

## Testing game theory

JÖRGEN W. WEIBULL\*  
DEPARTMENT OF ECONOMICS  
BOSTON UNIVERSITY

5 February 2004.

**ABSTRACT.** Experimentalists frequently claim that human subjects in the laboratory violate game-theoretic predictions. It is here argued that this claim is usually premature. The paper elaborates on this theme by way of raising some conceptual and methodological issues in connection with the very definition of a game and of players' preferences, in particular with respect to potential context dependence, interpersonal preference dependence, backward induction and incomplete information.

*JEL-codes:* A10, C70, C72, C90.

*Keywords:* game theory, experiments, equilibrium.

### 1. INTRODUCTION

An important development in economics is the emergence of experimental economics, and Werner Güth has been one of its pioneers. Moving from arm-chair theorizing to controlled laboratory experiments may be as important a step in the development of economic theory as it once was for the natural sciences to move from Aristotelian scholastic speculation to modern empirical science.<sup>1</sup>

The first experiments in game theory were carried out in the early fifties. However, a new wave of game experiments began in the mid seventies, and Güth, Schmittberger and Schwarze (1982) pioneered experimental work on so-called ultimatum bargaining situations. For surveys of such experiments, and for introductions to experimental game theory more generally, see Güth and Thietz (1990), Bolton and Zwick (1995),

---

\*This paper is a major revision of SSE WP 382, May 2000. I am grateful for helpful comments from Ana Ania, Geir Asheim, Kaushik Basu, Larry Blume, Vincent Crawford, Martin Dufwenberg, David Easley, Tore Ellingsen, Ernst Fehr, Jean-Michel Grandmont, Thorsten Hens, Jens Josephson, Sendhil Mullainathan, Rosemarie Nagel, Al Roth, Maria Saez-Marti, Larry Samuelson, Martin Shubik, Jon-Thor Sturlason, Sylvain Sorin, Fernando Vega-Redondo and Shmuel Zamir, and to seminar participants at presentation of various drafts of this paper

<sup>1</sup>The likelihood for success, however, may be smaller, in view of the complexity of human choice behavior and strategic interaction.

Kagel and Roth (1995), Zamir (2000), the special issue of the *Journal of Economic Theory* in 2002 devoted to experimental game theory, and the Camerer (2003) book.

The present note discusses some methodological and conceptual issues that arise when non-cooperative game theory is used for positive analysis of human strategic interaction. In particular, in the experimental literature, it has many times been claimed that certain game-theoretic solutions - such as Nash equilibrium and subgame perfect equilibrium - have been violated in laboratory experiments.<sup>2</sup> While it may well be true that human subjects do not behave according to these solutions in many situations, few experiments actually provide evidence for this. Especially in the early literature, experimentalists did not make any effort to elicit the subjects' preferences, despite the fact that these preferences constitute an integral part of the very definition of a game. Instead, it has been customary to simply assume that subjects care only about their own material gains and losses. In later studies, subjects' preferences were allowed to also depend on the "fairness" of the resulting vector of material gains and losses to all subjects. However, recent experiments, discussed below, suggest that even this is sometimes too restrictive — subjects' ranking of alternatives may depend on other parts of the game form, a phenomenon here called "context dependence."

In applications of non-cooperative game theory, the game is not only meant to represent the strategic interaction as viewed by the analyst, but also as viewed by the players — it is even frequently assumed that the game is common knowledge to the players. Indeed, a variety of epistemic models have been built in order to analyze the rationality and knowledge assumptions involved in game-theoretic analysis — under the classical interpretation that the game in question is played exactly once by rational players. The extent and exact form of knowledge assumed on behalf of the players varies across game forms, solutions, and on the epistemic model in question, see Tan and Werlang (1988), Reny (1993), Aumann and Brandenburger (1995), Aumann (1995), Ben-Porath (1997) and Asheim (2002). Being deductive, such epistemic models of games can of course not be empirically falsified as such, only their assumptions, which are known to be strong idealizations. So what can then be tested? One can test whether the theoretical predictions are at least approximately correct in environments which approximate the assumptions. Such testing is important, because this is the way game theory is used in economics and the other social sciences.<sup>3</sup>

This essay is somewhat discursive and philosophical, and contains no theorems. I hope, though, that it sheds some light on the very definition of a non-cooperative game, on the empirically relevant possibilities of context dependent preferences and

---

<sup>2</sup>The number of citations that could be made here is so large that any selection would be arbitrary.

<sup>3</sup>In recent years, evolutionary alternatives to these epistemic models have been developed. However, those models are not discussed here.

interpersonal preference dependence, on backward induction and on incomplete information modelling. The interested reader is recommended to read Levine (1998), Sprumont (2000), Ray and Zhou (2001) and Binmore *et al.* (2002) for other discussions of some of these, and related, issues (connections to these earlier studies are briefly commented below).

The discussion is organized as follows. Section 2 pins down some terminology and notation. In particular, a notion of “game protocol” is introduced. Section 3 applies this machinery to a class of very simple ultimatum bargaining situations. Section 4 discusses backward induction more in general, in particular how to reconcile it with context-dependent preferences. Section 5 discusses briefly interpersonal preference dependence. Section 6 shows how the model in Levine (1998) can be used to address the issues discussed in the two preceding sections, and section 7 concludes.

## 2. GAMES AND GAME PROTOCOLS

The present discussion is focused on a slight generalization of finite games in extensive form, as defined in Kuhn (1950,1953).<sup>4</sup> Such a game is a mathematical object that contains as its basic building block a directed tree, consisting of finitely many nodes (or vertices) and branches (or edges). A *play*  $\tau$  of the game is a “route” through the tree, starting at its initial node and ending at one of the *end nodes*  $\omega$ . A node  $k'$  is a *successor* of a node  $k$  if there is a play that leads first to  $k$  and then to  $k'$ . Moreover, each end-node is reached by exactly one play of the game, and each play reaches exactly one end-node. Let  $\Omega$  denote the set of end-nodes and  $T$  the set of plays. We then have  $|\Omega| = |T| < +\infty$ .

The set of non-end nodes is partitioned into player subsets, and each player subset is partitioned into information sets for that player role. In each information set, the number of outgoing branches from each node is the same, and the set of outgoing branches from an information set is divided into equivalence classes, the *moves* available to the player at that information set, such that every equivalence class contains exactly one outgoing branch from each node in the information set. A *choice* at an information set is a probability distribution over the moves available at the information set. In games with exogenous random moves, one of the players is “nature,” and all information sets for this “non-personal” player are singleton sets with fixed probabilities attached to each outgoing branch.

A *pure strategy* for a personal player role is a function that assigns one move to each of the role’s information sets. The *outcome* of a strategy profile is the probability distribution induced on the set  $\Omega$  of end-nodes, or, equivalently, on the set  $T$  of plays.

Since humans sometimes exhibit social preferences, that is their choices are in part

---

<sup>4</sup>See Ritzberger (2002) for a rigorous definition and analysis of finite extensive-form games.

driven by concerns for others, it is useful to allow for the possibility of *passive players* (or *dummy* players), that is, player roles with empty player sets, but where the player may be affected by the choices made by other, *active*, players. Relevant examples are the so-called dictator games, where one player is active and one is passive. More generally, an active player may be passive in certain subgames and yet influence active players' preferences in the subgame.

The ingredients described so far together make up a *game form*.

**2.1. Games.** A game form becomes a game when the (personal) player roles are endowed with preferences. More exactly, in standard non-cooperative game theory, each player  $i = 1, 2, \dots, n$  is assumed to have preferences over the unit simplex

$$\Delta(\Omega) = \Delta(T) = \left\{ p \in \mathbb{R}_+^{|\Omega|} : \sum_{i=1}^{|\Omega|} p_i = 1 \right\}$$

of lotteries over end-nodes, or, equivalently, plays, satisfying the von Neumann-Morgenstern axioms.<sup>5</sup> Hence, for each player  $i$  there exists a real-valued function  $\pi_i$  with domain  $\Omega$ , or  $T$ , such that player  $i$  prefers one lottery over another if and only if the expected value of the function  $\pi_i$  is higher in the first lottery than in the second. Such a function  $\pi_i : \Omega \rightarrow \mathbb{R}$  (unique up to a positive affine transformation) will here be called the *Bernoulli function* of player  $i$ . The number  $\pi_i(\omega)$  is usually called the *payoff* to player  $i$  at end node  $\omega$ .<sup>6</sup> If  $\Phi$  is a game form, then the pair  $\Gamma = (\Phi, \pi)$ , where  $\pi$  denotes a combined Bernoulli function  $\pi : \Omega \rightarrow \mathbb{R}^n$ , constitutes a finite extensive-form *game*.

**2.2. Game protocols.** In virtually all applications of game theory, including laboratory experiments, each play results in well-defined material consequences for the players. In applications to economics, and in most laboratory experiments, these material consequences are monetary gains or losses, in which case these are usually called *monetary payoffs* — not to be confounded with game theorists' definition of payoffs as Bernoulli function values.

In order to facilitate discussions of the effects of changed material or monetary payoffs, it is useful to introduce a name for game forms with specified material consequences. Hence, by a *game protocol* is here meant a pair  $(\Phi, \gamma)$ , where  $\gamma$  is a function that maps end-nodes  $\omega \in \Omega$  (or, equivalently, plays  $\tau \in T$ ) to material consequences

---

<sup>5</sup>Standard game theory can thus be criticized for its reliance on the von Neumann Morgenstern axioms — an empirically valid critique that will not be discussed here, though.

<sup>6</sup>By contrast, by a *payoff function* is usually meant the induced mapping from strategy profiles to Bernoulli function values.

$c \in C$ , for some set  $C$  rich enough to represent relevant aspects of the material consequences in question. If the material consequences are monetary gains and losses to the  $n$  players in the game form, then we may thus take  $C$  to be a subset of  $\mathbb{R}^n$ .<sup>7</sup>

The formal machinery of non-cooperative game theory does not require that a player's payoff value  $\pi_i(\omega)$  at an end node  $\omega$  be a function of the material consequences at that node. Indeed, two plays resulting in the same material payoffs to all players may well differ in terms of information sets reached, choices made and not made along the play, etc. — aspects that may be relevant for players' preferences and hence influence their Bernoulli functions. Standard game theory only requires the *existence* of a Bernoulli function  $\pi_i$  for each (personal) player  $i$ . Indeed, several laboratory experiments have convincingly — though perhaps not surprisingly for the non-economist — shown that human subjects' preferences are not driven only by their own monetary payoffs.<sup>8</sup>

### 3. MINI ULTIMATUM PROTOCOLS

A class of game protocols that have been much studied in the laboratory are those associated with *ultimatum bargaining protocols*. These two-player game protocols represent strategic interactions where the subject in role  $A$ , the *proposer*, makes a suggestion to the subject in role  $B$ , the *responder*, for how to split a fixed sum of money. The responder may accept or reject the proposal. If accepted, the sum is split as proposed. If rejected, both subjects receive nothing.

Figure 1 shows the extensive form of such a simplified strategic interaction, a *mini ultimatum protocol*, where 100 tokens are to be divided between two parties, a *proposer* and a *responder*. The proposer has only two choices, either to keep  $x$  tokens for herself, her *outside option*, or to offer the responder  $100 - y$  tokens. In the latter case, the responder has a binary choice, whether to accept or reject the division  $(y, 100 - y)$ . In case  $B$  rejects, both players receive zero tokens. Hence, the game form has three plays:  $T = \{\tau_1, \tau_2, \tau_3\}$ . In play  $\tau_1$ ,  $A$  chooses the division  $(x, 100 - x)$  and play stops at end node  $\omega_1$ . In play  $\tau_2$ ,  $A$  proposes the division  $(y, 100 - y)$ ,  $B$  accepts this and play stops at end node  $\omega_2$ . In play  $\tau_3$ ,  $A$  proposes  $(y, 100 - y)$ ,  $B$  rejects this and play stops at  $\omega_3$ . The numbers  $x$  and  $y$  are fixed and given, and known by the player subjects, where  $0 < x < y < 100$ . At the end of the experiment, tokens are exchanged for money, at a pre-set exchange rate.

---

<sup>7</sup>This is similar to consumer theory, where the consumption space is supposed to be rich enough to represent all relevant aspects of consumption alternatives.

<sup>8</sup>For prominent examples, see Güth, Schmittberger and Schwarze (1982), Binmore, Shaked and Sutton (1985), Ochs and Roth (1989), Fehr and Schmidt (1999), Bolton and Ockenfels (2000), and Brandts and Solà (2000), Binmore *et al* (2002), Falk *et al* (2003).

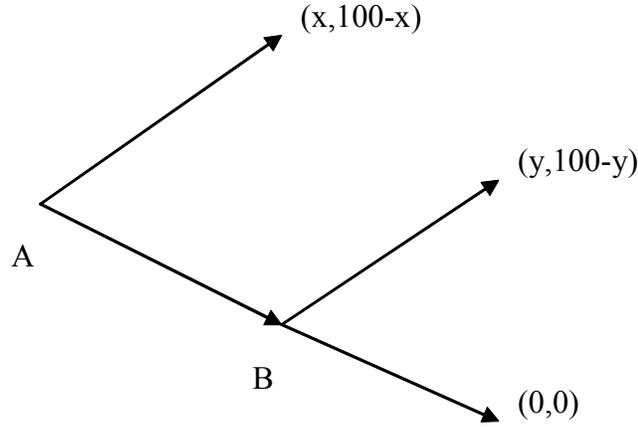


Figure 1: A mini ultimatum game protocol.

In the early experimental literature it was presumed that the payoff values  $\pi_i(\omega)$  to the subjects are monotone functions of their own monetary payoffs. Hence, in strategic interactions like this, it was claimed that non-cooperative game theory predict play  $\tau_2$ , namely, that  $A$  will propose the division  $(y, 100 - y)$  and that  $B$  will accept this. This is of course the unique subgame-perfect equilibrium of the game that defined by such preferences. The implicit hypothesis in this early literature is fivefold:

- (H1) the responder prefers play  $\tau_2$  over play  $\tau_3$ ,
- (H2) the responder is rational in the sense of playing optimally, according to his or her preferences,
- (H3) the proposer knows that H1 and H2 hold (or at least believes that they hold with a sufficiently high probability),
- (H4) the proposer prefers  $\tau_2$  over  $\tau_1$ , and
- (H5) the proposer is rational in the sense of acting optimally, in accordance with his or her knowledge and preferences.<sup>9</sup>

---

<sup>9</sup>In more complex games, hypotheses H2 and H5, which here seem innocuous, may actually be highly implausible. For example, in chess we know that H2 and H5 do not hold: no human player knows how to play optimally (presuming a strict preference for winning) from all game positions on the board.

A large number of laboratory experiments with more complex ultimatum bargaining situations of this sort have shown that many proposer subjects instead offer sizable shares to the responder, and that many responder subjects reject small shares if offered. In the present mini ultimatum game protocol, this corresponds to play  $\tau_1$ . Such findings were initially interpreted as rejections of the subgame perfection solution concept. What was rejected was the combined *preference-cum-knowledge* hypothesis H1-5 given above. This is not surprising, since neither hypothesis H1 nor H3 is true for all subjects.

In the present example, let  $\succeq_A$  be a proposer subject's preferences over the set  $\Delta(\Omega)$ , and let  $\succeq_B$  be a responder subject's preferences over the same set. For example, suppose  $x = 50$  and  $y = 90$ . A subject in player role  $A$  may then have the preference  $\tau_2 \succ_A \tau_1 \succ_A \tau_3$ , and the subject in role  $B$  may have the preference  $\tau_1 \succ_B \tau_3 \succ_B \tau_2$ .<sup>10</sup> Indeed, such preferences are consistent with many subjects' behavior in laboratory experiments. All games in the associated game class (that is, with compatible Bernoulli functions) have the 50/50 split, that is, play  $\tau_1$ , as the unique subgame perfect outcome.

In an experimental study of a variety of mini ultimatum protocols slightly more complex than the one in figure 1, Falk *et al.* (2003) found that the responder rejection rate (across 90 subjects) depends not only on the current offer they faced, but also on the "outside option" available to the proposer.<sup>11</sup> In the context of the present example:  $B$ 's ranking of plays  $\tau_2$  and  $\tau_3$  may well depend on the material consequences of play  $\tau_1$ .

As indicated above, this observation has implications for backward induction arguments: a change in one part of a game protocol may change players' preferences in another part of the game protocol, even if the first part cannot be reached from the second. This issue is addressed in the next section.

#### 4. BACKWARD INDUCTION

In a given game form  $\Phi$ , let  $K_0$  be the subset of nodes  $k$  such that (i)  $k$  is either a move by nature or  $\{k\}$  is an information set of a personal player, and (ii) no information set in  $\Phi$  contains both a successor node and a non-successor node to  $k$ . Each node  $k \in K_0$  is the initial node of a *subform*, a game form  $\Phi_k$ . For any such node  $k$ , let the associated *subprotocol* be defined as the game protocol  $(\Phi_k, \gamma_k)$ , where  $\gamma_k$  is the restriction of  $\gamma$  to the subset  $\Omega_k \subset \Omega$  of end nodes that succeed node  $k$ . So far, all well. The subtlety arises when we are to define subgames, since for any given subprotocol there are two distinct candidates claiming that name.

<sup>10</sup>Here each play is short-hand notation for the lottery that places unit probability mass on it.

<sup>11</sup>In Falk *et al.* (2002), also the "outside option" was subject to the responder's acceptance.

First, there is the following definition of a subgame: if  $\Gamma = (\Phi, \pi)$  is a game and  $k \in K_0$ , then the *subgame* at  $k$  is the game  $\Gamma_k = (\Phi_k, \tilde{\pi})$ , where  $\tilde{\pi}$  is the restriction of  $\pi$  to the subset  $\Omega_k$ . In other words, the Bernoulli-function values in  $\Gamma_k$  coincide with those in  $\Gamma$  at all end-nodes in  $\Omega_k$ . This is the *context-dependent* definition of a subgame. In this definition, players' preferences in the subgame, represented by  $\tilde{\pi}$ , may depend on parts of the full game protocol  $(\Phi, \gamma)$  outside the subprotocol  $(\Phi_k, \gamma_k)$  in question. For example,  $\tilde{\pi}$  may depend on choices available at along the unique play leading up the node  $k$  and/or on material payoffs at end-nodes not in  $\Omega_k$ . This approach treats the full game protocol as the relevant context for all players' decisions at all points in the game protocol.

A second candidate for the title of "subgame" at a node  $k \in K_0$  is the game  $\Gamma^\circ = (\Phi_k, \pi^\circ)$  that is obtained if the subprotocol  $(\Phi_k, \gamma_k)$  is played in isolation, that is, *beginning* at node  $k$  as the initial node and without the "context" of the rest of  $(\Phi, \gamma)$ . We will call  $\Gamma^\circ = (\Phi_k, \pi^\circ)$  the *isolated subgame* at  $k$ .

As an illustration of this distinction, consider the subform in figure 1 beginning at the node  $k$  where player  $B$  has to accept or reject the proposal  $(y, 100 - y)$ . Viewed in isolation, this is a one-player game protocol, where the unique active player ( $B$ ) has a binary choice of either (a) receiving  $100 - y$  tokens while  $y$  tokens are given to a passive player  $A$ , or (b) no tokens to any of the two players. I guess an overwhelming majority of subjects in this isolated game protocol would choose the first option. However, we know that many subjects in player role  $B$  in the full game protocol in figure 1 choose the second option. Taking their behavior as their revealed preference, this means that  $\tilde{\pi} \neq \pi^\circ$ . Indeed, it is an empirical question whether  $\pi^\circ = \tilde{\pi}$ . The above-mentioned observations in Falk et al. (2003), if taken as revealed preferences, show that  $\tilde{\pi} \neq \pi^\circ$ . Hence, the distinction between subgames and isolated subgames may be critical.

Since the full game protocol is supposed to represent the relevant decision context for the players' decision making, it is this author's opinion that backward induction should be applied to the full game protocol, with all players' preferences defined *in this protocol*. In particular, when applying subgame perfection, one should use the above definition of a subgame, and not that of an isolated subgame. Formally, a subgame perfect equilibrium is then a strategy profile that induces a Nash equilibrium on every subgame  $\Gamma_k = (\Phi_k, \tilde{\pi})$ .

However, much of the game-theoretic literature seems to ignore this distinction. For a recent example, see Ray and Zhou (2001), where it is implicitly assumed that  $\tilde{\pi} = \pi^\circ$ .<sup>12</sup> However, other researchers have provided experimental evidence against

---

<sup>12</sup>Likewise, Sprumont (2000) presumes preferences in normal-form games to be context independent in the same way: a player's ranking of pure strategies in a subset remains the player's ranking

backward induction when carried out in terms of the isolated subgames, that is by using  $\pi^o$  instead of  $\tilde{\pi}$ , see Binmore et al. (2002). Hence, what they reject is the combined hypothesis that the subjects' play is compatible with backward induction, as described here, *and* that their preferences satisfy  $\tilde{\pi} = \pi^o$ .

Note that the suggested approach — to base backward induction (and all other analysis) on the preferences in the full game protocol — is not a critique of “Kuhn’s algorithm,” the usual way of solving finite games of perfect information by way of successively replacing each final decision node in the game tree by an end node with a payoff vector that equals the expected payoff vector achieved by some optimal move by the player at the decision node in question. All that is suggested here is that the payoffs should then be Bernoulli function values as defined from players’ preferences in the full game protocol.

Context-dependence in preference formation may have many causes. It may be that human subjects in player roles have opinions about actions taken and not taken on the way to the information set in question, with or without regard to the possible intentions behind those actions. It may also be that players have social preferences that depend on options available to others outside the subprotocol in question. However, for the purposes of game-theoretic analysis it does not matter what the causes are, as long as players’ preferences in the full game protocol are well-defined. The analyst’s task to elicit the preferences of subjects in player roles of a given game protocol is in general not easy. It is particularly difficult if subjects have interpersonally dependent preferences, that is, if their rankings of outcomes depend in part on their expectations of other player subjects’ rankings of outcomes, which depend on those other subjects’ expectations of the others’ rankings etc. — the topic of the next two sections.

## 5. INTERPERSONAL PREFERENCE DEPENDENCE

The elicitation of players’ preferences raises a fundamental issue in the very definition of a game, namely whether a player’s preferences may depend on (knowledge of, or beliefs about) another player’s preferences, which in its turn may depend on (knowledge of, or beliefs about) the first player’s preferences etc. Such potential interpersonal preference dependence is disturbing since it makes the domain of preferences unclear — the game protocol is then not an exhaustive representation of the interactive situation — and yet such interdependence might realistically exist in some interactions.

---

in the reduced game in which these subsets are the full strategy sets.

**5.1. Example.** In order to illustrate this possibility, consider the game protocol in Figure 2 below.<sup>13</sup> There, a mini “dictator game” form follows upon the tossing of a fair coin deciding who of the two players should be the “dictator.” The game form has four plays:  $\tau_1$ , where  $A$  is the dictator and decides that they get 50 tokens each;  $\tau_2$ , where  $A$  is the dictator and decides that she will get 70 tokens while  $B$  will get only 10,  $\tau_3$  where  $B$  is the dictator and decides that they will get 50 tokens each; and, finally,  $\tau_4$ , where  $B$  is the dictator and decides that he will get 70 tokens while  $A$  will get only 10. (The token sum is thus 100 in the even splits and 80 in the uneven splits.)

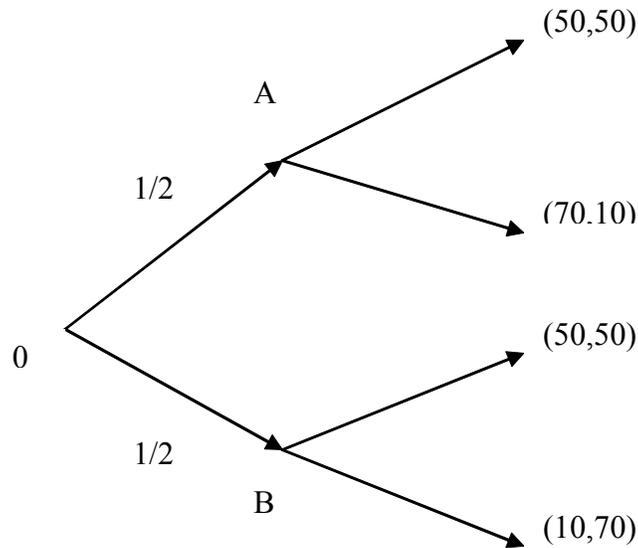


Figure 2: A mini random dictator game protocol.

Suppose there is one subject in each of the two player roles, and the experimentalist wants to elicit their preferences in order to identify the game they are playing. Suppose the experimentalist is able to find out each subject’s ordinal preference ranking of the four plays in the above game protocol. Suppose the subject in player role  $A$  ranks play  $\tau_1$  (being dictator and splitting 50/50) highest, while the subject in role

<sup>13</sup>I am grateful to Al Roth for suggesting this game protocol, which is simpler than the one I originally used.

$B$  ranks  $\tau_4$  first (being the dictator and keeping 70 for himself). Can the experimentalist then conclude that if the game (or more precisely, game class) were known by both subjects, the subjects would play according to their stated preferences? Not necessarily. Suppose for instance, that the subject in player role  $A$ , when learning about  $B$ 's "selfish" preference, changes her own ranking and now prefers to keep 70 tokens for herself if called upon to act as dictator.

Suppose now that the experimentalist anticipates this possibility of preference interdependence, and therefore instead proceeds as follows: the experimentalist asks subject  $A$  to state her ordinal ranking of the four plays for each of the 24 possible (strict) ordinal rankings that  $B$  may have, and likewise for subject  $B$ . Having done this, the experimentalists looks for matching preference orderings, that is a pair of orderings such that each subject's ordering applies to the other subject's ordering. This way, the experimentalist may end up with no matching pair, one matching pair, or several matching pairs. For instance, subject  $A$  may rank play  $\tau_1$  (being dictator and splitting 50/50) highest if  $B$  prefers play  $\tau_3$  (being dictator and splitting 50/50) over play  $\tau_4$ , and likewise for subject  $B$ . Presumably, two such subjects would not change their own ranking even when told the other's. However, the same two subjects may also have another matching pair of rankings, such as  $A$  ranking play  $\tau_2$  (being dictator and keeping 70 tokens for herself) highest if subject  $B$  prefers play  $\tau_4$  (being dictator and keeping 70 tokens for himself) over play  $\tau_3$ , and likewise for subject  $B$ . This corresponds to a behaviorally distinct class of games, so the experimentalist has ended up with two distinct game classes for one and the same game protocol and pair of subjects.

**5.2. Games of incomplete information.** Can this kind of preference interdependence be avoided by way of modelling the interaction as a game of incomplete information, and using the Harsanyi approach of transforming that game into a "meta game" of complete but imperfect information, with a common prior? The feasibility of this program seems to be an empirical question, for two reasons. First, it is an empirical question if such a game exists for given subjects in the player roles of a given game protocol, since interpersonal preference dependence may arise also in the resulting meta game of imperfect information: subjects may alter their own rankings of outcomes once they learn about the others' rankings.<sup>14</sup> Secondly, it is doubtful if human subjects will understand the so constructed meta game, since such a game is usually quite abstract, nor is it clear that they will agree on a common prior (or even understand what a prior is). In view of these difficulties, the game theorist may abandon direct preference elicitation, and work under the (falsifiable) hypothesis that

---

<sup>14</sup>This existence problem is distinct from the related existence problem analyzed in Mertens and Zamir (1985)

the subjects act *as if* they were players with hypothesized preferences in such a meta game. The subsequent section illustrates that route by way of applying the simple and yet rich incomplete-information model of interpersonal preference dependence suggested by David Levine (1998). As will be seen, a certain form of individual preference elicitation is possible even in that setting.

## 6. ALTRUISM-DRIVEN INTERPERSONAL PREFERENCE DEPENDENCE

In a two-player game protocol, such as those in figures 1 and 2, let  $x_A(\omega)$  and  $x_B(\omega)$  be the monetary payoffs to the two players,  $A$  and  $B$ , at each end-node  $\omega \in \Omega$ . The players are drawn from one and the same population (of, say, subject in an experiment). The type space is a subset  $\Theta$  of  $\mathbb{R}$ , and players' types are i.i.d. draws from a c.d.f.  $F : \Theta \rightarrow [0, 1]$ .<sup>15</sup> In the game protocol of the associated meta game, where "nature" first chooses the two players' types, an end node is a triplet  $(\omega, a, b) \in \Omega \times \Theta^2$ , where  $a$  and  $b$  are  $A$ 's and  $B$ 's types and  $\Omega$  is the set of end nodes in the underlying, or "basic," game protocol that neglects nature's draws (as in figures 1 and 2).

The Bernoulli functions of players  $A$  and  $B$  in this meta game, when  $A$  is of type  $a$  and  $B$  is of type  $b$ , are taken to be of the form

$$\pi_A(\omega, a, b) = x_A(\omega) + W(a, b)x_B(\omega) \quad (1)$$

and

$$\pi_B(\omega, a, b) = x_B(\omega) + W(b, a)x_A(\omega) \quad (2)$$

where  $W : \Theta^2 \rightarrow \mathbb{R}$  is a function that attaches a *relative weight* to the other player's material payoff (as compared with the unit weight attached to the player's own material payoff). A positive weight thus represents altruism towards the other player while a negative weight represents spite. Levine (1998) uses the following functional form:

$$W(\theta, \theta') = \frac{\theta + \lambda\theta'}{1 + \lambda} \quad \forall \theta, \theta' \in \Theta \quad (3)$$

where  $\lambda \in [0, 1]$  is a parameter that reflects interpersonal preference dependence: the weight  $W(\theta, \theta')$  placed on the other player's material payoff depends non-negatively on the other player's type  $\theta'$  and is more sensitive to the other player's type the larger  $\lambda$  is.

**6.1. Mini ultimatum games.** Let us first apply this approach to the game protocol in figure 1, for  $0 < x < y < 100$ . Suppose thus that "nature" first chooses the two players' types and reveals each player's type privately to the player in question.

---

<sup>15</sup>Levine (1998) sets  $\Theta = (-1, +1)$ , and assumes  $F$  to have finite support.

In the notation of figure 1, an outcome in the meta-game protocol is a triplet  $(\omega, a, b)$ , where  $\omega \in \{\omega_1, \omega_2, \omega_3\}$ ,  $a$  is  $A$ 's type and  $b$  is  $B$ 's type.

Let  $\lambda \in [0, 1]$ , and let  $F : \mathbb{R} \rightarrow [0, 1]$  be a c.d.f. with finite mean value  $\bar{\theta}$ . Player  $A$ 's Bernoulli function is then

$$\pi_A(\omega_1, a, b) = x + \frac{a + \lambda b}{1 + \lambda} (100 - x) \quad (4)$$

$$\pi_A(\omega_2, a, b) = y + \frac{a + \lambda b}{1 + \lambda} (100 - y) \quad (5)$$

and  $\pi_A(\omega_3, a, b) = 0$  for all  $a$  and  $b$ . Player  $B$ 's Bernoulli function is similarly defined:

$$\pi_B(\omega_1, a, b) = 100 - x + \frac{b + \lambda a}{1 + \lambda} x \quad (6)$$

$$\pi_B(\omega_2, a, b) = 100 - y + \frac{b + \lambda a}{1 + \lambda} y \quad (7)$$

and  $\pi_B(\omega_3, a, b) = 0$  for all  $a$  and  $b$ .

Note, however, the informational asymmetry in the two player's expectation formation at their respective decision nodes in figure 1. While player  $A$  forms an unconditional expectation of  $B$ 's type  $b$  when making her choice, player  $B$ , when making his choice, conditions his expectation of  $A$ 's type on the observation that  $A$  has chosen not to take her outside option  $(x, 100 - x)$ . The above description specifies a meta game of complete but imperfect information, defined by the following data:  $x, y, \lambda$  and  $F$ .

Suppose that we have laboratory data for given monetary payoffs  $x$  and  $y$ , and for human subjects who have been randomly and anonymously matched to play the two player roles. Let  $p = (p_1, p_2, p_3) \in \Delta(\{\omega_1, \omega_2, \omega_3\})$  be the empirical outcome in the underlying game protocol, that is, the observed population frequencies of the three plays of the game protocol in figure 1.<sup>16</sup> For the sake of clarity of exposition, suppose  $p_1 < 1$ , that is, at least one proposer subject has chosen not to take the outside option. What such outcomes  $p$  are consistent with play of a sequential equilibrium in the meta game?

In order to answer this question, first note that "reject" is sequentially rational for player  $B$  if and only if

$$b \leq -(1 + \lambda) \left( \frac{100}{y} - 1 \right) - \lambda \mathbb{E}_B[a], \quad (8)$$

---

<sup>16</sup>The subsequent equilibrium analysis would seem relevant if all subjects have had many learning rounds first, where they get familiar with the rules of the game and where they can learn about aggregate behaviors in the subject pool.

where  $\mathbb{E}_B [a]$  is  $B$ 's conditional expectation of  $A$ 's type  $a \in \Theta$ . A necessary condition for  $p$  to be compatible with sequential equilibrium is thus

$$F \left[ (1 + \lambda) \left( 1 - \frac{100}{y} \right) - \lambda \mathbb{E}_B [a] \right] = \frac{p_3}{1 - p_1}, \quad (9)$$

that is, the equilibrium probability of rejection should equal the empirical rejection frequency.

Secondly, it is sequentially rational for player  $A$  to forego the outside option  $(x, 100 - x)$  if and only if

$$a \leq (1 + \lambda) \frac{y - x - yq}{y - x + (100 - y)q} - \lambda \mathbb{E}_A [b], \quad (10)$$

where  $q$  is the equilibrium probability that  $B$  will reject the offer  $(y, 100 - y)$ . Hence, another necessary condition for the outcome  $p$  to be compatible with sequential equilibrium is

$$F \left[ \frac{(1 + \lambda)(y - x - yq)}{y - x + (100 - y)q} - \lambda \mathbb{E}_A [b] \right] = 1 - p_1, \quad (11)$$

where  $q = p_3 / (1 - p_1)$ . In other words, the equilibrium probability that  $A$  will not take the outside option should equal the empirical frequency of this event.

Thirdly, the consistency condition in the definition of sequential equilibrium requires that player  $A$ 's expectation of  $B$ 's type should equal the unconditional mean-value under the type distribution  $F$ ,

$$\mathbb{E}_A [b] = \bar{\theta}, \quad (12)$$

and that  $B$ 's expectation of  $A$ 's type, when called upon to make a move, should equal the conditional mean-value under  $F$ , given that  $A$ 's type  $a$  satisfies (10):

$$\mathbb{E}_B [a] = G \left[ \frac{(1 + \lambda)(y - x - yq)}{y - x + (100 - y)q} - \lambda \bar{\theta} \right], \quad (13)$$

where  $G : \mathbb{R} \rightarrow \mathbb{R}$  is the truncated mean-value function associated with  $F$ , defined by

$$G(t) = \frac{1}{F(t)} \int_{-\infty}^t s dF(s). \quad (14)$$

Indeed, given the meta-game data  $x, y, \lambda$  and  $F$ , the four equations (9) and (11)-(13) are together necessary and sufficient for an outcome  $p$  in the underlying game form to be compatible with sequential equilibrium in the meta game (granted  $0 < x < y < 100$  and  $p_1 < 1$ ).

As claimed above, this model is rich enough to allow for the possibility that the rejection rate  $q$  depends on the outside option. The game-theoretic link is that the outside option may influence  $B$ 's attitude to  $A$ , if  $A$  chooses not to take the outside option. To see this, note that equations (9) and (13) together imply

$$q = F \left[ - (1 + \lambda) \frac{100 - y}{y} - \lambda G \left( \frac{(1 + \lambda)(y - x - qy)}{y - x + (100 - y)q} - \lambda \bar{\theta} \right) \right], \quad (15)$$

a fixed-point equation in  $q$ . Given the meta-game data, the right-hand side defines a non-decreasing function of  $q$  that maps the closed unit interval into itself. Hence, by Tarski's fixed point theorem, there exists at least one fixed point. Moreover, we see in (15) that if the equilibrium rejection rate  $q$  is unique, then it is non-decreasing in  $x$  (and non-increasing in  $y$ ), since, at any given value of  $q$ , the right-hand side of the equation is increasing in  $(1 - q)y - x$ .<sup>17</sup> In particular, if the outside option is made "less generous" ( $x$  is increased), then the rejection probability increases. This monotonicity is in qualitative agreement with the empirical findings in Falk et al. (2003). In a slightly more complex game protocol, they find that the rejection rate to a given offer ( $y$ ) decreases when the proposer's outside option ( $x$ ) is decreased.<sup>18</sup> The fixed-point equation is illustrated in figure 3 below, drawn for  $\lambda = 0.5$  and  $y = 90$  when types are uniformly distributed on  $(-1, +1)$ . The straight line is the left-hand side of (15), and the curves represent the right-hand side for  $x = 20$  (the lowest curve),  $x = 50$  (the middle curve) and  $x = 70$  (the highest curve).

By varying the material payoffs  $x$  and  $y$  and recording the associated rejection rates,  $\tilde{q}(x, y)$ , equation (15) can be used to pin down  $\lambda$  and  $F$  for a given subject pool — under the hypothesis, of course, that the model is valid. In this sense, indirect aggregate preference elicitation is possible.

---

<sup>17</sup>The same comparative static property holds also for all "stable" equilibria in the case of non-uniqueness, where a fixed point is called "stable" if the right-hand side, as a function of  $q$ , intersects the diagonal from above.

<sup>18</sup>In Falk *et al.* (2002), the offer was  $y = 80$ , and the proposer's "outside options" (in their setting also subject to the responder's approval) were  $x = 50$  and  $x = 20$ , respectively. They also found low rejection rates when  $x \geq y$ , a case not analyzed here.

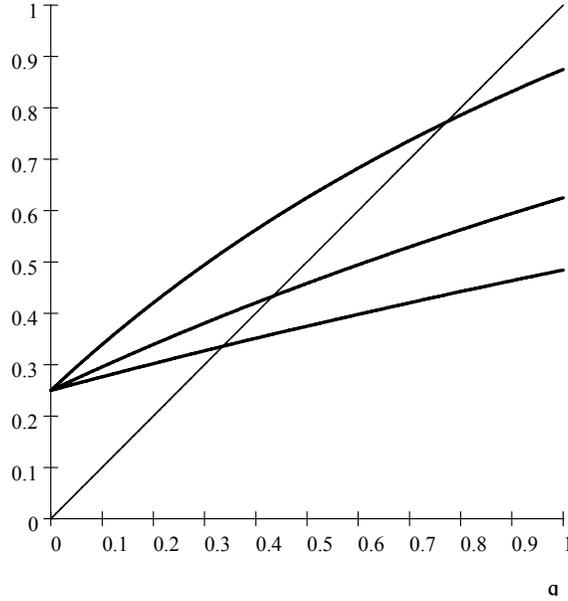


Figure 3: The fixed-point equation for the rejection rate.

However, also certain individual preference elicitation is possible. To see this, suppose estimates  $\tilde{\lambda}$  and  $\tilde{F}$  have been obtained as mentioned above. If records of individual subjects' actions have been kept, then further estimation and testing of the model at an individual level can be done by way of inequalities (8) and (10). For this purpose, let  $\tilde{\theta}$  denote the mean value associated with  $\tilde{F}$  and let  $\tilde{G}$  be the associated truncated mean-value function. According to the model and these estimates, a subject  $j$  of type  $\theta_j$ , when in the proposer role, does not take the outside option  $(x, 100 - x)$  if and only if

$$\theta_j \leq \left(1 + \tilde{\lambda}\right) \frac{y - x - y\tilde{q}(x, y)}{y - x + (100 - y)\tilde{q}(x, y)} - \tilde{\lambda}\tilde{\theta}, \quad (16)$$

and, when in the responder role, rejects the offer  $(y, 100 - y)$  if and only if

$$\theta_j \leq - \left(1 + \tilde{\lambda}\right) \frac{100 - y}{y} - \tilde{\lambda}\tilde{G} \left( \frac{\left(1 + \tilde{\lambda}\right) (y - x - y\tilde{q}(x, y))}{y - x + (100 - y)\tilde{q}(x, y)} - \tilde{\lambda}\tilde{\theta} \right). \quad (17)$$

For each subject  $j$ , Let  $\Theta_j \subset \Theta$  be the subset of parameter values  $\theta_j$  that satisfy these conditions for subject  $j$ , for all values of  $x$  and  $y$  in the experimental data. The set  $\Theta_j$  is either empty or a non-empty interval (determined by  $j$ 's choices for different values of  $x$  and  $y$ ). If  $\Theta_j = \emptyset$  for some subject  $j$ , then the model together with its estimates ( $\tilde{\lambda}$  and  $\tilde{F}$ ) is empirically rejected. If, however,  $\Theta_j \neq \emptyset$  for all

subjects  $j$ , then interval-valued estimates of all subjects' types have been obtained.<sup>19</sup> In the latter case it would be interesting to see whether such individual estimates have predictive power for the same subject's behavior in other game protocols.

**6.2. Mini random dictator games.** Turning to the game protocol in figure 2, suppose that “nature” not only chooses who will be the dictator, but also both players' types. In the notation of figure 2, an outcome in the meta-game protocol is thus a triplet  $(\omega, a, b)$ , where  $\omega \in \{\omega_1, \omega_2, \omega_3, \omega_4\}$ ,  $a$  is  $A$ 's type and  $b$  is  $B$ 's type. Let  $\lambda \in [0, 1]$ , and let  $F : \mathbb{R} \rightarrow [0, 1]$  be a c.d.f. with finite mean value  $\bar{\theta}$ . Player  $A$ 's Bernoulli function is then defined by

$$\pi_A(\omega_1, a, b) = \pi_A(\omega_3, a, b) = 50 + 50 \frac{a + \lambda b}{1 + \lambda} \quad (18)$$

$$\pi_A(\omega_2, a, b) = 70 + 10 \frac{a + \lambda b}{1 + \lambda} \quad (19)$$

$$\pi_A(\omega_4, a, b) = 10 + \frac{a + \lambda b}{1 + \lambda} 70 \quad (20)$$

and analogously for  $B$ .

Suppose, first, that both players' types are common knowledge. Then player  $A$ , if selected to be the dictator, will choose 50/50 if and only if  $a + \lambda b \geq (1 + \lambda)/2$ . Likewise, player  $B$ , if selected to be the dictator, will choose 50/50 if and only if  $b + \lambda a \geq (1 + \lambda)/2$ . Hence, each player's decision will depend in part on the other's type. In the context of the present example, it seems reasonable to constrain players' types to lie between zero and one, that is,  $\Theta = (0, 1)$ . Figure 4 shows how the unit square  $\Theta^2$  is divided into four regions by the two player's indifference lines (the diagram has been drawn for  $\lambda = 1/2$ ). Note, in particular, that there are type combinations  $(a, b)$  for which one player will choose 50/50 when selected to be the dictator, despite the fact (known by that player) that the other player, if called upon, would have chosen 70 for him- or herself. (These are the regions NW and SE of the intersection of the two indifference lines.)

---

<sup>19</sup>One further test is to see whether this collection of interval estimates, one for each subject, is consistent with  $\tilde{F}$  in the sense that, for every  $\theta \in \Theta$ , the number of subjects  $j$  with intervals  $\Theta_j$  that intersect  $(-\infty, \theta]$  approximates  $N\tilde{F}(\theta)$ , where  $N$  is the total number of subjects in the experiment.

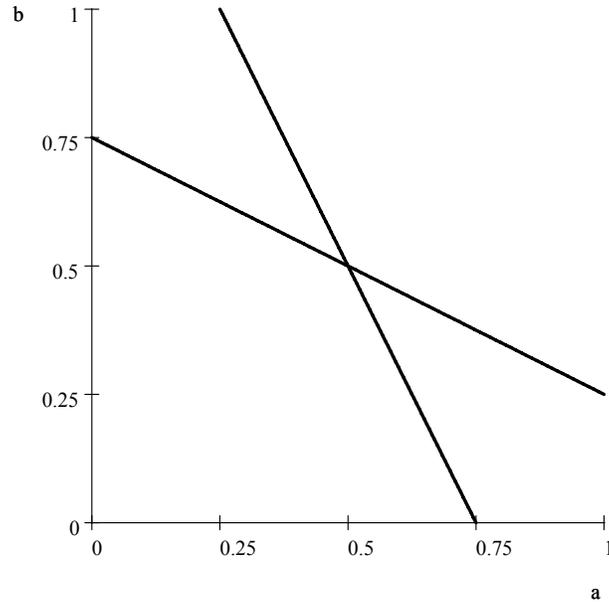


Figure 4: Players' indifference lines in the space of type pairs.

Secondly, suppose that each player's type is his or her private information. In sequential equilibrium, any of the two players, if selected to be the dictator, will choose 50/50 if and only if his or her own type  $\theta$  is at least  $(1 + \lambda)/2 - \lambda\bar{\theta}$ , where  $\bar{\theta}$  is the the mean-value under  $F$ . If  $\lambda = 0$ , then this condition is independent of the other's expected type, while for positive  $\lambda$ , the condition depends on the other's type and it is easier met the higher is the other's expected type.

## 7. CONCLUDING REMARKS

The methodological issues discussed here are relevant for an array of other game protocols than the few examples discussed here. For instance, one may ask if prisoners' dilemma game protocols indeed are prisoner dilemma games for all pairs of human players. For instance, in the light of the many experiments based on such protocols, it is not excluded that some individuals actually prefer the play  $(C, C)$  to the play  $(D, C)$  even in the role of player 1, although their own material payoff then is lower.<sup>20</sup>

A related empirical question is whether a repeated game protocol results in a repeated game, since the latter requires that preferences over plays are additively separable in the material payoffs in each round — a stringent requirement on preferences. Suppose, for example, that a prisoners' dilemma protocol is repeated ten

<sup>20</sup>Empirical support for this hypothesis has been found in preliminary experimental work by M. Kosfeld, E. Fehr and the author.

times in a laboratory setting and that the subjects are paid the sum of their material gains, after the last round. Will they behave as if they strived to maximize the sum of their material gains, even if they would exhibit such preferences in the one-shot prisoners' dilemma protocol?

It was shown above how Levine's (1998) model allows for certain interpersonal preference dependence in game protocols. In particular, it can explain why responders' rejection rate in ultimatum game protocols may depend on the proposer's outside options. While players' preferences in this approach are driven by altruism and spite, there may well be other reasons why some subjects reject small offers. A responder subject may, for example, want to punish the proposer's action, irrespective of the proposer's possible motives, because the action violates some "norm" supported by the responder. This would not be a case of interpersonal preference dependence, but of context-dependent preferences, as discussed in section 4. Indeed, Fehr and Gächter (2003) reported empirical evidence, in the context of a public-good provision game protocol, that suggests such explanations. Further analysis of preferences of this type seems highly relevant for our understanding of many social behaviors.

An even more basic issue, not discussed here, but yet of great importance for the relevance of game theoretic analysis for predictive purposes, is whether human subjects reason in a way that is consistent with any form of backward induction. Johnson *et al.* (2002) provide evidence that a significant fraction of human subjects in laboratory experiments do not even care to inform themselves of the material consequences in distant parts of the game protocol, although such information would be necessary for backward induction reasoning (and despite the fact that the subject can inform themselves at no other cost than that of touching a computer key).<sup>21</sup> (See Costa-Gomes *et al.* (2001) for similar evidence concerning normal-form games.) Such behaviors are clearly at odds with current game theory — another major challenge for future research.

To sum up: I would like to thank Werner Güth for his pioneering experimental studies of the predictive power of non-cooperative game theory. For economic theorists, the huge amount of experimental work done in the last two decades should be good news: although many theoretical presumptions have been challenged, new theoretical ideas can now be tested in the many laboratories around the world, hopefully leading us to better models of economic behavior.

---

<sup>21</sup>Indifference to material consequences at distant nodes makes sense if a subject does not care at all about material payoffs, or holds beliefs that they differ so little across plays that it is better to save effort by not hitting the computer key than to find out these material payoffs.

## REFERENCES

- [1] Asheim G. (2002): “On the epistemic foundation for backward induction”, *Mathematical Social Sciences* 44, 121-144.
- [2] Aumann R., (1995): “Backward induction and common knowledge of rationality”, *Games and Economic Behavior* 8, 6-19.
- [3] Aumann R. and A. Brandenburger (1995): “Epistemic conditions for Nash equilibrium”, *Econometrica* 63, 1161-1180.
- [4] Ben-Porath E. (1997): “Rationality, Nash equilibrium, and backwards induction in perfect information games”, *Review of Economic Studies* 64, 23-46.
- [5] Binmore K., A. Shaked and J. Sutton (1985): “Testing noncooperative bargaining theory: a preliminary study”, *American Economic Review* 75, 1178-1180.
- [6] Binmore K., J. McCarthy, G. Ponti, L. Samuelson and A. Shaked (2002): “A backward induction experiment”, *Journal of Economic Theory* 104, 48-88.
- [7] Bolton G. and R. Zwick (1995): “Anonymity versus punishment in ultimatum bargaining”, *Games and Economic Behavior* 10, 95-121.
- [8] Bolton G. and A. Ockenfels (2000): “ECR: A theory of equity, reciprocity and competition”, *American Economic Review* 90, 166-193.
- [9] Brandts J. and C. Solà (2000): “Reference points and negative reciprocity in simple sequential games”, *Games and Economic Behavior*, forthcoming.
- [10] Camerer C. (2003): *Behavioral Game Theory*. Princeton University Press (Princeton).
- [11] Costa-Gomes M., V. Crawford and B. Broseta (2001): “Cognition and behavior in normal-form games: an experimental study”, *Econometrica* 69, 1193-1235.
- [12] Falk A., E. Fehr and U. Fischbacher (2003): “On the nature of fair behavior”, *Economic Inquiry* 41, 20-26.
- [13] Fehr E. and S. Gächter (2003): “Altruistic punishment in humans”, *Nature* 415, 137-140.
- [14] Fehr E. and K. Schmidt (1999): “A theory of fairness, competition and cooperation”, *Quarterly Journal of Economics* 114, 817-868.

- [15] Glaser E., D. Laibson, J. Scheinkman and C. Soutter (2000): “Measuring trust”, *Quarterly Journal of Economics* (August), 811-846.
- [16] Güth W., R. Schmittberger and B. Schwarze (1982): “An experimental analysis of ultimatum bargaining”, *Journal of Economic Behavior and Organization* 3, 376-388.
- [17] Güth W. and R. Tietz (1990): “Ultimatum bargaining behavior: A survey and comparison of experimental results”, *Journal of Economic Psychology* 11, 417-449.
- [18] Johnson E., C. Camerer, S. Sen and T. Rymon (2002): “Detecting failures of backward induction: monitoring information search in sequential bargaining”, *Journal of Economic Theory* 104, 16-47.
- [19] Kagel J. and A. Roth (eds.) (1995): *The Handbook of Experimental Economics*. Princeton University Press, Princeton.
- [20] Kuhn H. (1950): “Extensive games”, *Proceedings of the National Academy of Sciences* 36, 570-576.
- [21] Kuhn H. (1953): “Extensive games and the problem of information”, *Annals of Mathematics Studies* 28193-216.
- [22] Levine D. (1998): “Modelling altruism and spitefulness in experiments”, *Review of Economic Dynamics* 1, 593-622.
- [23] Mertens J.-F. and S. Zamir (1985): “Formulation of Bayesian analysis for games with incomplete information”, *International Journal of Game Theory* 10, 619-632.
- [24] Ochs J. and A. Roth (1989): “An experimental study of sequential bargaining”, *American Economic Review* 79, 355-384.
- [25] Rabin M. (1993): “Incorporating fairness into game theory and economics”, *American Economic Review* 83, 1281-1302.
- [26] Ray I. and L. Zhou (2001): “Game theory via revealed preferences”, *Games and Economic Behavior* 37, 415-424.
- [27] Reny P. J. (1993): “Common belief and the theory of games with perfect information”, *Journal of Economic Theory* 59, 257-274.

- [28] Ritzberger K. (2002): *Foundations of Non-Cooperative Game Theory*. Oxford University Press (Oxford).
- [29] Sprumont Y. (2000): “On the testable implications of collective choice theories”, *Journal of Economic Theory* 93, 205-232.
- [30] Tan T. and S.Werlang (1988): “The Bayesian foundations of solution concepts of games”, *Journal of Economic Theory* 45, 370-391.
- [31] Zamir S. (2000): “Rationality and emotions in ultimatum bargaining”, mimeo., the Hebrew University and LEI/CREST (Paris).