric space of types. Let $\gamma : T \times T \to \mathcal{R}$ and $\omega : T \to \Omega$ be continuous functions and assume that ω is onto.[4] The collection $M = (T, \gamma, \omega)$ is called an interdependent preference model (IPM). The function ω describes a type's characteristic. A characteristic refers to those attributes that can be described or observed without reference to the type's behavior. For example, a characteristic could be a physical attribute, such as ethnicity. We refer to those aspects of a type that can only be described through behavior as *personality*.

Example 1: (Leaders and Followers) Player 1 must choose between a green shirt G and a red shirt R. The ranking of R versus G depends on both 1's and 2's color preference and on both players' responsiveness to each other's color preference. Suppose each player can be one of four possible personality types. (The set of types is identical for the two players). Two of these types are confident and always choose according to their own color preferences. The two remaining types are insecure and yield to the other player's color preference if that player is confident. We describe these preferences in the following table. In this example, all types have the same characteristic.

[4] Note that the ontoness entails no loss of generality since we can re-define Ω as $\omega(T)$.

	1	2	3	4
1	G	G	G	G
2	G	G	R	G
3	R	R	R	R
4	G	R	R	R

Table 1

The rows in Table 1 indicate the preference of a particular type of player 1 as player 2's type varies. Types 1 and 3 are confident types with different color preferences; types 2 and 4 are insecure. The interpretation is that insecure types adopt the color ranking of the opponent if (and only if) the opponent is confident.

Example 2: (Fixed Point Types) As in Example 1, player 1 must choose between a green shirt G and and a red shirt R. Type 1 prefers G if player 2 is also type 1 and R if player 2 is type 2. Type 2 prefers G if his partner is type 2 and R if his partner is type 1. All types have the same characteristic. The following table summarizes this IPM.

	1	2
1	G	R
2	R	G

Table 2

Our model includes the IPM of Example 1 but rules out the IPM of Example 2. Our criterion requires that types must be distinguishable through their behavior. To understand what this means, consider an experiment where a number of individuals whose interdependent preferences fit the description of Example 1 are repeatedly paired and required to make choices. In round 1, each agent is matched with various opponents. At this stage, agents have no information about their opponent's type. The preferences of types 1 and 3 are independent of the opponent's type; type 1 always chooses G and type 3 always chooses R. The preferences of types 2 and 4 depend on their beliefs about the opponent's type. When a type 2 or type 4 agent does not know his opponent's type, he may choose G or R. For example, if type 2 believes that his opponent is type 3 then he

chooses R; if type 2 believes that his opponent is type $1, 2$ or 4 then he chooses G. If there is enough variation over the agent's beliefs then types 2 and 4 sometimes choose G and sometimes choose R. In that case round 1 reveals the partition $\mathcal{D} = \{\{1\}, \{3\}, \{2, 4\}\}$ of the type space.

In round 2 agents are again repeatedly paired and required to make choices. In round 2 all agents know the outcomes from round 1 and therefore can identify the partition element $D \in \mathcal{D}$ that contains their opponent's type. A type 2 agent whose opponent is known to be in $\{2, 4\}$, will choose G while a type 4 agent in the same situation will choose R. Hence, observed behavior in round 2 will distinguish between types 2 and 4. Thus, all types are identified.

Now consider a similar experiment for Example 2. A type 1 agent who believes his opponent is of type 1 chooses G, while a type 1 agent who believes his opponent is of type 2 chooses R. Similarly, a type 2 agent who believes his opponent is of type 2 chooses G, while a type 2 agent who believes his opponent is of type 1 chooses R. Hence, if agents' beliefs vary sufficiently throughout round 1, both types of agents (1, 2) are observed making both choices (G, R). Therefore, round 1 reveals no information and we are left with the partition $\{\{1, 2\}\}$. In round 2, each agent is exactly in the same situation as in round 1. He knows that his opponent can be either type 1 or type 2. As a consequence, both types of agents (1, 2) again make both choices (G, R). Thus, subsequent behavior never reveals any information and therefore the agents' types cannot be identified.

Valid interdependent preference models are models where agents can figure out the opponent's type when given a sufficiently rich set of observations. A model is not valid if such identification is not possible. Observations consist of the characteristics of an agent (which is directly observable) and the set of preferences of the agent for a given history of observations. A valid model requires that types can be identified (1) *without prior knowledge* regarding the the opponent type (other than his characteristic) and (2) *without directly observing the agent's beliefs*. On the other hand, even if the model is valid, identifying an agent's type may require a rich set of observations. We assume that (3) *all possible preferences of the agent can be observed at each stage*. We can interpret this assumption as saying that the agents' beliefs vary sufficiently so that each preference is

eventually realized. To see the importance of this last assumption note that without a rich set of observations types may not be identifiable even in Example 1. For example, if in round 1 all agents of type 1 and 2 always respond with G, then it becomes impossible to distinguish between these two types.

Note that the IPM in Example 2 has a "fixed point" flavor: each type can be identified from his responses to the opponent provided that the opponent's type is already identified. While our model does not allow the IPM in example 2, it does permit the following version of the example: Suppose in Example 2 there are two possible characteristics (for example, "tall" and "short"). Suppose type 1 is tall whereas type 2 is short. Each type prefers G when he and his opponent share the same characteristic and prefers R otherwise. Now types can be identified from observed behavior because this behavior depends on an observable characteristic.

Next, we provide a formal criterion for an IPM to be valid. A decomposition (partition) $\mathcal{D}$ of T is a pairwise disjoint collection of subsets with $\bigcup_{D\in\mathcal{D}} D = T$. Let D^t denote the unique element of $\mathcal{D}$ that contains t. The decomposition $\mathcal{D}$ is non-trivial if there is $t \in T$ such that $D^t \neq \{t\}$. We use the (standard) notation

$$\gamma(t, D) := \{R \in \mathcal{R} | P = \gamma(t, t') \text{ for some } t' \in D\}$$

Definition: *The IPM M is valid if there does not exist a non-trivial decomposition $\mathcal{D}$ of T such that*

(i) $t, t' \in D \in \mathcal{D}$ implies $\omega(t) = \omega(t')$

(ii) $t' \in D^t \in \mathcal{D}$ implies $\gamma(t, D) = \gamma(t', D)$ for all $D \in \mathcal{D}$.

Validity requires that we cannot find a non-trivial decomposition of the type space such that types in each set are indistinguishable. It is easily checked that Example 1 is valid while Example 2 is not. For Example 2 the decomposition $\mathcal{D} = \{\{1, 2\}\}$ satisfies (i) and (ii) and hence types are indistinguishable.

Validity captures the idea that the process of learning through observations illustrated above does not terminate at a non-trivial decomposition. To understand the connection, first assume that the IPM is valid. Consider a process of learning from observations

as illustrated above. Let $\mathcal{D}$ be a non-trivial decomposition that describes the agent's information about his partner's type at the beginning of a particular round of interactions. In particular, an element of the decomposition $\mathcal{D}$ represents what is known about a group of subjects. Validity implies that for some $t' \in D^t$ we have $\omega(t) \neq \omega(t')$ or $\gamma(t, D) \neq \gamma(t', D)$ for some $D \in \mathcal{D}$. Hence, behavior in this round (or the characteristic) allows agents to distinguish between the types $t, t' \in D^t$. Validity implies that this process cannot terminate unless all types are identified.

Conversely, suppose the IPM is not valid. Let $\mathcal{D}$ be a non-trivial decomposition that satisfies (i) and (ii) of the definition. Conditions (i) and (ii) imply that neither the observed behavior nor the characteristics will reveal any further information about the type of the subjects. Once the informational situation corresponding to a non-trivial $\mathcal{D}$ is reached there are types t, t' that can never be distinguished. This implies that starting from any coarser decomposition $\mathcal{D}^*$ the decomposition following a round of interactions must be coarser than $\mathcal{D}$. Therefore, starting from a situation where agents have no information about each other's type, we can never reach a finer partition than $\mathcal{D}$, and therefore we can never distinguish all types.

3. The Canonical Type Space

This section identifies the canonical type space for valid interdependent preference models. The canonical type space depends on the set of alternatives A, the set of preferences $\mathcal{R}$ and a compact space of characteristics Ω. However, we suppress the dependence of the canonical type space space on $A, \mathcal{R}, \Omega$.

Let X, Z be compact metric spaces. Recall that $\mathcal{H}_Z$ denotes the set of all non-empty, closed subsets of Z. We let $\mathcal{C}(X, \mathcal{H}_Z)$ denote the set of all functions $f : X \to \mathcal{H}_Z$ such that $G(f) = \{(x, z) \in X \times Z \,|\, z \in f(x)\}$ is closed in $X \times Z$. We endow $\mathcal{C}(X, \mathcal{H}_Z)$ with the following metric: $d(f, g) = d_H(G(f), G(g))$ where d_H is the Hausdorff metric on the set of all nonempty closed subsets of $X \times Z$. We identify the function $f : X \to Z$ with the function $\bar{f} : X \to \mathcal{H}_Z$ such that $\bar{f}(x) = \{f(x)\}$ for all $x \in X$. It is easy to verify that such a function f is an element of $\mathcal{C}(X, \mathcal{H}_Z)$ if and only if f is continuous. Hence, we use $\mathcal{C}(X, Z) \subset \mathcal{C}(X, \mathcal{H}_Z)$ to denote the set of continuous functions from X to Z.

Let $\mathcal{H} = \mathcal{H}_{\mathcal{R}}$ denote the collection of non-empty, closed subsets of the set of preferences $\mathcal{R}$. We begin by defining a sequence of sets that represent a *system of interdependent preference hierarchies*.

Definition: *A collection of nonempty compact sets* $(\Theta_0, \Theta_1, \ldots)$ *is a system of interdependent preference hierarchies if* $\Theta_0 = \Omega$ *and*

$$\Theta_n \subset \Theta_{n-1} \times \mathcal{C}(\Theta_{n-1}, \mathcal{H})$$

for all $n \geq 1$.

The entry $\theta_0 \in \Theta_0$ specifies a characteristic. The entry θ_1 is a pair, $\theta_1 = (f_0, f_1)$ with f_0 a characteristic and $f_1 \in \mathcal{C}(\Theta_0, \mathcal{H})$. Hence, the function f_1 specifies for every characteristic a subset of preferences. More generally, $\theta_n = (f_0, \ldots, f_n)$ with $f_0 \in \Theta_0$ and $f_k \in \mathcal{C}(\Theta_{k-1}, \mathcal{H})$ for $k = 1, \ldots, n$. The function f_k specifies for each θ_{k-1} a set of preferences that represents the possible preferences for this entry in the hierarchy.

The entries $(f_0, f_1, \ldots, f_n) \in \Theta_n$ should be interpreted as a sequence of "reports" providing information about the agaent's type. The report θ_0 contains the characteristic. The report θ_1 specifies the characteristic together with a set of possible preferences for every possible characteristic of the opponent. The report θ_n contains the report θ_{n-1} together with a set of preferences for every possible report θ'_{n-1} of the opponent. For this interpretation to make sense, the entries $\theta_n \in \Theta_n$ must satisfy a consistency requirement.

Consider any $\theta_{n+1} = (f_0, \ldots, f_{n+1}) \in \Theta_{n+1}$. Consistency requires that the functions f_{n+1} and f_n act consistently. This means that for any $\theta'_n = (\theta'_{n-1}, f'_n)$ we must have $f_{n+1}(\theta'_n) \subset f_n(\theta'_{n-1})$, i.e., adding f'_n to the opponent's report θ'_{n-1} must imply a smaller set of possible preferences for θ_{n+1}. Consistency also requires that f_n contains all the available information, that is, $f_n(\theta'_{n-1})$ does not contain any preference that is sure to be removed in the next round. Therefore, consistency requires that $f_n(\theta'_{n-1})$ is the union of the sets $f_{n+1}(\theta'_n)$ taken over all the possible continuations θ'_n of θ'_{n-1} (i.e., $\theta'_n = (\theta'_{n-1}, f'_n)$ for some f'_n.)

Definition: *The system of interdependent preference hierarchies* $(\Theta_0, \Theta_1, \ldots)$ *is consistent if for all* $n \geq 1$ *and for all* $(f_0, f_1, \ldots, f_{n+1}) \in \Theta_n$

$$f_n(\theta'_{n-1}) = \bigcup_{\{f'_n | (\theta'_{n-1}, f'_n) \in \Theta_n\}} f_{n+1}(\theta'_{n-1}, f'_n) \qquad (*)$$

for all $\theta'_{n-1} \in \Theta_{n-1}$.

Note that a consistent system of interdependent preferences has the feature that for every $\theta_n \in \Theta_n$ there is f_{n+1} such that $(\theta_n, f_{n+1}) \in \Theta_{n+1}$. Hence, every report θ_n has a feasible continuation. To see this, take any $(f'_0, \ldots, f'_n, f'_{n+1}) \in \Theta_{n+1}$ and note that $f'_n(\theta_{n-1})$ must be non-empty by definition. If $\{f_n | (\theta_{n-1}, f_n) \in \Theta_n\}$ were empty then $(*)$ would imply that $f'_n(\theta_{n-1})$ is also empty.

Next, we define types and components of the canonical type space. For a given consistent system of interdependent preference hierarchies $(\Theta_0, \Theta_1, \ldots)$ we define a type to be a sequence $(f_0, f_1, \ldots)$ with the property that $(f_0, \ldots, f_n) \in \Theta_n$. To qualify as a component of the canonical type space, Θ must satisfy an additional property. Every type must generate a unique preference when confronted with any other type in the component Θ. This means that for every pair of types $(f_0, f_1, \ldots), (f'_0, f'_1, \ldots)$ it must be the case that $f_n(f'_0, \ldots, f'_{n-1})$ converges to a singleton as $n \to \infty$. To simplify the notation, let $\theta(n) = (f_0, f_1, \ldots, f_n)$ denote the $n-$truncation of the sequence $\theta = (f_0, f_1, \ldots)$.

Definition: *Let* $(\Theta_0, \Theta_1, \ldots)$ *be a consistent sequence of interdependent preference hierarchies. Let* $\Theta := \{\theta \in \Theta_0 \times \prod_{n=1}^{\infty} \mathcal{C}(\Theta_{n-1}, \mathcal{H}) \mid \theta(n) \in \Theta_n\}$. *Then* Θ *is a component of interdependent types if* Θ *is compact and for all* $\theta = (f_0, f_1, \ldots) \in \Theta$

$$\textstyle\bigcap_{n \geq 0} f_{n+1}(\theta'(n)) \text{ is a singleton}$$

for all $\theta' \in \Theta$.

The canonical type space is the union of all the components of interdependent types. Let $\mathcal{I}$ denote the set of all components of interdependent types. The set

$$\mathcal{F} = \bigcup_{\Theta \in \mathcal{I}} \Theta$$

is the *canonical interdependent preference type space* or simply the canonical type space. Note that each element $\theta \in \mathcal{F}$ belongs to a unique component $\Theta \in \mathcal{I}$. Hence, $\mathcal{I}$ is a decomposition (or partition) of $\mathcal{F}$.

For any $\Theta \in \mathcal{I}$, let $\psi : \Theta \times \Theta \to \mathcal{R}$ denote the function that specifies the preference of type θ when the opponent is type θ'. Hence,

$$\psi(\theta, \theta') := \bigcap_{n \geq 0} f_{n+1}(\theta'(n)) \text{ for } (f_0, f_1, \ldots) = \theta$$

Further, we define the function $\phi : \Theta \to \Omega \times \mathcal{C}(\Theta, \mathcal{R})$ that specifies for every type $\theta \in \Theta$ the characteristic of the type θ and the mapping used by type θ to assign a preference to opponent types. Hence,

$$\phi(\theta) := (f_0, \psi(\theta, \cdot))$$

Theorem 1 shows that ψ is continuous and ϕ is a homeomorphism from Θ to $\phi(\Theta)$.

Theorem 1: *The function ψ is continuous and ϕ is a homeomorphism from Θ to $\phi(\Theta)$.*

An immediate consequence of Theorem 1 is that any component $\Theta \in \mathcal{I}$ is an interdependent preference model as defined in section 2. For $\Theta \in \mathcal{I}$ define $M^\Theta = (\Theta, \psi, \omega)$ where $\omega(\theta) = \theta_0$. Since Θ is compact (by definition) and ψ is continuous it follows that M^Θ is an IPM. Our main theorem, Theorem 2 proves the converse: any valid IPM corresponds to a component $\Theta \in \mathcal{I}$.

Two IPM's $M = (T, \gamma, \omega)$, $M' = (T', \gamma', \omega')$ are isomorphic if there exists a homeomorphism $\iota : T \to T'$ such that $\omega(t) = \omega'(\iota(t))$ and $\gamma(s,t) = \gamma'(\iota(s), \iota(t))$ for all $s, t \in T$.

Theorem 2: *An interdependent preference model $M = (T, \gamma, \omega)$ is valid if and only if it is isomorphic to a component of the canonical type space.*

4. Reciprocity

In this section, we identify a class of interdependent preference models that capture the idea that agents reciprocate. We consider a simple setting where outcomes are pairs of prizes, denotes $A \subset [0,1]^2$. We assume that every preference (on A) can be described by a single real number $\beta \in (-\infty, 1)$ where $(a_1, a_2) \succeq (a_1', a_2')$ if and only if

$$\beta u(a_2) + (1-\beta)u(a_1) \geq \beta u(a_2') + (1-\beta)u(a_1') \tag{1}$$

where u is a strictly increasing utility function. We assume that the set of admissible preferences is a compact interval

$$\mathcal{R}^o = [\beta_*, \beta^*], \beta_* < \beta^* < 1 \tag{2}$$

An example is a situation where a reward is given either to the individual, to his opponent, or to neither of the two agents. In this case, the set A represents the lotteries over the three possible outcomes. Let $(a_1, a_2) \in A$ and assume that a_1 represents the probability the agent receives the prize while a_2 represents the probability that the opponent receives the prize. In that case, every continuous preference that satisfies the independence axiom and is strictly monotone in the own-prize probability can be represented by a single parameter $\beta \in (-\infty, 1)$ where

$$(a_1, a_2) P (a_1', a_2') \text{ if and only if } \beta a_2 + (1-\beta) a_1 \geq \beta a_2' + (1-\beta) a_1' \tag{1'}$$

Hence, in this case the restriction to a one-dimensional set of preferences corresponds to the standard expected utility hypothesis and a straightforward monotonicity assumption.

Of course, other interpretations of the set A are possible. For example $A = [0,1]^2$ can be interpreted as the set of money transfers to the two agents. For this interpretation the assumed set of preferences is clearly more restrictive.

A higher β means that the individual is willing to give up more in exchange for a fixed benefit hence, higher β's correspond to kinder types. A negative β means that the agent is spiteful, that is, he is better off if the other player receives a smaller reward. A positive β means that the player is altruistic, that is, he is better off if the other player receives a higher reward.

Below, we provide assumptions that identify the following simple model of reciprocity. A type is a sequence of non-negative numbers $t = (1, t_1, ..., t_n, ...)$ with $t_n \in [0, \bar{t}_n]$ where $\bar{t}_n \geq 0$ and $\bar{t}_n = 0$ implies $\bar{t}_{n+1} = 0$. (The first coordinate is normalized to 1). When a player has type t and his opponent has type t' then the resulting preference is $\gamma(t, t')$ where

$$\gamma(t, t') = b_0 + \sum_{n=1}^{\infty} t_n t'_{n-1} b_n$$

where $(b_1, \ldots)$ is a sequence of positive numbers and $b_0 \geq \beta_*$.[5] The parameters of a type $t = (1, t_1, \ldots)$ have a simple interpretation: the parameter t_1 measures the minimal weight β for all opponent types. Note that for type $t = (1, t_1, \ldots)$ the preference β satisfies $\beta \geq t_1 + b_0$ irrespective of the type of the opponent. Hence, t_1 measures the *kindness* of a type. The parameter t_2 measures how a type responds to the kindness of the opponents. The preference β for an opponent with kindness t_1' satisfies $\beta \geq t_1 + t_2 \cdot t_1' + b_0$. Therefore, t_2 measures the types response to kindness or his *first order reciprocity*. The parameter t_3 represents the agent's *second order reciprocity* which measures the type's response to the opponent's first order reciprocity. More generally, the parameter t_{n+1} measures n-th order reciprocity which is the response to the opponent's parameter t_n'.

The type $\bar{t} = (1, \bar{t}_1, \ldots)$ is the kindest type, that is, type $\bar{t}$ puts more weight on the opponent's payoff than any other type (irrespective of the opponent's type). Furthermore, type $\bar{t}$ reciprocates more than any other type. By contrast, the type $\underline{t} = (1, 0, 0, \ldots)$ is the least kind or meanest type. Type $\underline{t}$ puts the same (minimal) weight on the utility of his opponent irrespective of the opponent's type. All other types are in between those two extreme types. Note that if $\bar{t}_n = 0$ then the $n-1$st order reciprocity of all types is zero. In that case our assumption on $\bar{t}$ implies that $\bar{t}_{n'} = 0$ for all $n' > n$ and hence reciprocity of all higher orders are zero.[6]

In this section, we assume that all types have the same characteristic. We therefore omit ω from the description of an IPM. For the remainder of this section, an IPM is a pair (T, γ). Next, we define IPM's that have the representation described above.

Definition: *The IPM (T, γ) has a reciprocity representation if there are sequences $b = (b_0, b_1, \ldots)$ and $\bar{t} = (1, \bar{t}_1, \ldots)$ with $t_n, b_n \geq 0$ for all $n \geq 1$ and $t_n = 0 \Rightarrow t_{n+1} = 0$ such that*

(i) $T = \{t \in I\!R^\infty | t_0 = 1, t_n \in [0, \bar{t}_n]\}$

(ii) $\gamma(t, t') = b_0 + \sum_{n=1}^\infty t_n t'_{n-1} b_n$

(iii) $\beta_* \leq b_0 \leq b_0 + \sum_{n=1}^\infty \bar{t}_n \bar{t}_{n-1} b_n \leq \beta^*$.

[5] The sequence $(b_1, \ldots)$ satisfies $b_0 + \sum_{n=1}^\infty \bar{t}_n \bar{t}_{n-1} b_n \leq \beta^*, b_0 \geq \beta_*$

[6] This follows from the assumption that $\bar{t}_n = 0$ implies $\bar{t}_{n+1} = 0$.

Recall that $\gamma(t,s)$ is the weight type t places on the payoff of his opponent if the opponent is type s. If type t places a larger weight on the opponent than type t' irrespective of the opponent's type then we say that t is kinder than t'. Hence, $t \in T$ kinder than t' if

$$\gamma(t,s) \geq \gamma(t',s) \text{ for all } s \in T$$

. We denote the "kinder than" relationship with $\succeq$ and note that $\succeq$ is transitive. If (T,γ) has a reciprocity representation then $t = (1, t_1, \ldots)$ is kinder than $t' = (1, t'_1, \ldots)$ if and only if $t_n \geq t'_n$ for all $n \geq 1$.

To obtain the a reciprocity representation, we make the following assumptions.

(i) (*Reciprocity*) All types reciprocate, i.e., all types reward kinder behavior by others.

(ii) (*Linearity*) Types respond to other types in a linear fashion. Suppose a type t behaves like a convex combination of the types t' and t'' . In that case, all types will respond to t by the corresponding convex combination of their responses to t' and t''. In a richer model with uncertainty over types we can interpret this assumption as risk neutrality with respect to the opponent's type.

(iii) (*Kindest and meanest type*) There is a kindest type $\bar{t}$ and a meanest type $\underline{t}$. The behavior of $\underline{t}$ does not depend on the opponent's type. The type $\bar{t}$ offers the highest reward for kindness of other types.

(iv) (*Minimal Richness*) The type space contains every linear and reciprocating type that is "in between" the kindest and meanest type. This assumption implies a minimal richness of the type space.

(v) (*Min-Validity*) Recall that validity of an IPM requires that there be no non-trivial partition of the type space with the property that the types in each partition element have the same set of preferences for each partition element. As we argue in section 2, validity can be interpreted as saying that types can be identified from behavior. Min-validity is a stronger version of this assumption. It requires that there be no non-trivial partition with the property that the types in each partition element have the same minimal β for each partition element. Hence, we can identify types if we observe the *lower bound* of

the preferences (i.e., the lower bound on β) assigned to each opponent of unknown type instead of the *set of preferences* assigned to each opponent (of unknown type).

Theorem 3 below shows that when an IPM satisfies the assumptions described above ((i)-(v)) then this IPM has a reciprocity representation. To give a formal statement of our assumptions we require the following definitions.

Suppose that for $s, s' \in T$ and $\lambda \in [0,1]$ there is $t \in T$ with $\gamma(t,\cdot) = \lambda\gamma(s,\cdot) + (1-\lambda)\gamma(s',\cdot)$. In that case, type t's response to the opponent's type is a convex combination of the responses of s and s'. For a *linear type* the response to t must be a linear combination of the responses to s and s'. Therefore, if $\hat{t}$ is a linear type then $\gamma(\hat{t},t) = \lambda\gamma(\hat{t},s) + (1-\lambda)\gamma(\hat{t},s')$. More generally, we say that a function $f : T \to \mathcal{R}^o$ is *linear* if

$$\gamma(t,\cdot) = \lambda\gamma(s,\cdot) + (1-\lambda)\gamma(s',\cdot) \text{ implies}$$
$$f(t) = \lambda f(s) + (1-\lambda)f(s')$$

Hence, type t is linear if $\gamma(t,\cdot)$ is a linear function.

Suppose that $s \succeq s'$, that is, type s' is kinder than type s'. If type t is reciprocating then he must put more weight on the payoff of an opponent type s than on the payoff of an opponent with type s', i.e., $\gamma(t,s) \geq \gamma(t,s')$. More generally, we say that a function $f : T \to \mathcal{R}^o$ is reciprocating if

$$s \succeq s' \text{ implies } f(s) \geq f(s')$$

Hence, type t is reciprocating if $\gamma(t,\cdot)$ is a reciprocating function.

Next, we formally state the assumptions on $M = (T,\gamma)$ that identify a reciprocity model.

Definition: *The IPM (T,γ) is a reciprocity model if*

(i) There exists $\bar{t}, \underline{t} \in T$ such that $\gamma(\bar{t},\cdot) \geq \gamma(t,\cdot) \geq \gamma(\underline{t},\cdot)$ for all t. Moreover, $\gamma(\bar{t},t') - \gamma(\bar{t},\underline{t}) \geq \gamma(t,t') - \gamma(t,\underline{t})$ and $\gamma(\underline{t},\cdot)$ is constant.

(ii) All types are linear and reciprocating.

(iii) For any linear and reciprocating function $f : T \to \mathcal{R}^o$ with $\gamma(\bar{t},\cdot) \geq f \geq \gamma(\underline{t},\cdot)$ and $\gamma(\bar{t},t') - \gamma(\bar{t},\underline{t}) \geq f(t') - f(\underline{t})$ there is a type t with $\gamma(t,\cdot) = f$.

(iv) There does not exist a non-trivial decomposition $\mathcal{D}$ of T such that each $D \in \mathcal{D}$ is closed and $t, t' \in D$ implies and $\min \gamma_{\hat{t} \in D}(t, \hat{t}) = \min_{\hat{t} \in D} \gamma(t', \hat{t})$ for all $D \in \mathcal{D}$.

Recall that the models $M = (T, \gamma)$ and $M' = (T', \gamma')$ are isomorphic if there is a homeomorphism $\iota : T \to T'$ such that $\gamma(t, t') = \gamma'(\iota(t), \iota(t'))$.

Theorem 3: *The IPM M is a simple reciprocity model if and only if M is isomorphic to an IPM that has a simple reciprocity representation.*

Theorem 3 says that a simple reciprocity representation obtains if and only if the assumptions (i)-(iv) in the definition of a reciprocity model are satisfied.

Next, we provide a simple example of a reciprocity model that addresses the findings from ultimatum bargaining games.

Example 4: This example considers an ultimatum bargaining game with interdependent types. We assume that the IPM has a reciprocity representation. We will assume that players are uncertain about the opponent's type. The support of the player's beliefs consists of two types:

$$t = (1, 0, \ldots), t' = (0, {}^{19}\!/_{26}, 0 \ldots)$$

Hence, type t is more generous than type t'. On the other hand, type t' reciprocates but type t does not. Let

$$u(x) = x^{.87}$$

and

$$\gamma(t, t') = t_1 + t_2 t'_1 - 1/2$$

The game is a standard ultimatum bargaining game with private information regarding the other player's type. Player 1 must choose $x \in [0, 1]$ and player 2 must either accept of reject this offer. If x is accepted, player 1 receives the material reward x and player 2 receives $1 - x$. The prior probability of type t is $3/4$ and types are distributed independently.

This game has a unique equilibrium. A type t player 1 chooses $x = 1/2$ while type t' player 1 chooses $x \approx 1/5$. A type t player 2 accepts all offers while a type t' player 2 accepts $1/2$ and rejects $1/5$.

Now consider the following variation of the game. Player 1's offer is determined by a roll of the dice (either 1/2 or 1/5). In that case, player 2 accepts both offers irrespective of type.

Example 4 captures three key features found in experiments. (See Blount (1995)). First, players choose offers x that are significantly below $x = 1$. In fact, many players choose $x = 1/2$. Second, some player 2 types reject offers even if they entail positive rewards for themselves. Third, the response of player 2 changes if the offer comes from a randomization device. In particular, player 2 is less inclined to reject an offer from a randomization device than an offer that is chosen by player 1. In our model, the reason for this is clear: if the offer is chosen by player 1 then player 2 can infer the type of player 1. This leads player 2 to reject the low offer. On the other hand, if the randomization device chooses the offer then player 2 cannot make any inference about the type. As a result he will not reject the offer.

5. Relation to the Literature on Psychological Games

Geanakoplos, Pearce and Stacchetti (1989) (henceforth GPS) introduce psychological games to capture phenomena related to interdependent preferences. To illustrate their approach, we consider the "bravery game" described in GPS .

The Bravery Game as a Psychological Game: There are two players but only player 1 chooses an action.[7] Player 2 has beliefs about the behavior of player 1 and these beliefs affect his payoff. The payoff of player 1 depends on his beliefs about the beliefs of player 2. The b^* column of the bimatrix below describes payoffs to the two players conditional on player 2 believing (and player 1 knowing that 2 believes) that player 1 will be bold, while the t^* column describes payoffs conditional on player 2 believing that player 1 will be timid.

	t^*	b^*
t	3, 1	0, 0
b	2, 2	1, 4

[7] In the GPS treatment there is a collection of agents in the role of player 2. For simplicity, we assume that there is only a single player 2.

Psychological Game

As noted by GPS, this game has two pure strategy equilibria and one mixed equilibrium. One pure strategy equilibrium is for player 1 to choose the bold action. Given that player 1 is choosing b, in equilibrium, player 2 assigns probability 1 to b, and hence b^* is the relevant column. Conditional on b^*, it is indeed optimal for player 2 to choose b. The other pure strategy equilibrium is for player 1 to choose the timid action. Again, given t^*, it is optimal for 1 to choose t. Finally, note that the only mixture between b^* and t^* that makes player 1 indifferent between b and t is the fifty-fifty mixture. Hence, the only mix strategy equilibrium entails player 1 choosing t^* with probability .5 and b^* with probability .5.

In a psychological game a player's payoff depends directly on the player's beliefs. As GPS (1989, pg. 61) put it, *"a player's payoffs depend not only on what everybody does but also on what everybody thinks."* Of course, even in standard game theory, a player's payoff depends on his beliefs since these beliefs are used to compute his expected utility. But it is clear that GPS mean something more than this when they refer to the direct dependence of payoffs on beliefs: *"Hence, we argue in many cases the psychological payoffs associated with a terminal node are endogenous, in the same sense as equilibrium strategies are."* Thus, their view of expected utility theory is not one based on the revealed preferences over lotteries but rather on taking expectations of *psychological payoffs*. Indeed, the payoffs in the matrix above are not observable payoffs but psychological payoffs parameterized by the player's beliefs in that game.

The discussion in GPS reveals three distinct roles for beliefs in psychological games: First, beliefs have the standard interpretation and describe a player's predictions of the opponent's play (or of the opponent's beliefs in the case of higher order beliefs). Second, beliefs play a role similar to that played by types in our model. Note that in the game above, both player 1 and player 2 have different preferences over player 2's actions depending on player 2's beliefs (or player 1 beliefs about player 2's beliefs).[8] Finally, the dependence

[8] GPS often use beliefs as proxies for more permanent personality attributes. They describe the motives of player 1 in the bravery game as follows: "His payoff depends not only what he does but also on what he thinks his friends think of his character (that is, on what he thinks they think he will do)." It is easier to see that player 1 might be happier if he thinks that his friends believe that he will do the right thing; the argument for why player 1 would be less inclined to do the right think otherwise is much less clear. This problem does not arise if we model player 1's character with personality types; one that cares about what others think, and one that does not.

on beliefs may be a shortcut for describing the payoff consequences of the response of an opponent. For example, player 1 may be concerned about an unfavorable response by player 2 if he does not meet player 2's expectations. In a standard game, this would be captured by allowing player 2 to choose a punishment or a reward after player 1 has made his choice. In the GPS interpretation, player 1 may internalize the potential punishment even if player 2 has no opportunity to carry out the punishment and even if player 1 never gets to verify his assessment of player 2's beliefs.

Next, we show that the model of interdependent types can capture the phenomena that psychological games try to address. We illustrate two interpretations of the bravery game with interdependent types below. In the first interpretation player 2 cares about player 1's action and player 1 may care about player 2's welfare. In the second interpretation, player 2 cares about the type of player 1 and player 1 tries to signal his type.

The Bravery Game with Interdependent Types I: Player 1 must choose between a timid (t) and a bold (b) action. Player 2 prefers player 1 to be bold. Player 1 is either an "altruistic" type (a), or a "selfish" type (s). The altruistic type prefers b over t in order to make player 2 happy while the selfish prefers t to b. Payoffs are described by the table below.

	s	a
t	3, 1	0, 0
b	2, 2	1, 4

Game with interdependent preferences

Assume that the prior probability of player 2 being type a is α and the prior probability of player 2 being type s is $1-\alpha$. Then, if player 1 does not wish to disappoint player 2 (which happens with probability α) he chooses b otherwise he chooses t. Note that "disappoint" here does not mean "act contrary to prediction" but rather "make unhappy." Hence, by choosing the appropriate distribution of types we can replicate any of the equilibrium outcomes of the psychological game above.

The Bravery Game with Interdependent Preferences II: Player 1 must choose between a bold and a timid action. Observing player 1's action, 2 chooses a reward (r) or a

punishment (p). Player 1 may be strong or weak. The probability of a weak type is $\frac{1}{3}$. For the weak type, choosing the timid action is a dominant strategy. Therefore, we will ignore the strategy of the weak type. Player 1's payoffs are

	r	p
t	2	0
b	1	-1

Player 1's payoffs

Player 2 would like to reward the strong type and punish the weak type. Player 2's payoff is

	s	w
r	2	-1
p	-2	1

Player 2's payoffs

It is easy to verify that there are three equilibrium outcomes. One equilibrium is for the strong type of player 1 to choose b and for player 2 to choose r if and only if 1 chooses b. There is also a class of equilibria (all leading to the same outcome) where the strong type chooses t and player 2 chooses r whenever player 1 chooses t. Finally, there is an equilibrium where the strong type mixes 50-50; player 2 rewards player 1 if player 1 chooses bold and mixes 50-50 if 1 chooses timid. Note that as in GPS – given the set of equilibrium responses for player 2 – it is indeed the case that a strong type of player 1 wants to be bold if player 2 expects him to be bold and wants to be timid if he is expected to be timid.

Like the model of psychological games, our model also enlarges the set of payoff relevant parameters. Payoffs in our model depend on "what kind of person" the opponent is as well as the profile of actions. We introduce interdependent preference types to capture the effect of the opponent's character. Note, however, that a player's interdependent preference type can be identified *independently of the particular game*. The types in our model can be inferred from a player's behavior in other contexts. As we have argued in section 2, valid interdependent types can be identified from observations much like the

parameters describing an agent's risk aversion. In contrast, the payoff functions in GPS are *game specific* and depend on an *unobservable parameter* – the player's beliefs in that game. Therefore, even repeated observations of play will not reveal the player's payoff function because beliefs are unobservable. Note that equilibrium analysis and, more generally, the analysis of incentives requires that we are able to identify the player's payoffs.

The bravery game illustrates that the direct dependence of payoffs on beliefs is unnecessary. The next example illustrates that parameterizing payoffs by beliefs may not be sufficient to capture the consequences of interdependent preferences. In particular, the new roles of beliefs in psychological games may interacts with their standard role as predictions. Therefore, it may be difficult to separate the different roles of beliefs, as illustrated in the following 2-person game:

	A	B
a	9, 8	5, 7
b	5, 5	9, 6

Table 3

The numbers in table 3 above denote monetary payoffs to the players. The game in table 3 is strategically equivalent to a penny matching game. However, a decision by player 1 to choose b is unambiguously bad for player 2. If player 1 enjoys reducing the reward to player 2, then he may choose b over a. Alternatively, player 1 may choose b over a even if he has a more benevolent attitude towards 2 simply because he predicts that 2 will choose B and prefers increasing his own reward by 4 to increasing player 2's reward by 1. Suppose we are trying to capture the following situation: Player 2 prefers to reduce his own monetary reward from 6 to 5 in order to reduce his opponent's reward from 9 to 5 if the opponent chooses b because he benefits from reducing the payoff of other players. However, player 2 prefers the $(9, 6)$ outcome if the opponent chooses b simply because he expects B. It is difficult to see how this situation can be captured by belief-dependent preferences, especially within the context of an equilibrium theory.

In contrast, our model can capture this situation by endowing player 1 with two types t, t'. The utility function of type t is spiteful, i.e., decreasing in the reward to player 2

while the utility function type t' is altruistic, i.e., increasing in the reward to player 2. Player 2's utility function is decreasing in the reward to player 1 if player 1 is spiteful and increasing if player 1 is altruistic. In our approach, a type's utility functions represents a stable *personality*. A personality can be determined by observing preferences in contexts where strategic issues can be disentangled from considerations of altruism and reciprocity.

6. Appendix

Let Z be a compact metric space. For any sequence $A_n \in \mathcal{H}_Z$, let

$\underline{\lim} A_n = \{z \in Z \,|\, z = \lim z_n \text{ for some sequence } z_n \text{ such that } z_n \in A_n \text{ for all } n\}$

$\overline{\lim} A_n = \{z \in Z \,|\, z = \lim z_{n_j} \text{ for some sequence } z_{n_j} \text{ such that } z_{n_j} \in A_{n_j} \text{ for all } j\}$

Let X be a metric space and $p : X \to \mathcal{H}_Z$. We say that p is Hausdorff continuous if it is a continuous mapping from the metric space X to the metric space $\mathcal{H}_Z$. Note that if p is a Hausdorff continuous mapping from X to $\mathcal{H}_Z$ then $p \in \mathcal{C}(X, \mathcal{H}_Z)$. However, the converse is not true.

Lemma 1: *Let X, Y, Y', Z be nonempty compact metric spaces, $q \in \mathcal{C}(X \times Y, Z)$, $p \in \mathcal{C}(Y', \mathcal{H}_Y)$, and $r \in \mathcal{C}(Y', Y)$. Then, (i) $A_n \in \mathcal{H}_Z$ converges to A (in the Hausdorff topology) if and only if $\underline{\lim} A_n = \overline{\lim} A_n = A$. (ii) $x_n \in X$ converges to x implies $q(x_n, B)$ converges to $q(x, B)$ for all $B \in \mathcal{H}_Y$. (iii) If $q^*(x, y') = q(x, p(y'))$ for all $x \in X, y' \in Y'$ then $q^* \in \mathcal{C}(X \times Y', \mathcal{H}_Z)$. (iv) If r is onto, then $r^{-1} \in \mathcal{C}(Y, \mathcal{H}_{Y'})$.*

Proof: Part (i) is a standard result. See Brown and Pearcy (1995).

(ii) Suppose $x_n \in X$ converges to x. Let $z_{n_j} \in q(x_{n_j}, B)$ such that $\lim z_{n_j} = z$. Hence, $z_{n_j} = q(x_{n_j}, y_{n_j})$ for some $y_{n_j} \in B$. Since B is compact, we can without loss of generality assume y_{n_j} converges to some $y \in B$. Hence, the continuity of q ensures $z = q(x, y)$ and therefore $z \in q(x, B)$ proving that $\overline{\lim} q(x_n, B) \subset q(x, B)$. If $z \in q(x, B)$, then there exists $y \in B$ such that $z = q(x, y)$. Since q is continuous, we have $z = \lim q(x_n, y)$. Hence, $q(x, B) \subset \underline{\lim} q(x_n, B)$. Since, $\underline{\lim} q(x_n, B) \subset \overline{\lim} q(x_n, B) \subset q(x, B)$, we conclude $\underline{\lim} q(x_n, B) = q(x_n, B) = \overline{\lim} q(x_n, B)$ as desired.

(iii) Suppose (x_n, y'_n) converges to (x, y) and $z_n \in q^*(x_n, y'_n)$ converges to z. Pick $y_n \in p(y'_n)$ such that $q(x_n, y_n) = z_n$. Since Y is compact, we can assume that y_n converges

to some y. Since $p \in \mathcal{C}(Y', \mathcal{H}_Y)$, we conclude that $y \in p(y')$ and since q is continuous, $q(x,y) = z$. Therefore, $z \in q^*(x,y')$, proving that $q^* \in \mathcal{C}(X \times Y, \mathcal{H}_Z)$.

(iv) The continuity and ontoness of r ensures that r^{-1} maps Y into $h_{Y'}$. Assume that y_n converges to y, $y'_n \in r^{-1}(y_n)$ and y'_n converges to y'. Then, $r(y'_n) = y_n$ for all n and by continuity $r(y') = y$. Therefore, $y' \in r^{-1}(y)$ as desired. □

Lemma 2: *Let X and Z be compact metric spaces. Suppose $p_n \in \mathcal{C}(X, \mathcal{H}_Z)$ and $p_n(x) \subset p_{n+1}(x)$ for all $n \geq 1$, $x \in X$. Let $p(x) := \bigcap_{n \geq 1} p_n(x)$ and assume $p(x)$ is a singleton for all $x \in X$. Then, (i) p is continuous and (ii) p_n converges to p.*

Proof: Obviously, $\bigcap_{n \geq 1} G(p_n) = G(p)$. Since $p_n \in \mathcal{C}(X, \mathcal{H}_Z)$ and X, Z are compact, so is $G(p_n)$. Therefore $G(p)$ is compact (and therefore closed) as well. Since p is a function and both X, Z are compact, the fact that p has a closed graph implies that p is continuous.

To prove (ii), it is enough to show that if G_n is a sequence of compact sets such that $G_{n+1} \subset G_n$ then G_n converges (in the Hausdorff topology) to $G := \bigcap_n G_n$. If not, since G_1 is compact, we could find $\epsilon > 0$ and $y_n \in G_n$ converging to some $y \in G_1$ such that $d(y_n, G) > \epsilon$ for all n. Hence, $d(y, G) \geq \epsilon$ and therefore there exists K such that $y \notin G_K$ for all $n \geq K$. Choose $\epsilon' > 0$ such that $\min_{y' \in G_K} d(y', y) \geq \epsilon'$ and K' such that $n \geq K'$ implies $d(y_n, y) < \epsilon'/2$. Then, for $n \geq \max\{K, K'\}$ we have $d(y_n, y) \geq \epsilon'$ and $d(y_n, y) < \epsilon'/2$, a contradiction. □

We say that $p_n \in \mathcal{C}(X, \mathcal{H}_Z)$ converges to $p \in \mathcal{C}(X, \mathcal{H}_Z)$ uniformly if for all $\epsilon > 0$, there exists N such that $n \geq N$ implies $d(p_n(x), p(x)) < \epsilon$.

Let X be an arbitrary set and Z be a compact metric space. Given any two functions p, q that map X into $\mathcal{H}_Z$, let $d^*(p,q) = \sup_{x \in X} d(p(x), q(x))$ where d is the Hausdorff metric on $\mathcal{H}_Z$.

Lemma 3: *(i) If $p_n \in \mathcal{C}(X, \mathcal{H}_Z)$ converges to $p \in \mathcal{C}(X, Z)$, then p_n converges to p uniformly; that is, $\lim_n d^*(p_n, p) = 0$. (ii) The relative topology of $\mathcal{C}(X, Z) \subset \mathcal{C}(X, \mathcal{H}_Z)$ is the topology of uniform convergence.*

Proof: Let $\lim p_n = p \in \mathcal{C}(X, Z)$. Then, p is continuous and since X is compact, it is uniformly continuous. For $\epsilon > 0$ choose a strictly positive $\epsilon' < \epsilon$ such that $d(x, x') < \epsilon'$

implies $d(p(x), p(x')) < \epsilon$. Then, choose N so that $d_H(G(p), G(p_n)) < \epsilon'$ for all $n \geq N$. Hence, for $n \geq N$, $x \in X$ and $z \in p_n(x)$, we have $x' \in X$ such that $d(x, x') < \epsilon'$ and $d(p(x'), z) < \epsilon'$. Hence, $d(p(x), z) \leq d(p(x'), z) + d(p(x'), p(x)) < 2\epsilon$ as desired.

Next, we will show that p_n converges to p uniformly implies $G(p_n)$ converges to $G(p)$ in the Hausdorff metric. This, together with (i) will imply (ii). Consider any sequence p_n converging uniformly to p. Choose N such that $n \geq N$ implies $d(p_n(x), p(x)) \leq \epsilon$. Hence, for $n \geq N$, $(x, z) \in G(p_n)$ implies $d((x, z), (x, p(x)) < \epsilon$, proving $\overline{\lim} G(p_n) \subset G(p) \subset \underline{\lim} G(p_n)$. □

For $\theta_n \in \Theta_n$ and $n \geq 0$, let

$$\Theta(\theta_n) = \{\theta' \in \Theta \,|\, \theta'(n) = \theta_n\}$$

Lemma 4: *Let $\hat{\theta} \in \Theta \in \mathcal{I}$ with $\hat{\theta} = (f_0, f_1, \ldots)$ and $\phi(\hat{\theta}) = (f_0, f)$. Then, for all $n \geq 1$ and $\theta_{n-1} \in \Theta_{n-1}$, $\bigcup_{\theta \in \Theta(\theta_{n-1})} f(\theta) = f_n(\theta_{n-1})$.*

Proof: Let $P \in f_n(\theta_{n-1})$. Since the sequence $\{\Theta_n\}$ is consistent, we may choose $\theta_n \in \Theta_n(\theta_{n-1})$ so that $P \in f_{n+1}(\theta_n)$. Repeat the argument for every $k > n$ to obtain $\theta = (\theta_{n-1}, g_n, g_{n+1}, \ldots) \in \Theta$ such that $\phi(\hat{\theta})(\theta) = P$. Hence, $f_n(\theta) \subset \bigcup_{\theta \in \Theta(\theta_{n-1})} f(\theta) = f_n(\theta_{n-1})$. That $\bigcup_{\theta \in \Theta(\theta_{n-1})} f(\theta) \subset f_n(\theta_{n-1})$ follows from the definition of f and the fact that $f_{n+1}(\theta) \subset f_n(\theta)$ for all n and all $\theta \in \Theta$. □

Lemma 5: *Let X, Y be compact metric spaces and Z be an arbitrary metric space. Let $q : X \times Y \to Z$ and let the mapping p from X to the set of functions from Y to Z be defined as $p(x)(y) := q(x, y)$. Then, $q \in \mathcal{C}(X \times Y, Z)$ if and only if $p \in \mathcal{C}(X, \mathcal{C}(Y, Z))$.*

Proof: Assume q is continuous. Since $X \times Y$ is compact, q must be uniformly continuous. Hence, for all $\epsilon > 0$ there exists $\epsilon' > 0$ such that $d((x, y), (x', y')) < \epsilon'$ implies $d(q(x, y), q(x', y')) < \epsilon$. In particular, $d(x, x') < \epsilon'$ implies $d(q(x, y), q(x', y)) < \epsilon$ for all $y \in Y$. Hence, $d(x, x') < \epsilon'$ implies $d(p(x), p(x')) < \epsilon$, establishing the continuity of p. Next, assume that p is continuous and let $\epsilon > 0$. To prove that q is continuous, assume $(x^k, y^k) \in X \times Y$ converges to some $(x, y) \in X \times Y$. The continuity of p ensures that for

some $K \in I\!N$, $k \geq K$ implies $d(p(x^k), p(x)) \leq \epsilon$. Since $p(x)$ is continuous, we can choose K so that $d(p(x)(y^k), p(x)(y)) < \epsilon$ for all $k \geq K$ as well. Hence,

$$d(p(x^k)(y^k), p(x)(y)) \leq d(p(x^k)(y^k), p(x)(y^k)) + d(p(x)(y^k), p(x)(y)) < 2\epsilon$$

□

Lemma 6: *Let X be compact and Z be an arbitrary metric space. Suppose $p \in \mathcal{C}(X, Z)$ is one-to-one. Then, p is a homeomorphism from X to $p(X)$.*

Proof: It is enough to show that $p^{-1} : p(X) \to X$ is continuous. Take any closed $B \subset X$. Since X is compact, so is B. Then, $(p^{-1})^{-1}(B) = p(B)$ is compact (and therefore closed) since the continuous image of a compact set is compact. Hence, the inverse image of any closed set under p^{-1} is closed and therefore p^{-1} is continuous. □

Lemma 7: *Let X, Y be a compact metric spaces. For $p \in \mathcal{C}(X, \mathcal{H}_Y)$, let $\bar{d}(p) = \max_{x \in X} \max_{y,z \in p(x)} d(y, z)$. Then, (i) $p_n \in \mathcal{C}(X, \mathcal{H}_Y)$ converges to $p \in \mathcal{C}(X, \mathcal{H}_Y)$ implies $\limsup \bar{d}(p_n) \leq \bar{d}(p)$. (ii) $p, q, p', q' \in \mathcal{C}(X, \mathcal{H}_Y)$ and $p(x) \subset p'(x), q(x) \subset q'(x)$ for all $x \in X$ implies $d(p, q) \leq \max\{d(p', q') + \bar{d}(p'), d(p', q') + \bar{d}(q')$.*

Proof: Since $X \times Y$ is compact (i) is equivalent to the following: $p_n \in \mathcal{C}(X, \mathcal{H}_Y)$ converges to $p \in \mathcal{C}(X, \mathcal{H}_Y)$, $\lim \bar{d}(p_n) = \alpha$ implies $\alpha \leq \bar{d}(p)$. To prove this, choose $x_n \in X$ and $y_n, z_n \in p_n(x_n)$ such that $d(y_n, z_n) = \bar{p}_n$. Without loss of generality, assume (x_n, y_n, z_n) converges to (x, y, z). Since p_n converges to p, for all $\epsilon > 0$, there exists N such that for all $n \geq N$, there exists (x'_n, y'_n) and $(\hat{x}_n, \hat{z}_n)$ such that $d((x'_n, y'_n), (x_n, y_n)) < \epsilon$ and $d((\hat{x}_n, \hat{z}_n), (x_n, z_n)) < \epsilon$. Hence, we can construct a subsequence n_j such that $x'_{n_j}, \hat{x}_{n_j}$ both converge to x, y'_{n_j} converges to y, $\hat{z}_{n_j}$ converges to z, and $y'_{n_j} \in p(x'_{n_j}), \hat{z}_{n_j} \in p(\hat{x}_{n_j})$ for all n_j. Since $p \in \mathcal{C}(X, \mathcal{H}_Y)$ we conclude $y, z \in p(x)$. But $\alpha = \lim \bar{p}_n = \lim d(y_n, z_n) = d(y, z)$. Hence, $\alpha \leq \bar{d}(p)$.

(ii) Let $(x, z) \in G(p), (\hat{x}, \hat{z}) \in G(p')$. Then,

$$d((x, z), (\hat{x}, \hat{z})) \leq \min_{(\hat{x}, \hat{y}) \in G(q')} d((x, z), (\hat{x}, \hat{y})) + \bar{d}(p')$$

Therefore,

$$\min_{(\hat{x},\hat{z})\in G(q)} d((x,z),(\hat{x},\hat{z})) \leq d(p',q') + \bar{d}(q')$$

and a symmetric argument shows that

$$\min_{(x,z)\in G(p)} d((x,z),(\hat{x},\hat{z})) \leq d(p',q') + \bar{d}(p')$$

Therefore, $d(p,q) \leq \max\{d(p',q') + \bar{d}(p'), d(p',q') + \bar{d}(q')\}$. □

6.1 Proof of Theorem 1:

We first show that ϕ is continuous. Consider any sequence $\theta^k = (f_0^k, f_1^k, \ldots) \in \Theta$ such that $\lim \theta^k = \theta = (f_0, f_1 \ldots) \in \Theta$. Let $\phi(\theta) = (f_0, f)$ and $\phi(\theta^k) = (f_0^k, f^k)$ for all k. Let $\theta^k = (f_0^k, f_1^k, \ldots)$, $\theta = (f_0, f_1, \ldots)$ and $\epsilon > 0$. By Lemma 2 f_n converges to f and therefore by Lemma 7(i) there exists N such that $\bar{d}(f_N) < \epsilon$. Since $f_N^k \to f_N$ Lemma 7(i) implies that there exists K' such that for $k \geq K'$, $\bar{d}(f_N^k) \leq 2\epsilon$. Finally, there is K'' such that $d(f_N^k, f_N) \leq \epsilon$ for $k > K''$. Let $K = \max\{K', K''\}$. Lemma 7(ii) now implies that $d(f_n^k, f_n)) \leq 3\epsilon$, for all $n \geq N$ and $k \geq K$. Therefore $d(f^k, f) \leq 3\epsilon$ for all $k \geq K$. This shows that ϕ is continuous.

Next, we prove that ϕ is one-to-one. Pick any $(f_0, f_1, \ldots), (g_0, g_1, \ldots) \in \Theta$. Let $(f_0, f) = \phi(f_0, f_1, \ldots)$ and $(g_0, g) = \phi(g_0, \ldots)$. If $f_0 \neq g_0$, then clearly $(f_0, f) \neq (g_0, g)$. Hence, assume $f_0 = g_0$. Then, there exists a smallest $n \geq 1$ and $\theta_{n-1} \in \Theta_{n-1}$ such that $g_n(\theta_{n-1}) \neq f_n(\theta_{n-1})$. By Lemma 4, $\bigcup_{\theta' \in \Theta(\theta_{n-1})} f(\theta') \neq \bigcup_{\theta' \in \Theta(\theta_{n-1})} g(\theta')$ and hence $f \neq g$ as desired.

Since ϕ is continuous and one-to-one and Θ is compact, it follows from Lemma 6 that ϕ is a homeomorphism from Θ to $\phi(\Theta)$. The continuity of ψ follows from the compactness of Θ and Lemma 5. □

6.2 Proof of Theorem 2:

We say that $\mathcal{D}$ is strongly continuous if the function $\sigma : X \to \mathcal{D}$ defined by $\sigma(x) = D^x$ is an element of $\mathcal{C}(X, \mathcal{H}_X)$.

Let $M = (T, \gamma, \omega)$ be an IPM. Define the sequence of decompositions $\mathcal{D}_n$ on T as follows:

$$D_0^t = \{t' \in T | \omega(t') = \omega(t)\}$$

and $\mathcal{D}_0 = \{D_0^t | t \in T\}$. For $n \geq 1$ we define inductively

$$D_n^t := \{t' \in D_{n-1}^t | \gamma(t', D) = \gamma(t, D) \text{ for all } D \in \mathcal{D}_{n-1}\}$$

and $\mathcal{D}_n = \{D_n^t | t \in T\}$. Let $\mathcal{D} = \left\{ \bigcap_n D_n^t \,\middle|\, t \in T \right\}$ and note that $\mathcal{D}$ is a decomposition of T.

Step 1: Each $\mathcal{D}_n$ is continuous.

The proof is by induction. Assume that t_k converges to t, $\hat{t}_k \in D_0^{t_k}$ and $\hat{t}_k$ converges to $\hat{t}$. Then, $\omega(\hat{t}) = \lim \omega(\hat{t}_k) = \lim \omega(t_k) = \omega(t)$. Hence, $\hat{t} \in D_0^t$, proving the strong continuity of $\mathcal{D}_0$. Assume that $\mathcal{D}_n$ is satisfies strong continuity. Hence, every $D \in \mathcal{D}_n$ is compact. Assume that t_k converges to t, $\hat{t}_k \in D_{n+1}^{t_k}$ and $\hat{t}_k$ converges to $\hat{t}$. Hence, $\hat{t}_k \in D_n^{t_k}$ and by the strong continuity of $\mathcal{D}_n$, we have $\hat{t} \in D_n^t$. Pick any $D \in \mathcal{D}_n$ and $P \in \gamma(\hat{t}, D)$. By, Lemma 1(ii), we have $P_n \in \gamma(\hat{t}_n, D) = \gamma(t_n, D)$ such that $\lim P_n = P$. Then, by Lemma 1(iii), we have $P \in \gamma(t, D)$, proving that $\gamma(\hat{t}, D) \subset \gamma(t, D)$. A symmetric argument ensured that $\gamma(\hat{t}, D) = \gamma(t, D)$, establishing that $\hat{t} \in D_{n+1}^t$ and proving the strong continuity of $\mathcal{D}_{n+1}$. This concludes the proof of step 1.

Step 2: M is isomorphic to some $\Theta \in \mathcal{I}$ if and only if $\mathcal{D} = \{\{t\} \,|\, t \in T\}$.

Let $\Theta_0 = \Omega = \omega(T)$, $f_0^t = \omega(t)$ for all $t \in T$. Define $\iota_0(t) := f_0^t$ and define inductively $f_n^t : \Theta_{n-1} \to \mathcal{H}, \Theta_n, \iota_n : T \to \Theta_n$ as follows:

$$f_n^t(\theta_{n-1}) = \gamma(t, \iota_{n-1}^{-1}(\theta_{n-1}))$$

$$\iota_n(t) = (\iota_{n-1}(t), f_n^t)$$

$$\Theta_n = \iota_n(T)$$

Finally, let

$$\Theta = \{(f_0, f_1, \ldots) \,|\, (f_0, f_1, \ldots, f_n) \in \Theta_n \text{ for all } n \geq 0\}$$

$$\iota(t) = (f_0, f_1, \ldots) \text{ such that } (f_0, f_1, \ldots, f_n) = \iota_n(t) \text{ for all } n.$$

We will prove inductively that Θ_n are nonempty, compact, ι_n is continuous and onto for every n. Clearly, this statement is true for $n = 0$. Suppose it is true for n. Then, by Lemma 1 parts (iii) and (iv), $\iota_{n+1} \in \mathcal{C}(T, \Theta_{n+1})$ and Θ_{n+1} is compact.

Next, we show that $\iota : T \to \Theta$ is onto. Pick any $(f_0, f_1, \ldots)$ such that $(f_0, f_1, \ldots, f_n) \in \Theta_n$ for all n. Then, for all n, there exists $t_n \in T$ such that $\iota_n(t_n) = (f_0, f_1, \ldots, f_n)$. Take

t_{n_j}, a convergent subsequence of t_n converging to some $t \in T$. For all n and $n_j > n$, $\iota_n(t_{n_j}) = (f_0, f_1, \ldots, f_n)$. Hence, the continuity of ι_n ensures that $\iota_n(t) = (f_0, f_1, \ldots, f_n)$ for all n, establishing that $\iota(T) = \Theta$.

Next, we prove that $\iota_n(t) = \iota_n(s)$ if and only if $D_n^t = D_n^s$. To see this, note that for $n = 0$, the assertion is true by definition. Suppose, it is true for n. Then, if $s \in D_{n+1}^t$, we have $s \in D_n^t$ and $\gamma(t, D_n) = \gamma(s, D)$ for all $D \in \mathcal{D}_n$. Hence, $f_{n+1}^t = f_{n+1}^s$ and therefore, by the inductive hypothesis, $\iota_{n+1}(t) = \iota_{n+1}(s)$. Conversely, if $\iota_{n+1}(t) = \iota_{n+1}(s)$, then $f_{n+1}^t = f_{n+1}^s$ and $i_n^t = f_n^s$. Therefore,by the inductive hypothesis, $s \in D_{n+1}^t \in \mathcal{D}_{n+1}$.

It follows that $\Theta_{n-1}(\theta_{n-2}) = \iota_{n-1}(D_{n-2}^s)$ for s such that $\iota_{n-2}(t) = D_{n-2}^s$. Therefore,

$$f_n^t(\theta_{n-2}) = \gamma(t, D_{n-2}^s) = \bigcup_{s' \in D_{n-2}^s} \gamma(t, D_{n-1}^{s'}) = \bigcup_{\theta'_{n-1} \in \Theta_{n-1}(\theta_{n-2})} f^t(\theta'_{n-1})$$

proving that $\{\Theta_n\}$ satisfies the consistency condition.

Let $\sigma_n : T \to \mathcal{H}$ be defined by $\sigma_n(t) = D_n^t$ for $D_n^t \in \mathcal{D}$ and for $t \in T$, $f^t : \Theta \to \mathcal{H}$ be defined by $f^t(\theta) = \bigcap_{n \geq 1} f_n^t(\theta)$.

Assume that M is $\mathcal{D} = \{\{t\} \,|\, t \in T\}$. Since, $\iota_n(t) = \iota_n(s)$ if and only if $D_n^t = D_n^s$, we conclude that ι is one-to-one. Also, for every $t \in T$, $f_n^t(\theta) = \gamma(t, D_n^s)$ for s such that $\iota(s) = \theta$. Hence, $f^t(\theta) = \gamma(t, s)$. To complete the proof of the validity condition, note that by Lemma 2(ii) f_n^t converges to $\gamma(t, \cdot)$.

Finally, to prove that ι is a homeomorphism, we prove that ι is continuous and appeal to Lemma 6. Let $\Pi_0 = \Omega = \omega(T)$, $\pi_0(t) = \omega(t)$ for all $t \in T$ and define the collection of sets Π_n and functions $\pi_n : T \to \Pi_n$ for $n \geq 1$ as follows:

$$\pi_n(t) = (\pi_{n-1}(t), g_n^t) \text{ such that}$$
$$g_n^t(s) = \gamma(t, D_{n-1}^s) \text{ for all } s \in T$$
$$\Pi_n = \pi_n(T) \text{ for all } t \in T$$

Also, let

$$\pi(t) = (\omega, g_1, \ldots) \text{ such that } (\omega, g_1, \ldots, g_n) = \pi_n(t) \text{ for all } n.$$
$$\Pi = \pi(T)$$

Clearly, π and each π_n are onto. Similarly, repeating the arguments made above to establish the nonempty and compactness of Θ_n and continuity ι_n proves that each Π_n is nonempty

and compact and each π_n is continuous. Henceforth, for any $t \in T$ such that $\pi(t) = (\omega, g_1, \ldots)$ we write g_n^t to denote the corresponding g_n.

Consider t_k converging to t and let $\iota(t_k) = (f_0^k, \ldots)$ for all k and $\iota(t) = (\omega, f_1, \ldots)$. By Lemma 2, $g^k = \lim_n g_n^k$ and $g = \lim g_n$ are well-defined and continuous. Similarly, let $\pi(t_k) = (g_0^k, \ldots)$ for all k and $\pi(t) = (\omega, g_1, \ldots)$. Hence, $\bar{d}(g) = 0$. Therefore by Lemmas 3(i), 7(i) there exists N such that $d(g, g_N) < \epsilon$ and $\bar{d}(g_N) < \epsilon$. By Lemma 7(i) we can choose K so that $\bar{d}(g_N^k) \leq 2\epsilon$ for all $k \geq K$. Therefore, by Lemma 7(ii), $d(g_n^k, g_n)$ for all $k \geq K$, $n > N$. Since each π_n is continuous, we can choose this K so that for all $k \geq K$, $n \leq N$, $d(g_n^k, g^k) < \epsilon$, proving that $d(\pi(t_k), \pi(t)) \leq 2\epsilon$ for all $k \geq K$ and establishing the continuity of π.

Since g_n converges to g, the continuity of π implies that for any two subsequences of natural numbers n_j, k_j both converging to ∞, $g_{k_j}^{n_j}$ converges to g. Recall that d^* is the sup metric. It follows from Lemma 3(i) that $g_{k_j}^{n_j}$ converges to g in the sup metric d^* as well. Hence, for any $\epsilon > 0$, there exits N such that $k \geq N$, $n \geq N$, $d^*(g_n^k, g) < \epsilon$. Since each ι_n is continuous, we can choose $K > N$ large enough so that $d(f_n^k, f_n) < \epsilon$ for all $n \leq N$. Hence,

$$d(f_n^k, f_n) \leq d^*(f_n^k, f_n) = d^*(g_n^k, g_n) \leq d^*(g_n^k, g) + d^*(g, g_n^k) = \epsilon$$

proving the continuity of ι. Note that

$$\psi(\iota(t), \iota(s)) = \bigcap_{n \geq 1} f_n^t(\iota(s)) = \lim \gamma(t, D_n^s) = \gamma(t, s)$$

Hence Θ is isomorphic to M as desired.

Next we will show that if M is isomorphic to some Θ, then the function ι defined above is the isomorphism. Let $\hat{\iota} : T \to \Theta$ be an isomorphism and $\hat{\iota}_n$ denote the $n-$the coordinate function of $\hat{\iota}$. Recall that ι defined above satisfies the property

$$\iota_n(t) = \iota_n(s) \text{ if and only if } D_n^t = D_n^s \qquad (*)$$

Note that this property uniquely identifies the function ι. That is, if $\hat{\iota}$ is any function that also satisfies $(*)$, $\hat{\iota} = \iota$. To see this note that if $\iota\hat{\iota}_0$ satisfies $(*)$ then obviously, $\hat{\iota}_0 = \omega = \iota_0$.

Then, a simple inductive step yields the desired conclusion. To see that $\hat{\iota}$ satisfies $(*)$, note that since it is a isomorphism, we have $\omega = \hat{\iota}_0$ and hence $(*)$ is satisfied for $n = 0$, Suppose it is satisfied for n. Then, suppose $\hat{\iota}_{n+1}(t) = \hat{\iota}_{n+1}$. Since $\hat{\iota}$ is an isomorphism, we conclude $f^t_{n+1} = f^s_{n+1}$. Then, the inductive hypothesis yields $D^t_{n+1} = D^s_{n+1}$. Conversely, suppose $D^t_{n+1} = D^s_{n+1}$. Then, $\gamma(t, D_n) = \gamma(s, D_n)$ for all $D_n \in \mathcal{D}_n$. Since, $\hat{\iota}$ is an isomorphism, we conclude $\psi(\hat{\iota}(t), \hat{\iota}(D_n)) = \psi(\hat{\iota}(t), \hat{\iota}(D_n))$ for all $D_n \in \mathcal{D}_n$. Which, by the inductive hypothesis, yields $\hat{\iota}_{n+1}(t) = \hat{\iota}_{n+1}(s)$.

Suppose $s \in D^t_n \in \mathcal{D}_n$ for all n. Since ι is an isomorphism, we have

$$f^t_n(\iota(D_n)) = \psi(\iota(t), \iota(D_n)) = \gamma(t, D_n) = \gamma(s, D_n) = \psi(\iota(s), \iota(D_n)) = f^s_n(\iota(D_n))$$

for all $n, D_n \in \mathcal{D}_n$. By $(*)$, we have $\iota(t) = \iota(s)$. Since ι is one-to-one, we conclude $s = t$. This concludes the proof of step 2.

Step 3: M is valid if and only if $\mathcal{D} = \{\{t\} \,|\, t \in T\}$.

If $\mathcal{D} \neq \{\{t\} \,|\, t \in T\}$, the $\mathcal{D}$ challenges M and hence M is not valid. Suppose M is not valid and hence there exists a continuous decomposition $\mathcal{D}^*$ that challenges M. Then, $\mathcal{D}^*$ is a refinement of $\mathcal{D}$; that is, $D^*_t \in \mathcal{D}^*$ and $D_t \in \mathcal{D}_t$ implies $D^*_t \subset D_t$. To see this note that since $\mathcal{D}^*$ challenges M it is a refinement of $\mathcal{D}^0$. Moreover, if $\mathcal{D}^*$ is a refinement of $\mathcal{D}^k$ then $\mathcal{D}^*$ is a refinement of $\mathcal{D}^{k+1}$. Then last assertion follows from the fact that for $t' \in D^*_t \in \mathcal{D}^*$,

$$\gamma(t, D^k) = \bigcup_{D \in \mathcal{D}^*, D \subset D^k} \gamma(t, D) = \bigcup_{D \in \mathcal{D}^*, D \subset D^k} \gamma(t', D) = \gamma(t', D^k)$$

Hence, $\mathcal{D}^*$ is a refinement of $\mathcal{D}^k$ for all $k \geq 0$. Hence, $D^*_t \in \mathcal{D}^*$ implies

$$D^*_t \subset \left\{ \bigcap_k D^k_t \,\middle|\, t \in T \right\} = D_t \in \mathcal{D}$$

This concludes the proof of step 3.

That steps $1 - 3$ prove the theorem is obvious. □

6.3 Proof of Theorem 3

Lemma 8: *If $(T, \gamma) \in \mathcal{M}$, then γ is continuous.*

Proof: Let $(t,t') \in T \times T$. For any $\epsilon > 0$, choose N so that $\sum_{n=N}^{\infty} a_n t_n^* t_{n-1}^* < \epsilon/4$. Choose $\epsilon' > 0$ so that $d(t,\hat{t}) < \epsilon'/4$, $d(t',\hat{t}') < \epsilon'/4$ implies $a_1|t_1 - \hat{t}_1| + \sum_{n=2}^{N-1} b_n |t_n t'_{n-1} - \hat{t}_n \hat{t}'_{n-1}| < \epsilon/2$. Clearly, such an ϵ exists. Then, for any $(\hat{t},\hat{t}')$ within an ϵ' ball of (t,t'), we have

$$\begin{aligned}|\gamma(t,t') - \gamma(\hat{t},\hat{t}')| &\leq b_1|t_1 - \hat{t}_1| + \sum_{n=2}^{\infty} b_n |t_n t_{n-1} - \hat{t}_n \hat{t}'_{n-1}| \\ &= \epsilon/2 + \sum_{n=N}^{\infty} b_n |t_n t'_{n-1} - \hat{t}_n \hat{t}'_{n-1}| \leq \epsilon/2 + 2\sum_{n=N}^{\infty} a_n t_n^* t_{n-1}^* < \epsilon\end{aligned}$$

Hence, γ is continuous.

Lemma 9: *Let $M \in \mathcal{M}$ then M is a min-valid IPM.*

Proof: That each $M \in \mathcal{M}$ is an IPM follows from Lemma 8.

Next, we show that $(T,\gamma) \in \mathcal{M}$ is valid. Recall the sequence of decompositions $\mathcal{D}_n$ and the limiting decomposition $\mathcal{D}$ defined in the proof of Theorem 2. In Step 3 of the proof of Theorem 2, it is established that the IPM is valid if and only if $\mathcal{D} = \{\{t\} \,|\, t \in T\}$. Hence, to prove that (T,γ) is valid, it is enough to show that $D_n^t \subset \{s \in T \,|\, s_1 = t_1, \ldots, s_{n-1} = t_{n-1}\}$ for all $n > 1$.

Let $t^0 = (0,0,\ldots)$ and let $t^n = (t_1^*, \ldots, t_n^*, 0, 0, \ldots)$ for $n \geq 1$. First, we show that $D_1^{t^*} = \{t^*\}$, $D_1^{t^0} = \{t^0\}$ and $D_1^{t^1} = \{t^1\}$. Since $\gamma(t^*,t^*) \in \gamma(t,T)$ if and only if $t = t^*$ it follows that $D_1^{t^*} = \{t^*\}$. Since $\gamma(t,T) = \{b_0\}$ if and only if $t = t^0$ and $\gamma(t,T) = \{b_0 + t_1^* b_1\}$ if and only if $t = t^1$ it follows that $D_1^{t^0} = \{t^0\}$ and $D_1^{t^1} = \{t^1\}$. We will show by induction that $D_n^{t^n} = \{t^n\}$ for all n. Suppose it is true for $n \geq 1$. By the induction hypothesis $D_n^{t^n} = \{t^n\}$ and by the argument above $D_n^{t^*} = \{t^*\}$. Note that $\gamma(t,t^*) = \gamma(t,t^n) = b_0 + b_1, t_1^* + \ldots + b_{n+1} t_{n+1}^* t_n^*$ if and only if $t = t^{n+1}$. Hence, it follows that $D_{n+1}^{t^{n+1}} = \{t^{n+1}\}$. Next, we will prove by induction that $D_n^t \subset \{s \in X \,|\, s_1 = t_1, \ldots, s_{n-1} = t_{n-1}\}$. Note that $\gamma(s,t^0) = \gamma(t,t^0)$ implies $s_1 = t_1$. Therefore, $D_2^t \subset \{s \in X \,|\, s_1 = t_1\}$. Hence, the assertion is true for $n = 2$. Suppose it is true for an arbitrary $n \geq 2$. Then, by the inductive hypothesis, $\gamma(s,t^n) = \gamma(t,t^n)$ and $s \in D_n^t$ implies $s_k = t_k$ for all $k \leq n$. Therefore, $s \in D_{n+1}^t$ implies $t_1 = s_1, \ldots, t_n = s_n$.

Next, we show that M is min-valid if $M \in \mathcal{M}$. Let $\mathcal{D}$ be any non-trivial decomposition with D closed for all $D \in \mathcal{D}$. Define $\hat{n}$ to be the largest integer such that for $n < \hat{n}$

$$t_n = t'_n$$

for all $t \in D$ and for all $D \in \mathcal{D}$. Let $t = (1, t^*_1, \ldots, t^*_{\hat{n}-1}, 0, \ldots)$. Let D^* be the element of $\mathcal{D}$ that contains t and note that $t' \succeq t$ for all $t' \in D^*$ by the definition of $\hat{n}$. Further, by the definition of $\hat{n}$, there is $D \in \mathcal{D}$ with $t^1, t^2 \in D$ such that $t^1_{\hat{n}} > t^2_n$ (and $t^1_n = t^2_n$ for $n < \hat{n}$). It follows that

$$\gamma(t^1, t) = \min \gamma_{t' \in D^*}(t^1, t') > \gamma(t^2, t) = \min_{t' \in D^*} \gamma(t^1, t')$$

This shows that M is min-valid. □

Proof of Theorem 4: It remains to show that every min-valid simple reciprocity model is in $\mathcal{M}$. In this proof, we denote with $\bar{t}$ the kindest type (denoted t^* in the text) and with $\underline{t}$ the meanest type (denoted t_* in the text. Let $a_0 = \gamma(\underline{t}, \underline{t})$. By Assumption 1 (ii) $\gamma(\underline{t}, \cdot)$ is equal to the constant function a_0.

Let a_1, ρ^1, t^1 and t^1_s be defined as follows:

(1) $a_1 = \gamma(\bar{t}, \underline{t}) - a_0$

(2) $\rho^1(t) = \frac{1}{a_1}(\gamma(t, \underline{t}) - a_0)$ if $a_1 \neq 0$; otherwise $\rho^1(t) = 0$.

(3) $\gamma(t^1, t) = \gamma(\bar{t}, \underline{t})$ for all $t \in T$

(4) $\gamma(t^1_s, t) = \gamma(s, \underline{t})$ for all $t \in T$.

For $n > 1$, we define a_n, ρ^n, t^n, t^n_s inductively as follows:

(1') $a_n = \gamma(\bar{t}, t^{n-1}) - \gamma(\bar{t}, \underline{t})$

(2') $\rho^n(t) = \frac{1}{a_n}(\gamma(t, t^{n-1}) - \gamma(t, \underline{t}))$ if $a_{n-1} \neq 0$; otherwise, $\rho^n(t) = 0$ for all $t \in T$.

(3') $\gamma(t^n, t) = a_n \rho^{n-1}(t) + a_0$ for all $t \in T$.

(4') $\gamma(t^n_s, t) = \sum_{i=2}^n a_i \rho^i(s) \rho^{i-1}(t) + a_1 \rho^1(s) + a_0$ for all $s \in T, t \in T$ or equivalently:

(4") $\gamma(t^n_s, t) = \sum_{i=1}^n \rho^i(s) \gamma(t^i, t) + [1 - \sum_{i=1}^n \rho^i(s)] \gamma(\underline{t}, \underline{t})$ for all $s \in T, t \in T$

Our next step is to prove that

(i) $a_n \geq 0$, $\rho^n(\bar{t}) \in \{0, 1\}$, $\rho^n(\bar{t}) = 1$ iff $a_n > 0$, $\rho^n(\underline{t}) = 0$, and $\rho^n(t) \in [0, 1]$ for all $t \in T$,

(ii) For $n > 1$, $\rho^i(t^n) = 0$ for all $i < n$. Moreover, $a_{n-1} = 0$ implies: $a_n = 0$, $\rho^n(t) = 0$, $\gamma(t^n, t) = a_0$ and $\gamma(t^n_s, t) = \gamma(t^{n-1}_s, t)$ for all $t \in T$.

(iii) $t^n \in T$ and $t^n_s \in T$ are well defined for all n.

(iv) ρ^n is linear, continuous, and reciprocating.

(v) $s \succeq t^n_s$ for all $s \in T$.

(vi) $\rho^n(t^n) = \rho^n(\bar{t})$ and $\rho^n(t^n_s) = \rho^n(s)$.

We will prove the assertions (simultaneously) by induction. That (i) holds for $n = 1$ follows from Assumption 1(i).

To see that t^1, t^1_s are well defined, note that $\gamma(\bar{t}, t) \geq \gamma(t^1, t) \geq \gamma(\underline{t}, t)$ and $\gamma(\bar{t}, t) \geq \gamma(t^1_s, t) \geq \gamma(\underline{t}, t)$ since γ is reciprocating and $\gamma(\underline{t}, \cdot)$ is constant. Further, $\gamma(\bar{t}, t) - \gamma(\bar{t}, \underline{t}) \geq \gamma(t^1, t) - \gamma(t^1, \underline{t}) = 0$ and $\gamma(t^1_s, t) - \gamma(t^1_s, \underline{t}) = 0$. Since $\gamma(t^1, \cdot), \gamma(t^1_s, \cdot)$ are constant on T, they are linear, reciprocating and continuous. Therefore Assumption 1(iii) implies that $t^1, t^1_s \in T$ are well defined.

Continuity of γ implies that ρ^1 is continuous. That ρ is linear follows from the fact that (T, γ) is convex. That ρ is reciprocating is obvious and hence (iv) follows.

For $n = 1$, (v) follows immediately from reciprocity and (vi) follows from the definitions of t^1 and t^1_s.

Next, we prove (ii) for $n = 2$. The first part follows from the definitions. If $a_1 > 0$, there is nothing left to prove. Otherwise, consider the decomposition $\{D\}$ of T consisting of a single element $D = \{T\}$. Since $a_1 = 0$, Assumption 1(i) implies $\min_{s \in D} \gamma(t, s) = a_0$ for all $t \in T$. Therefore, by min-validity D must be a singleton: $\bar{t} = t^1 = \underline{t}$. Hence, $a_2 = 0$ and the rest of (ii) follows.

Suppose (i)$-$(vi) hold for all $n \leq N$ for some $N \geq 1$ and let $n = N + 1$. The argument for (i) is essentially identical to the case of $n = 1$. For (ii), note that the case of $n = 2$ has already been dealt with. For $n > 2$, the first assertion of (ii) follows by induction. If $a_{n-1} = 0$ then by (2), $\rho^n(t) = 0$ and therefore $\gamma(t^n, t) = a_0 = \gamma(\underline{t}, t)$ for all $t \in T$. Hence, $t^n = \underline{t}$ and therefore (1) implies $a_n = 0$. Then, (4) implies $t^n_s = t^{n-1}_s$ for all $s \in T$. This proves (ii). For $n > 1$, the proofs of (i) and (iv) follow from the inductive hypothesis and the arguments used in the case of $n = 1$.

First, note that ρ^i is linear, continuous and reciprocating and since $a_n \geq 0$ this implies the same properties for the right hand side of (3') and (4'). By the inductive hypothesis and the fact that $\bar{t}$ is reciprocating, we have $\gamma(\bar{t}, s) \geq \gamma(\bar{t}, t_s^{n-1})$. Then (4'') and the linearity of $\gamma(\bar{t}, \cdot)$ imply

$$\begin{aligned}\gamma(\bar{t}, s) \geq \gamma(\bar{t}, t_s^{n-1}) &= \sum_{i=1}^{n-1} \rho^i(s)\gamma(\bar{t}, t^i) + (1 - \sum_{i=1}^{n-1} \rho^i(s))\gamma(\bar{t}, \underline{t}) \\ &= \sum_{i=2}^{n} a_i \rho^{i-1}(s) + \gamma(\bar{t}, \underline{t})\end{aligned}$$

Since $\rho^i(s) \in [0,1], a_i \geq 0$ for all $i \leq n$, we have

$$\begin{aligned}\gamma(\bar{t}, s) &\geq \gamma(\bar{t}, t_s^{n-1}) \geq \gamma(t^n, s) \geq \gamma(\underline{t}, s) \\ \gamma(\bar{t}, s) &\geq \gamma(\bar{t}, t_s^{n-1}) \geq \gamma(t_t^n, s) \geq \gamma(\underline{t}, s) \\ \gamma(\bar{t}, s) - \gamma(\bar{t}, \underline{t}) &\geq \gamma(\bar{t}, t_s^{n-1}) - \gamma(\bar{t}, \underline{t}) \geq \gamma(t^n, s) - \gamma(t^n, \underline{t}) \geq \gamma(\underline{t}, s) - \gamma(\underline{t}, \underline{t}) \\ \gamma(\bar{t}, s) - \gamma(\bar{t}, \underline{t}) &\geq \gamma(\bar{t}, t_s^{n-1}) - \gamma(\bar{t}, \underline{t}) \geq \gamma(t_t^n, s) - \gamma(t_t^n, \underline{t}) \geq \gamma(\underline{t}, s) - \gamma(\underline{t}, \underline{t})\end{aligned}$$

Proving that $t^n, t_s^n \in T$ for all $s \in T$.

By the inductive hypothesis and the fact that t is reciprocating, we have $\gamma(t, s) \geq \gamma(t, t_s^{n-1})$. Then (4') and the linearity of $\gamma(t, \cdot)$ imply

$$\begin{aligned}\gamma(t, s) \geq \gamma(t, t_s^{n-1}) &= \sum_{i=1}^{n-1} \rho^i(s)\gamma(t, t^i) + (1 - \sum_{i=1}^{n-1} \rho^i(s))\gamma(t, \underline{t}) \\ &= \sum_{i=2}^{n} a_i \rho^i(t)\rho^{i-1}(s) + \gamma(t, \underline{t}) = \gamma(t_t^n, s)\end{aligned} \tag{$*$}$$

proving (v). As in the case of $n = 1$, (vi) follows from the definitions of t^n, t_s^n.

Let $N = \min\{n \,|\, a_n = 0\}$ if the set $\min\{n \,|\, a_n = 0\}$ is nonempty and set $N = \infty$ otherwise. Set $t_n^* = \sqrt{a}_n$ for all n such that $1 < n < N$, $b_0 = a_0$, $b_1 = \sqrt{a}_1$, $b_n = \frac{\sqrt{a_n}}{\sqrt{a_{n-1}}}$ for all n such that $1 < n < N$. Define $\iota : T \to T$ as follows: $\iota_i(t) = \sqrt{a}_n \rho^n(t)$ for all $t \in T$ and $\iota(t) = (\iota_1(t), \iota_2(t), \ldots)$.

By reciprocity and (v), we have $\gamma(\bar{t}, \bar{t}) \geq \gamma(\bar{t}, t_{\bar{t}}^{n-1})$ for all n. Then, (vi) and $(*)$ implies $\gamma(\bar{t}, t_{\bar{t}}^n) = \sum_{i=1}^{n-1} a_i$. Since $a_n \geq 0$ for all n, we conclude $\sum_{i=1}^{\infty} a_i < \infty$ therefore $\lim a_n = 0$. This, together with the fact that each $\rho^n \leq 1$ is continuous implies ι is continuous. Define

the following decomposition d of T: $s \in D^t$ if and only if $\iota(t) = \iota(s)$. The continuity of ι ensures that each $D \in d$ is closed.

Let t_s be defined as follows: $\gamma(t_s, t) = \sum_{i=2}^{\infty} a_i \rho^i(s)\rho^{i-1}(t) + \gamma(s, \underline{t})$ for all $t, s \in T$. Note that $\gamma(t_s^n, \cdot)$ converges to $\gamma(t_s, \cdot)$, so the compactness of T and the continuity of γ implies that $t_s \in T$. Since $s \succeq t_s^n$ for each n it follows that $s \succeq t_s$. From (vi), we conclude that $\iota(t_s) = \iota(s)$. Also, equation $(*)$, implies $\gamma(s, t_t^{n-1}) = \gamma(t_s^n, t_t^{n-1})$. Hence, the continuity of γ ensures $\gamma(s, t_t) = \gamma(t_s, t_t)$. Then, reciprocity implies $\min_{t' \in D^s} \gamma(s, t') = \gamma(s, t_t) = \gamma(t_s, t_t)$ for all $s \in D \in \mathcal{D}$. It follows from min-validity that each $D \in d$ is a singleton. That is, $s = h_s$ for all $s \in T$. $\square$

References

1. Battigalli, P. and Siniscalchi, M (2003): "Rationalization and Incomplete Information," *Advances in Theoretical Economics*, Vol. 3 No. 1, Article 3.

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

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

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

5. Charness, G, and Rabin, M. (2002), "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, 117(3), 817-869.

6. Cox, J. C., Friedman, D. and Gjerstad, S. (2004): "A Tractable Model of Reciprocity and Fairness", mimeo, University of California, Santa Cruz.

7. Dufwenberg, M. and G. Kirchsteiger (1999): "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, 47(2), 268-298.

8. Falk A. and Fishbacher U. (1999): "A Theory of Reciprocity", Working paper No. 6, University of Zurich.

9. Falk A., E. Fehr and U. Fishbacher (2000): "Testing Theories of Fairness - Intentions Matter", working paper no. 63. University of Zurich.

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

11. Geanakoplos J., D. Pearce and E. Stacchetti (1980): "Psychological Games and Sequential Rationality" Games and Economic Behavior, 1, pp. 60-80.

12. Levine, D, (1998): "Modelling Altruism and Spitefulness in Game Experiments," *Review of Economic Dynamics*, 7, 348-352.

13. Mertens J.F. and S. Zamir, (1985): "Formulation of Bayesian Analysis for Games with Incomplete Information," *International Journal of Game Theory*, 14, 1-29.

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

15. Segal, U. and J. Sobel, (2004):"Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings," Discussion Paper, University of California, San Diego.

16. Sobel, J., (2004) "Interdependent Preferences and Reciprocity", mimeo, University of California, San Diego.