Testing game theory

Jörgen W. Weibull[*]
Stockholm School of Economics
and
Research Institute of Industrial Economics
Stockholm, Sweden.

SSE WP 382, May 2000. Revised 16 April 2002.

Abstract. Experimentalists frequently claim that human subjects in the laboratory violate such game-theoretic solutions as Nash equilibrium and subgame perfect equilibrium. It is here argued that this claim is usually premature. What have been rejected are certain joint hypotheses concerning subjects' preferences, rationality and knowledge. This note discusses conceptual and methodological aspects of non-cooperative game theory in its epistemic interpretation. An alternative "empirical" interpretation is outlined, and an associated empirical equilibrium hypothesis is formulated.

JEL-codes: A10, C70, C72, C90.
Keywords: game theory, experiments, equilibrium.

[*Incomplete and preliminary*]

1. Introduction

An important current development in economics is the emergence of experimental economics. 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

[*]The author thanks CEPREMAP, Paris, for its hospitality in February 1998, when part of this paper was written. I am grateful for helpful comments from 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, to seminar participants at the Arne Ryde Symposium in Experimental Economics, at Harvard University, and at Université de Montreal.

empirical science.[1] The first experiments in game theory were carried out in the early fifties. A new wave of game experiments began in the mid seventies, and in the early eighties Güth, Schmittberger and Schwarze (1982) pioneered experimental work on ultimatum bargaining games. See Zamir (2000) for a recent discussion of such experiments, and see e.g. Kagel and Roth (1995) for surveys of experimental game theory more generally. The present note discusses some methodological and conceptual issues when applying and experimentally testing non-cooperative game theory, viewed as a positive model of human strategic interaction. The discussion suggests certain weaknesses both in the current practice of experimentalists and in non-cooperative game theory itself. A new experimental agenda, which emphasizes observability and falsifiability, is outlined, leading to a falisifiable "empirical equilibrium hypothesis."

In the experimental literature, it has many times been claimed that certain well-known game-theoretic solutions, such as Nash equilibrium and subgame perfect equilibrium, have been violated in laboratory experiments.[2] While it may well be true that human subjects actually do not behave according to these solutions in many situations, few experiments actually provide evidence for this. While experimentalists usually make efforts to carefully specify to the subjects the *game form* or, as will be defined below, the "game protocol" of the interaction in question, they usually do not make much effort to find the subjects' preferences, despite the fact that these preferences constitute an integral part of the very definition of a game. Instead, it is customary to simply hypothesize subjects' preferences. In the early literature, it was thus hypothesized 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. This narrow approach to preferences contrast with the formal machinery of non-cooperative game theory which does not impose any such restrictions on preferences.

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 (note assumptions like "the game is common knowledge to the players"). The extent and exact form of the latter varies across game forms, solutions, and on the interpretation or "meta model" in which the game is "embedded." There seem to be essentially two broad classes of such interpretations, one rationalistic and static, the other boundedly rationalistic and dynamic/evolutionary. Already in his Ph.D. dis-

[1] The likelihood for success, however, may be smaller, in view of the complexity of human decision making.

[2] The number of citations that could be made here is so large that any selection would be arbitrary.

sertation, John Nash suggested two distinct interpretations of non-cooperative game theory, along these same lines, see Nash (1950). In one interpretation, which for many years has been standard, players are assumed to be "rational," the game is played exactly once, and all relevant aspects of the game, including other players' preferences and rationality, are common or mutual knowledge among the players. Nash did not specify the exact form of rationality and knowledge (or beliefs) that would lead to a given Nash equilibrium. However, later research has provided exact epistemic conditions which together are sufficient for Nash equilibrium play and for subgame perfect equilibrium play, see e.g. Tan and Werlang (1988), Blume, Brandenburger and Dekel (1991), Reny (1993), Aumann and Brandenburger (1995), Aumann (1995), Ben-Porath (1997) and Asheim (2000). In general, these sufficient epistemic conditions are tailored to a class of games and a solution concept.

The second interpretation, which Nash called the "mass-action interpretation," is close in spirit to current models of social learning and evolution in games. For each of the n player roles in the game, there is a large population of individuals with identical preferences. An n-tuple of individuals, one from each such player population, is randomly and recurrently drawn to play the game. These individuals are not necessarily well-informed about the game, but base their strategy choice on observed past play.

These two interpretations, or "embedding models" can of course not be empirically falsified as such, only their assumptions, which we already from the outset know are strong idealizations. So what can 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 how game theory is used in economics and the other social sciences. In many cases it is not even possible to assure that the exact theoretical assumptions hold. For instance, "players' knowledge" in practice usually has to be replaced by some form of "information provided to the actors," and "common knowledge among the players" by some form of "public information provided to the actor" etc. Such "operational approximations" of course fall short of the theoretical assumptions. Hence, there is indeed plenty of opportunity for operationally approximated game-theoretic models to be empirically falsified.

Current evolutionary and learning models make weaker knowledge and rationality assumptions than the epistemic ones. In particular, no knowledge about other players' preferences is assumed. Instead, one usually assumes recurrent play and some form of social learning or adaptation, based on empirical observations of behaviors and outcomes. If such an evolutionary or learning model is to be tested empirically, then the laboratory setting should approximate the setting in the particular model at hand. However, a discussion of evolutionary and learning models falls outside the scope of the present paper.

Instead, the epistemic interpretation is here discussed at some length, and is contrasted with a new, still preliminary, "empirical" interpretation. Unlike the epistemic, but like the evolutionary interpretation, this interpretation imagines a finite population of individuals (laboratory subjects), one for each player role, and the game protocol is played recurrently between randomly matched individuals, one from each population. However, unlike standard evolutionary and learning models, individuals in the same player population may here differ with respect to preferences, beliefs and expectations. The accompanying equilibrium concept, called *empirical equilibrium*, is not a property of a strategy profile but of an *outcome of play* - a frequency distribution over the set of plays.

The rest of the paper is organized as follows. Section 2 provides some terminology, notation and definitions. Section 3 presents a simple example. Section 4 concerns backward induction, and section 5 discusses the possibility that one player's preferences may depend on (knowledge or beliefs about) another player's preferences. Section 6 suggests some experimental procedures in connection with testing of the epistemic approach. Section 7 develops the above-mentioned "empirical" interpretation, and section 8 concludes with a discussion of avenues for further research, and gives a brief discussion of the difference between empirical equilibrium and Eyster's and Rabin's (2000) "cursed" equilibrium.

2. Preliminaries

2.1. Extensive forms and protocols. The present discussion is focused on finite games in extensive form, as defined in Kuhn (1950,1953). Such a game is a mathematical object that contains as its basic building block a directed tree. A *play* τ of the game is a "route" through the tree, starting at its initial node (or "root") and ending at exactly one of the (finitely many) end nodes ω (or "leaves"). The finite set of intermediate nodes is partitioned into player subsets, and each player subset is partitioned into information sets for that player. One of the players may be "nature," and all information sets for this "non-personal" player are singleton sets with probabilities attached to each outgoing branch from the node in question. Probabilities are assigned to all nature's moves. There is a one-to-one relation between *plays* τ and *end-nodes* ω: each end-node is reached by exactly one play of the game, and each play reaches exactly one end-node. Letting Ω denote the set of end-nodes and T the set of plays, we thus have $|\Omega| = |T| < +\infty$.

The ingredients described so far belong to what is usually called the *game form* of a game. A game form becomes an extensive-form *game* if one attaches a vector of real numbers to each end-node of the tree, each such vector containing n components, one for each of the n personal players. These vectors are called the *payoff vectors*, and the collection of all the i'th components, one from each of the $|T|$ end-nodes, together are supposed to represent the i'th player's preferences. More exactly, each player is assumed to have complete and transitive preferences over the unit simplex

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

of lotteries over end-nodes, or, equivalently, over plays. These preferences are assumed to satisfy the von Neumann-Morgenstern axioms, implying the existence of a player-specific real-valued function π_i with domain Ω, or T, for each personal player i, such that player i prefers one lottery over another if and only if the first lottery gives a higher expected value to the function π_i than the second.

For definiteness, and without loss of generality, the domain of π_i will be taken to be the set T of plays - rather than the set Ω of end-nodes (which is usually taken to be the domain). The function $\pi_i : T \to \mathbb{R}$ will be called the *Bernoulli function* (or von Neumann-Morgenstern function) of player i. If Φ is a game form, then the pair $\Gamma = (\Phi, \pi)$, where π denotes the combined Bernoulli function $\pi : T \to \mathbb{R}^n$, constitutes an extensive-form *game*.

Let S_i denote the set of pure strategies for player role i, with $S = \times_i S_i$ denoting the set of pure-strategy profiles. Likewise, $\Delta(S_i)$ denotes the unit simplex of mixed strategies for player i, and $\square(S) = \times_i \Delta(S_i)$ denotes the polyhedron of mixed-strategy profiles. The *path* induced in the tree by a mixed-strategy profile σ is the subset of nodes which are reached with positive probability when σ is played, and an information set h is said to be *on the path* of a strategy profile σ if some node $x \in h$ is reached with positive probability under σ. Likewise, the probability distribution induced on the set Ω of end-nodes, or; equivalently over the set T of plays, by a strategy profile σ will be called its *outcome*. Hence, and outcome is a point $p \in \Delta(T)$. Moreover, since each mixed-strategy profile induces an outcome, one may compute the expected value of each player's Bernoulli function under any mixed-strategy profile. This defines the player's *payoff function* $u_i : \square(S) \to \mathbb{R}$, which maps each mixed-strategy profile to the associated mathematical expectation of the player's Bernoulli function.

Not every analysis of a game requires preferences over lotteries. It is sometimes sufficient to assume that every player i has a complete and transitive binary preference ordering $\succeq_i$ defined directly on the finite set T of plays, rather on the infinite set $\Delta(T)$ of lotteries over plays.[3] This defines what one could call an *ordinal game*.[4] The binary relation $\succeq_i$ will here be called the i'th player's *ordinal preferences* over plays.

[3] This ordinal approach is insufficient, though, for analyses involving players' subjective uncertainty about other players' unobserved moves - even if all players use pure strategies and there are no random moves by "nature." However, an ordinal approach can be extended to handle such uncertainty without necessarily invoking the von Neumann-Morgenstern axioms.

[4] Osborne and Rubinstein (1994) develop such an approach.

In virtually all applications of game theory, including laboratory experiments, players (or subjects) receive material, usually monetary, gains and losses after each play of the game. A game form Φ, with specified such material consequences, will here be called a *game protocol.* Since some experiments keep the game form constant while varying these consequences of play, the following terminology and notation is sometimes convenient. A *game protocol* is a pair (Φ, γ), where $\gamma : T \rightarrow C$, for some set C of material consequences. Clearly the material consequences of play in a game form Φ influence the players' (subjects') preferences over plays. Hence, we will here depart formally, though not substantially, from traditional non-cooperative game theory by viewing preferences as defined for a game protocol rather than for a game form. Consequently, we will here think of a game as a triplet (Φ, γ, π), where Φ is a game form, γ a consequence mapping, and π as a combined Bernoulli function for the personal players in the game form Φ.

The formal machinery of non-cooperative game theory does not require that a player's payoff value at an end node depend only on the material consequences. Indeed, two plays resulting in the same vector of material consequence may well differ in terms of information sets reached, choices made etc., aspects that may be relevant for players' preferences.[5] The formal machinery only requires the *existence* of a Bernoulli function π_i with domain T (or, more conventionally, Ω) for each player i.

Indeed, a large number of experiments have convincingly - though perhaps not surprisingly for the non-economist - shown that human subjects are not solely motivated by their own monetary gains.[6] This led researchers to postulate payoff functions which allow for a trade-off between own material gains and "fairness", see e.g. Fehr and Schmidt (1999) and Bolton and Ockenfels (2000a). However, even these more general payoff function forms are sometimes still too special, since they require that each payoff value depend *only* on the vector of material consequences at that end node. Falk *et al* (1999), Binmore *et al* (1999), Bolton and Ockenfels (2000b) and Brandts and Solà (2000) give experimental evidence that human subjects also care about the choices made along a play up to the end node, see below.[7] This last observation has implications for backward induction arguments, see below.[8]

[5] For example, you may not want to accept stolen money offered to you, or money offered by someone who in other ways has harmed another person or violated your moral norms.

[6] For early contributions, see e.g. Roth, Malouf and Murnigham (1981), Güth, Schmittberger and Schwarze (1982), Binmore, Shaked and Sutton (1985), and Ochs and Roth (1989).

[7] While psychological game-theory (see Geanakoplos, Pearce and Stacchetti (1989), Rabin (1993) and Dufwenberg and Gneezy (2000)) generalizes the notion of a game by explicitly allowing players' preferences to also depend on their expectations, the present discussion sticks to the classical approach in game theory in which preferences are defined over the set of plays, or, equivalently, the set of end-nodes.

[8] I am grateful to Sylvain Sorin for pointing this out.

2.2. Epistemic interpretations. At least two notions of "rationality" are used in non-cooperative game theory. The most common , which here will be called *rationality*^{1} (or "Savage rationality"), dictates avoidance of strategies that are not the best reply to any strategy profile in the game. By contrast, *rationality*^{2} (or "strict dominance rationality") dictates avoidance of strictly dominated strategies. Evidently rationality^1 implies rationality^2. As is well known, the converse is generally true only in two-player games, see Pearce (1984). Common knowledge of rationality^1 implies play of *rationalizable* strategies (Pearce (1984)), while common knowledge of rationality^2 implies play of strategies that are not iteratively strictly dominated. Rationalizability is a coarsening of Nash equilibrium. Non-cooperative game theory also provides many refinements of Nash equilibrium. The most commonly used refinement in extensive-form games being *subgame perfection* (Selten (1965)). For precise sufficient epistemic conditions for Nash equilibrium and subgame equilibrium, see e.g. Tan and Werlang (1988), Reny (1993), Aumann (1995), Brandenburger and Aumann (1995), Ben-Porath (1997) and Asheim (2000).[9]

3. Example

A class of game protocols that have been much studied in the laboratory are those associated with the so-called *ultimatum bargaining game* - though this is not a game in the theoretical sense. 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 given and known 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 a simple such strategic interaction, a *mini ultimatum-game protocol*, where the proposer has only two choices, either to offer the responder 50% (and keep 50% for himself), or to offer the responder 10% (and keep 90% for himself). The responder does not have the possibility to reject the 50/50 split in this game protocol, but she can choose whether to accept or reject the 90/10 split, if proposed. Hence, each player role has two pure strategies. This game form has three plays: $T = \{\tau_1, \tau_2, \tau_3\}$. In play τ_1, player A proposes the equal split, and play stops at end node ω_1. In play τ_2, A proposes the 90/10 split, B accepts this, and play stops at end node ω_2. In play τ_3, finally, A proposes the 90/10 split, B rejects this, and play stops at end node ω_3.

[9] Sufficient conditions for equilibrium play in epistemic models of games of imperfect information are typically based some form of mutually known "recommendation" or "expectation." In order to test the predictions of these models, the experimentalist thus needs some mechanism that makes a "recommendation" or "expectation" publicly known to the subjects. Brandts and MacLeod (1995) report results from experiments with recommended play. An alternative, "subjectivistic," approach is suggested in Kalai and Leherer (1995).

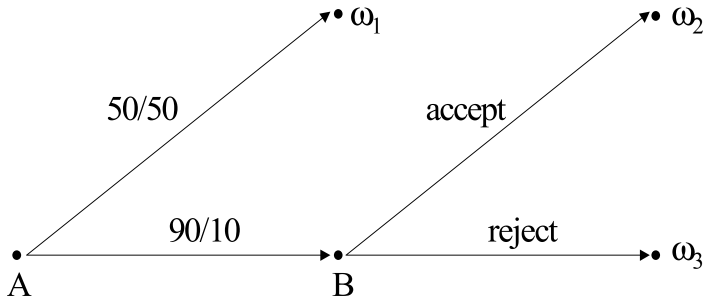

Figure 1: A mini ultimatum-game form.

In the early experimental literature it was presumed that the payoff values to the subjects are their own monetary gains. Hence, in strategic interactions like this, play τ_2 was predicted, i.e., that A proposes the 90/10 split and B accepts. This is of course the unique subgame-perfect equilibrium of the game that results if preferences are such. The implicit hypothesis in this early literature is fivefold:[10]

(H1) the responder prefers play τ_2 over play τ_3,

(H2) the responder is rational in the sense of playing according to his or her preferences,

(H3) the proposer role knows that H1 and H2 hold (or at least believes that they hold with a sufficiently high probability),

(H4) the proposer prefers τ_2 over τ_1, and

(H5) the proposer is rational in the sense of acting in accordance with his or her knowledge and preferences.

A large number of laboratory experiments with ultimatum bargaining have shown that many proposer subjects instead offer a sizable share to the responder, and that many responder subjects reject small shares. In the present mini ultimatum-game form, this corresponds to play τ_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. But since hypothesis

[10] 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.

H1 is not generally true, the combined hypothesis must be false, and the whole exercise seems superfluous. From the viewpoint of game theory, it would have been more interesting to see experimental findings concerning subjects' preferences, thereby suggesting what the *game* played might be.

In the present example, let $\succeq_A$ be a proposer subject's preferences over the set T, and let $\succeq_B$ be a responder subject's preferences over the same set. Associated with this simple game form, and assuming strict preferences, there are potentially as many as 36 ordinal games.[11] For example, a subject in player role A may have the preference $\tau_2 \succ_A \tau_1 \succ_A \tau_3$, and the subject in role B may have the preference $\tau_1 \succ_B \tau_3 \succ_B \tau_2$.[12] The associated ordinal game would have the 50/50 split as its unique subgame perfect outcome, and hence play τ_1 would not be evidence against subgame perfection but for subgame perfection. Another possibility is that A has preferences $\tau_1 \succ_A \tau_2 \succ_A \tau_3$. Irrespective of B's preferences, the unique subgame perfect outcome would again be the play τ_1, etc.

In an experimental study of a "mini ultimatum-game" protocol not very different from the one in Figure 1, Falk *et al* (1999) found evidence in support of the hypothesis that the recipient's rejection rate depends on the alternative choices available to the proposer along the play in question. In the present example, this means that B's relative ranking of plays τ_2 and τ_3 may depend on the material consequence of play τ_1.[13] If the alternative choice available to the proposer instead had been, say, a 100/0 split, then the evidence in Falk *et al* (1999) suggests an increase in the population share of responder subjects who would prefer τ_2 over τ_3.[14] As indicated above, this observation has far-reaching implications for backward induction arguments: for many subjects in roles A and B in the game protocol in Figure 1, the subgame beginning at B's node is not independent of the full game protocol: a change of the material consequences of play τ_1 may affect the preferences of player B. Hence, the usual practice in game theory of analyzing a subgame as if were independent of its context (the remaining truncated, or "pruned" tree) is erroneous: the subgame may change if the something in the rest of the game is changed. The backward induction argument has to be made with reference to a fixed and given context game.

[11] More exactly, there are 6 complete, transitive and irreflexive orderings of the three plays, and hence $36 = 6^2$ ordinal strict preference profiles in this game form.

[12] There is experimental evidence that many subjects have such preferences.

[13] This raises an interesting issue of whether such responder behavior violates independence of irrelevant alternatives. This is a subtle question because of the interlinkage between choices in games: the choice *not to* offer the equal split is *part of* both play τ_2 and play τ_3.

[14] Such evidence can also be viewed as a violation of subgame consistency, see Binmore et al (1999).

4. Backward induction

In a given game form Φ, let X_0 be the subset of nodes x in Φ such that x either is a move by nature, or $\{x\}$ is an information set of a personal player such that no information set in Φ contains both a successor node x' and a non-successor node x'' to x. For the purpose of this discussion, let us call this (non-empty) subset X_0 the set of *clean branching points*, and the part of the game form that begins at such a node x the subform with root x. Likewise, let a *subgame protocol* be defined as the game protocol (Φ_x, γ_x) obtained from (Φ, γ) when play starts at a clean branching point x of Φ. Suppose the number of personal players in Φ is n, and suppose that $\pi : T \longrightarrow \mathbb{R}^n$ is their joint Bernoulli function. A *subgame* of the game $\Gamma = (\Phi, \gamma, \pi)$ is a game $\Gamma_x = (\Phi_x, \gamma_x, \pi^x)$, where π^x is the restriction of π to the subset $T_x \subset T$ of (full) plays $\tau \in T$ that contain x.[15] In particular, π^x does *not*, as is usual assumed, have as its domain the set of plays *beginning* at the root x of the subform Φ_x - these truncated plays start at x, and thus do not contain the (shared) history leading up to x from the root of the full tree. Since players' preferences π^x in the subgame may depend on the choices made along the way up to its initial node x, the subgame Γ_x is in general not identical with the game $\Gamma' = (\Phi_x, \gamma_x, \pi')$ that is obtained if the subgame protocol is played in isolation, that is *beginning* at node x without the "history" in Φ leading up to x.

The observations in Falk et al (1999) show that this distinction is sometimes crucial. However, a restricted form of backward induction is still sound. One simply has to first define the full game protocol and all players' preferences *in this protocol.* Once this has been done, one may formally define, for instance, subgame perfect equilibrium as a strategy profile σ that induces a Nash equilibrium on every subgame $\Gamma_x = (\Phi_x, \gamma_x, \pi^x)$, for all $x \in X_0$, where $\pi^x : T_x \longrightarrow \mathbb{R}^n$ is the restriction of π, the combined Bernoulli function in the full game Γ (and not the combined Bernoulli function in the game that is obtained by starting play with x as the *initial* node). It is an empirical question whether "Kuhn's algorithm" is valid: Are players' preferences unaffected if a subgame protocol (Φ_x, γ_x) is replaced by an end node to which is assigned the material consequences, or a lottery over such, of one of the Nash equilibria of the subgame $\Gamma_x = (\Phi_x, \gamma_x, \pi^x)$?[16]

[15] The set of players in the subgame should be the same as in the full game, even if not all players have a move in the subgame, since otherwise preferences may be affected. Hence, the notion of a game has to be extended so as to include "passive" players if needed.

[16] Actually, Kuhn's algorithm can be interpreted in a more abstract sense: the initial node x of a subgame is replaced by a terminal node which has such material consequences that all players' preferences in the remaining ("pruned") game form can be represented by the same payoff values as before, at all end nodes not following x, and by the payoff value that correspond to a Nash equilibrium in the subgame. It is an open question, however, if this is possible.

For the sake of illustration, consider the unique proper subform of the game form in figure 1. This subform begins at the node x where player B has to accept or reject the proposal 90/10. Viewed in isolation, this is a one-player game protocol, where the unique player (B) has a binary choice of either (a) receiving 10 dollars and the remaining 90 dollars be given to a passive player A, or (b) receiving 0 dollars and no money is given to the passive player A. I guess most subjects in this one-player game protocol will choose the first option. However, we also know that many subjects in player role B in the full game protocol in figure 1 choose the second option. Suppose, moreover, that the preferences of player B in the full game protocol ranks play τ_3 (reject) before τ_2 (accept). Kuhn's algorithm presumes that, without affecting A's preferences, we may replace B's decision node by a terminal node with material consequences 0/0.

A distinct, but for analytical and predictive purposes related issue is whether human decision makers' preferences meet the dynamic consistency requirement implicit in the very definition of an extensive-form game. Do individuals rank the plays in a game protocol in the same way at each of their (individual) information sets? Certain experimental evidence suggests a negative answer, see Ainslie (1992) for a survey over empirical evidence that many subjects have hyperbolic rather than exponential time preferences, and thus rank future options differently depending on the date at which they make their ranking. In this case, one can construct extensive-form game protocols in which subjects's preferences will depend on the current information set at which they are. For example in a game protocol concerning savings decisions, it may well be that many subjects will at their first decision node prefer more savings at later decision nodes, but when one of these later decision nodes have been reached will prefer less savings at that node, see e.g. Laibson (1997) and Harris and Laibson (2001). This calls for a revision of the notion of an extensive-form game, allowing for the possibility that a player's preferences depend on the information set.[17]

At an even more basic level, one may ask if human subject reason in a way that is consistent with backward induction. Johnson et al (2000) give "hard" empirical laboratory evidence that suggest a negative answer in many game protocol. By way of cleverly designed software, these researchers were able to show that many subjects in certain sequential bargaining protocols do not even consider later decision nodes when making decisions at earlier decision nodes.

[17] Theoretical possibilities to relax this and certain other limitations of the current definition of the extensive form are being investigated in a joint research project with Alos Carlos-Ferrer and Klaus Ritzberger.

5. Interpersonal preference dependence

Even if we correctly identify the game associated with the game protocol in figure 1, for a particular pair of subjects, and this game indeed has a unique subgame-perfect equilibrium, non-cooperative game theory does still not provide any prediction. Embedded in an epistemic model, such a prediction assumes (at least) that player A knows (or has almost correct beliefs about) B's ranking of plays τ_2 and τ_3.

The identification 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 theoretically disturbing since it makes the domain of preferences unclear, and yet such interdependence might realistically exist in some interactions. This is the case, if, for instance, a subject's ranking of plays in a game protocol with monetary payoffs depends on whether or not another subject is (known or believed to) be "generous" or "cheap." One subject, I, may be willing to share money with another subject, J, if I knows that J would (prefer to) share money with I at a similar decision node for J, while I may not be willing to share money with J if I would know that J had preferred not to share money with I. If the subjects' preferences are mutually dependent in this way, then we face a problem of self-reference in the very definition of a game.

From a theoretical viewpoint, such potential preference-interdependence calls for a representation in the style of the usual Harsanyi transformation of a game of incomplete information into a game of incomplete but imperfect information. In order to handle interpersonal preference dependence in this way, the type space has to be chosen rich enough so that a type can be identified as a combination of a preference in the game protocol, and a belief about others' preferences in the protocol. From a predictive viewpoint this approach meets certain difficulties, however. First, the resulting game might have a large set of equilibria. Second, the game may become so abstract that human subjects would find it hard to state their preferences in that "meta game." Moreover, it does not seem clear that one can guarantee that the preferences in the so constructed meta game do not exhibit interpersonal dependencies, hence potentially leading to an infinite regress of meta games with higher and higher type spaces. This route of investigation has substantial theoretical interest, I think, but falls outside the scope of the present investigation.

Instead, a partial analysis of interpersonal preference dependence is here sketched, not in order to "solve" this problem, but rather in terms of a given game protocol associated with a given game form or game protocol. For this limited purpose, consider a finite game protocol (Φ, γ) with n personal players. Let $\Delta = \Delta(T)$ as before denote the set of lotteries over the plays in Φ. Without loss of generality, normalize all Bernoulli functions to take non-negative values summing to one; $\pi_i(\tau) \geq 0$ for all $\tau \in T$, and $\sum_{\tau \in T} \pi_i(\tau) = 1$. We may thus identify each Bernoulli function π_i with a point in the same unit simplex Δ in $\mathbb{R}^{|T|}$. Suppose each player i has von Neumann-Morgenstern preferences over the set Δ of lotteries over T, given any (hypothetical) profile $\pi'_{-i} = (\pi'_j)_{j \neq i} \in \Delta^{n-1}$ of normalized Bernoulli functions $\pi_j : T \to \mathbb{R}$ for all other players j. Let $\pi_i \in \Delta$ be the normalized Bernoulli function for player i's von Neumann-Morgenstern preferences, given π'_{-i}. This defines a function $\varphi_i : \Delta^{n-1} \to \Delta$, where the Bernoulli function $\pi_i = \varphi_i\left(\pi'_{-i}\right)$ represents player i's preferences over plays when faced with other players with Bernoulli functions π'_{-i}.[18] For any given hypothetical preference profile $\pi' \in \Delta^n$, let $\pi = \varphi(\pi') \in \Delta^n$, where $\varphi : \Delta^n \to \Delta^n$ is defined by combining the player-specific mappings φ_i.[19] If this combined mapping φ is continuous, then, by Brouwer's fixed-point theorem, there exists at least one Bernoulli-function profile $\pi^* \in \Delta^n$ such that $\pi^* = \varphi(\pi^*)$. Such a function profile will be called *interpersonally consistent.* For each player i, π_i^* is a Bernoulli function for that player, when facing other players with Bernoulli functions π_j^*, for all $j \neq i$. The triplet (Φ, γ, π^*) constitutes an extensive-form *game* with interpersonally consistent payoffs. Such a game, and only such a game, can be common knowledge among the players. We have established the following proposition.

Proposition 1. *Let (Φ, γ) be a game protocol, and suppose i's preferences in this protocol are given by $\pi_i = \varphi_i\left(\pi'_{-i}\right)$, for all $\pi'_{-i} \in \Delta^{n-1}$, where $\varphi_i : \Delta^{n-1} \to \Delta$ is continuous. Then there exists at least one game (Φ, γ, π^*) where the payoff profile $\pi^* : T \to \mathbb{R}^n$ is interpersonally consistent.*

Note that the proposition does not exclude the possibility of multiple games associated with one and the same game protocol. In order to illustrate this possibility, consider the game protocol in Figure 2 below.[20] There, a mini "dictator game" form is played after the tossing of a coin deciding who of the two players should be the "dictator." The game form has four plays: τ_1, where A is the dictator and proposes 50/50; τ_2, where A is the dictator and proposes 90 for herself and 10 for B; τ_3, where B is the dictator and proposes 50/50; and, finally, τ_4, where B is the dictator and

[18] This approach short-cuts the potential infinite regress that may arise if a player i does not only care about player j's Bernoulli function π_j, but also about how j arrived at his or her Bernoulli function, *i.e.* if i cares about the whole mapping g_j, not only its current value $\varphi_j(\pi_{-j})$, see remark below.

[19] Expand the domain of each φ_i to Δ^n by letting $\varphi_i(\pi'_1, .., \pi'_n)$ depend only on π'_{-i}.

[20] I am grateful to Alvin Roth for suggesting this game protocol, which is simpler than the one I originally suggested.

proposes 90 for herself and 10 for B ("90/10" in the diagram thus refers to 90 to the proposer, who may be A or B, and 10 to the responder).

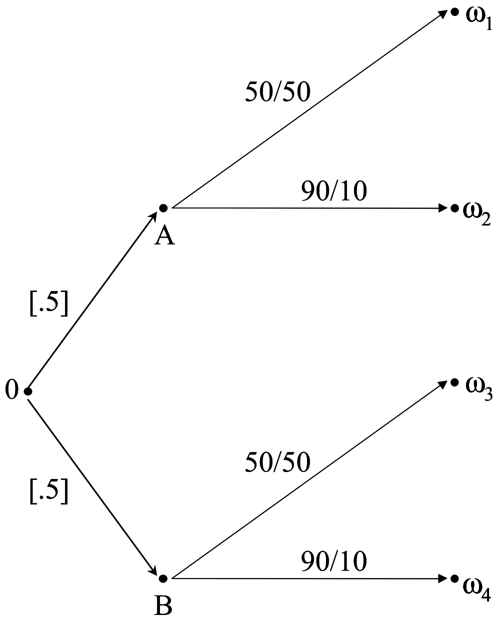

Figure 2: A mini dictator-game form with randomly assigned dictator.

Suppose the subject in player role A ranks highest the play τ_1 (A being the dictator and proposing fair division) if and only if B's most preferred play is τ_3 (B being the dictator and proposing fair division). Suppose also that A ranks highest the play τ_2 (maximal monetary gain for A) if and only if B' most preferred play is τ_4 (maximal monetary gain for B). In other words, each subject is generous if the other is generous and selfish if the other is selfish. With such subjects, this game protocol has (at least) two games, each with interpersonally consistent preferences. In the first game, the unique subgame perfect outcome is the monetary outcome $(50, 50)$ for sure, while in the second the unique subgame perfect equilibrium outcome is a two point probability distribution, yielding monetary consequences $(90, 10)$ or $(10, 90)$, with equal probability.

In order to illustrate the opposite possibility, that of non-existence of a game, consider any two-player game protocol where the two players as siblings, with player 1 being older than player 2, and where 2 admires 1, but 1 wants to differ from 2. If told player 2's preference ordering, player 1 always has another preference ordering, while if told player 1's preference ordering, player 2 has the same ordering as player 1. In such a situation it is possible that no game exists.

Remark 1: Continuity of the mapping φ was seen to be sufficient for the existence, but not uniqueness, of interpersonally consistent preferences. By the contraction mapping theorem, uniqueness is guaranteed if $\varphi : \Delta^n \to \Delta^n$ is a contraction.

Remark 2: The present approach sidesteps the potential infinite regress in preference formation, mentioned above. For it assumes that i's Bernoulli function be a function of j's Bernoulli function π_j, not of j's preference formation rule φ_j leading to π_j. For example, suppose the subject in player role A in the game form in figure 2 is asked to state his ranking of the four plays, not given B's (hypothetical) ranking of plays, but given B's (hypothetical) *conditional* ranking rule φ_B of plays. One of B's many (hypothetical) conditional rankings places τ_3 in top if A places τ_1 in top, and otherwise B places τ_2 in top. This defines π_A as a function ψ_A of φ_B. Next, suppose B would know ψ_A, potentially leading to π_B as a function χ_B of ψ_A, etc. ad infinitum. A theoretically satisfactory treatment of this problem might be found in the style of the Harsanyi approach mentioned above.

Interpersonal preference dependence is an empirical question, in principle open to testing. I expect that in many, if not most, game protocols relevant for economic analysis, preferences are *not* interpersonally dependent. In this respect, ultimatum bargaining and dictator game protocols, which are popular for laboratory experiments, seem to constitute "worst case scenarios" for controlled testing of standard epistemic non-cooperative game theory.

6. Testing epistemic game theory

6.1. Identifying the game. If the experimentalist's task is restricted to identifying a game which represents the interaction of given human subjects in the player roles of a given game protocol, without having the subjects play the game together, then subjects do not have any incentives to misrepresent their true preferences.[21] The experimentalist's task is then to find an interpersonally consistent payoff profile over the set of plays, or, alternatively, an interpersonally consistent profile of ordinal preferences of the set of plays. In principle, this can be done by means of an extension of the usual revealed-preference approach. Instead of asking subjects which play they would *choose* out of any subset of plays (like choosing among consumption baskets), the experimentalist has to ask which play the subject would *prefer to happen* - since a play in general also depends on other subjects' choices.

In practice, how can this be done? Suppose we have an extensive game form Φ with n player roles, and with material consequences $\gamma : T \to C$. The analyst may

[21] Even if subjects do not have any incentives to misrepresent their preferences, it may easily be the case that they also lack incentives to truthfully report their preferences. An important challenge for the experimentalist is thus to provide subjects with incentives to truthfully report their preferences. Here lessons from experimental psychology seem relevant.

then elicit the ordinal preference ordering $\succsim_i$ over T, of each subject, in his or her player role i, for $i = 1, 2, ..., n$, given any hypothetical preference profile $\succsim_{-i}$ for the subjects in the other player roles. There are finitely many such hypothetical profiles. This way, the analyst can identify each subject's ordinal preference ranking, for each possible preference ordering that the subjects in the other player roles might have. If there exists a preference profile $\succsim^*$ such that each subject i has preferences $\succsim_i^*$ over T, given $\succsim_{-i}^*$, then $(\Phi, \succsim^*)$ constitutes an ordinal game. It may be harder to find Bernoulli functions, though, since the set of potential such profiles constitutes a continuum. An alternative, less theoretically stringent but in many cases perhaps sufficient approach, would be to first ask each subject to rank the plays, with no information about the other subject's ranking, then make these rankings public to all subjects, and then ask the subjects if their rankings have changed in the light of this information. If no subject changes his or her ranking, then these are (at least approximately) interpersonally consistent preferences for those subjects in that game protocol. If at least one of the subject changes his or her ranking when informed about the others' rankings, then this is empirical evidence for the hypothesis that preferences in games may be interdependent. The experimentalist could then make the subjects' new rankings public, and ask them if they now want to change their rankings, etc. until no change occurs (or give up if rankings keep changing).[22]

There is yet another difficulty in identifying preferences. Subjects' preferences may depend on what they know or believe about the other subjects' personality, income, gender etc. For example, if the proposer subject in Figure 1 would know that the responder subject is a criminal, a hero, a close relative, a wealthy or poor person, a man or woman, old or young, then this could influence the subject's preferences over plays (even if the other subject's preferences were known and kept constant). If this is possibly the case, then such background information should be part of the information provided to the subjects, if the game is intended for predictions in such environments. Indeed, this seems to be done in part in many experiments by way of telling the subjects from what population pool the subjects have been drawn.

If a player role has multiple information sets in the game form, then one needs to test for dynamic consistency of the subject's preferences. More precisely, for the subject in such a player role, and for each of his or her information sets in the game, one should test whether the subject, once one of his or her information sets has been reached by play, ranks the plays through the information set in the same way as the subject ranked them before play (at the initial node of the game form).

If the game has been successfully identified, and no dynamic inconsistency found,

[22] In case the game is to be played after this identification, then the analyst faces the issue of subjects' strategic misrepresentation of their own preferences, see next section.

then the experimentalist has found a game that represents the interaction at hand. In this case, the analyst may apply game-theoretic solution concepts to the identified game.

As a first shot at the issue of game identification, it would be interesting if experimentalists, after having made an experiment in the currently customary way, would ask the subjects to rank all possible plays in the game form, ask them whether their rankings depend on (their knowledge of or beliefs about) the rankings of the subjects in the other player roles, why they have played the way they did, etc. This could guide future experiments involving game identification and solution testing.

6.2. Solution concepts. Suppose that the experimentalist wants to test a given game-theoretic solution in a game that represents the interaction of given human subjects in the player roles of a given game form. If the subjects in the identification phase expect the game to be played by themselves in the player roles they have, and against "opponent subjects" who have been informed about their preferences, then they may have incentives to misrepresent their true preferences. In order to test game theoretic solutions, the experimentalist thus has to find some incentive-compatible scheme for preference revelation in the game protocol, such that the game can be made publicly known to the subjects without giving them incentives to misrepresent their preferences in the preceding game identification. Some approaches to counter this potential source of preference distortion will be briefly discussed. Note, however, that in some interactions such incentives may be insignificant or absent, in which case the following suggestions are unnecessarily cautious; the experimentalist can then just go ahead along the lines suggested in the preceding subsection.

Hypothetical play. Present a game protocol to *one* subject and assign a player role to this subject. Assign hypothetical payoff (Bernoulli function) values to each play (or end node) of the game, for all *other* player roles. Then identify the subject's preferences, given this game protocol and hypothetical preference profile for the other players. This defines a game, and hence allows the application of game-theoretic solution concepts. In the spirit of Selten's so-called strategy method (Selten (1967), one could ask the subject questions, at each of his or her information sets, such as: "If play would reach this information set, and the other players had the preferences represented by their payoff values, what choice would you then make?" Anticipation in the game identification stage of such "hypothetical play" should not have a distorting effect on the subject's incentives to truthfully report his or her preferences.

Play against computerized clones.[23] Allocate one subject to each player role in a given game form. Inform each subject that he or she will not play against the

[23] This approach seems less suitable for game protocols where subjects have interdependent preferences.

other subjects, but against computer "clones" on their behalf. Each clone will be programmed according to the "parent" subjects' preferences, that the experimentalist has elicited, but independently of all other subject's stated preferences. Hence, each clone's play will be independent of the subject's stated preferences in the game identification phase. This way, a game can been identified, and game-theoretic solutions may be applied. The analyst may let each subject play in his or her player role, against these computer clones replacing the other subjects, and let the "parent" subjects obtain the material rewards stipulated in the game protocol. It is possible that subjects' preferences will be different when they know that they will play against computers rather than human subjects - even if they know that the. However, note that unlike some such experiments, subjects will receive material rewards from their clones' play, and this will be known by the subjects.

Play under incomplete information. A third approach, does not lead to a test of solutions in the same game as before, but avoids the potential incentive to misrepresent one's preferences. This approach is restrictive, it presumes that all subjects' preferences are interpersonally independent (which can be empirically tested). Take $m > 1$ subjects for each of the n player roles in a given game form. Identify each subject's preferences over the plays of the game. Then draw at random one subject from each of the n groups to his or her player role. Inform each of the drawn subjects of the *preference distribution* in each of the other groups, excluding the subject drawn to play the role in question (hence, a distribution over $m - 1$ subjects in each of the other player populations). This way, no *playing* subject's preferences is revealed to the other *playing* subjects. Therefore, subjects should not have incentives to misrepresent their preferences (just as in the Clark-Groves mechanism) and yet each subject obtains relevant statistical information about the likely preferences of each opponent subject (the "type" distribution), more relevant the larger m is and the more homogeneous each group is. In the most fortunate case for the experimenter, all subjects in each player subpopulation have the same preferences (or sufficiently close to induce the same behavior). In this case, we in practice have obtained a publicly known game of *complete* information, without incentives for untruthful preference announcements. With heterogeneous preferences, this kind of experiment can be used as a test of solution concepts for the corresponding *game of incomplete information.* For example, in the game protocol in Figure 1, standard epistemic game theory models would predict that the responder subject will play optimally according to his or her preferences, while the proposer subject will maximize his or her expected payoff (Bernoulli function), subject to the inferred distribution of optimal responses for a randomly drawn responder subject.[24]

[24] Falk et al (1999) find evidence pointing in the direction of such optimizing behavior.

6.3. Conjectures. The author's conjecture is that, if a game has been identified and made publicly known to all subjects, then a vast majority of these subjects will play the subgame perfection equilibrium in such simple game forms as the one in figure 1. In slightly more complex game forms, however, violation of subgame perfection and also of weaker solution concepts will presumable not be infrequent. For instance, experimentalists have presented large groups of subjects - the readers of a certain newspaper - with the following "beauty-contest game protocol." Without observing each others' choices, each subject has to choose a nonnegative number not exceeding 100. The subject(s) with the number closest to 2/3 of the average of all chosen numbers share a pre-specified monetary prize (by means of a fair lottery among the winners), and the other subjects receive nothing, see e.g. Nagel (1995) and Bosch *et al* (2000). Let us say that individuals with the best guess "win." It turns out that subjects' choices vary over a wide range of numbers, despite the fact that the number zero is the unique Nash equilibrium strategy in the game that results *if* every subject prefers "winning alone" over "shared winning" over "not winning".[25] Indeed, some subjects reported that they had come to the conclusion that zero is the unique "equilibrium guess" but had nevertheless chosen a positive number, since they (correctly) believed that not all others will play the equilibrium strategy. (The equilibrium strategy was not a winning strategy in any of these experiments.)

Suppose that the above experiment would be carried out in the laboratory, but with a fixed and known number $n > 2$ of subjects, and with all subjects' preferences identified. Suppose moreover that the resulting game were made public information to all the n subjects. Suppose that the experimentalist found that every subject prefers winning alone over shared winning over not winning. Then the game's unique Nash equilibrium would be that everybody chooses the number zero. However, in view of Nagel's *et al*'s findings, one can expect that in the first few rounds of recurrent play of such a game form, subjects' guesses are far off the Nash equilibrium (given the expected preferences), while their guesses tend to converge to Nash equilibrium over time. Nagel has also carried out the above experiment, with similar results, in the case $n = 2$, although that game, with the same preferences as mentioned above, has exactly one dominant strategy for each player, namely the number zero. This number is a best reply to *all* choices of one's opponent. To choose a positive number thus violates cautious rationality^1, but not rationality^1.[26]

[25] The question of preferences is an empirical question also in this game form. The stated preference hypothesis is violated if, for example, some subject prefers that another subject wins. Such preferences have been observed in these experiments - the subjects in question were close relatives (Nagel, personal communication).

[26] In a similar spirit, Søvik (2000) gives experimental evidence concerning human subjects' "depth of reasoning" in the sense of the number of iterated eliminations of strictly dominated strategies.

Many experiments concern ultimatum game forms, but also many other game forms have been studied in the laboratory, such as so-called dictator "games" and trust "games" (see e.g. Glaser *et al* (2000)). The methodological critique raised here applies also to those experiments. Another popular probing stone for game theory has been the repeated play of the prisoners' dilemma. In this case, the present critique applies at two levels. First, it is an empirical question whether the stage game indeed is of the prisoners' dilemma variety. Do all subjects in player role 1, say, prefer the play (D, C) over the play (C, C) in the one-shot interaction? Second, it is an empirical question if the repetition of this interaction represents a *repeated* game in the usual game-theoretic sense. The latter requires, among other things, that preferences over plays are additive functions of the monetary gains in all rounds. In particular, this precludes the possibility that players' preferences also depends on the sequencing *per se* of "defections", "retaliations", "punishments" etc.

7. Empirical game theory - a rough outline

An "empirical" interpretation of game theory is here outlined, an interpretation that (a) does not presume that the players have any knowledge of unobservables (others' preferences, knowledge or rationality), (b) allows preferences to depend on observations of others' play, and (c) presumes a certain degree of rationality. The interpretation differs from both the rationalistic and the evolutionary interpretations in a few respects. It takes as given a game protocol (Φ, γ), and it defines equilibrium as a property of an outcome in such a protocol, not as a property of a strategy profile in a game.

For each player role in the game form there is a finite population of individuals (in the laboratory: subjects). Unlike Nash's mass action interpretation and the evolutionary interpretation, each such player populations may be heterogeneous with respect to preferences. Moreover, individuals' preferences may depend both on the game protocol at hand, and on observations of the outcome of others' play (observations that may serve as a basis for beliefs about others' preferences). The game protocol is played recurrently between randomly matched individuals, and these are informed of play in the recent past. When called upon to play, the individual commits to a mixed strategy in the game form at hand, before play, and has to stick to that strategy when the game is played.

7.1. Definition. Consider an n-player game protocol (Φ, γ), and, for each player role i in Φ, a finite population A_i of individuals, where all populations are disjoint. When the game is played, one individual is drawn at random from each population i, with statistical independence and equal probability for all individuals within each population. We call such a triplet $\Psi = (\Phi, \gamma, A)$, where $A = (A_1, A_2, ..., A_n)$ a *population game protocol.*

Definition: For any $\varepsilon > 0$, an outcome $p \in \Delta(T)$ constitutes an *ε-precise empirical equilibrium* of the population game protocol $\Psi = (\Phi, \gamma, A)$ if there for every player role i and individual $a \in A_i$ exists a mixed-strategy profile $\sigma^a \in \square(S)$ such that:

(A) σ_i^a is a best reply for individual a in role i to σ^a,

$$\sigma_i^a \in \arg \max_{\sigma_i \in \Delta(S_i)} u_i^a \left(\sigma_i, \sigma_{-i}^a\right), \tag{1}$$

(B) its outcome is within distance ε from p, and

(C) the induced aggregate behavior $\mu \in \square(S)$, defined by

$$\mu_i = \frac{1}{|A_i|} \sum_{a \in A_i} \sigma_i^a \qquad \forall i,$$

has its outcome within distance ε from p.

Here $u_i^a : \square(S) \to \mathbb{R}$ is the payoff function associated with a Bernoulli function $\pi_i^a : T \to \mathbb{R}$ that represents a's preferences, in his or her player role i in the game protocol (Φ, γ), and given (some empirical observation of) p. The strategy profile σ^a thus represents both a's behavior, σ_i^a, and a rationalizing expectation of others' behaviors (in a random matching), σ_{-i}^a. In other words: the defining property of an empirical equilibrium is that every individual should play a best reply, in his or her player role, against some expectation of others' behaviors which is approximately consistent with the given outcome p, and the population aggregate of these best replies should also be approximately consistent with the given outcome.

The concept of empirical equilibrium is closely related to that of Nash equilibrium. Suppose p is the outcome of a Nash equilibrium σ^* in a game $\Gamma = (\Phi, \gamma, \pi)$, then p is a precise empirical equilibrium, that is, an ε-precise empirical equilibrium for $\varepsilon = 0$, of the population game protocol $\Psi = (\Phi, \gamma, A)$, where all individuals in population i have the same preferences π_i, given Φ, γ and p (each population could, for instance, consist of one individual).

7.2. Experimental testing. Here an experimental setting for testing of empirical equilibrium is outlined.

Step 1. The experimentalist informs all subjects about the game protocol, in such a way that it is clear to each subject that also the other subjects have been likewise informed. The experimentalist divides the subject pool into equally large subpopulations, say each of size m, and assigns each subpopulation to a player role in the game form. The experimentalist informs each subject which player role that subject will have in the game protocol, i.e., which player population the subject belongs. This should ensure that the game protocol is practically speaking ”common knowledge” in the total population A of subjects.

Step 2. The game protocol is played with random matching in $L+K$ rounds. In each round, all mn subjects are randomly matched into m groups, where each group consists of n subject, one subject from each player population. Each such group plays the game protocol once. The numbers L and K are chosen beforehand by the experimentalist. The first L rounds are “learning rounds,” and will not be used for testing, and the subsequent K rounds are recorded for testing purposes. The matchings are set up in such a way that all matchings in each round are equally probable, and such that there is statistical independence between all matchings. Hence, during these $L+K$ rounds, each subject plays the game protocol exactly $L+K$ times, each time in the same player role, and against a randomly drawn $(n-1)$-tuple of subjects in the other player roles. After each round, each subject is informed about the full play that he or she just took part in.

Step 3. After the $L+K$ plays, the analysts makes a statistical test of the hypothesis that aggregate play (suitably represented) in the last K rounds is stationary. If stationarity is rejected, then that ends the experiment, and no testing is done. If, by contrast, the stationarity hypothesis was not rejected, then testing proceeds as outlined below, under the presumption that aggregate play in the last K rounds is stationary.

Step 4. The experimentalist calculates the empirical distribution of play, or the *empirical outcome*, $\tilde{p}$, in the last K rounds, informs each subject of $\tilde{p}$, and then asks each subject to choose a strategy (mixed or pure) for one additional round of play, against randomly drawn opponents, just as in the preceding $L+K$ rounds. These choices will be the strategies σ_i^a, and their aggregate defines $\mu \in \square(S)$ They will be implemented (in order to make the subjects' choices payoff relevant), one random matching for each subject, executed as follows: for a given subject $a \in A_i$, play σ_i^a against a (computer based) realization of μ.

Step 5. Elicit each subject's preferences, directly or indirectly by way of restricting preferences to some pre-specified class. The first approach in principle allows for testing of the empirical equilibrium hypothesis *per se*, while the second only allows for testing the empirical equilibrium hypothesis in conjunction with some preference hypothesis. More exactly, in the indirect approach, the experimentalist postulates a (restrictive) candidate class of payoff functions. The test then consists of finding (by way of computer calculations), for each subject a some payoff function in the corresponding class such that the subject's strategy choice σ_i^a is consistent with maximization against some belief $\sigma^a \in \square(S)$ which is consistent with the empirical outcome. For falsification to be possible, it is necessary that there exist a non-empty subset of outcomes $P \subset \Delta(T)$ which are incompatible with optimality in the class.

The empirical equilibrium hypothesis $EEH(\varepsilon, \delta)$: The empirical outcome $\tilde{p}$ constitutes an ε-precise empirical equilibrium for at least the fraction δ of all subjects a.

8. Concluding remarks

The outlined concept of empirical equilibrium differs from Eyster's and Rabin's (2000) "cursed equilibrium" concept. The latter relaxes the standard notion of Bayesian equilibrium in incomplete-information games by allowing for the possibility that players underestimate the informational content of other players' actions. Their approach maintains all other assumptions of the usual epistemic approach to game theory, and therefore differs dramatically from the present approach.

The difference can be illustrated by means of their introductory example, where a buyer may purchase a used car from a seller at a predetermined price of $1,000. The seller knows whether the car is a "lemon" or not. If a lemon, then it is worth $0 to both, while if not a lemon, then it is worth $2,000 to the seller and $3,000 to the buyer, where both cases, lemon and non-lemon, are equally likely. The seller and buyer are risk neutral and purely selfish. The two parties simultaneously announce whether they want to trade or not, and the car is sold if and only if both parties announce that they wish to trade. In this simultaneous-move incomplete-information game, the seller has a weakly dominant strategy, namely to sell if and only if the car is a lemon. If the seller uses this strategy, then the buyer should clearly not buy. This is the unique perfect Nash equilibrium, and there exists no Nash equilibrium in which trade takes place with positive probability. However, a sufficiently "cursed" buyer may buy a lemon.[27] A "fully cursed" buyer believes that a car for sale is a lemon with probability 1/2, the prior probability for the car to be a lemon. Hence, the payoff that the buyer (incorrectly) expects from announcing "buy" is $500, while the buyer (correctly) expects a "pass" to result in $0 for sure. The fully cursed buyer thus announces "buy," and, in this cursed equilibrium, either gets nothing or a lemon. This outcome is radically different from the Nash equilibrium outcome - no trade - and it is also incompatible with empirical equilibrium. For the fully cursed equilibrium induces two plays, namely τ_1, "a lemon for sale, a willing buyer and trade," and τ_2, "a non-lemon not for sale, a willing buyer and no trade," respectively, each play with probability 1/2. Given any outcome which assigns positive probability to the first play, "buy" is a suboptimal strategy (given the presumed preferences), and hence no such outcome is an empirical equilibrium.

The discussion in the preceding sections call for new experiments. In particular, it would be valuable to see experiments where subjects' preferences were identified -

[27] Eyster and Rabin (2000) parametrize the expectational cursing from "no curse" (Bayesian Nash equilibrium) to "fully cursed equilibrium."

even without play of the resulting game. In the many game forms that have been used in experiments up to date: how do subjects, when allocated to a particular player role, information set etc., rank the set of plays? Do their rankings depend on their beliefs or knowledge of the rankings of the other player subjects? If they do, can then a game be identified, with preferences that are interpersonally compatible? If such a game can be identified, and made public information to all subjects, do subjects play according to standard game theoretic solution concepts?[28] At a first stage in such a research program, it might be advisable to at least initially restrict the domain of game forms to such where subjects' preferences are interpersonally independent.

More generally, it seems that standard epistemic game theory works best, as a predictive tool, in game protocols where subjects' preferences are interpersonally independent, and in that subclass of such games where subjects have no incentive to strategically misrepresent their preferences. The first condition should be met in many games of interest for economics, such as strategic market interactions. However, in those settings, the second condition is usually not met; in many cases a subject does have strategic incentives to misrepresent his or her preferences. For instance, in a Cournot oligopoly interaction, it benefits a profit maximizing manager to make other managers believe that he or she is not profit maximizing, but, say, sales maximizing. However, in other cases, and in certain institutional settings for the Cournot interaction, other players' preferences can more or less be taken for granted. The empirical equilibrium approach (as well as the evolutionary game theory approach) render both these conditions irrelevant; other players' preferences are simply not known.

An avenue for further theoretical elaboration of the empirical equilibrium concept is to refine the belief requirements, to not only be, as here, consistent with observed behaviors but also consistent with some general principles concerning others' preferences and rationality.

It would also be interesting to see whether human subjects' preferences in some decision situations exhibit dynamic inconsistency. If this is the case, then the very definition of an extensive form would need to be relaxed accordingly. Another aspect of extensive-form games that could be tried empirically is simultaneity. In the formal definition of a game, time *per se* is presumed to have no relevance. Is it indeed the case that human subjects' behavior are unaffected by temporal aspects that the extensive form treat as irrelevant? For example, if subject A moves before subject B in a "Battle-of-the-Sexes" interaction, but subject B is uninformed of A's move, will this not bias play towards the end-node with the highest monetary reward for

[28] A novel experimental approach is taken in Dufwenberg and Gneezy (2000). This study is focused on subjects' expectations about each other's play and of the others' expectations of play, in a given game form.

subject A, despite the fact that these decisions are modelled as simultaneous moves in non-cooperative game theory?

References

[1] Ainslie G. (1992): *Picoeconomics,* Cambridge University Press, Cambridge UK.

[2] Asheim G. (2000): "On the epistemic foundation for backward induction", mimeo., Oslo University.

[3] Aumann R., (1995): "Backward induction and common knowledge of rationality", *Games and Economic Behavior* 8, 6-19.

[4] Aumann R. and A. Brandenburger (1995): "Epistemic conditions for Nash equilibrium", *Econometrica* 63, 1161-1180.

[5] Ben-Porath E. (1997): "Rationality, Nash equilibrium, and backwards induction in perfect information games", *Review of Economic Studies* 64, 23-46.

[6] Binmore K., A. Shaked and J. Sutton (1985): "Testing noncooperative bargaining theory: a preliminary study", *American Economic Review* 75, 1178-1180.

[7] Binmore K., J. McCarthy, G. Ponti, L. Samuelson and A. Shaked (1999): "A backward induction experiment", mimeo.

[8] Björnerstedt J. and J. Weibull (1996): "Nash equilibrium and evolution by imitation", in K. Arrow *et al* (eds), *The Rational Foundations of Economic Behaviour.* MacMillan Press Ltd, London.

[9] Blume L., Brandenburger A. and E. Dekel (1991): "Lexicographic probabilities and choice under uncertainty", *Econometrica* 59, 61-79.

[10] Bolton G. and A. Ockenfels (2000a): "ECR: A theory of equity, reciprocity and competition", *American Economic Review* 90, 166-193.

[11] Bolton G. and A. Ockenfels (2000b): "A stress test of fairness measures in models of social utility", mimeo., Penn State University.

[12] Bosch-Domènech A., J. García-Montalvo, R. Nagel and A. Satorra (2000): "One, two, (three), infinity: newspaper and lab beauty-contest experiments", mimeo., Universitat Pompeu Fabra.

[13] Brandts J. and B. MacLeod (1995): "Equilibrium selection in experimental games with recommended play", *Games and Economic Behavior* 11, 36-63.

[14] Brandts J. and C. Solà (2000): "Reference points and negative reciprocity in simple sequential games", *Games and Economic Behavior*, forthcoming.

[15] Camerer C. and T.-H. Ho (1999): "Experience-weighted attraction learning in normal form games", *Econometrica* 67, 827-874.

[16] van Damme E. (1987): *Stability and Perfection of Nash Equilibria*, Springer Verlag (Berlin).

[17] Dufwenberg M. and U. Gneezy (2000): "Measuring beliefs in an experimental lost wallet game", *Games and Economic Behavior 30, 163-182.*

[18] Eyster E. and M. Rabin (2000): "Cursed equilibrium", mimeo., University of California at Berkeley.

[19] Falk A., E. Fehr and U. Fischbacher (1999): "On the nature of fair behavior", mimeo., Zürich University.

[20] Geanakoplos J., D. Pearce and E. Stacchetti (1989): "Psychological games and sequential rationality", *Games and Economic Behavior* 1, 60-79.

[21] Glaser E., D. Laibson, J. Scheinkman and C. Soutter (2000): "Measuring trust", *Quarterly Journal of Economics (*August*)*, 811-846.

[22] Güth W., R. Schmittberger and B. Schwarze (1982): "An experimental analysis of ultimatum bargaining", *Journal of Economic Behavior and Organization* 3, 376-388.

[23] Harris C. and D. Laibson (2001): "Dynamic choices of hyperbolic consumers", *Quarterly Journal of Economics* 69, 935-958.

[24] Harsanyi J. (1967-8): "Games with incomplete information played by Bayesian players", *Managemenent Science* 14, 159-182, 320-334, 486-502.

[25] Johnson E., C. Camerer, S. Sen and T. Rymon: "Detecting failures of backward induction: monitoring information search in sequential bargaining", Columbia School of Business, mimeo.

[26] Kagel J. and A. Roth (eds.) (1995): *The Handbook of Experimental Economics*, Princeton University Press (Princeton, NJ).

[27] Kalai E. and E. Lehrer (1995): "Subjective games and equilibria", *Games and Economic Behavior* 8, 123-163.

[28] Kuhn H. (1950): “Extensive games”, *Proceedings of the National Academy of Sciences* 36, 570-576.

[29] Kuhn H. (1953): “Extensive games and the problem of information”, *Annals of Mathematics Studies* 28193-216.

[30] Laibson D. (1997): “Golden eggs and hyperbolic discounting”, *Quarterly Journal of Economics* 62, 443-479.

[31] Mitzkewitz M. and R. Nagel (1993): “Experimental results on ultimatum games with incomplete information”, *International Journal of Game Theory* 22, 171-198.

[32] Nagel R.-M. (1995): “Unraveling in guessing games: an experimental study”, *American Economic Review* 85, 1313-1326.

[33] Nash J. (1950): “Non cooperative games”, PhD thesis, Department of Mathematics, Princeton University.

[34] Ochs J. and A. Roth (1989): “An experimental study of sequential bargaining”, *American Economic Review* 79, 355-384.

[35] Osborne M. and A. Rubinstein (1994): *A Course in Game Theory*. MIT Press (Cambridge, MA).

[36] Pearce D. (1984): “Rationalizable strategic behavior and the problem of perfection”, *Econometrica* 52, 1029-1050.

[37] Rabin M. (1993): “Incorporating fairness into game theory and economics”, *American Economic Review* 83, 1281-1302.

[38] Radner R. (1980): “Collusive behavior in noncooperative epsilon-equilibria of oligopolies with long but finite lives”, *Journal of Economic Theory* 22, 136-154.

[39] Reny P. J. (1993): “Common belief and and the theory of games with perfect information”, *Journal of Economic Theory* 59, 257-274.

[40] Roth A., M. Malouf and J. Murnighan (1981): “Sociological versus strategic factors in bargaining”, *Journal of Economic Behavior and Organization* 2, 153-177.

[41] Roth A. and I. Erev (1995): “Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term”, *Games and Economic Behavior* 8, 164-212.

[42] Rubinstein A. (1991): "Comments on the interpretation of game theory", *Econometrica* 59, 909-924.

[43] Segal U. and J. Sobel (2001): "Tit for tat: Foundations of preferences for reciprocity in strategic settings", mimeo.

[44] Selten R. (1965): "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit", *Zeitschrift frür die gesamte Staatswissenschaft* 12, 301-324.

[45] Selten R. (1967): "Die Strategiemethode zur Erforschung des eigeschrankt Rationalen Verhaltens im Rahmen Eines Oligopolexperiments", in Sauerman H. (ed.), *Beitrage zur Experimentellen Wirtschaftsforschung*, J.C.B. Mohr Publishing (?).

[46] Søvik Y. (2000): "Strength of dominanec and depth of reasoning", mimeo., Oslo University.

[47] Tan T. and S.Werlang (1988): "The Bayesian foundations of solution concepts of games", *Journal of Economic Theory* 45, 370-391.

[48] Zamir S. (2000): "Rationality and emotions in ultimatum bargaining", mimeo., the Hebrew University and LEI/CREST (Paris).