The Canonical Type Space for Interdependent Preferences[†]

Faruk Gul

and

Wolfgang Pesendorfer

Princeton University

March 2006

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 the kindness of other agents.

[†] 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 for those around him. For example, a person's preference over consumption goods may depend on how members of his peer group value these goods. Alternatively, the individual's inclination to charitable giving may depend on the personalities 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 modeling 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. A 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 personality be identified 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 preference profiles 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. Hence, round n specifies a set of preference profiles 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 all types have the same characteristic. Round 1 specifies the possible preference profiles of a type (for any opponent type). Let

$$\theta_1 = \{(S,S),(G,G)\},\ \theta_1' = \{(S,S)\}$$

be the possible round 1 statements. In this case, both players have the same preference in each contingency. The statement θ_1' implies the corresponding type always leads to the selfish profile while θ_1 implies that either profile is possible. Note that round 1 identifies a unique preference profile for the type corresponding to θ_1'. Hence, round 1 identifies the selfish type.

Round 2 specifies a preference profile for every round 1 statement. For example, suppose there are three distinct round 2 statements:

$$\begin{aligned}
\theta_2' &: \quad \theta_1' \to \{(S,S)\}; \theta_1 \to \{(S,S)\} \\
\theta_2 &: \quad \theta_1' \to \{(S,S)\}; \theta_1 \to \{(S,S),(G,G)\} \\
\theta_2'' &: \quad \theta_1' \to \{(S,S)\}; \theta_1 \to \{(G,G)\}
\end{aligned}$$

The statement θ_2' is the round 2 statement of the selfish type. The statement θ_2 corresponds to types who prefer the selfish action when the opponent is the selfish type but whose preference remains undetermined otherwise. The statement θ_2'' identifies a type who leads to the generous profile unless the opponent is selfish.

We require that the hierarchy of preference statements eventually lead to a unique preference profile for each contingency. The collection of interdependent preference models that can be described in this way is our canonical type space. When the set of preferences is finite there must exist a finite number of rounds after which any type can determine his preference given any opponent type.

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 profile if type t faces type t'. In addition, there is a continuous function ω that describes the characteristic of each type. Our main result (Theorem 2) identifies a simple condition (*validity*) that is necessary and sufficient for (T, Γ, ω) to be a component of the canonical type space. Validity can be interpreted as a consistency requirement that ensures that types can be distinguished in a model where players only know their own preference parameters (Theorem 3).

Our canonical type space provides a foundation for valid IPMs that is analogous to the Mertens and Zamir (1985)/Brandenburger and Dekel (1993) foundations for informational (Harsanyi) types. Mertens-Zamir and Brandenburger-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 and Brandenburger-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]

As an application, we present a definition of reciprocity and identify a class of interdependent preference models that is characterized by reciprocity. 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

[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.

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 consider 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 $r \in [a, a+d]$, where r is the weight the agent puts on the opponent's payoff and $a < a+d \leq 1$.[2] The preference r maximizes

$$ru(a_2) + (1-r)u(a_1)$$

for some fixed u. The parameter r depends on the player's type and on his opponent's type. Higher r's correspond to nicer preferences. Suppose a type is a numbers $x \in [0,1]$. The function γ specifies the weight r that type x gives the payoff of an opponent type y.

$$\gamma(x,y) = a + bx + cy + dxy \tag{1}$$

[2] Note that a may be negative. A preference $r < 0$ 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.

where $0 < b, 0 < c, d + c > 0, b + d > 0$. The number x measures the *kindness* of a type. When $x > y$ then type x is nicer to z than y for all z, i.e., x is *kinder* than y. A type *reciprocates* if he is nicer to a kinder opponent. For the parameter values indicated above all types reciprocate.

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 type x for player 1. Since all types are reciprocating, player 2 will put more weight on his opponent's payoff if he believes that the opponent's x 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.

The formula (1) above, generalizes a related model by Levine (1998). Levine's formulation corresponds to the case where $d = 0$. Our Theorem 4 gives conditions on the type space that identify formula (1) above.

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.

[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.

The models based on GPS require a strategic context to define reciprocity or other personality traits. 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 kind person is someone who is nice to the opponent irrespective of their type. An reciprocating agent is someone who is nicer to a kinder opponent, etc. The advantage of our model is that types can be identified independently of the particular strategic context.

2. Interdependent Types

In this section, we define interdependent preference models as consisting of an abstract type space, a map that associates each pair of types with a preference profile and a map that associates each type with a characteristic. We provide an interpretation of a type as a sequence of conditional preference statements and construct the corresponding canonical type space. We show in Theorem 1 that each component of the canonical type space can be interpreted as an interdependent preference model.

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 subset of X with the Hausdorff topology.

We assume that there are two players. Each player is described by a type that captures all relevant information about the player. Each pair of types gives rise to a preference profile. In addition, types may be distinguished by a 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 the ethnicity or the height of a type. We refer to those aspects of a type that can only be described through behavior as the type's *personality*.

Let T denote the compact metric space of types and let Ω denote a compact metric space of characteristics. The characteristic of a type is captured by function $\omega : T \to \Omega$. We assume that ω is continuous and onto.[4] Let $\Gamma : T \times T \to \mathcal{R} \times \mathcal{R}$ be the function that associates with each pair of types (t, t') the corresponding pair of preferences (R, R'). We assume that Γ is continuous and satisfies the following symmetry requirement

$$\Gamma(t, t') = (R, R') \text{ implies } \Gamma(t', t) = (R', R) \tag{S}$$

The symmetry requirement (S) ensures that the preference profile depends only on the types of players (and not on their names). Condition (S) implies that we can express Γ as

$$\Gamma(t, t') = (\gamma(t, t'), \gamma(t', t))$$

where $\gamma : T \to \mathcal{R}$ is a continuous function. We call $M = (T, \gamma, \omega)$ an *interdependent preference model* (IPM). When all types have the same characteristic, we ignore ω and write $M = (T, \gamma)$ instead of $M = (T, \gamma, \omega)$.

Example 1: (Levels of Politeness) Two agents stand in front of an open door. Each agent either prefers to go (first) or to wait and let the other person go first. We denote the former preference with g and the latter preference with w. There is a finite set of types $T = \{1, \ldots, k\}$ and types can be ordered by their politeness. Moreover, politeness begets politeness, that is, all types are (weakly) more inclined to be polite if the opponent is a more polite type. To capture this, assume that when type t faces opponent type t' he prefers to wait if $t + t' > k$ and he prefers to go when $t + t' \leq k$. In this example, all types have the same characteristic. The function γ is defined as follows:

$$\gamma(t, t') = \begin{cases} w & \text{if } t' + t > k \\ g & \text{if } t' + t \leq k \end{cases} \tag{2}$$

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

To illustrate our interpretation of types, we offer an alternative description of the γ in equation (2) as a hierarchy of preference statements: In round 1 each type reports the set of possible preference profile. Hence, type 1 reports $\theta_1^1 = \{(g,g)\}$, types $2,\ldots,k$ report $\theta_1^2 = \{(g,g),(w,w)\}$. Note that round 1 identifies types 1. In round 2 each type reports the set of possible preference profiles for every possible round 1 statement of the opponent. Let θ_2^1 denote the round 2 statement of type $2 \leq t \leq k-1$. Then,

$$\theta_2^1(\theta_1) = \begin{cases} \{(g,g)\} & \text{if } \theta_1 = \theta_1^1 \\ \{(g,g),(w,w)\} & \text{if } \theta_1 = \theta_1^2. \end{cases}$$

Let θ_2^2 denote the round 2 statement of type k. Then,

$$\theta_2^2(\theta_1) = \begin{cases} \{(g,g)\} & \text{if } \theta_1 = \theta_1^1 \\ \{(w,w)\} & \text{if } \theta_1 = \theta_1^2. \end{cases}$$

Note that the round 2 statement identifies type k. Similarly, in round 3 each agent specifies the set of possible preferences for each possible statement of the opponent in the previous two rounds, i.e., for each pair (θ_1,θ_2). Continuing in this fashion it is easy to see that all types are identified after at most $k-1$ rounds. Note that in this example, all types have the same characteristic. In the general case where different types have different characteristics, there would be a round 0 in the hierarchy of preference statements for reporting characteristics. Then, an agent's subsequent reports would also be a function of his opponent's announced characteristic.

Our construction of the canonical type space generalizes the example above. The canonical type space depends on the set of alternatives A, the set of preferences $\mathcal{R}$ and a compact space of characteristics Ω. For notational simplicity 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 $\mathcal{H} = \mathcal{H}_{\mathcal{S}}$ denote the collection of non-empty, closed subsets of the set of preference profiles $\mathcal{S} = \mathcal{R} \times \mathcal{R}$. 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, $\theta_1 = (f_0, f_1)$ such that f_0 is a characteristic and f_1 is an element of $\mathcal{C}(\Theta_0, \mathcal{H})$. Hence, the function f_1 specifies for every characteristic a subset of preference profiles. More generally, $\theta_n = (f_0, \ldots, f_n)$ where $f_0 \in \Theta_0$ 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} the set of possible preference profiles given the information that has been revealed by round k.

The entries $\theta_n \in \Theta_n$ must satisfy certain consistency requirements. To understand the first consistency requirement, consider any $\theta_{n+1} = (f_0, \ldots, f_{n+1}) \in \Theta_{n+1}$. The first requirement is that f_{n+1} should not contradict what was revealed by f_n and that f_n should reveal whatever is already known by round n. The first part of this statement 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}. The second part requires that $f_n(\theta'_{n-1})$ not contain any preference that is sure to be removed in the next round. Hence, 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}.

To understand the second consistency requirement, let $\theta_n = (\theta'_{n-1}, f_n)$ and $\theta'_n = (\theta'_{n-1}, f'_n)$. The second part of consistency says that the round n statements of the two individuals cannot be incompatible. That is, $f_n(\theta'_{n-1})$ must contain a preference profile (R, R') such that (R', R) is contained in $f'_n(\theta_{n-1})$.

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

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

(ii) For all $(\theta'_{n-1}, f'_n) \in \Theta_n$ *there is* $(R, R') \in \mathcal{S}$ *such that* $(R, R') \in f_n(\theta'_{n-1})$ *and* $(R', R) \in f'_n(\theta_{n-1})$.

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; if $\{f_n | (\theta_{n-1}, f_n) \in \Theta_n\}$ were empty then (i) would imply that $f'_n(\theta_{n-1})$ is 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 $\Psi : \Theta \times \Theta \to \mathcal{S}$ denote the function that specifies the preference profile when the player is type θ and the opponent is type θ'. Hence,

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

Requirement (ii) in the definition of consistency ensures that the function ψ satisfies the following symmetry condition.

$$\Psi(\theta, \theta') = (R, R') \text{ implies } \Psi(\theta', \theta) = (R', R) \qquad (S)$$

If Ψ satisfies (S) we say that Ψ is *symmetric*. We define $\phi : \Theta \to \Omega \times \mathcal{C}(\Theta, \mathcal{S})$ to be the function that specifies for every type $\theta \in \Theta$ the characteristic of the type θ and the mapping θ uses to assign preferences profile 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 symmetric 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. Note that for a symmetric Ψ there is a $\psi : \Theta \times \Theta \to \mathcal{R}$ such that

$$\Psi(\theta, \theta') = (\psi(\theta, \theta'), \psi(\theta', \theta))$$

Corollary: *Let Θ be a component of the canonical type space. Then $M^{\Theta} = (\Theta, \psi, \omega)$ is an IPM.*

Note that Θ is compact (by definition) and ψ is continuous by Theorem 1. Therefore, it follows that M^{Θ} is an IPM. Theorem 2 in section 4 provides a converse to the Corollary

above. It characterizes all those IPM's that correspond to some component of the canonical type space.

3. Valid Models

Not every IPM represents a component of the canonical type space. To qualify as a component of the canonical type space, types must be uniquely identified by the hierarchy of preference statements described in the previous section. The following is an example of an IPM that does not satisfy this property.

Example 2: (Fixed Point Types) As in Example 1, agents have preferences about who goes first through an open door. Type 1 prefers to go first (g) if the opponent is type 1 and prefers to wait (w) if the opponent is type 2. Type 2 prefers g if the opponent is type 2 and w if the opponent is type 1. All types have the same characteristic. The following table summarizes this IPM.

	1	2
1	g, g	w, w
2	w, w	g, g

Table 2

We can construct the hierarchy of preference statements for this example as follows. In round 1 both types report $\{(g, g), (w, w)\}$. Given that types are indistinguishable in round 1, higher order statements will simply repeat the first order statement $\{(g, g), (w, w)\}$. Hence, the interdependent preference hierarchy cannot distinguish between types 1 and 2.

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. However, if there is no a priori distinction between types 1 and 2 then it is not possible to explain the difference between the two types in terms of contingent preference statements.

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. Type 1 has the preference g if the opponent is tall and the preference w if the opponent is short.

Type 2 has the preference w if the opponent is tall and g if the opponent is short. Now types are non-circular fashion because they differ in how they respond to a (self-evident) characteristic.

Below, we introduce the notion of validity to rule out interdependent preference models with circular types. Our main theorem (Theorem 2) shows that valid IPM's correspond to components of the canonical type space.

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 $D^t \neq \{t\}$. Let $M = (T, \gamma, \omega)$ be an IPM and recall that $\Gamma(t, t') = (\gamma(t, t'), \gamma(t', t))$. We use the (standard) notation

$$\Gamma(t, D) := \{\Gamma(t, t') |\, t' \in D\}$$

Definition: *The IPM $M = (T, \gamma, \omega)$ 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 $D \in \mathcal{D}$.

Hence, if M is valid only if 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 Example 2 is not valid.

Our main theorem, Theorem 2 shows that 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.*

Validity captures the idea that the process of refining the set of possible types through preference statements does not terminate at a non-trivial decomposition. Our next objective is to demonstrate that a model is valid if and only if types can be identified without an a priori understanding of types.

Consider an IPM where all types have the characteristic. Hence, we write $M = (T, \gamma)$ and ignore condition (i) in the definition of validity. Suppose that every type knows his own personality but does not know the personality of the opponent and that players have no a priori way to identify or refer to types. Under these circumstances, we can represent a type's epistemic state by a *knowledge hierarchy*. In our non-probabilistic setting, the appropriate model to describe a type's knowledge is a hierarchy of possibilities. Since each type considers every opponent type possible, the preference profile and opponent type combinations that type t considers possible are $\rho(t) = \{(\Gamma(t, t'), t' \,|\, t' \in T\}$. Then, we can define the knowledge hierarchy of each type can be defined as follows:

$$m_1(t) = \{\Gamma(t, t') \,|\, t' \in T\}$$

represents the set of preference profiles that type t considers possible. For $n > 1$ inductively define

$$m_n(t) = \{(s, m_{n-1}(t') \,|\, (s, t') \in \rho(t)\}$$

The sequence $m(t) := (m_1(t), m_2(t), \ldots)$ represents the knowledge of the epistemic type t, i.e., it represents a hierarchy of possibilities that captures what the agent knows, what he knows the opponent knows, what he knows the opponent knows he knows, and so on.

We refer to the sequence $m(t)$ as the *epistemic representation* of type t. If $m(t) = m(t')$ for $t \neq t'$, then a different label is given to types with the same epistemic representation. Note that two types with the same epistemic representation behave in the same way and therefore those types are indistinguishable. We say that an IPM has identifiable types if each type $t \in T$ has a distinct epistemic representation.

Definition: *Let $M = (T, \gamma)$ be an IPM and let $T_\infty = \{m(t) \,|\, t \in T\}$ be the corresponding epistemic types. The IPM M has identifiable types iff $m : T \to T_\infty$ is one to one.*

Theorem 3 says that validity and identifiability amount to the same restriction.

Theorem 3: *The types of an IPM M are identified if and and only if M is valid.*

Proof: See Appendix.

The definition of $m(t)$ is taken from the recent paper by Mariotti, Meier and Piccione (2004) (MMP) who provide Mertens-Zamir and Brandenburger-Dekel type foundations for

possibility structures. In a possibility structure, a type t is identified with a set of parameter (for example, preference profile) and opponent type pairs that t considers possible. Formally, let S be any compact metric space. A *possibility structure* on S is a pair (T, ρ^*) such that $\rho^* : T \to \mathcal{H}_{S \times T}$ is a continuous one-to-one mapping. It is easy to see that IPMs can be viewed as a subclass of possibility structures. For $M = (T, \gamma)$, let $S = \mathcal{S}$, $\rho^*(t) = \{\rho(t)\}$, where ρ is as defined above, and note that (T, ρ^*) is a possibility structure. Theorem 3 shows that only valid IPM's correspond to well-defined possibility structures in the sense that each type has a distinct hierarchical representation.[5]

In a general IPM (T, γ, ω), where different types may have different characteristics, an agent's epistemic type will consist of a pair (t, ω_0): the t will denote his own type and ω_0 will be his opponent's (observable) characteristic. Then, the function m will map epistemic types to their epistemic representations; that is, $m : T \times \Omega \to T_\infty$. In this case, Theorem 3 will still hold provided we modify the definition an IPM with identifiable types to be one where distinct t's yield distinct functions $m(t, \cdot)$. Hence, validity is equivalent to the requirement that distinct types must have distinct knowledge hierarchies *in some state of the world*.

There are two differences between possibility structures and IPMs. The first difference is the interpretation. A possibility structure describes agents' knowledge hierarchies regarding some underlying parameters. In contrast, an interdependent preference model describes an agent's response to the characteristic and personality of his opponent. The second difference is that in an IPM the types of the two agents uniquely determine the value of the parameter. A possibility structure allows for but does not require this property.

[5] MMP provide a universal possibility structure (T^u, ρ^u) that contains every other possibility structure (T, ρ) in the sense that the hierarchical representation corresponding to (T, ρ) is a subset of T^u. The universal possibility structure (T^u, ρ^u) is identified (as defined above). Moreover, each possibility structure can be mapped to an identified subset of the universal possibility structure. However, an arbitrary possibility structure may not be identified.

4. Reciprocity

Reciprocity describes a personality that is nicer to kinder opponents. We analyze reciprocity in a simple setting where every type has the same characteristic and preferences can be described by a single real number $r \in \mathcal{R}_o$. We assume that $\mathcal{R}_o$ is compact and that higher values of r are *nicer* preferences. A *simple empathy model (SEM)* is an IPM (T, γ) such that $\gamma : T \times T \to \mathcal{R}_o$. Since all types have the same characteristic, we omit the function ω.

Let $M = (T, \gamma)$ be an SEM. If type $t \in T$ responds to every opponent type with a nicer preference than type $t' \in T$, we say that t is *kinder than* t'. Formally, t is kinder than t' if

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

Let $\succeq$ denote the "kinder than" relationship and note that $\succeq$ is a transitive binary relation on T.

Definition: *(i) The SEM (T, γ) is complete if t is kinder than t' ($t \succeq t'$) or t' is kinder than t ($t' \succeq t$) for all $t, t' \in T$. (ii) The SEM (T, γ) is reciprocating if $t \succeq t'$ implies $\gamma(t'', t) \geq \gamma(t'', t')$ for all $t'' \in T$.*

A *reciprocity model* is a reciprocating SEM in which types are described by a compact set of reals and higher types represent strictly kinder types.

Definition: *An SEM (K, γ) is a reciprocity model if K is a compact subset of the reals, γ is non-decreasing in both arguments and $\gamma(x, \cdot) = \gamma(y, \cdot)$ implies $x = y$.*

In a reciprocity model (K, γ), each level of kindness is represented by exactly one type. If $x > y$ then $\gamma(x, \cdot) \geq \gamma(y, \cdot)$ and $\gamma(x, \cdot) \neq \gamma(y, \cdot)$ and therefore x is strictly kinder than y. Theorem 4 below establishes that an SEM is isomorphic to a reciprocity model if and only of it is valid, complete, and reciprocating.

Theorem 4: *The SEM (T, γ) is isomorphic to a reciprocity model if and only if it is valid, complete, and reciprocating.*

Proof: See Appendix.

Since every IPM is isomorphic to itself, Theorem 4 also shows that every reciprocity model is valid, complete, and reciprocating. To gain intuition for Theorem 4, note that completeness together with standard utility representation arguments ensure that the types' kindness can be represented by a real-valued function. Then, reciprocity yields that two equally kind types are treated the same way by all other types. Hence, $\Gamma(t,\cdot) = (\gamma(t,\cdot),\gamma(\cdot,t)) = (\gamma(t',\cdot),\gamma(\cdot,t')) = \Gamma(t',\cdot)$ whenever t and t' are equally kind. Then, validity implies that two distinct types cannot be equally kind. Lemma 9 in the appendix completes the argument by showing that reciprocity models are valid.

4.1 The Linear Reciprocity Model

The *linear reciprocity model* has $[0,1]$ as its type space and γ of the form

$$\gamma(x,y) = a + bx + cy + dxy$$

with $0 \leq b, 0 \leq c, 0 \leq d+b, 0 \leq d+c$. Theorem 4.1 below demonstrates that the linear reciprocity model obtains if the following convexity and linearity assumptions are satisfied.

Definition: *The reciprocity model (K,γ) is convex if $x,y \in K$ and $\lambda \in [0,1]$ implies there exists $z \in K$ such that $\gamma(z,\cdot) = \lambda\gamma(x,\cdot)+(1-\lambda)\gamma(y,\cdot)$. The reciprocity model (K,γ) is linear if*

$$\gamma(z,\cdot) = \lambda\gamma(x,\cdot) + (1-\lambda)\gamma(y,\cdot) \text{ implies}$$
$$\gamma(\cdot,z) = \lambda\gamma(\cdot,x) + (1-\lambda)\gamma(\cdot,y)$$

Convexity means that if f and f' are possible personality types (i., there exist $x,y \in K$ such that $f = \gamma(x,\cdot)$ and $f' = \gamma(y,\cdot)$) then so is $\lambda f + (1-\lambda)f'$ for any $\lambda \in (0,1)$. Note that convexity of a reciprocity model does not refer to the convexity of the set K. Rather it refers to the convexity of the set (of functions) $\{\gamma(x,\cdot) \,|\, x \in K\}$. Without convexity, the linearity assumption has no force. Linearity means that if type z behaves like a $\lambda,(1-\lambda)$ convex combination of types x,y then the response of all other types to z is a $\lambda,(1-\lambda)$ convex combination of their responses to x and y. We say that reciprocity model is (K,γ) is *nontrivial* if K contains more than one element.

Theorem 4.1: *A nontrivial reciprocity model is isomorphic to a reciprocity model* $([0,1],\gamma)$ *such that*

$$\gamma(x,y) = a + bx + cy + dxy$$

where $b \geq 0, c \geq 0, b+d \geq 0, c+d \geq 0$ *and* $(b,c,d) \neq (0,0,0)$ *if and only if it is convex and linear.*

Proof: See Appendix.

It is easy to verify that $([0,1],\gamma)$ with γ of the form described in Theorem 4.1 is a reciprocity model. We will refer to reciprocity models that satisfy convexity and linearity as linear reciprocity models and identify them with the γ's described in Theorem 4.1. The model presented in Levine (1998) corresponds to the subclass of convex-linear reciprocity models for which $d = 0$.

Linearity and convexity ensure that γ can be identified by four parameters: a, b, c, and d, where $\gamma(0,0) = a$, $\gamma(1,0) = a + b$, $\gamma(0,1) = a + c$, and $\gamma(1,1) = a + b + c + d$. The comparison of b and c determines whether it is more important to be kind or to face a kinder opponent; if b is large relative to c, it means that all types respond strongly to small increases in the kindness of their opponents, while if c large than b it means that one type has to be significantly kinder than another to be treated significantly nicer by opposing players. The parameter d reveals the complementarities between reciprocity and kindness. If $d > 0$ then kinder types reciprocate more than less kind types whereas for $d < 0$ kinder types reciprocate less.[6]

Application to Ultimatum Bargaining: Considers an ultimatum bargaining game with players who have interdependent preferences. Let $(c_1, c_2) \in [0,1]^2$ be the outcome of the bargaining game. Each $r \in \mathcal{R}_o$ corresponds to the preference $\succeq_r$ such that $(c_1, c_2) \succeq_r (c_1', c_2')$ if and only if

$$ru(c_2) + (1-r)u(c_1) \geq ru(c_2) + (1-r)u(c_1')$$

[6] This interpretation is appropriate provided that the real numbers identified with preferences can be interpreted as cardinal quantities; that is, provided that $r_4 - r_3 = r_2 - r_1$ can be meaningfully interpreted as "r_4 is just as nicer than r_3 as r_2 is nicer than r_1." The same is needed for interpreting the linearity assumption.

where $u(c) = c^{.87}$ for $c \in [0,1]$. Consider the linear reciprocity model $M = ([0,1], \gamma)$ where

$$\gamma(x,y) = -1/2 + x + {}^{19}\!/_{26}y - {}^{19}\!/_{26}xy$$

Note that for the specified parameters kinder types reciprocate less ($d < 0$).

To address the experimental evidence on the ultimatum bargaining game, we assume that players are uncertain about the opponent's interdependent preference type. For simplicity, we assume that the support of the player's beliefs consists of two types $x \in \{0,1\}$, the prior probability of type 1 is 3/4 and types are distributed independently.

The game is a standard ultimatum bargaining game with private information regarding the other player's type. Player 1 must choose $c_2 \in [0,1]$ and player 2 must either accept of reject this offer. If c_2 is accepted, player 1 receives the material reward $c_1 = 1 - c_2$ and player 2 receives c_2.

This game has a unique equilibrium. A type 1 player 1 chooses $c_2 = 1/2$ while type 0 player 1 chooses $c_2 \approx 1/5$. A type 1 player 2 accepts all offers while a type 0 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 his type.

This example captures three features of the experimental evidence (see Blount (1995)). First, players choose offers that are above $c_2 = 0$. Second, some player 2 types reject offers even if they entail positive rewards for themselves. Third, player 2's response changes when 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 by player 1. In our model, the reason for this is clear: if player 1 makes the offer, then the low offer enables player 2 to infer that player 1 is an unkind type. Hence, player 2 rejects a low offer from player 1. When the randomization device makes the offer then player 2 cannot infer player 1's type from his offer. As a result, player 2 does not reject any offer.

Note that our model predicts that players who make generous offers (type 1) are least likely to reject proposals while types who make the least generous offers (type 0) are most likely to reject an offer. This is a consequence of the assumption that the model is *complete*,

i.e., that types can be ranked according to their kindness. If players who make generous offers are likely to reject offers this suggests that types cannot be ranked according to their kindness.[7]

4.2 The Binary Reciprocity Model

Next, we consider reciprocity models with two actions; the nicer action 1 and the mean action 0. There are two preferences; the nice preference (1) strictly prefers the nice action to the mean action, while the mean preference (0) corresponds to the opposite ranking. Hence, $\mathcal{R}_o = \{0, 1\}$.

We say that a reciprocity model (K, γ) is binary if $\{\gamma(x, y) \,|\, x, y \in K\} = \mathcal{R}_o = \{0, 1\}$. Note that this definition implies that a binary reciprocity model is nontrivial. Theorem 4.2 characterizes all binary reciprocity models. For any two integers k, l define

$$G(k, l) = \begin{cases} 1 & \text{if } k > l \\ 0 & \text{if } k \leq l \end{cases}$$

Theorem 4.2: *A reciprocity model is isomorphic to* $(\{1, \ldots, k\}, \gamma)$ *where*

$$\gamma(i, j) = G(i + j, k + G(n, i)) \tag{3}$$

for some $k \in \{2, 3, \ldots\}, n \in \{1, \ldots, k + 1\}$ *if and only if it is binary.*

Proof: See Appendix.

To gain intuition for Theorem 4.2, note that, by compactness, a binary reciprocity model must have a finite number of types $\{1, \ldots, k\}$. Assume, without loss of generality higher types are kinder. Since there are only two preferences, each type is identified by the set of types to whom he is not nice; hence, let $K^i = \{j \in K \,|\, \gamma(i, j) = 0\}$. Validity and reciprocity ensure that $i \neq j$ implies $K^i \neq K^j$. Reciprocity also ensures that K^i is either empty or of the form $\{1, \ldots, j\}$ for some j. Hence, there are k types identified with one of $k + 1$ sets of the form $K^i = \{\{1, \ldots, j_i\}$ where $j_i < j_{i+1}$ and $j_i = 0$ is interpreted as $K^i = \emptyset$. Let n be the smallest $j \in \{1, \ldots, k\}$ such that $K^j = \{1, \ldots, k - j\}$ if such a type j exists. Then, we must have $K^i = \{1, \ldots, k - i\}$ for all $i \geq j$ and $K^i = \{1, \ldots, k + 1 - i\}$ for

[7] We thank Larry Samuelson for a related observation.

$i < j$. If no such j exists, set $n = k+1$ and note that we must have $K^i = \{1, \ldots, k+1-i\}$ for all i. Hence,

$$\gamma(i,j) = \begin{cases} 1 & \text{if } [i+j > k+1 \text{ and } i < n] \text{ or } [i+j > k \text{ and } i \geq n] \\ 0 & \text{otherwise} \end{cases}$$

This proves the theorem.

Note that Example 1 (levels of politeness) in the introduction corresponds to the class of binary reciprocity models with $n = 1$. When $n = 1$, $G(n,i) = 0$ for all $i = 1, \ldots, k$. Hence, every type wants to be polite to the most polite type and the most polite type wants to be polite to every type.

Example 3: (Leaders and Followers) The binary reciprocity model can also capture situations where types differ in how confident they are about their preferences. Suppose the set $\mathcal{R}_o = \{0,1\}$ refers to the color of a shirt, say red (0) or green (1). For this interpretation 'nice' is replaced with 'prefers green' and 'mean' is identified with prefers 'red.' In this context, reciprocity represents the social pressure to have preferences like the other player. Let $k = 4$, $n = 3$, and suppose type 1 always chooses the red shirt while type 4 always chooses the green shirt. Types 1 and 4 are confident types whose preferences are unaffected by the social environment. Type 2 chooses the red shirt if the opponent is type $t \in \{1,2,3\}$ but chooses the green shirt if the opponent is type 4. Type 3 chooses the green shirt if the opponent is type $t \in \{2,3,4\}$ but chooses the red shirt if the opponent is type 1. Types 2 and 3 are insecure types whose preference is affected by the peer group. Hence,

$$\gamma(i,j) = \begin{cases} 1 & \text{if } [i+j > 5 \text{ and } i < 3] \text{ or } [i+j > 4 \text{ and } i \geq 3] \\ 0 & \text{otherwise} \end{cases}$$

This formula corresponds to equation (3) for $k = 4$ and $n = 3$.

In the following example, we embed a binary reciprocity model into an overlapping generations model. The example illustrates two features of reciprocity models. First, when players discover that the fraction of generous agents has increased they respond with a further increase in generosity. This pattern distinguishes reciprocity from altruism. As Rabin (1998) points out, a model of altruism could not generate this pattern because increased giving by others would reduce or leave unchanged the incentive for giving. Second, the

preferences in our model depend on the underlying distribution of types and - as a result - past behavior reveals relevant information about opponent's personality and therefore influences behavior in the current period. In belief-based models such as Geanakoplos, Pearce and Stacchetti (1989) and Rabin (1993), preferences depend on play's beliefs in the current period and therefore, past behavior is often irrelevant.

Example 4: (Cycles of Generosity) Suppose there are two types, a generous type who always takes the generous action (irrespective of the opponent type) and a selfish type who takes the generous action if the opponent is a generous type and the selfish action if the opponent is a selfish type. To analyze a situation with uncertainty, we assume that the selfish type takes the generous action if and only if the probability that the opponent is generous is greater than or equal to $\alpha \in (0,1)$.

Consider the following overlapping-generations game. Each period a continuum of players are born and all players live for two periods. A player born in period t is matched with a player born in period $t-1$ and takes either the generous or the selfish action. Players observe the fraction of players who take the selfish action in the previous period but cannot observe individual actions.

We assume that the distribution of types evolves according to a symmetric Markov process with two states. In state H each player is independently assigned the generous type with probability $h > 1/2$ and in state L each player is independently assigned the generous type with probability $1-h$. The transition probabilities are as follows:

	1	2
1	π	$1-\pi$
2	$1-\pi$	π

Table 2

We assume that $\pi > 1/2$ and

$$\pi h + (1-\pi)(1-h) > \alpha > 1/2 \tag{4}$$

Furthermore, we assume that the initial state (in period 0) is L. Players of generation 0 take no action. At the end of periods $t = 1, 2, \ldots$, players observe the fraction of players

who take each action. Players know the initial state but cannot observe the the state in subsequent periods or the actions of individual players.

Generous types in this game always take the generous action. Selfish types take the generous action if they believe their opponent is a generous type with probability greater than or equal to α. (We assume, for simplicity, that ties are broken in favor of the generous action.) The Nash equilibrium of this game has the following structure:

Let λ_t be the fraction of players who take the generous action in period t. If $\lambda_t = 1-h$ then the selfish types take the selfish action in period $t+1$. If $\lambda_t = h$ then the selfish players take the generous action in periods $t+1, t+2, \ldots t+\tau$. Hence, after observing $\lambda_t = h$ *all* players take the generous action for the next τ consecutive periods and then the selfish type reverts back to the selfish action in period $t+\tau+1$. Note that during the τ periods where all players are taking the generous action, no new information is being revealed. Let π_k be the probability that the society is in state H, k periods after it was observed that h fraction of the population took the generous action. Hence,

$$\pi_{k+1} = \pi_k \pi + (1 - \pi_k)(1 - \pi) \tag{5}$$

The first order difference equation (5), together with the initial condition $\pi_1 = \pi$ yields

$$\pi_k = 1/2 + 2^{k-1}(\pi - 1/2)^k \tag{6}$$

for $k = 1, \ldots$. Then τ satisfies

$$\pi_\tau \cdot h + (1 - \pi_\tau) \cdot (1 - h) \geq \alpha > \pi_{\tau+1} \cdot h + (1 - \pi_{\tau+1}) \cdot (1 - h)$$

That is, τ is the largest integer k such that $\pi_k h + (1 - \pi_k)(1 - h) \geq \alpha$. Our parameter restrictions together with equation (6) imply that τ is well-defined.

In period 1, players know that the probability that the opponent is a generous type is $\pi(1 - h) + (1 - \pi)h$. Equation (4) ensures that this probability is less than α. Hence, all selfish types take the selfish action and all generous types take the generous action. If the state in period 1 is H, then the fraction of players who take the generous action is h; if the state is L then this fraction is $1 - h$. Hence, players in period 2 can infer the state and act

accordingly. If the state was H then *all players* take the generous action. But if all players take the generous action the state is not observable. In this case, selfish players continue to take the generous action for τ periods and then revert back to the selfish action. For a generic set of parameters values (i.e., $\pi_k \neq \alpha$ for all $k = 1, \ldots$) the outcome described above is the unique equilibrium outcome.

Note that if the fraction of agents that choose the generous action increases from $1 - h$ to h, the the following period all agents take the generous action. Hence, finding out that population is generous causes an increase in generosity. The higher level of generosity persists for τ periods and then the population's generousity is tested again. Observed behavior is relevant because it sometimes reveals the personality distribution in the population.

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.[8] 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

Psychological Game

[8] 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.

As GPS observe, 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 utilities in the matrix above are not observable payoffs but psychological payoffs that depend on 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).[9] Finally, the dependence on

[9] 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 easy to see that player 1 might be happier if he thinks that his friends believe that he will do the right thing; why player 1 would be less inclined to do the right thing 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.

beliefs may be a shortcut for describing the payoff consequences of the opponent's response. 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. 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 timid and a bold 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	1	-1
p	-1	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 GPS model, our model 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 personality. 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 3, 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 a particular play of that game.

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: *(i) Each $\mathcal{D}_n$ is continuous. (ii) M is valid if and only if $\mathcal{D} = \{\{t\}|t \in T\}$.*

Proof: (i) 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 part (i).

If $\mathcal{D} \neq \{\{t\} \,|\, t \in T\}$, the $\mathcal{D}$ is a non-trivial partition with properties (i) and (ii) in the definition of validity. Therefore, 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

$$D_t^* \subset \Big\{ \bigcap_k D_t^k \,\Big|\, t \in T \Big\} = D_t \in \mathcal{D}$$

This concludes the proof of step 1. □

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

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

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

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

Let

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

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

Henceforth, for any $t \in T$ such that $\iota(t) = (f_0, f_1, \ldots)$ we write f_n^t to denote the corresponding f_n. We also define the functions $g_n^t : T \to \mathcal{H}$ as

$$g_n^t(s) = \Gamma(t, D_{n-1}^s)$$

Fact 1: *For all n, the functions ι_n are onto and continuous and the sets Θ_n are non-empty and compact.*

Proof: 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), $\iota_{n+1} \in \mathcal{C}(\Theta_n, \mathcal{H})$ and Θ_{n+1} is compact. The functions ι_n is onto by definition. □

Fact 2: *(i) The function ι is onto. (ii) $\iota_n(t) = \iota_n(s)$ if and only if $D_n^t = D_n^s$. (iii) $f_n^t(\iota_{n-1}(s)) = g_n^t(s)$.*

Proof: Next, we show that $\iota : T \to \Theta$ is onto. Pick $(f_0, f_1, \ldots)$ 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, $s \in D_{n+1}^t \in \mathcal{D}_{n+1}$.

Part (iii) follows from part (ii) and the definitions of g_n^t, f_n^t. □

Fact 3: *If M is valid then (i) $g^t = \lim g_n^t$ is well defined and continuous and (ii) $d\,(g_n^{t_n}, g) \to 0$ if $t_n \to t$ as $n \to \infty$.*

Proof: Part (i) follows from Lemma 2. For part (ii) fix $\epsilon > 0$ and note that by Lemmas 3(i), 7(i) there exists N such that $d(g^t, g^t_N) < \epsilon$ and $\bar{d}(g^t_N) < \epsilon$. By Lemma 1(ii) $g^{t_n}_N \to g^t_N$. By Lemma 7(i) we can choose K so that $\bar{d}(g^{t_k}_N) \leq 2\epsilon$ for all $n \geq K$. Therefore, by Lemma 7(ii), $d(g^{t_n}_n, g_n) < 3\epsilon$ for all $n > \max\{K, N\}$. It follows that $d(g^{t_n}_n, g) < 4\epsilon$ for all $n > \max\{K, N\}$ as desired. □

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

Fact 2(ii) implies that $\Theta_{n-1}(\theta_{n-2}) = \iota_{n-1}(D^s_{n-2})$ for s such that $\iota_{n-2}(t) = D^s_{n-2}$. Therefore,

$$f^t_n(\theta_{n-2}) = \Gamma(t, D^s_{n-2}) = \bigcup_{s' \in D^s_{n-2}} \Gamma(t, D^{s'}_{n-1}) = \bigcup_{\theta'_{n-1} \in \Theta_{n-1}(\theta_{n-2})} f^t(\theta'_{n-1})$$

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

Let $f^t : \Theta \to \mathcal{H}$ be defined by

$$f^t(\theta) = \bigcap_{n \geq 1} f^t_n(\theta)$$

Assume that M is valid and hence $\mathcal{D} = \{\{t\} \,|\, t \in T\}$. Since, $\iota_n(t) = \iota_n(s)$ if and only if $D^t_n = D^s_n$ (Fact 2(ii)), we conclude that ι is one-to-one. For $\theta = (g_0, g_1, \ldots)$ we let $\theta(n)$ be defined as $(g_0, \ldots, g_n)$. By Fact 2, $f^t_n(\theta(n-1)) = g^t_n(s) = \Gamma(t, D^s_n)$ for $theta = \iota(s)$. It follows that $f^t(\theta) = \Gamma(t, s)$ and therefore f^t is a singleton.

To prove that ι is a homeomorphism we prove that ι is continuous and appeal to Lemma 6. Consider t_k converging to t. By Fact 3 it follows that for any two subsequences of natural numbers $n(j)$, $k(j)$ both converging to ∞, $g^{t_{k(j)}}_{n(j)}$ converges to g^t. Recall that d^* is the sup metric. It follows from Lemma 3(i) that $g^{t_{k(j)}}_{n(j)}$ converges to g^t in the sup metric d^* as well. Hence, for any $\epsilon > 0$, there exits N such that $k \geq N$, $n \geq N$, $d^*(g^{t_k}_n, g) < \epsilon$. Since each ι_n is continuous, we can choose $K > N$ large enough so that $d(f^k_n, f_n) < \epsilon$ for all $n \leq N$. Hence,

$$d(f^k_n, f_n) \leq d^*(f^k_n, f_n) = d^*(g^k_n, g_n) \leq d^*(g^k_n, g) + d^*(g, g_n) \leq 2\epsilon$$

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 $\hat{\iota}_0$ 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 $f_{n+1}^t = f_{n+1}^s$. Then, the inductive hypothesis yields $D_{n+1}^t = D_{n+1}^s$. Conversely, suppose $D_{n+1}^t = D_{n+1}^s$. 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 $s \in D_n^t \in \mathcal{D}_n$ for all n. Since ι is an isomorphism, we have

$$f_n^t(\iota(D_n)) = \psi(\iota(t), \iota(D_n)) = \Gamma(t, D_n) = \Gamma(s, D_n) = \psi(\iota(s), \iota(D_n)) = f_n^s(\iota(D_n))$$

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. □

Theorem 1 and Step 1 imply that any component of the canonical types space is a valid IPM. Steps 1 and 2 imply that any valid IPM is isomorphic to a component of the canonical type space. □

6.3 Proof of Theorem 3

Lemma 8: *The function m_n is continuous for all $n \geq 1$.*

The proof is by induction. For $n = 1$, continuity amounts to showing that if t_n converges to t then $\Gamma(t_n, T)$ converges to $\Gamma(t, T)$ in the Hausdorff topology. This follows easily from Lemma 1(i) and the continuity of γ. To prove the inductive step, we will show that if $q : T \to Y$ is continuous, then $r : T \to \mathcal{H}(S \times Y)$ defined by $r(t) = \{(s, q(\hat{t})) \,|\, (s, \hat{t}) \in \rho(t)\}$ is also continuous. By Lemma 1(i), we need to show that if t_n converges to t then (i) for any convergent sequence $(s_n, y_n) \in r(t_n)$, $\lim(s_n, y_n) \in r(t)$ and (ii) for all $(s, y) \in r(t)$ there exists a convergent sequence $(s_{n_j}, y_{n_j}) \in r(t_{n_j})$ such that $\lim(s_{n_j}, y_{n_j}) \in r(t)$.

To prove (i), let $y_n = q(\hat{t}_n)$ for all n. Since $\hat{t}_n$ is in the compact set $q(T)$, it has a convergent subsequence. Without loss of generality, assume that this subsequence is the sequence itself. Hence, $\lim(s_n, y_n) = (\lim s_n, q(\lim \hat{t}_n))$. It follows from the continuity of ρ that $(\lim s_n, \lim \hat{t}_n) \in \rho(t)$ proving that $(\lim s_n, q(\lim \hat{t}_n)) \in r(t)$.

To prove (ii), let $(s, y) \in r(t)$. Hence, $(s, q(\hat{t})) \in r(t)$ for some $\hat{t} \in T$. Since ρ is continuous, there exists a subsequence t_{n_j} and $(s_{n_j}, \hat{t}_{n_j}) \in \rho(t_{n_j})$ (hence, $(s_{n_j}, q(\hat{t}_n j)) \in r(t_{n_j})$) such that $\lim(s_{n_j}, \hat{t}_{n_j}) \in \rho(t)$. Since q is continuous $q(\lim \hat{t}_{n_j}) = y$. Therefore, $\lim(s_{n_j}, q(\hat{t}_{n_j})) \in r(t)$ completing the proof of (ii). □

Let d be any decomposition of T such that for all $t \in T$, $\Gamma(t', D) = \Gamma(t, D)$ for all $D \in d$ and $t' \in D^t$. To prove that m is one-to-one implies M is valid, we will show that $m(t') = m(t)$ for all $' \in D^t$ and all $t \in T$. Note that $m_1(t) = \Gamma(t, T) = \bigcup_{D' \in d} \Gamma(t, D') = \bigcup_{D' \in d} \Gamma(t', D') = \Gamma(t', T) = m_1(t')$ whenever $t' \in D^t$. Next, assume that $m_n(t) = m_n(t')$ for all $t' \in D^t$ and $t \in T$. To complete the proof by induction, we will show that $m_{n+1}(t) = m_{n+1}(t')$ for all $t' \in D^t$ and $t \in T$. Suppose $(s, m_n(\hat{t})) \in \rho(t)$. That is, $\Gamma(t, \hat{t}) = s$. Then, there exists $\bar{t} \in D^{\hat{t}}$ such that $\Gamma(t', \bar{t}) = s$ and hence $(s, m_n(\bar{t})) \in \rho(t')$. By the inductive hypothesis, $m_n(\bar{t}) = m_n(\hat{t})$ and therefore $(s, m_n(\hat{t})) \in \rho(t')$ as desired.

To prove the converse, consider the decomposition $d = \{m^{-1}(t_\infty) \,|\, t_\infty \in T_\infty\}$. We will show that $\Gamma(t', D) = \Gamma(t, D)$ for all $D \in d$ and $t' \in D^t$. Suppose $(s, \hat{t}) \in \rho(t)$ and $m(t') = m(t)$. Then, for every n, there exists $\bar{t}_n \in T$ such that $m_n(\bar{t}_n)) = m_n(\hat{t})$ and $(s, \bar{t}_n) \in \rho(t)$. Since T is compact, we can assume without loss of generality that $\bar{t}_n$ converges to some $\bar{t}$. Since Γ is continuous, $(s, \bar{t}) \in \rho(t')$. Since each m_n is continuous $m_n(\bar{t}) = \lim_{k \geq n} m_n(\bar{t}_k) = m_n(\hat{t})$ for all $n \geq 1$. Therefore, $(s, \bar{t}) \in \Gamma(t')$ and $\bar{t} \in D^{\hat{t}}$ as desired. □

6.4 Proof of Theorem 4:

Lemma 9: *Let K be any compact subset of the reals. Let $\gamma : K \times K \to \mathbb{R}$ be weakly increasing in both arguments and continuous. Assume that $\gamma(x,\cdot) = \gamma(z,\cdot)$ implies $\gamma(\cdot,x) = \gamma(\cdot,z)$. Then, (K,γ) is a valid IPM if and only if for every $x, z \in K$, $\gamma(x,\cdot) = \gamma(z,\cdot)$ implies $x = z$.*

Proof: Suppose all the assumptions of the lemma are satisfied and there exists $x \neq z$ such that $\gamma(x,y) = \gamma(z,y)$ for all $y \in K$. Then, define the decomposition $\mathcal{D}$ as follows: for $y \notin \{x,z\}$, $D^y = \{y\}$ and $D^x = D^z = \{x,z\}$. It follows that $\gamma(x,\cdot) = \gamma(z,\cdot)$ and therefore $\gamma(\cdot,x) = \gamma(\cdot,z)$ and hence $\Gamma(w,D) = \Gamma(w',D)$ for all $w \in K$, $w' \in D^w$, and $D \in \mathcal{D}$. Hence, (K,γ) is not valid.

Next, suppose that (K,γ) is not valid. Then, there exists a decomposition $\mathcal{D}$ of K such that (i) there is $D \in \mathcal{D}$ with $x, z \in D$ such that $x \neq z$, (ii) $\Gamma(w,D) = \Gamma(w',D)$ for all $w \in K$, $w' \in D^w$, and $D \in \mathcal{D}$.

Let $\bar{D}$ denote the closure of $D \in \mathcal{D}$. The continuity of Γ ensures that

$$\Gamma(w,\bar{D}_2) = \Gamma(w',\bar{D}_2) \qquad (*)$$

for all $w, w' \in \bar{D}_1$ with $D_1 \in \mathcal{D}$ and $D_2 \in \mathcal{D}$. To see this, take $w, w' \in \bar{D}_1$ and $y \in \bar{D}_2$. By definition, there exists a sequence $(w_n, w'_n, y_n) \in D_1 \times D_1 \times D_2$ converging to (w,w',y). Moreover, there exists $y'_n \in D_2$ such that $\Gamma(w_n,y_n) = \Gamma(w'_n,y'_n)$ for all n. Since $\bar{D}_2$ is compact, y'_n has a convergent subsequence that converges to some $y' \in \bar{D}_2$. Assume, without loss of generality, that this subsequence is y'_n itself. Then, the continuity of Γ ensures $\Gamma(w,y) = \Gamma(w',y')$ as desired.

The weak monotonicity of γ in both arguments together with $(*)$ implies

$$\Gamma(\max \bar{D}_1, \max \bar{D}_2) = \Gamma(\min \bar{D}_1, \max \bar{D}_2) = \Gamma(\min \bar{D}_1, \min \bar{D}_2)$$

Then, monotonicity of Γ ensures $\gamma(w,y) = \gamma(w',y)$ for all $y \in K$ whenever $w, w' \in \bar{D}$, in particular, for $w = x$ and $w' = z$. □

Lemma 9 establishes that every reciprocity model is valid. Clearly, every reciprocity model is complete and reciprocating. Suppose that the SEM (T,γ) is isomorphic to some

reciprocity model (K, γ'). Then, since (K, γ') is valid, complete, and reciprocating, so is (T, γ).

To prove the converse, suppose (T, γ) is a valid, complete, and reciprocating SEM. Hence, $\succeq$, the kinder than relation is a preference relation. The continuity of γ yields the continuity of $\succeq$. Since T is a compact metric space, it separable and hence there exists a continuous real-valued function $x : T \to \mathbb{R}$ that represents $\succeq$. Let $K := x(T) = \{x(t) \,|\, t \in T\}$. Let $D^t = \{t' \in T \,|\, x(t') = x(t)\}$ and $\mathcal{D} = \{D^t \,|\, t \in T\}$. Clearly, $\mathcal{D}$ is a decomposition of T such that $\gamma(t', \cdot) = \gamma(t, \cdot)$ for all $t' \in D^t$. It follows from reciprocity that $\gamma(\cdot, t') = \gamma(\cdot, t)$ for all $t' \in D^t$, and therefore $\Gamma(t, D) = \Gamma(t', D)$ for all $t \in T$, $t' \in D^t$ and $D \in d$. Hence, validity implies each D^t is a singleton and therefore x is one-to-one. Then, the compactness of T ensures that K is compact and that x is a homeomorphism. Define, $\gamma(w, y) = \gamma(x^{-1}(w), x^{-1}(y))$ for all $w, y \in K$. Since x and γ are continuous, so is γ'. It follows that (K, γ') is a SEM and isomorphic to (T, γ). Since x represent $\succeq$ and (T, γ) is reciprocating, γ' is weakly increasing in both arguments. Finally, Lemma 9 and the fact that (T, γ) is valid imply that for every $x, z \in K$, $\gamma(x, y) = \gamma(z, y)$ implies $x = z$, proving that (K, γ') is a reciprocity model. □

6.5 Proofs of Theorems 4.1 and 4.2

Proof of Theorem 4.1:

Let (K, γ') be a convex-linear reciprocity model. Let $x_\lambda \in K$ be such that

$$\gamma(x_\lambda, \cdot) = \lambda\gamma(\max K, \cdot) + (1 - \lambda)\gamma(\min K, \cdot)$$

By convexity, such an x_λ exits for every $\lambda \in [0, 1]$. Since (K, γ') is a reciprocity model, $\gamma'(x, \cdot) = \gamma'(z, \cdot)$ implies $\gamma'(\cdot, x) = \gamma'(\cdot, z)$, and therefore $\Gamma'(x, \cdot) = \Gamma'(z, \cdot)$. Then, validity yields $x = z$. Hence, the mapping $\pi : \lambda \to x_\lambda$ is one-to-one. The continuity of γ' ensures that it is also continuous. Obviously, $x_1 = \max K$ and $x_0 = \min K$. Then, since $[0, 1]$ is a connected set and π is a continuous function its image must be connected. Therefore, π is onto. It follows that $\pi : [0, 1] \to K$ is a homeomorphism. Let $a = \gamma'(\pi(0), \pi(0))$,

$b+c+d = \gamma'(\pi(1),\pi(1)) - a$, $b = \gamma'(\pi(1),\pi(0)) - a$, and $c = \gamma'(\pi(0),\pi(1)) - a$. We claim that the IPM $([0,1],\gamma)$ where

$$\gamma(x,y) = a + bx + cy + dxy$$

is isomorphic to (K,γ'). By construction,

$$\begin{aligned}
\gamma'(\pi(1),\pi(1)) &= a+b+c+d = \gamma(1,1)\\
\gamma'(\pi(1),\pi(0)) &= a+b = \gamma(1,0)\\
\gamma'(\pi(0),\pi(1)) &= a+c = \gamma(0,1)\\
\gamma'(\pi(0),\pi(0)) &= a = \gamma(0,0)
\end{aligned}$$

Then, by the definition of π,

$$\gamma'(\pi(x),\pi(y)) = x\gamma'(\pi(1),\pi(y)) + (1-x)\gamma'(\pi(0),\pi(y))$$

Since π is onto, linearity implies

$$\begin{aligned}
&\gamma'(\pi(x),\pi(y)) = xy\gamma'(\pi(1),\pi(1)) + x(1-y)\gamma'(\pi(1),\pi(0))\\
&+ (1-x)y\gamma'(\pi(0),\pi(1)) + (1-x)(1-y)\gamma'(\pi(0),\pi(0)) = \gamma(x,y)
\end{aligned}$$

proving that π is an isomorphism form $([0,1],\gamma)$ to (K,γ').

To see that the conditions on the parameters b, c, and d are met, note $\frac{\partial\gamma}{\partial x} = b + dy$ and $\frac{\partial\gamma}{\partial y} = c + dx$. Reciprocity of (K,γ') together with the fact that π is increasing ensures that both of these partial derivatives are nonnegative for all values of x, y. The desired restrictions follow.

Verifying that every reciprocity model of this form is a convex-linear reciprocity model is straightforward and omitted. □

Proof of Theorem 4.2:

Let (K,γ') be a binary reciprocity model. Assume without loss of generality that $\gamma'(x,y) \in \{0,1\}$ for all $x, y \in K$ and consider the mapping $\pi : K \to \mathcal{C}(K,S)$ defined by $\pi(x)(y) = \Gamma'(x,y) = (\gamma'(x,y),\gamma'(x,y))$. Since Γ' is continuous, so is π. Hence, $\pi(K)$ is compact. Since $S = \{r_1,r_2\} \times \{r_1,r_2\}$, this means $\pi(K)$ is finite. Validity ensures

that π is one-to-one. Hence, K is finite and contains at least two elements. Then, set $K = \{x_1, \ldots, x_k\}$, where $x_{i+1} \succeq x_i$ for all $i = 1, \ldots k-1$. Let $K_i = \{x_1, \ldots, x_i\}$ and $K_0 = \emptyset$. It follows from reciprocity that for each x_i there exists some $j = 0, \ldots, k$ such that $\gamma'(x_i, y) = 0$ if $y \in K_j$ and $\gamma'(x_i, y) = 1$ if $y \in K \backslash K_j$. Let $\pi(x_i)$ denote this j. Reciprocity implies that if $\pi(x_i) = \pi(x_l)$ then $\gamma'(\cdot, x_i) = \gamma'(\cdot, x_l)$, and hence $\Gamma'(x_i, \cdot) = (\gamma'(x_i, \cdot), \gamma'(\cdot, x_i)) = (\gamma'(x_l, \cdot), \gamma'(\cdot, x_l)) = \Gamma'(x_l, \cdot)$. Then, by validity, $x_i = x_l$. It follows that the mapping $\pi : K \to \{0, \ldots, k\}$ is one-to-one. Since K has k elements, there exists a unique $n \in \{0, \ldots, k\}$ such that $n \notin \pi(K)$. Note that for $i > n$, $\gamma'(x_i, x_j) = 1$ if and only if $i + j > n$, while for $i \leq k$, $\gamma'(x_i, x_j) = 1$ if and only if $i + j > n + 1$. Hence,

$$\gamma'(x_i, x_j) = G(i + j + G(i, n), k + 1)$$

where $G(i, j) = 1$ if $i > j$ and $G(i, j) = 0$ otherwise. Define $\gamma : \{1, \ldots, k\} \to \{0, 1\}$ as follows: $\gamma(i, j) = \gamma'(x_i, x_j)$. Hence, $\iota(x_i) = i$ defines an isomorphism from (K, γ') to $(\{1, \ldots, k\}, \gamma)$

Verifying that every reciprocity model of this form is a binary reciprocity model is straightforward and omitted..

□

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. Brandenburger A. and E. Dekel (1993): Hierarchies of Beliefs and Common Knowledge, *Journal of Economic Theory* , 59, 1993, 189-198.

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

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

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

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

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

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

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

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

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

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

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

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

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