INTROSPECTION AND EQUILIBRIUM SELECTION IN 2x2 MATRIX GAMES*

Gonzalo Olcina and Amparo Urbano**

WP-AD 93-01

* We would like to thank an anonymous referee for helpful comments. The usual disclaimer applies.

** Universitat de València.

Editor: **Instituto Valenciano de Investigaciones Económicas, S.A.**
Primera Edición Mayo 1993.
ISBN: 84-482-0176-0
Depósito Legal: V-1518-1993
Impreso por KEY, S.A., Valencia.
Cardenal Benlloch, 69, 46021-Valencia.
Printed in Spain.

INTROSPECTION AND EQUILIBRIUM SELECTION IN 2x2 MATRIX GAMES

Gonzalo Olcina and Amparo Urbano

ABSTRACT

Game theory lacks an explanation of how players' beliefs are formed and why they are in equilibrium. This is the reason why it has failed to make significant advances with the problem of equilibrium selection even for quite simple games, as 2x2 games with two strict Nash equilibria.

Our paper models the introspection process by which the selected equilibrium is achieved in this class of games. Players begin their analysis with imprecise priors, obtained under weak restrictions formulated as Axioms. For a large class of reasoning dynamics we obtain as the solution the risk dominant Nash equilibrium.

EN BLANCO

I. INTRODUCTION.

Equilibrium analysis has dominated game theory from its beginnings. But game theory lacks an explanation of how players' beliefs are formed and why they are accurate. As a consequence, it has failed to make significant advances with the problem of equilibrium selection even for quite simple games as 2X2 matrix games with two strict Nash equilibria. In fact, both strict equilibria satisfy the strongest equilibrium refinements, for example, strategic stability in the sense of Kohlberg and Mertens(1986). This state of the problem strongly contrasts with the fact that common sense dictates an "obvious way to play", at least in some of these games. Namely, the, so called, games of "common interest". Moreover, common sense seems supported by experimental evidence (Van Huyck et al.1988).

Our paper models the introspective process by which the selected equilibrium - the solution - is achieved in this class of games. Players begin their analysis with imprecise priors- sets of priors- obtained under very weak restrictions formulated as Axioms. These imprecise priors reflect the global preliminary incentives of the players in the game situation. For a large class of reasoning dynamics players will achieve an equilibrium in beliefs that corresponds to the risk dominant Nash equilibrium (NE) in the sense of Harsanyi and Selten.

There exist some justifications of this equilibrium for the class of 2X2 symmetric "coordination" games in the recent literature on evolutive games (for example, Kandori et al. 1991). Our model is also a justification

of this particular equilibrium but in the context of a one-shot play interpretation of the NE. In particular, through a boundedly rational model of preplay introspection. Moreover, our results are stronger in the sense that we isolate the risk dominant NE not only for "coordination" games but also for 2X2 games with a conflict of interest, that is, Battle of the Sexes type games. Additionally, we prove that in non-generic 2X2 games our theory selects in a quite natural way a Perfect equilibrium as the solution of the game.

Aumann (1987) gives Bayesian foundations to equilibrium analysis and proves that if there is common knowledge of Bayesian rationality among the players and common priors, the players must coordinate in beliefs in a Nash equilibrium. As many authors have remarked, Savage's theory is entirely and exclusively a consistency theory. And it says nothing about how players come to have the beliefs ascribed to them. In this sense, Aumann's result is, at most, also a consistency result. But, in game theory the question of where beliefs come from cannot sensibly be ignored. In other words, where do these common priors come from ?, and, if there are several equilibria, in which one will the players coordinate ?

One possible explanation is experience by learning as in the models on 2X2 games of Hendon et al.(1990) and Eichberger et al.(1990) among others. Players play the same game repeatedly and base their expectations about current play on their past observations. If you obtain convergence in the limit to a Nash equilibrium (NE), you have an explanation that justifies equilibrium analysis.

But learning does not give an answer to the old conventional problem of game theory: if rational and intelligent players face a one-shot non-cooperative game without pre-play communication (i.e. a contest), what will they play? In other words, can we predict, as theorists, a solution for the game ?

In this eductive context (Binmore) the "merits" of Nash equilibrium derive from an old argument, due to Von Neumann and Morgensten, that can be summarized as follows : if a game has a unique solution (when played by rational and intelligent players), then this solution must be an equilibrium. Therefore, the special status of equilibrium analysis in traditional game theory is conditional on the game having a unique solution.

The conclusion seems obvious: we have to model the introspective process by which rational players arrive at beliefs about one another.

Any theory on rational introspection has to model explicitely both, the players' initial prior formation and their reasoning dynamics. Although both evolutive and introspective models share the dynamic feature, any initial condition ("prior") is possible in the former and the analyst is interested in the stability properties of the rest points of the evolutive dynamics or in global convergence results as in Kandori et al.(1991).

But in an eductive context, even if players are boundedly rational, they will not be so naïve to assign, for example, positive probability to strongly dominated strategies. Neither will they ignore the underlying

symmetries of the game. In other words, intelligent players will be uncertain, except in very simple games as the one-shot Prisoners' Dilemma, about what their initial priors must be, but they understand that, whatever they are, they have to satisfy some few weak properties.

There has been already some few attempts in the literature to model this thought process. Among others, Harsanyi and Selten's theory is probably the best known of the Bayesian approaches to the problem of equilibrium selection. Their tracing procedure seeks to model the manner in which Bayesian players will reason their way to an equilibrium.

But, there are, at least, two problems with this work. The first one is related to the formation of initial priors in order to begin the analysis, or what Harsanyi and Selten call the players' preliminary theory. In particular, they assume that players begin their analysis with a unique and common prior that is common knowledge among them. This implies to assume implicitely that there is a unique rational preliminary analysis of the game.

Instead of that, we think that we have to accept the possibility of imprecise priors, i.e. sets of initial priors. In other words, instead of assuming a unique rational way of forming priors, we only require very weak restrictions or properties that rational priors should satisfy.

The second problem, is that we think that a "general" theory has to obtain convergence of expectations for a large class of introspective dynamics.

In this paper we give a solution to these problems for the class of 2X2 games with two strict NE. The main idea of our work is that, although in this class of games all strategies are rationalizable, initial priors can be "bounded" more than rationalizability suggests. One possible way to attack this issue might be to consider that players, when forming a preliminary theory and once they eliminate non-rationalizable strategies, are facing a decision problem under uncertainty with "complete ignorance" about the states' likelihood, where the states are their opponents' actions. We could then apply some of the classical criteria for this context in Decision Theory: maximin, Savage's minimax regret, Hurwicz's Index, Principle of Insufficient Reason ...

However, our approach is somewhat different. We look for weak requirements that must be satisfied by the players' criteria to form preliminary decisions. Basic requirements on players' rationality and about their consciousness that they are going to play a one-shot non-cooperative game without pre-play communication.

The first requirement is that a rational player cannot get confused by the particular "labelling" of the game . In other words, changes in strategically irrelevant aspects of the game cannot make him change his preliminary theory.

In second place, given that we want a purely noncooperative theory (based on considerations of individual rationality), players will only employ the best reply structure of the game in order to elaborate their preliminary

theory. That is, we take a strict methodological individualism viewpoint and hence we do not consider collective rationality criteria, such as payoff dominance. And if it is obtained in the solution of some particular games is just as a byproduct of the players' individual rational analysis[1].

Finally, we think that all preliminary theory for 2X2 games should satisfy a very weak requirement of payoff monotonicity.

A first result that we obtain, interesting on its own, is that these requirements, formulated as axioms, yield two reasonable criteria on prior formation. In fact, these criteria, that reflect the relatioship between the incentives to play the strategies and their assignment of prior probability by the players, could perfectly be the "primitives" of our theory. Applied to 2X2 games, they define a set of initial priors -a common imprecise prior - that is a proper subset of the set of rationalizable mixed strategies.

The main result of our paper for 2X2 games with two strict pure NE and one mixed NE, is the following. If the players begin their introspective analysis with preliminary theories defined by these imprecise priors and their reasoning dynamics are smooth enough, they will achieve an equilibrium

---

(1) Note that these first two requirements eliminate the possibility of exogenous expectations influencing the players' analysis (focal points, conventions, ...). Our approach, like most of game theory, depends solely on the assumption of endogenous expectations, i.e., based only in internal factors of the game.

in beliefs that corresponds to the risk-dominant NE (in the terminology of Harsanyi and Selten). In case that there is no risk-dominant NE, they will achieve the mixed NE as an equilibrium in beliefs.

Finally, we show, for non-generic 2X2 games, how our theory eliminates weakly dominated strategies from the solution without "exogenous" assumptions such as, completely mixed priors, "trembles", and so on.

The rest of the paper is organized as follows. Section II presents notation and some preliminaries. The core of the paper are sections III and IV. In section III we formulate as axioms the requirements about the players' preliminary decision criteria and we deduce formally from these axioms the two basic criteria that define the set of initial priors. In section IV we define a class of introspective dynamics and we obtain the convergence results for 2X2 games with two strict NE. We divide the analysis in two subclasses : common interest games and games with a conflict of interest, i.e. Battle of the Sexes type games.

In section V we present the convergence to a perfect equilibrium in non-generic 2X2 games. Finally, section VI, is dedicated to final comments on results and possible future extensions.

II. NOTATION AND PRELIMINARIES.

A 2X2 matrix game G will be represented by:

$$G = (S_i, u_i)_{i=1,2} = (S , u)$$

where i = 1,2 are the players; $S_i = \{s_i, s_i'\}$ is player i's set of pure strategies or actions and $u_i : S_1 X S_2 \Rightarrow \mathbb{R}$, is player i's utility or payoff function. As usual, $S = S_1 X S_2$, and $u = u_1 X u_2$. The elements of G are assumed to be common knowledge among the players.

We can also represent the game G in matrix form as :

1		s_2	2	s_2'
	s_1	a_{11}, b_{11}		a_{12}, b_{12}
	s_1'	a_{21}, b_{21}		a_{22}, b_{22}

Figure II-1

Let M_i be the set of player i's mixed strategies, i.e. probability distributions on S_i.

$$M_i = \{ q_i , q_i' \in \mathbb{R}_+ : q_i + q_i' = 1 \}$$

We can think of a mixed strategy of player j as the probability q_j that he assigns to his first pure strategy s_j (and $(1-q_j)$ to his second pure strategy s_j'). Therefore, we can represent the set of j's mixed strategies as the interval [0,1]. As usual, we can extend the domain of the payoff functions, calculating the expected payoff to each player for a given mixed strategy combination.

Notice that we can also interpret $(q_1,1-q_1)$, for example, as the conjecture or belief of player 2 on player 1's pure strategies.

Let $R_i(q_j)$ denote the set of i's best-replies to j's mixed strategy q_j for $j\neq i$, i.e.

$$R_i(q_j) := \underset{q_i \in [0,1]}{\operatorname{argmax}}\, u_i(q_i,q_j)$$

By compactness of [0,1] and continuity of u_i, $R_i(q_j)$ is non-empty for all q_j. R_i is a correspondence, $R_i:[0,1]\Rightarrow[0,1]$. We construct the vector best-reply correspondence by:

$$R(q) := R_1(q_2) \times R_2(q_1)$$

The best-reply structure of a 2X2 game G consists of the pairs $([0,1], R_i)_{i=1,2}$ given as above. Note that different payoff functions u_i may give rise to the same best-reply structure.

As can be easily proved, one transformation in the matrix of Figure II-1 that preserves the best-reply structure is the one consisting in adding a constant to a column of player 1's payoffs and/or to a row of player 2's payoffs.

In particular, let us define the following numbers:

$$\begin{aligned} a_1 &= a_{11} - a_{21} \\ a_2 &= b_{11} - b_{12} \\ b_1 &= a_{22} - a_{12} \\ b_2 &= b_{22} - b_{21} \end{aligned} \qquad \text{(II-1)}$$

Then, the game of Figure II-1 has the same best-reply structure as the following one :

1 \ 2	s_2	s'_2
s_1	a_1, a_2	0,0
s'_1	0,0	b_1, b_2

Figure II-2

Therefore, for any concept which is best-reply invariant, as for example the set of NE , any 2X2 game G can be diagonalized and analyzed as the game shown in Figure II-2. Let us classify the three subclasses of generic 2X2 games presented as in Figure II-2, as it is usual in the literature.

1) G has exactly one NE in mixed strategies if and only if :

$$\text{either } [\ a_1<0, b_1<0 \text{ and } a_2>0, b_2>0\ ]$$
$$\text{or } [\ a_1>0, b_1>0 \text{ and } a_2<0, b_2<0\ ]$$

2) G has exactly one strict NE in pure strategies iff :

$$\text{either } a_1.b_1 < 0 \quad \text{or} \quad a_2.b_2 < 0.$$

3) G has exactly two strict NE in pure strategies and one mixed NE iff:

$$\text{either } a_i>0,\ b_i>0,\ i=1,2$$
$$\text{or} \quad a_i<0,\ b_i<0,\ i=1,2$$

This gives a characterization of all generic 2X2 games, i.e. with $a_i, b_i \neq 0$, i=1,2. In between these three subclasses lie all non-generic 2X2 games where one or more of the a_i, b_i are zero.

Our main interest here is to analyze 2X2 games with two strict NE. Hence, let us concentrate in case 3). Assume, without loss of generality, $a_i, b_i > 0$, for i=1,2.

Player i's best-reply correspondence $R_i(.)$ is, for $i \neq j$:

$$R_i(q_j) = \begin{cases} [1] \text{ if } q_j > b_i/a_i + b_i \\ [0,1] \text{ if } q_j = b_i/a_i + b_i \\ [0] \text{ if } q_j < b_i/a_i + b_i \end{cases} \qquad \text{(II-2)}$$

The two strict NE are [1,1] i.e.(s_1, s_2),and [0,0] i.e.(s_1', s_2'). And, $(b_2/a_2+b_2 , b_1/a_1+b_1)$ is the mixed NE.

Definition II-1.

The stability set $B(s_i)$ of player i's pure strategy s_i is the set of all mixed strategies of his opponent j for which s_i is a best-reply.

We list some standard properties of the B(.) in 2X2 games with two strict NE.

Lemma II-1.

For all $i \in \{1,2\}$,

a) $B(s_i)$ and $B(s_i')$ are closed subsets of [0,1].

b) $B(s_i) \cup B(s_i') = [0,1]$

c) $B(s_i) \cap B(s_i') = \{ b_i/a_i+b_i \}$

Proof: Follows trivially from the definitions and from (II-2). ■

Using Lemma II-1 is obvious that we can easily compare the stability sets' size of different strategies by means of their Lebesgue measure. We will use this possibility in the next section in order to define a relation between players' strategies.

Definition II-2.

A renaming of players and actions in a game G = (S,u) is a system of mappings one-to-one that relates G to another game G'=(S',u'). The old names of players and actions in G are replaced by new names in G'.

Therefore, a renaming consists of a mapping β from the players set of G onto the player set of G'and, for each player i a mapping f_i that maps his strategy set S_i onto β(i)'s strategy set $S'_{\beta(i)}$ in G'. All these mappings are one-to-one.

The payoffs of the players are Von Neumann-Morgensten utilities. Since these utilities are determined only up to positive linear transformations and since interpersonal comparisons are irrelevant in a context of a purely non-cooperative situation, a game G remains essentially unchanged if each player's payoff is subjected to a different positive linear transformation.

Definition II-3.

Two games G =(S,u) and G'=(S,u') with the same set of pure strategy combinations S are equivalent if constants $\alpha_i>0$ and γ_i can be found for every i, such that,

$$u'_i(s)=\alpha_i u_i(s)+\gamma_i$$

holds for every $s \in S$ and every player i.

Definition II-4.

An isomorphism from a game G=(S,u) to G'=(S',u') is a system $f=(f_i)_{i=1,2}$ *of one-to-one mappings* f_i *of i's strategy set* S_i *in G onto* $\beta(i)$*'s strategy set in G', such that the following conditions are satisfied:*

1-. The mapping β *is a one-to-one mapping of the player set of G onto the player set of G'.*

2-. The payoff functions u and u' are related as :

$$u'_{\beta(i)}(f(s)) = \alpha_i . u_i(s) + \gamma_i \quad \text{for every } i \text{ and every } s \in S,$$

$$\text{with constants } \alpha_i > 0 \text{ and } \gamma_i .$$

where f(s) is the combination $s' \in S'$ *whose components are related to those of S as follows :*

$$s'_k = f_i(s_i) \text{ with } k = \beta(i) \text{ for } \forall\ i.$$

Obviously, an isomorphism involves any combination of renaming with a system of positive linear transformations. It can be easily seen that an isomorphism preserves best-reply relations and carries NE into NE,i.e. the best-reply structure and the set of NE of G are invariant with respect to isomorphisms[(2)].

---

(2) For more detailed definitions on isomorphisms the reader can consult Harsanyi and Selten (1988).

III. AXIOMS ON PRELIMINARY THEORIES AND THE COMMON IMPRECISE PRIOR.

If a player i tries to figure out an initial prior- a preliminary theory- over the actions of his opponent j , then he has to think about what could be a "reasonable" decision criteria for j. But these criteria will also be reasonable for him. Alternatively, we can imagine a rational outside observer who shares the players' common knowledge and thinks about rational preliminary theories for them in the game situation. It is clear that rational and intelligent players must form their opinions in the same way as the outside observer.

We do not assume that there is a unique rational way to form initial expectations over the players' behavior. Instead of that, we impose some weak requirements on the players' criteria when forming initial priors.

Definition III-1.

A preliminary theory for player i is a probability distribution over his actions, i.e. a $q_i \in [0,1]$, *that tells him to play action* s_i *with probability* q_i.

Next we state three axioms that reflect properties that both players expect to be satisfied by the preliminary theories that they may possibly have.

AXIOM 1. *Invariance with respect to isomorphisms.*

Let f be an isomorphism from G to G'.Then the preliminary theory for player i tells him to play $f(s_i)$ with probability q_i in G' if and only if it tells him to play s_i with probability q_i in G.

AXIOM 2. *Best-reply invariance.*

Let G'=(S,u') be a game with the same best-reply structure as G=(S,u). Then the preliminary theory for player i tells him to choose s_i with probability q_i in G' if and only if it tells him to choose s_i with probability q_i in G.

AXIOM 3. *Payoff monotonicity.*

Let q_i be the preliminary theory of player i in G. Let M be the maximum possible payoff according to action s_i in G. Suppose another game G' where M'>M and nothing else changes with respect to G, then the preliminary theory q_i' in G' should satisfy $q_i' \geq q_i$.

Interpretation of the Axioms:

Axiom 1 implies that rational players cannot get confused by the labelling of the game in a world without exogenous expectations. Strategically irrelevant aspects of the game have no influence in their decisions. In particular, invariance with respect to renaming may be considered a symmetry property. A very powerful implication is that in symmetric games players will expect symmetric solutions. As we said before, isomorphisms preserve best-reply relations.

Axiom 2 implies that only the best-reply structure is relevant in decision making in a purely non-cooperative context. This is the only information that the players have in a world without exogenous expectations, about the global preliminary incentives of rational players, i.e. players that optimize against subjective conjectures.

It is well-known in the literature, that best-reply invariance conflicts with payoff dominance. But, we believe that there is no room for payoff dominance considerations in a purely individualistic rational theory. Note also that two classical decision criteria do not satisfy this axiom. Namely, the maximin criterium and the one based on Hurwicz's index of optimism-pessimism.

Axiom 3 seems obvious if we think in a game against Nature with "complete ignorance" about the states' likelihood. In our opinion, it remains valid in a strategic context, as a reasonable property for "preliminary theories" of the players. It is worth insisting that not only Axiom 3 but all of them, have to be applied to preliminary decision criteria and therefore to the initial priors with which players begin their reasoning processes. Players start this process in a state of "complete ignorance", once they have eliminated non-rationalizable strategies. Obviously, this will change during their introspective analysis and their decisions at the end of the process, when they eliminate all the strategic uncertainty, might not satisfy Axiom 3.

Notice also, that within the classical decision criteria only Savage's criterium and the one based on the Principle of Insufficient Reason (PIR) satisfy the three Axioms.

Definition III-2.

The set of initial priors consistent with Axioms 1,2 and 3 will be called the common imprecise prior (CIP).

Now we prove that if the players' preliminary theories satisfy the three axioms already defined, then the players' CIP is the one defined by the Criteria that we present next. Let us define first an incentive-dominance relation between pure strategies. We compare the stability sets'size of different strategies by means of their Lebesgue measure. Let l(A) be the Lebesgue measure of $A \subset \mathbb{R}_+$.

Definition III-3.

The pure strategy s_i incentive dominates for player i the strategy s'_i if $l(B(s_i))$ is strictly greater than $l(B(s'_i))$.

Roughly speaking, s_i incentive-dominates s'_i for player i if s_i is a best-reply for "more" mixed strategies of his opponent than s'_i is. If both stability sets have equal size then we say that there is no incentive dominance between s_i and s'_i.

We formulate now two weak criteria on initial prior formation. These criteria reflect the relationship between the incentives to play the strategies and their assignment of prior probability by the players.

CRITERIUM 1. When forming a preliminary theory, both players expect that no one of them assigns more probability to his incentive dominated strategy. That is, for i=1,2.

$$\text{if } l(B(s_i)) > l(B(s_i')) \quad \text{then} \quad q_i \geq 1-q_i.$$

$$\text{if } l(B(s_i)) < l(B(s_i')) \quad \text{then} \quad q_i \leq 1-q_i.$$

$$\text{if } l(B(s_i)) = l(B(s_i')) \quad \text{then} \quad q_i = 1-q_i.$$

This is a very weak and intuitive requirement about players'initial priors. Notice that we do not require as Harsanyi and Selten do , that on the unique initial prior the probability weights are proportional to the pure strategies'stability sets, i.e. the bicentric prior. Obviously, this particular prior satisfies Criterium 1.

Next, we want a criterium that reflects the idea that players, using their common knowledge about the game, will also realize that their incentive dominance relations are probably different in relative "intensity". For example, suppose that strategy s_1 incentive dominates s_1' for player 1 and s_2' incentive dominates s_2 for player 2. But suppose also that $l(B(s_1))$ is relatively higher with respect to $l(B(s_1'))$ than $l(B(s_2'))$ is with respect to $l(B(s_2))$. Then, both players, when forming a preliminary theory, should expect that the probability weight assigned by player 1 to s_1 will be greater than the weigth assigned to s_2' by player 2. Roughly speaking, both players know that player 1 has relatively more "motives" to play s_1 than player 2 has to play s_2'.

CRITERIUM 2. Suppose, without loss of generality, that $l(B(s_i)) > l(B(s_i'))$ and $l(B(s_j')) > l(B(s_j))$.

When forming a preliminary theory, both players will expect that:

if $l(B(s_i)) - l(B(s'_i)) > l(B(s'_j)) - l(B(s_j))$ then $q_i \geq 1-q_j$.

if $l(B(s_i)) - l(B(s'_i)) < l(B(s'_j)) - l(B(s_j))$ then $q_i \leq 1-q_j$.

if $l(B(s_i)) - l(B(s'_i)) = l(B(s'_j)) - l(B(s_j))$ then $q_i = 1-q_j$.

for $i,j = 1,2$, $i \neq j$.

It is clear that combining these two criteria we obtain a set of initial priors. The next proposition establishes our first result.

PROPOSITION III-1.

Let G be a 2X2 game with two strict NE. If the players' preliminary decision theories satisfy Axioms 1,2 and 3, then the set of admissible initial priors- the common imprecise prior- is defined by Criteria 1 and 2.

Proof: In order to avoid notation (indexes) in the proof, let us represent an initial prior on player 1's actions as (x, 1-x) and similarly, (y, 1-y) for player 2, i.e., $x:=q_1$ and $y:=q_2$.

Every game in the class of 2X2 games with two strict NE is as in Figure II-1. Without loss of generality, we assume that all the entries of this bimatrix are positive[3]. As we saw before, any such a game has the same best-reply structure as the game of Figure II-2.

---

(3) It is well-known that we can assume that all payoffs in a bimatrix game are positive. Just add the adequate positive constant to all the payoffs of the original game.

Let us, in the matrix of Figure II-2, multiply 1's payoffs by $1/b_1$ and 2's payoffs by $1/a_2$, yielding :

$$\begin{array}{c} \\ s_1 \\ s'_1 \end{array} \begin{array}{c} \begin{array}{cc} s_2 & s'_2 \end{array} \\ \begin{bmatrix} a,1 & 0,0 \\ 0,0 & 1,b \end{bmatrix} \end{array} \qquad \text{where} \quad a = a_1/b_1 \quad b = b_2/a_2$$

Again, this game has the same best-reply structure as the original one.

First of all we deduce Criterium 1. Assume a=1. Then we have a game G:

$$\begin{array}{cc} & \begin{array}{cc} & 2 \\ & s_2 \end{array} \quad s'_2 \\ \begin{array}{c} 1 \\ \\ \end{array} & \begin{array}{c} s_1 \\ s'_1 \end{array} \begin{bmatrix} 1,1 & 0,0 \\ 0,0 & 1,b \end{bmatrix} \end{array}$$

Apply the following renaming of strategies and interchange of players.

$$\beta(1)=2 \qquad \beta(2)=1$$

$$f_1(s_1)=s'_2 \qquad f_2(s_2)=s'_1$$

$$f_1(s'_1)=s_2 \qquad f_2(s'_2)=s_1$$

We obtain the following game G':

$$\begin{array}{cc} & \begin{array}{cc} & 1 \\ & s_2 \end{array} \quad s'_2 \\ \begin{array}{c} 2 \\ \\ \end{array} & \begin{array}{c} s_1 \\ s'_1 \end{array} \begin{bmatrix} b,1 & 0,0 \\ 0,0 & 1,1 \end{bmatrix} \end{array}$$

Axiom 1 implies that x=1-y'. But because of the symmetry of these games, 1-y'=1-x. (Notice that the individual decision problems for player 1 in G and in G' are identical.) Therefore, x=1-x in G.

Now, assume a>1. This change results in a new game G'' where the only thing that has changed with respect to G is that the highest possible payoff for 1 according to action s_1 has been increased. Therefore, by Axiom 3, $x'' \geq x$, and this in turn implies, $x'' \geq 1-x''$.

Given that game G'' has the same best-reply structure as the one of Figure II-2, if a>1, i.e. $a_1 > b_1$, then $x \geq 1-x$.

On the other hand, straightforward calculation yields :

$$l(B(s_1)) > l(B(s_1')) \quad \text{if and only if } a_1 > b_1.$$

Therefore, combining the two above expressions, if $l(B(s_1)) > l(B(s_1'))$ then $x \geq 1-x$.

In like fashion we can deduce that :

$$\text{if } l(B(s_1)) < l(B(s_1')) \text{ then } x \leq 1-x.$$

$$\text{if } l(B(s_1)) = l(B(s_1')) \text{ then } x = 1-x.$$

A similar argument runs for player 2.

In order to deduce Criterium 2 we distinguish between two subclasses of the games under analysis. Namely, what we call games with a conflict of interest and games with common interest.

A game with a conflict of interest is characterized by :

$$a_i > b_i \qquad i,j = 1,2$$

$$a_j < b_j \qquad i \neq j$$

Without loss of generality, suppose that $a_1 > b_1$ and $a_2 < b_2$. Let us make the same transformation as above :

$$\begin{bmatrix} a,1 & 0,0 \\ 0,0 & 1,b \end{bmatrix}$$

where

$a = a_1/b_1 > 1$

$b = b_2/a_2 > 1$

Suppose a = b. Call this game G. Now, apply the same renaming of strategies and interchange of players as above and obtain a new game G'. By Axiom 1, x = 1-y'. But, given that a = b, by symmetry, 1-y'= 1-y. Therefore, x = 1-y.

Now, for a > b and applying Axiom 3 we obtain that x'≥x. But, player 2's decision problem does not change. Therefore, 1-y = 1-y', and then x' ≥ 1-y'.

$$a > b \text{ implies that } a_1/b_1 > b_2/a_2.$$

On the other hand, straightforward calculation yields again that :

$$l(B(s_1)) - l(B(s_1')) > l(B(s_2')) - l(B(s_2)) \text{ iff } a_1/b_1 > b_2/a_2.$$

Therefore, by the above expressions, we have proved that ,

if $l(B(s_1)) - l(B(s_1')) > l(B(s_2')) - l(B(s_2))$, then $x \geq 1-y$, in a game with a conflict of interest such that, $a_1 > b_1$ and $b_2 > a_2$.

A similar argument proves that :

$$\text{if } a_1/b_1 < b_2/a_2 \text{ then } x \leq 1-y \text{ , and}$$

$$\text{if } a_1/b_1 = b_2/a_2 \text{ then } x = 1-y.$$

A game with common interest is characterized by $a_i \geq b_i$, i=1,2.

Let us divide player 1's payoffs by $1/b_1$ and 2's payoffs by $1/b_2$. We obtain :

$$\begin{bmatrix} a,b & 0,0 \\ 0,0 & 1,1 \end{bmatrix}$$

where

$a=a_1/b_1 \geq 1$

$b=a_2/b_2 \geq 1$

Assume a = b. Then, this game carries a symmetry that implies x = y. Now, suppose a > b, i.e. $a_1/b_1 > a_2/b_2$, then by Axiom 3 a similar argument as the previous one, implies $x \geq y$.

Notice that,

$$l(B(s_1))- l(B(s_1')) > l(B(s_2))- l(B(s_2')) \text{ iff } a_1/b_1 > a_2/b_2.$$

Therefore, we conclude that :

$$\text{if } l(B(s_1))- l(B(s_1')) > l(B(s_2))- l(B(s_2')), \text{ then } x \geq y.$$

In case this inequality goes in the other direction, then $x \leq y$, and if it is a strict equality, then x = y. ■

Comment : As we will confirm later on, Criterium 2 is specially relevant in the case of games with a conflict of interest. In this case, incentive dominance relations between players' strategies point in different directions and, therefore, it is important for them to make the kind of comparisons implied by Criterium 2. On the other hand, this point is not so important in games of common interest and in fact, we will see that our theory also works nicely in this case only with Criterium 1.

IV. INTROSPECTIVE DYNAMICS : RESULTS FOR 2X2 GAMES WITH TWO STRICT NASH EQUILIBRIA.

IV-1. INTROSPECTIVE DYNAMICS.

In our theory, rational players will start with a common imprecise prior. Namely, the one obtained in the previous section. The CIP reflects their "state of confusion" or initial predictive uncertainty, but the players can reduce it through introspection, before they play the non-cooperative game just once.

For any point in the common imprecise prior (CIP) the players may do the following thought experiment : if this were the theory held by both of us, what would we play ? The only reasonable answer is to play a best-reply. But then, players can update the initial point prior and follow in the same way the thought process.

Players should do this introspection for all the set of initial priors, i.e. the CIP, because they have no a priori reason to expect one particular point prior in the set. Notice that rational players will only stop the introspection and make a firm decision when the process reaches a fixed point.

The question is, does this process converge ? The answer is yes. From any point in the CIP, the reasoning process reaches the same unique fixed point, namely, the risk dominant NE. This result is robust for any kind of introspective dynamics with a property that roughly expressed says: players update their beliefs smoothly enough in the direction of best-reply.

In particular, we employ a particular dynamic in order to prove our results. But our claim is that the theory works for a much widely class of dynamics that satisfy the above property. For example, both the beta distributions used as priors by Eichberger et al.(1991) and the generalized version of fictitous play by Hendon et al.(1990), when appropiatedly interpreted as reasoning processes and applied to the CIP, yield the same convergence result. At the end of this section we will state formally the characterization of this quite large class of dynamics.

The simple introspective dynamics we are going to use is the following "small learning process".

Let $q^0 = (q_1^0, q_2^0)$ be any initial point prior in the CIP. For $\varepsilon \in (0,1)$,

$$\text{(IV-a)} \qquad q_i^{t+1} = (1-\varepsilon) q_i^t + \varepsilon . \delta(s_i^t) \qquad i=1,2.$$

$$\text{(IV-b)} \qquad \text{where } s_i^t = r_i(q_j^t) \qquad j \neq i.$$

and, where $\delta(x)$ is the Dirac measure that assigns all the weight to x, and $r_i(q_j^t)$ is some selection from player i's best-reply correspondence $R_i(q_j)$.

Notice that each player will run this updating process for him and his opponent.

IV-2. GAMES WITH A CONFLICT OF INTEREST.

Let us analyze, first of all, the case of 2X2 games with two strict NE with a conflict of interest, i.e. Battle of the Sexes type games.(We work, making use of Best-reply Invariance, with games as those of Figure II-2.)

Without loss of generality, assume that $a_1>b_1$, $a_2<b_2$ and $a_1.a_2>b_1.b_2$.

This game has two strict NE, namely, (s_1,s_2) [or $(q_1=1,q_2=1)$] and (s_1',s_2') [or $(q_1=0,\ q_2=0)$], and a mixed NE $\bar{q}=(\bar{q}_1,\bar{q}_2)$ where : $\bar{q}_1=b_2/a_2+b_2$ and $\bar{q}_2=b_1/a_1+b_1$.

From Section III is evident that the common imprecise prior is defined by :$q_1 \geq 1-q_1$, $1-q_2 \geq q_2$, $q_1 \geq 1-q_2$.

That is, $q_1 \geq 1/2$, $q_2 \leq 1/2$, and $q_1+q_2 \geq 1$.

Graphically, the best-reply structure of a typical game in this class looks as the following diagram :

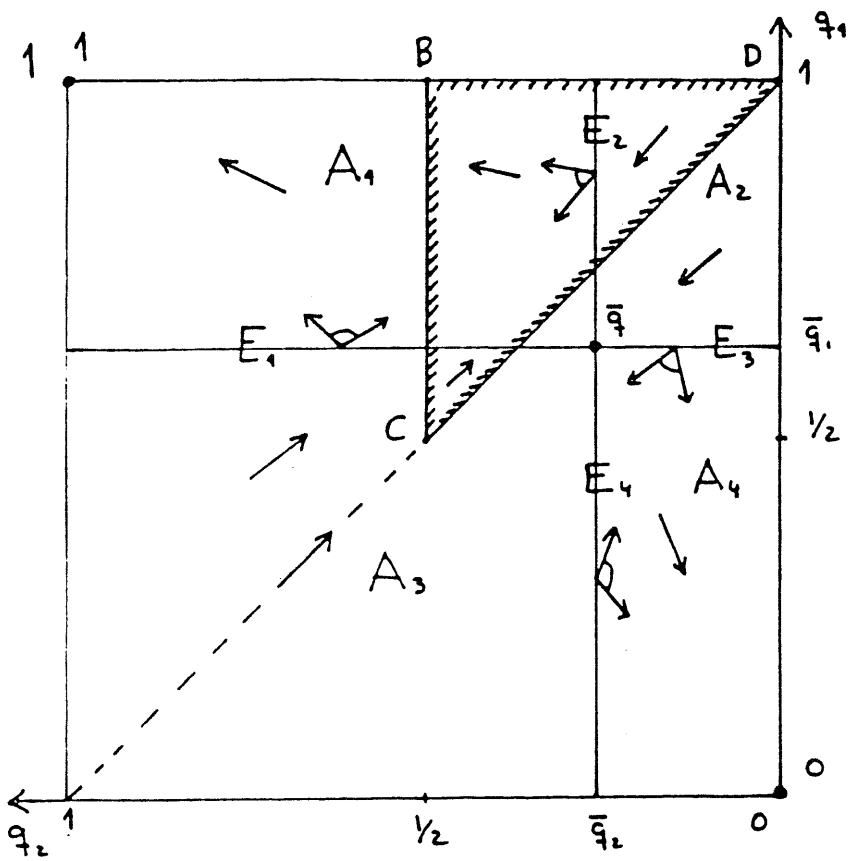

FIGURE IV-1

Notice that the rectangle A_1 is the stability set of the strategy combination (s_1,s_2) , A_2 of (s'_1,s_2) , A_3 of (s_1,s'_2) and A_4 of (s'_1,s'_2). In the segment E_1 player 1's best-reply is s_1 and player 2 is indifferent between his pure strategies. In E_2 player 2's best-reply is s_2 and player 1 is indifferent, and so on.

The arrows in the diagram point from mixed strategy combinations to their vector best-replies. The two strict NE are points (1,1) and (0,0) and the mixed NE is point $\bar{q}$. The common imprecise prior is the triangle BCD.

It is quite obvious that the phase diagram of a dynamics that updates beliefs smoothly in the direction of best reply would look as the arrows in Figure IV-1. Graphical intuition suggests that any point in region BCD will converge to (1,1), i.e. to the risk dominant NE in the terminology of Harsanyi and Selten. We prove formally this intuition.

PROPOSITION IV-1

Let G be a 2X2 game with a conflict of interest. For any prior vector q^0 *in the common imprecise prior as initial condition and for any selection r(q) from R(q), the dynamics defined by IV-a and IV-b, with* ε *sufficiently small, will converge to the risk dominant NE.*

Proof: Without loss of generality, assume that the risk-dominant NE is q = (1,1). We can express the reasoning dynamics in the following way, for i=1,2.

$$(1) \quad q_i^t = (1-\varepsilon)q_i^{t-1} + \varepsilon.\delta(r_i(q_j^{t-1})) \qquad j\neq i$$

Notice that we can also express it as:

$$(2) \quad q_i^t = (1-\varepsilon)^t q_i^0 + \varepsilon \cdot \sum_{k=0}^{t-1} \delta(r_i(q_j^k))(1-\varepsilon)^{(t-1-k)}$$

Any prior vector $q^0=(q_1^0,q_2^0)$ in the CIP satisfies:

$$q_1^0 \geq 1/2, \; q_2^0 \leq 1/2, \; q_1^0+q_2^0 \geq 1.$$

We partition the CIP in three regions and we are going to prove that from any of them we converge to the NE (1,1). See Figure IV-2.

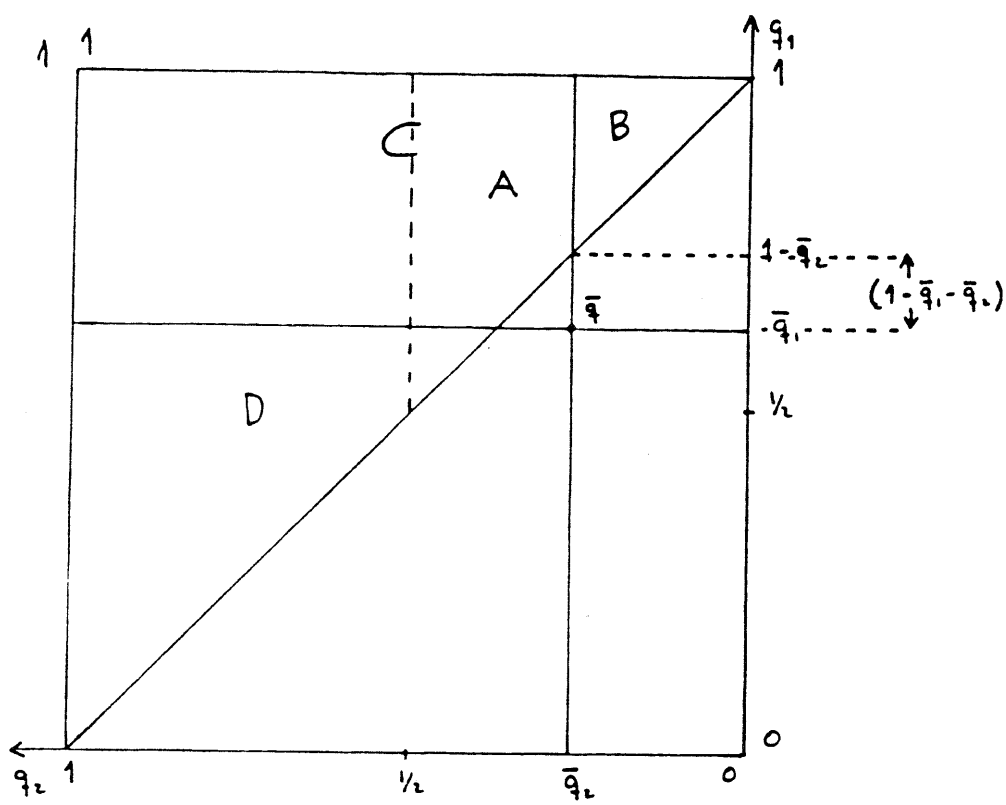

FIGURE IV-2

i) Add to the inequalities that define the CIP the following ones:

$q_1^0 \geq \bar{q}_1$, and $q_2^0 \geq \bar{q}_2$. Call the resulting subset of the CIP, region A.

Let us begin with some q^0, such that $q_1^0 > \bar{q}_1$ and $q_2^0 > \bar{q}_2$.Notice that these priors of the CIP form a proper subset of the stability set of the strategy combination (1,1). From (1) :

$$(3) \quad q_i^t - q_i^{t-1} = \varepsilon.[\delta(r_i(q_j^{t-1})) - q_i^{t-1}]$$

Therefore,

$$q_i^t \geq q_i^{t-1} \text{ as long as } q_i^{t-1} > \bar{q}_i \text{ for } i=1,2.$$

Hence, the reasoning sequence stays in (1,1)'s stability set and both players best-reply is always (1,1).(Because $(1,1) \in B[(1,1)]$.)

Substituing in (2), for i=1,2.

$$q_i^t = (1-\varepsilon)^t q_i^0 + \varepsilon.(1-(1-\varepsilon)^t/\varepsilon) =$$
$$= (1-\varepsilon)^t q_i^0 + (1-(1-\varepsilon)^t)$$

and, obviously, $q_i^t \to 1$ as $t \to \infty$.

Suppose now, that $q_2^t = \bar{q}_2$ and $q_1^t > \bar{q}_1$. Then player 2's best-reply is $R_2(q_1^t)=1$ and player 1 is indifferent, i.e. $R_1(q_2^t)=[0,1]$. But, for ε sufficiently small, whatever selection he uses from his best-reply correspondence, $q_1^{t+1} > \bar{q}_1$ and $q_2^{t+1} > \bar{q}_2$. Therefore we are back in the above case.

A similar argument applies for $q_1^t = \bar{q}_1$ and $q_2^t > \bar{q}_2$.

Therefore, whenever the reasoning process starts at a point prior in region A or the sequence of point beliefs generated by it hits region A in some finite T, the process will converge to (1,1) as desired.

ii) Add to the inequalities that define the CIP the following one: $q_2^0 < \bar{q}_2$. Notice that in this region -let us call it B- $q_1^0 > \bar{q}_1$.

We are going to prove that any reasoning sequence in this region will hit the stability set of (1,1) in finite time, for ε sufficiently small.

If the sequence stays in region B, we have for all t, $R_1(q_2^t) = 0$ and $R_2(q_1^t) = 1$.

Therefore, substituing in (2) :

$$q_1^t = (1-\varepsilon)^t q_1^0$$
$$q_2^t = (1-\varepsilon)^t q_2^0 + 1-(1-\varepsilon)^t$$

Notice that as t increases q_2^t is increasing and q_1^t is decreasing.

Notice, also, from (3), that:

$$|q_i^t - q_i^{t-1}| \leq \varepsilon \qquad \text{for } i=1,2.$$

That is, ε is an upper bound for the change in the probability that players assign to their first pure strategy in each round of the introspective process.

Call C the intersection of $E_1 \cup A_1 \cup E_2$ of Figure IV-1 with the triangle with vertices (0,1), (1,1) and (1,0). D the intersection of this triangle with A_3 and B,as before (the intersection of this triangle with A_2). See Figure IV-2.

Now, we prove that for $(q_1^t, q_2^t) \in B, q_1^t + q_2^t \geq 1$, given that $q_1^0 + q_2^0 \geq 1$.

$$q_1^t + q_2^t = (1-\varepsilon)^t (q_1^0 + q_2^0) + 1 - (1-\varepsilon)^t =$$

$$= 1 + (1-\varepsilon)^t (q_1^0 + q_2^0 - 1)$$

But, as $(q_1^0 + q_2^0 - 1) \geq 0$, then $q_1^t + q_2^t \geq 1$.

A similar argument runs for region D, as can be easily seen.

Therefore, the sequence never goes below the diagonal from (0,1) to (1,0), i.e. for all t, $q^t \in B \cup C \cup D$.

We want to show that there is T such that $q^T \in C$.

Suppose not, i.e. $\forall$ t, $q^t \in B \cup D$.

We know that $\bar{q}_1 + \bar{q}_2 < 1$, therefore $1 - \bar{q}_1 - \bar{q}_2 > 0$. Let $\varepsilon \leq (1 - \bar{q}_1 - \bar{q}_2)$.

The reasoning sequence must contain q_s from D because B is a subset of the stability set of strategy combination (0,1) and it is not possible that q^t for all t is contained in B because $(0,1) \notin B$.(For a similar reason the sequence must contain q_s from B.)

Therefore, it is possible to find T, such that, $q^T \in B$ and $q^{T+1} \in D$. But, then, we must have :

$$q_1^T > \inf\{q_1 | \ni q_2:(q_1,q_2)\in B\}=1-\bar{q}_2$$
$$q_1^{T+1} < \sup\{q_1 | \ni q_2:(q_1,q_2)\in D\}=\bar{q}_1$$

implying, $|q_1^{T+1}-q_1^T| > (1-\bar{q}_1-\bar{q}_2)$.

But, this contradicts : $|q_1^{T+1}-q_1^T| \leq \varepsilon \leq (1-\bar{q}_1-\bar{q}_2)$.

Therefore, we have proved that there is a T such that $q^T \in C$. In other words, for ϵ sufficiently small, the sequence must hit the stability set of strategy combination (1,1).

From there on, the arguments of case i) yields convergence to (1,1).

iii) Add to the inequalities that define the CIP the following one: $q_1^o < \bar{q}_1$. Notice that in this region, $q_2^o > \bar{q}_2$.

A completely similar argument as in case ii) proves convergence to (1,1). ■

Therefore, players starting with a common imprecise prior will arrive, after introspective reasoning, to an equilibrium in beliefs that corresponds

to the risk dominant NE. And, obviously, they will play it in the one-shot game.

It is worth insisting that this result only depends on the players having an introspective dynamics that satisfy some properties to be defined in Section IV-5. In particular, the result does not change if each player employs in his analysis a different dynamics within this class. The reason is obvious: the only important thing to guarantee is convergence in the "head" of each individual player.

Notice also that our theory is much less demanding, with respect to players' capabilities, that it seems at first. Even a player that does not carry over the calculations for the whole set of initial common priors and for some reason (laziness, "impulse", ...) only does the calculations for one or a few point priors from the region, will also quickly arrive to the risk dominant NE as an equilibirum in beliefs.

IV-3- GAMES WITH COMMON INTEREST

We have called 2X2 games with two strict NE with common interest the ones that satisfy :$a_i \geq b_i$, i =1,2. Let us analyze the case where $a_i > b_i$, i=1,2.

Without loss of generality, suppose that, $a_1/b_1 > a_2/b_2$.

In this games the common imprecise prior is defined by :

$$q_1 \geq 1-q_1,\ q_2 \geq 1-q_2 \text{ and } q_1 \geq q_2.$$

Graphically:

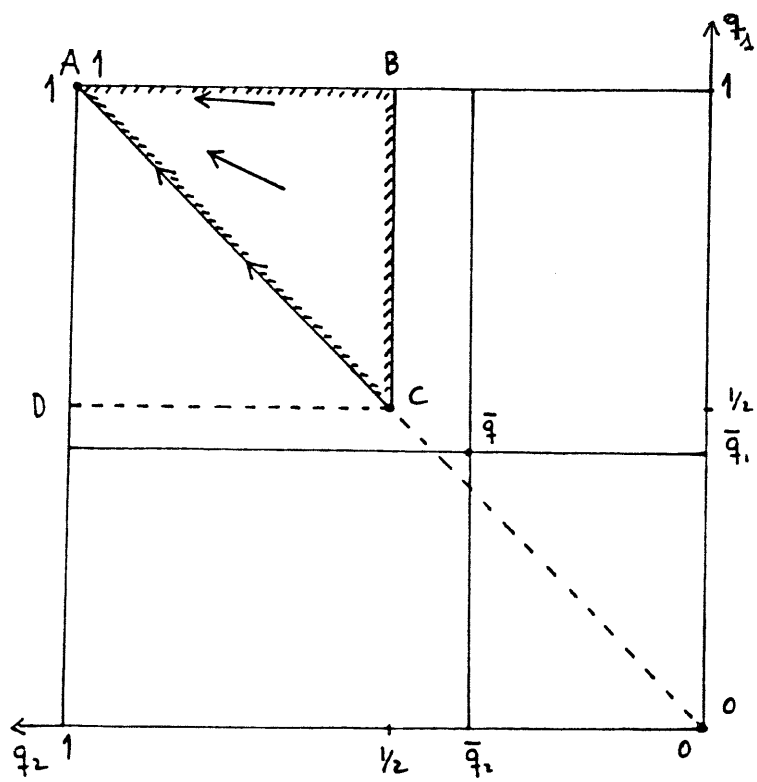

FIGURE IV-3

Notice that in these games the mixed NE $\bar{q}$ is always below the diagonal from (0,1) to (1,0), i.e. $\bar{q}_1+\bar{q}_2 < 1$.

This implies, as can be seen in the diagram - Figure IV-3 -, that the common imprecise prior - the rectangle ABC - is completely included in the stability set of (1,1), i.e. the risk dominant (and payoff dominant) NE.

PROPOSITION IV-2.

Let G be a 2X2 game with common interest. For any prior vector q^o in the common imprecise prior as initial condition, the dynamics defined by IV-a and IV-b, will converge to the risk dominant Nash equilibrium.

Proof: Just apply the arguments of case i) on the proof of Proposition IV-1.

Comment : Notice that for this result we do not need Criterium 2 of Section II. The rectangle ADCB is the region defined by Criterium 1 alone. And it is also completely included in the stability set of (1,1).

Therefore, in games of common interest,even players that do not make "incentiveness" comparisons within them (Criterium 2) will coordinate in the risk dominant NE in a one-shot game.

Notice also, that in this subclass of 2X2 games players do not need to update beliefs "smoothly". They just can apply best-reply. We think this is a good feature of our theory. In games with common interest where common sense and experimental evidence dictates a straightforward solution, our theory selects this solution also in a straightforward manner.

IV-4. SOME PARTICULAR CASES.

We have left apart intentionally some particular cases in order to illustrate some features of our theory.

The first case will be games with a conflict of interest, but in which $a_1/b_1 = b_2/a_2$.

Notice that our theory will see as equivalent any such game with a corresponding symmetric Battle of the Sexes, obtained dividing 1's payoffs by b_1 and 2's payoffs by a_2.

In this case the mixed NE $\bar{q}$ is on the diagonal from (0,1) to (1,0), i.e. $\bar{q}_1+\bar{q}_2= 1$.

The common imprecise prior will be defined by :

$$q_1 \geq 1-q_1,\ 1-q_2 \geq q_2 \text{ and } q_1 = 1-q_2.$$

Alternatively, $q_1 \geq 1/2$, $q_2 \leq 1/2$ and $q_1+q_2= 1$.

Graphically , the common imprecise prior is the segment CA.

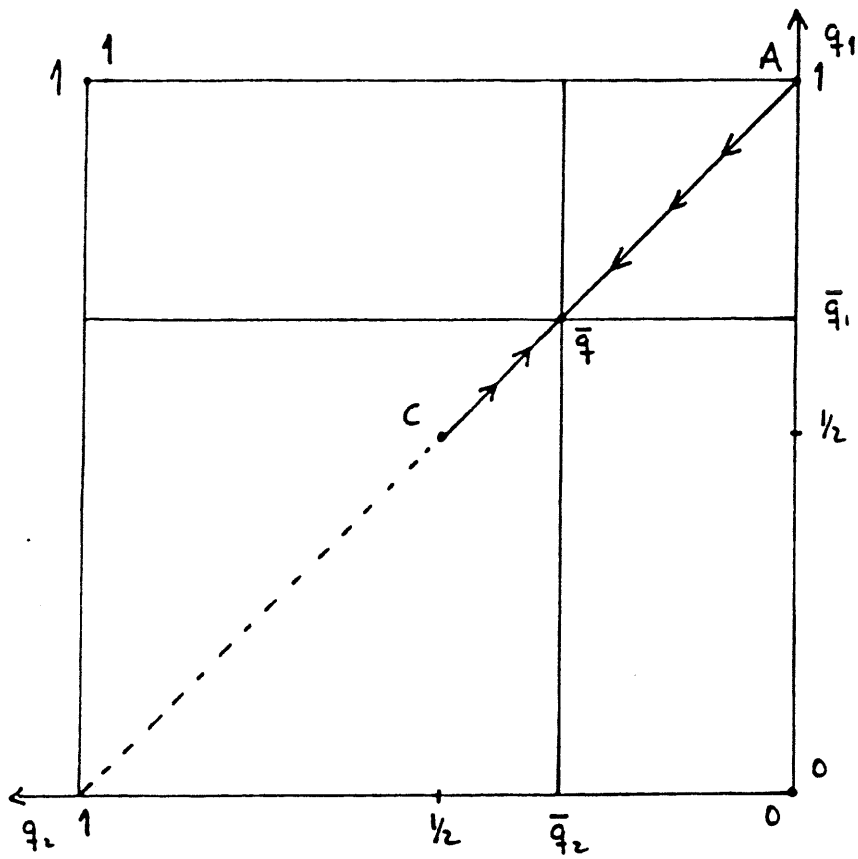

FIGURE IV-4

Assume the following selection $r^*(q)$ from the best-reply correspondence. Let $q = (q_1,q_2)$ where, for example, $q_2 = \bar{q}_2$, then $r^*(q_1,\bar{q}_2) = (q_1,R_2(q_1))$ where $R_2(q_1)$ is 1 if $q_1>\bar{q}_1$ and 0 if $q_1<\bar{q}_1$. In words, when the players are indifferent (with the currently held theory), then they do not change their beliefs.

It can be easily proved[4] that, with this selection function, our introspective dynamics yield convergence to an equilibrium in beliefs that corresponds to the mixed NE $\bar{q}$ (as illustrated in the diagram.)

We claim that this is a quite natural selection for a reasoning process. In this kind of process there is no play in real time, there are no "observations", things just happen in players' heads.(Fictitous play). Therefore, if, given the current beliefs, player j is completely indifferent between his pure strategies, it seems quite sensible not to change q_j. Any other selection would be to some extent arbitrary because it implies updating in some particular direction without any justification. Notice, also, that this particular selection implies that players can recognize equilibria, i.e. they can recognize when their beliefs are at equilibrium.

With another selection (or tie-breaking rule), we obtain convergence to some NE, but to which one of them strongly depends on the selection function used by the players in their computations. It is in this sense that we can say that mixed NE, when isolated as solution, seems more "unstable" than pure NE.

Notice that this kind of games with a conflict of interest, correspond to games where, in Harsanyi and Selten's terminology, there is no risk dominance relation between the two strict pure NE. Although, in our terminology, there is incentive dominance between the players'strategies.

---

(4) We do not prove it just not to be redundant. The proof employs similar arguments to those of Proposition IV-1.

Let us analyze briefly another particular case within the class of games with common interest.

Assume $a_1=b_1$ and $a_2=b_2$.

Notice, that in our context, any such game should be equivalent for our players to the game :

$$\begin{array}{c} \\ s_1 \\ s'_1 \end{array} \begin{array}{c} \begin{array}{cc} s_2 & s'_2 \end{array} \\ \begin{bmatrix} 1,1 & 0,0 \\ 0,0 & 1,1 \end{bmatrix} \end{array}$$

This is the prototype of a pure coordination game. Again, it has two strict NE, namely, (s_1,s_2) and (s'_1,s'_2) and a mixed NE $\bar{q}$ = (1/2 , 1/2).

If we apply our theory we obtain that the common imprecise prior, in this particular game, "collapses" into a point prior.

$$q_1 = 1-q_1,\ q_2 = 1-q_2,\ \text{i.e. } q_1 = 1/2,\ q_2 = 1/2.$$

Therefore, the unique prior consistent with Axioms 1,2 and 3 of Section III is the "centroid" ($q_1 = 1/2$, $q_2 = 1/2$) and there is no place for introspective dynamics, because players are completely indifferent between their strategies when they hold these beliefs. In other words, (1/2, 1/2) is already the equilibrium in beliefs and, obviously, corresponds to the mixed NE.

Notice that also in this case, there is no risk dominance between the two strict pure NE. And again, our theory selects the mixed NE as the equilibrium in beliefs in the players' minds.

It also seems striking that there is no place for introspection (reasoning process) in this game. The reason is quite clear: there is no incentive-dominance between players' strategies. This is the formal cause behind the fact that the CIP collapses to a point prior which is already an equilibrium in beliefs. In other words, in our context (a one-shot game without preplay communication) players cannot reduce their uncertainty through introspection and they have to stand with their initial priors. Notice also, that in the symmetric games analyzed in this subsection, the Axiom of Invariance with respect to isomorphisms is very powerful. In some sense, players should expect symmetric priors and obtain symmetric solutions.

In a more informal discussion, this seems also very reasonable. Remember that we are analyzing a one-shot non-cooperative game without preplay communication and without any room for exogenous expectations (any symmetry-breaking convention). In this context, what would had been surprising is a theory that, for example in the pure coordination game, obtains one of the strict pure NE. Obviously, were the players to play the game repeatedly, probably they will coordinate in some strict NE using some precedent. For example, if, by chance, they play one of them, then they will keep on playing it.

In conclusion,the results on one-shot 2X2 games with two strict NE and one mixed NE can be summarized as follows :if the players begin their analysis with preliminary theories defined by the common imprecise prior and their reasoning processes update beliefs smoothly in the direction of

best-reply, they will achieve an equilibrium in beliefs that is the risk dominant NE. In case that there is no risk dominant strict NE, they will achieve the mixed NE as an equilibrium in beliefs.

IV-5. A GENERAL CLASS OF INTROSPECTIVE DYNAMICS.

As can be easily checked, all our results of convergence are satisfied by any linear reasoning dynamics as:

$$q^{t+1} = (1-\alpha_t)q^t + \alpha_t.\delta(r(q^t)) \qquad \alpha_t \in (0,1)$$

where the subindex in α_t, reflects that this weight can depend on t.

The only properties that these reasoning processes should satisfy are:

a) $\lim_{T\to\infty} \prod_{t=1}^{T} (1-\alpha_t) = 0$

b) $\alpha_t \to 0$ as $t \to \infty$

Property a) implies that if there is T such that $r_i(q^t) = \begin{cases} 0 \\ 1 \end{cases}$ for all $t \geq T$, then $\lim_{t\to\infty} q_i^t = \begin{cases} 0 \\ 1 \end{cases}$.

In words, the reasoning function has to place enough weight in what , repeatedly, seems the best-reply.

Property b) guarantees that, eventually, the process becomes "smooth" enough. In particular, it implies a decreasing upper bound in the change of players' beliefs during the process.

V. OTHER 2X2 GAMES.

In this section we make some comments on the other two subclasses of generic 2X2 games and, specially, on non-generic 2X2 games. As our purpose is to illustrate some interesting features of our theory in these cases, we will take as "primitives" Criteria 1 and 2.

The reader can confirm that our theory isolates a unique solution for all these games. Obviously, in generic 2X2 games with a unique NE (mixed or in pure strategies), the solution coincides with this equilibrium.

Notice that generic 2X2 games with a unique NE in pure strategies are solvable by rationalizability (or, equivalently, by succesive elimination of strongly dominated strategies). Therefore, common knowledge of rationality is enough to obtain the NE as the solution of the game[5].

As we have explained our theory enters at work within the set of rationalizable strategies. It is in this situation where rational players (who are going to play a contest) are facing a decision problem under uncertainty with initial "complete ignorance". But, in any case, it is easy to check that our Criteria , applied to the whole game replicates the introspective process of rationalizability.

---

(5) In fact, it is enough that both players are rational and both know that they are rational.

A nice feature of our theory is that it always isolates a Perfect equilibrium (PE) as the solution of all non-generic 2X2 games in a quite natural way.

It is well-known that in two-person games all perfect equilibrium is an undominated equilibrium and viceversa. Where, an equilibrium strategy combination $s=(s_1,s_2)$ is said to be undominated if each player equilibrium strategy itself is undominated (in the weak sense). Typically, in most of the non-generic 2X2 games, one or both players have a weakly dominated strategy. This gives rise to the coexistence of PE and imperfect NE, in which the players assign positive probability to their weakly dominated strategies.

A weakly dominated strategy in 2X2 games has stability set with Lebesgue measure zero. This fact has two consequences. First, weak dominance implies incentive dominance. And second, the weakly dominated strategy of, say, player i is always "more dominated in incentives" than the incentive dominated strategy of player j (in case he has one).

Therefore, when applying Criteria 1 and 2, these two facts shape the CIP in a particular way, guaranteeing always convergence to a PE. (given our class of dynamics).

The remarkable fact is that we need not "exogenous" assumptions to obtain this result, such as "perturbing" the game("mistakes"). Neither do we need, assumptions, without justification, about players beginning their introspection with completely mixed initial priors, i.e. $q^0>>0$, $q^0=(q_1^0,q_2^0)$,

where $q_1^0 > 0$ and $q_2^0 > 0$. In fact, in the CIP there are always non-completely mixed initial point priors.

Let us see, first of all, an example, before we present this result as a Proposition.

		2	
		s_2	s'_2
1	s_1	1,1	1,1
	s'_1	2,0	-1,-2

(s'_1, s_2), i.e. (0,1) is the unique PE , but also there is a continuum of imperfect NE : $\{ q_1 = 1,\ 0 \leq q_2 < 2/3 \}$. The CIP, as can be easily computed is characterized by : $\{ q_1^0 \geq 1/2,\ q_2^0 \geq 1/2,\ q_2^0 \geq q_1^0 \}$

Graphically, is the triangle ABC of Figure V-1.

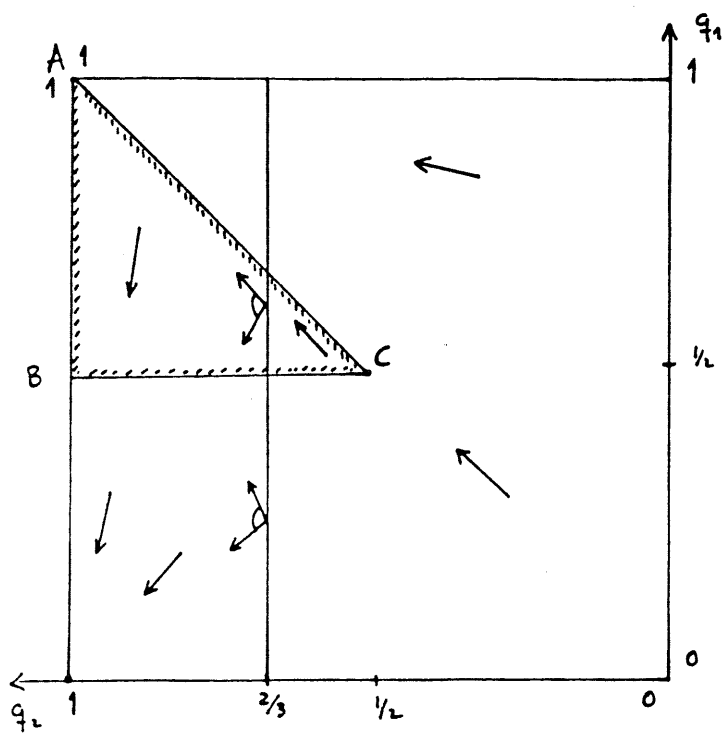

FIGURE V-1

Arguments as the ones used in previous sections yield convergence to the PE, (0,1).

As can be easily seen in the example, there are some basic relations that are driving the result. On one hand, note that none imperfect NE belongs to the CIP. On the other hand, all non-completely mixed priors in the CIP are characterized by $q_2 = 1$, i.e. both players assign probability zero to 2's weakly dominated strategy.

These basic features are general as we prove in the following Proposition.

PROPOSITION V-1.

Let G be a 2X2 non-generic game. If players begin their analysis using Criteria 1 and 2 and employ a reasoning dynamics within the class defined in Section IV-5, then the solution of G is a perfect equilibrium.

Proof : In a 2X2 game, if any player i has a weakly dominant strategy, the PE are the equilibria in which i plays this strategy with probability one. Therefore, it is enough to prove that no player assigns positive probability to his weakly dominated strategy in the solution of the reasoning process.

Assume, without loss of generality, that player i's pure strategy s_i weakly dominates s_i'. Hence, in the PE, $q_i = 1$. Furthermore, all imperfect NE will satisfy that $q_i < 1$.

Assume, in first place, that player j has no weakly dominant strategy. The CIP is characterized in the following two ways, depending on the incentive-dominance relations between j's pure strategies.

$$\text{either } \{ q_i \geq 1/2,\ q_j \geq 1/2,\ q_i \geq q_j \} \qquad (1)$$

$$\text{or } \{ q_i \geq 1/2,\ q_j \leq 1/2,\ q_i+q_j \geq 1 \} \qquad (2)$$

(There is a third possibility, where there is no incentive dominance between j's strategies ; $\{ q_i \geq 1/2,\ q_j = 1/2 \}$. As this case is quite trivial we leave its proof to the reader.)

First, we prove that the imperfect NE do not belong to the CIP.

The set of imperfect NE will satisfy : $\{ q_i < 1,\ q_j = 0 \}$, or, alternatively, $\{ q_i < 1,\ q_j = 1 \}$.

In the first case, if the CIP is as in (1) there is a contradiction between $q_j \geq 1/2$ and $q_j = 0$. If the CIP is as in (2), the contradiction is now with $q_i+q_j \geq 1$.

In the second case, if the CIP is as in (1) the contradiction appears with $q_i \geq q_j$ and if the CIP is as in (2) with $q_j \leq 1/2$.

Therefore, in all the possible cases the imperfect NE do not belong to the CIP. This implies that players cannot get caught on in an imperfect NE, at the beginning of the process.

Now, we prove that from any initial point prior in the CIP there is convergence to a PE.

i) From all completely mixed priors in the CIP there is convergence to $q_i = 1$.

Suppose that $q^{t-1} >> 0$. With our general class of dynamics, q^t is a strictly convex combination of q^{t-1} and $r(q^{t-1})$ for any t. Therefore, since $r_i(q_j^{t-1}) \geq 0$, if $q_i^{t-1} > 0$ then $q_i^t > 0$. Hence, given that we are assuming that $q^0 >> 0$. this implies that $q^t >> 0$ for any t.

It is well-known that a weakly dominated strategy is never a best-reply to a completely mixed belief. Therefore, as s_i weakly dominates s_i' for player i, $s_i' \notin R_i(q_j^t)$, for any t, because $q_j^t > 0$. So, $r_i(q_j^t) = 1$ for all t.

Therefore, because of Property a) of our class of linear dynamics, $q_i^t \to 1$ as $t \to \infty$.

ii) The set of non-completely mixed priors of the CIP is characterized as follows : $[q_i^0 = 1, q_j^0 \geq 1/2]$, or $[q_i^0 = 1, q_j^0 \leq 1/2]$.

In the first case, player 1's best reply is always $q_i = 1$ because $q_j > 0$. Player 2's best reply is his pure strategy in the PE. In any case, it is clear that there is convergence to the PE (within our class of dynamics).

A similar argument applies to the second case, when $q_j^0 > 0$. If $q_j^0 = 0$, player i is indifferent between his pure strategies. But as player j has a pure best-reply that is again his strategy in the PE, the reasoning process will move towards it.

(We leave to the reader to check the easy case where player j has also a weakly dominated strategy.) ■

Remark : Notice that, again as in Section IV, Criterium 2 is necessary for our results only if the incentive-dominance relations of the players point in opposite directions. But, if they point in the same direction, all we need is Criterium 1.

VI. CONCLUSIONS.

In this paper we give an explanation of how players' beliefs are formed in a game situation and why they can be accurate. Players form imprecise priors using the only information they have when playing a contest : the best-reply structure. Starting from a situation of complete predictive uncertainty, they can reduce it through introspection from this set of priors.

We show how, in 2X2 games, this model yields a solution, i.e. players achieve a state of predictive certainty. This solution coincides with the risk-dominant Nash equilibrium in 2X2 games with two strict Nash equilibria. A merit of our model, seen as a boundedly rational model of preplay introspection, is that it obtains the solution with very weak restrictions on rational prior formation and for a large class of reasoning dynamics.

Our results are stronger than similar findings obtained in the evolutive literature. We offer a justification of the risk dominant NE not only for common interest games but also for games with a conflict of interest. Moreover, our theory highlights the different difficulties for the players in order to reach their conclusions in these two subclasses of games.

A possible future extension of our analysis could be to consider decision theories that explicitly take account of the uncertainty generated by set-valued priors. In other words, decision theory under "knightian" uncertainty, as in Dow and Werlang (1992) and Gilboa and Schmeidler (1989, 1991) among others. With this approach the updating of players' beliefs would be less "Bayesian" than in the introspective process envisaged in this paper.

EN BLANCO

REFERENCES

AUMANN, R (1987): "Correlated Equilibrium as an Expression of Bayesian Rationality", *Econometrica*, 55, pp. 1-18.

BERNHEIM, B.D. 81984): "Rationalizable Strategic Behavior", *Econometrica*, 52, pp. 1007-1029.

BINMORE, K. (1991): "DeBayesing Game Theory", Working Paper.

DOW, J. & WERLANG, S. (1992): "Uncertainty Aversion, Risk Aversion and the Optimal Choice of Portfolio", *Econometrica,* 60, pp. 197-204.

EICHBERGER, J., HALLER, H. & MILNE, F (1990): "Naive Bayesian Learning in 2×2 Matrix Games", *CentER* Discussion Paper, 9071.

GILBOA, I. & SCHMEIDLER, D. (1989): "Maxmin Expected utility with Non-Unique Priors", *Journal of Mathematical Economics*, 18, pp. 141-153.

GILBOA, I. & SCHMEIDLER, D. (1991): "Updating Ambigous Beliefs", Discussion Paper, 924, Northwestern University.

HARSANYI, J.C. (1975): "The Tracing Procedure", *International Journal of Game Theory,* 4, pp. 61-94.

HARSANYI, J.C. & SELTEN, R. (1988): **A General Theory of Equilibrium Selection in Games**, Cambridge, Mass., MIT Press.

HENDON, E., JACOBSEN, H. NIELSEN, M. & SLOTH, B. (1990): "A Learning Process for Games", Discussion Paper 90-20, Institute of Economics, University of Copenhagen.

KANDORI, M., MAILATH, G. & ROB, R. (1991): "Learning, Mutation and Long Run Equilibria in Games". Mimeo, Princeton University.

KOHLBERG, E. & MERTENS, J.F. (1986): "On the Strategic Stability of Equilibria", *Econometrica*, 54, pp. 1003-1037.

LUCE, R.D. & RAIFFA, H. (1957): **Games and Decisions**, New Yourk: Wiley.

VAN HUYCK, J., GILLETTE, A. & BATTALIO, R. (1988): "Credible Assignments in Non-Cooperative Games", Texas A&M University Working Paper.

PUBLISHED ISSUES

FIRST PERIOD

1 "A Metatheorem on the Uniqueness of a Solution"
T. Fujimoto, C. Herrero. 1984.

2 "Comparing Solution of Equation Systems Involving Semipositive Operators"
T. Fujimoto, C. Herrero, A. Villar. February 1985.

3 "Static and Dynamic Implementation of Lindahl Equilibrium"
F. Vega-Redondo. December 1984.

4 "Efficiency and Non-linear Pricing in Nonconvex Environments with Externalities"
F. Vega-Redondo. December 1984.

5 "A Locally Stable Auctioneer Mechanism with Implications for the Stability of General Equilibrium Concepts"
F. Vega-Redondo. February 1985.

6 "Quantity Constraints as a Potential Source of Market Inestability: A General Model of Market Dynamics"
F. Vega-Redondo. March 1985.

7 "Increasing Returns to Scale and External Economies in Input-Output Analysis"
T. Fujimoto, A. Villar. 1985.

8 "Irregular Leontief-Straffa Systems and Price-Vector Behaviour"
I. Jimenez-Raneda / J.A. Silva. 1985.

9 "Equivalence Between Solvability and Strictly Semimonotonicity for Some Systems Involving Z-Functions"
C. Herrero, J.A. Silva. 1985.

10 "Equilibrium in a Non-Linear Leontief Model"
C. Herrero, A. Villar. 1985.

11 "Models of Unemployment, Persistent, Fair and Efficient Schemes for its Rationing"
F. Vega-Redondo. 1986.

12 "Non-Linear Models without the Monotonicity of Input Functions"
T. Fujimoto, A. Villar. 1986.

13 "The Perron-Frobenius Theorem for Set Valued Mappings"
T. Fujimoto, C. Herrero. 1986.

14 "The Consumption of Food in Time: Hall's Life Cycle Permanent Income Assumptions and Other Models"
F. Antoñazas. 1986.

15 "General Leontief Models in Abstract Spaces"
T. Fujimoto, C. Herrero, A. Villar. 1986.

16 "Equivalent Conditions on Solvability for Non-Linear Leontief Systems"
J.A. Silva. 1986.

17 "A Weak Generalization of the Frobenius Theorem"
J.A. Silva. 1986

18 "On the Fair Distribution of a Cake in Presence of Externalities"
A. Villar. 1987.

19 "Reasonable Conjetures and the Kinked Demand Curve"
L.C. Corchón. 1987.

20 "A Proof of the Frobenius Theorem by Using Game Theory"
B. Subiza. 1987.

21 "On Distributing a Bundle of Goods Fairly"
A. Villar. 1987.

22 "On the Solvability of Complementarity Problems Involving Vo-Mappings and its Applications to Some Economic Models"
C. Herrero, A. Villar. 1987.

23 "Semipositive Inverse Matrices"
J.E. Peris. 1987.

24 "Complementary Problems and Economic Analysis: Three Applications"
C. Herrero, A. Villar. 1987.

25 "On the Solvability of Joint-Production Leontief Models"
J.E. Peris, A. Villar. 1987.

26 "A Characterization of Weak-Monotone Matrices"
J.E. Peris, B. Subiza. 1988.

27 "Intertemporal Rules with Variable Speed of Adjustment: An Application to U.K. Manufacturing Employment"
M. Burgess, J. Dolado. 1988.

28 "Orthogonality Test with De-Trended Data's Interpreting Monte Carlo Results using Nager Expansions"
A. Banerjee, J. Dolado, J.W. Galbraigth. 1988.

29 "On Lindhal Equilibria and Incentive Compatibility"
L.C. Corchón. 1988.

30 "Exploiting some Properties of Continuous Mappings: Lindahl Equilibria and Welfare Egalitaria Allocations in Presence of Externalities"
C. Herrero, A. Villar. 1988.

31 "Smoothness of Policy Function in Growth Models with Recursive Preferences"
A.M. Gallego. 1990.

32 "On Natural Selection in Oligopolistic Markets"
L.C. Corchón. 1990.

33 "Consequences of the Manipulation of Lindahl Correspondence: An Example"
C. Bevía, J.V. LLinares, V. Romero, T. Rubio. 1990.

34 "Egalitarian Allocations in the Presence of Consumption Externalities"
C. Herrero, A. Villar. 1990.

SECOND PERIOD

WP-AD 90-01 "Vector Mappings with Diagonal Images"
C. Herrero, A.Villar. December 1990.

WP-AD 90-02 "Langrangean Conditions for General Optimization Problems with Applications to Consumer Problems"
J.M. Gutierrez, C. Herrero. December 1990.

WP-AD 90-03 "Doubly Implementing the Ratio Correspondence with a 'Natural' Mechanism"
L.C. Corchón, S. Wilkie. December 1990.

WP-AD 90-04 "Monopoly Experimentation"
L. Samuelson, L.S. Mirman, A. Urbano. December 1990.

WP-AD 90-05 "Monopolistic Competition : Equilibrium and Optimality"
L.C. Corchón. December 1990.

WP-AD 91-01 "A Characterization of Acyclic Preferences on Countable Sets"
C. Herrero, B. Subiza. May 1991.

WP-AD 91-02 "First-Best, Second-Best and Principal-Agent Problems"
J. Lopez-Cuñat, J.A. Silva. May 1991.

WP-AD 91-03 "Market Equilibrium with Nonconvex Technologies"
A. Villar. May 1991.

WP-AD 91-04 "A Note on Tax Evasion"
L.C. Corchón. June 1991.

WP-AD 91-05 "Oligopolistic Competition Among Groups"
L.C. Corchón. June 1991.

WP-AD 91-06 "Mixed Pricing in Oligopoly with Consumer Switching Costs"
A.J. Padilla. June 1991.

WP-AD 91-07 "Duopoly Experimentation: Cournot and Bertrand Competition"
M.D. Alepuz, A. Urbano. December 1991.

WP-AD 91-08 "Competition and Culture in the Evolution of Economic Behavior: A Simple Example"
F. Vega-Redondo. December 1991.

WP-AD 91-09 "Fixed Price and Quality Signals"
L.C. Corchón. December 1991.

WP-AD 91-10 "Technological Change and Market Structure: An Evolutionary Approach"
F. Vega-Redondo. December 1991.

WP-AD 91-11 "A 'Classical' General Equilibrium Model"
A. Villar. December 1991.

WP-AD 91-12 "Robust Implementation under Alternative Information Structures"
L.C. Corchón, I. Ortuño. December 1991.

WP-AD 92-01 "Inspections in Models of Adverse Selection"
I. Ortuño. May 1992.

WP-AD 92-02 "A Note on the Equal-Loss Principle for Bargaining Problems"
C. Herrero, M.C. Marco. May 1992.

WP-AD 92-03 "Numerical Representation of Partial Orderings"
C. Herrero, B. Subiza. July 1992.

WP-AD 92-04 "Differentiability of the Value Function in Stochastic Models"
A.M. Gallego. July 1992.

WP-AD 92-05 "Individually Rational Equal Loss Principle for Bargaining Problems"
C. Herrero, M.C. Marco. November 1992.

WP-AD 92-06 "On the Non-Cooperative Foundations of Cooperative Bargaining"
L.C. Corchón, K. Ritzberger. November 1992.

WP-AD 92-07 "Maximal Elements of Non Necessarily Acyclic Binary Relations"
J.E. Peris, B. Subiza. December 1992.

WP-AD 92-08 "Non-Bayesian Learning Under Imprecise Perceptions"
F. Vega-Redondo. December 1992.

WP-AD 92-09 "Distribution of Income and Aggregation of Demand"
F. Marhuenda. December 1992.

WP-AD 92-10 "Multilevel Evolution in Games"
J. Canals, F. Vega-Redondo. December 1992.

WP-AD 93-01 "Introspection and Equilibrium Selection in 2x2 Matrix Games"
G. Olcina, A. Urbano. May 1993.

WP-AD 93-02 "Credible Implementation"
B. Chakravorti, L. Corchón, S. Wilkie. May 1993.

WP-AD 93-03 "A Characterization of the Extended Claim-Egalitarian Solution"
M.C. Marco. May 1993.