MASSACHUSETTS
LIBRARIES
INSTITUTE of TECHNOLOGY

Digitized by the Internet Archive
in 2011 with funding from
Boston Library Consortium Member Libraries

http://www.archive.org/details/rationalbehavior00deke

HB31
.M415
no471

DEWEY
MAR 16 1988

working paper
department
of economics

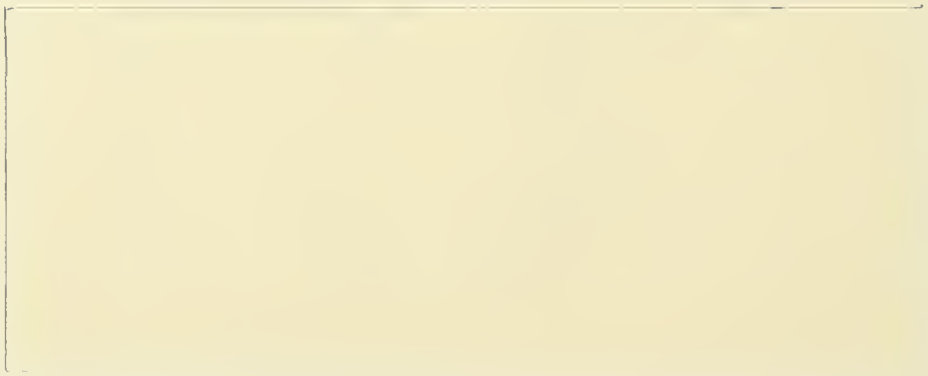
RATIONAL BEHAVIOR WITH PAYOFF UNCERTAINTY

Eddie Dekel
and
Drew Fudenberg

Number 471 October 1987

massachusetts
institute of
technology

50 memorial drive
cambridge, mass. 02139



RATIONAL BEHAVIOR WITH PAYOFF UNCERTAINTY

Eddie Dekel
and
Drew Fudenberg

Number 471 October 1987

M.I.T. LIBRARIES
MAR 16 1966
RECEIVED

Rational Behavior with Payoff Uncertainty

Eddie Dekel and Drew Fudenberg*

October 1987

ABSTRACT

The iterated deletion of weakly dominated strategies has been advanced as a necessary requirement for 'rational' play. However, this requirement relies on the assumption that the players have no doubts about their opponents payoffs. We show that once such doubts are introduced, all that can be justified by an appeal to rationality is one round of deletion of weakly dominated strategies, followed by iterated deletion of those which are strongly dominated. This extends the Fudenberg, Kreps and Levine (1987) study of the robustness of Nash equilibrium refinements to the robustness of solution concepts based only on common knowledge of rationality. Our results also clarify the relationship between various notions of what it means for payoff uncertainty to be 'small'.

* University of California, Berkeley and Massachusetts Institute of Technology, respectively. This work was begun while the second author was at the University of California, Berkeley. Financial support from the Miller Institute and the NSF is gratefully acknowledged.

2009749

Introduction

Nash equilibrium and its refinements describe situations with little or no "strategic uncertainty," in the sense that each player *knows* and is *correct* about the beliefs of the other players' regarding how the game will be played. While this will sometimes be the case, it is also interesting to understand what restrictions on predicted play can be obtained using only the assumption that it is common knowledge that the players are rational. Bernheim (1984) and Pearce (1984) have argued that these restrictions are captured by the concept of rationalizability. A more general notion is that of iterated deletion of strongly dominated strategies, which is equivalent to correlated rationalizability.[1] While (correlated) rationalizability may be appropriate for generic normal form games it has been argued that it does not capture all the implications of 'rationality' in non-trivial extensive forms (Bernheim (1984, Section 6(b)), Pearce (1984, Section 4)) For example in games of perfect information the only solution consistent with common knowledge of rationality might seem to be that given by backwards induction.

Recently, Fudenberg, Kreps and Levine (1987, henceforth referred to as FKL) have argued that the Nash equilibrium refinements are not 'robust' in the following sense. Extensive form refinements succeed in restricting the set of outcomes by rejecting some out of equilibrium play as unreasonable. Now the way a player should respond to a deviation by his/her opponents depends on how he expects the opponents to play subsequently. If the observed play to date is not consistent with the players initial understanding of the game, one plausible inference is that the reason for the deviation is that the deviator's payoffs are different than had originally been supposed. FKL model these inferences by supposing that players entertain small ex-ante doubts about their opponents' payoffs. They then characterize the sets of equilibria which can "justified" (made to satisfy strong equilibrium refinements) by allowing for different classes of such doubts.

---

[1] Correlated rationalizability, in contrast to rationalizability, does not impose the restriction that each player believes the other players' strategy choices are independent. The relationship between these rationalizability concepts, formal definitions of common knowledge of rationality, and equilibrium solution concepts is discussed in Aumann (1987), Brandenburger and Dekel (1987), and Tan and Werlang (1984).

The question of what players should infer from behavior they did not expect to occur is not restricted to equilibrium analysis: Rosenthal (1981), Reny (1985), Basu (1985) and Binmore (1987) discuss this issue in the context of solution concepts based on common knowledge of rationality alone. In this paper we adopt the FKL explanation that the reason for the unexpected play is that the payoffs are different than had been supposed. Thus we are led to characterize the implications of small uncertainties about the payoffs for the predictions that can be based on the assumption of 'rational' play. We maintain that the assumption of payoff uncertainty is, if anything, more apt here than in the equilibrium context. This is because correlated rationalizability and its refinements assume that the payoffs are common knowledge, but allow the players to have imprecise and inconsistent beliefs (inconsistent in the sense that they may disagree) about each others play. Yet in many situations with substantial strategic uncertainty, the common knowledge of payoffs assumption is suspect as well.

There are two modeling issues which need to be considered in order to achieve our objectives. First, a sharp notion for the implications of rational behavior must be given for the games with small doubts. We chose the notion of iterated deletion of weakly dominated strategies[2] since it clearly incorporates the intuitive objectives of rationality postulates.[3] The second modeling issue is related to the assumption of consistency. In the rationalizability approach to modeling strategic uncertainty players are allowed to have inconsistent beliefs about each others' strategies. Hence it seems natural to allow them to have inconsistent doubts about each others' payoffs, so we consider both the case of inconsistent and consistent beliefs.

These modeling issues emphasize the fact that the key question in evaluating the robustness of various solution concepts is which sequence of games are to be considered good approximations of a given game. Section 2 introduces our model and explains the notions of convergence we consider. Briefly, we say that a sequence of games converges weakly to a limit if each game in the sequence has the same "physical extensive form," so that the only difference between the games is

[2] This is similar to the use of strict equilibrium by FKL.

[3] The relationship between backwards and forwards induction (two primary notions of rationality) and weak dominance is discussed in Kohlberg and Mertens (1986).

in the beliefs about the payoffs, and moreover almost all types have almost the same payoffs as in the limit game. The sequence converges strongly if almost all types have *exactly* the same payoffs.

Section 3 proves our main result: The closure of iterated weak dominance with respect to the strong convergence described above is the set we call 1*WIZ*. This set is computed first by deleting the weakly dominated strategies, and then continuing with iterated deletion of strongly dominated strategies.[4] The intuition for this result is the following: Each player knows his/her own payoffs, and so by our rationality postulate will not choose a weakly dominated strategy. In order to do a second round of deletion players must know that all the others will not choose certain strategies. A small amount of payoff uncertainty cannot alter strong dominance relationships, but can break weak ones, so that after the first round we can only proceed with the iterated deletion of strongly dominated strategies. This result suggests reconsidering the intuition that since anything may occur iterated deletion of weakly dominated strategies is appropriate. The point is that if the reason that anything might occur is because of uncertainty about the payoffs, then iterated weak dominance goes to far.

Section 4 shows that weak convergence yields the set $\overline{1WIZ}$ which is the closure of 1*WIZ* with respect to extensive form payoff perturbations. To facilitate comparisons with FKL, Section 4 also considers the closure of a slightly more restrictive version of iterated weak dominance, namely the iterated deletion of strategies that are never strict best replies. Section 5 discusses the alternative interpretation of the robustness program in terms of how players interpret strategies which were unexpected, and how the two interpretations relate to our two definitions of convergence. Furthermore, using the notions of lexicographic beliefs derived in Blume(1986) and Brandenburger and Dekel (1986) it is argued that the distinction between the two notions of convergence is analogous to the difference between perfect and sequential equilibrium. Section 6 gives some examples to help explain the 1*WIZ* solution concept.

---

[4] In two person games this coincides with Bernheim's (1984) extension of trembling hand perfection to the context of rationalizability. For *n* person games this differs from Bernheim's notion by allowing for correlation -- cf. footnote 1.

2. Perturbations, Elaborations, and Convergence

Since this paper examines some implications of "small" amounts of payoff uncertainty, a crucial issue is to consider what forms of uncertainty are small. This is formalized by using different definitions for the convergence of sequences of games. A basic premise throughout the paper is that the physical extensive form (who moves when and the players' information regarding their opponents' actions) is common knowledge, and the only doubts the payers entertain (other than those explicitly specified in the given extensive form) are about each others' payoffs. More precisely, we begin with a finite I player game of perfect recall, E. This game E prescribes a game tree Y with representative nodes y, terminal nodes z, information sets H, and a utility function $u^i \in U \equiv \{f \mid f:Z \to R\}$ for each player i.

Following Harsanyi (1967-68), we model the idea that the players have doubts about the payoffs by considering "elaborations" $\tilde{E}$ of E, in which nature randomly chooses a utility function u^i for each player, and then an extensive form with the same structure as E is played. Each player's beliefs about the true payoffs, about his/her opponents' information, etc. is summarized by the players' type $t^i \in T^i$. We assume that each player i is informed of his/her utility function, and receives no information regarding the other players' utility function. (This is called "personal types" in FKL.) Therefore we can identify T^i with U. The game tree $\tilde{Y}$ of $\tilde{E}$ has one copy of Y for each possible choice by Nature, which is denoted by $t \in T \equiv \prod_i T^i$. If player i has a move at node y of Y then (s)he has a move at (y, t) for all $t \in T$. Similarly i's information at node y is just $H^i(y) \times \{t^i\}$.

The beliefs of each player i are derived from a prior p^i on the set T, which determines conditional beliefs $\nu^i(\cdot \mid t^i)$ on the set $T^{-i} \equiv \prod_{j \neq i} T^j$ of the other players types and marginal beliefs μ^i on T^i. For technical reasons the measures p^i and ν^i are assumed to have finite support. The set of pure strategies of player i in game E is denoted S^i. Player i's mixed strategies are denoted by $\sigma^i \in \Sigma^i \equiv \Delta(S^i)$, and beliefs over S^{-i} are denoted by $\sigma^{-i} \in \Delta(S^{-i})$, where $\Delta(X)$ is the set of probabil-

ity measures over X.

In general in this paper we will be considering games E and sequences of elaborations of E, denoted $\tilde{E}_n$. To distinguish between the strategy sets, utility functions, etc. in the elaborations $\tilde{E}_n$ and the game E, we add a $\tilde{}$ and a subscript of n to the appropriate symbol, eg. $\tilde{S}_n^i$ denotes the pure strategies of i in $\tilde{E}_n$. To further emphasize the distinction the game may be added in explicitly as an argument, eg. $IW^i(\tilde{E}_n)$ denotes the set of strategies of i which survive iterated deletion of weakly dominated strategies in the game $\tilde{E}_n$. When discussing a particular elaboration $\tilde{E}_n$ it will occasionally be necessary to refer to the utility functions or strategy choice of a player in a particular version of the game, that is when each player is of a particular type. This is done by including the type explicitly as an argument, eg. $\tilde{u}_n^i(t^i, t^{-i})$ denotes the utility function for player i when i is of type t^i. Since this utility does not depend on the types of the other players we will drop t^{-i} from the notation. Finally, the support of the limit of the beliefs μ_n^i is denoted by $m^i \subset T^i$, and the support of the limit of the beliefs $\nu_n^i(\cdot \mid t^i)$ will be denoted by $m^i(t^i) \subset T^{-i}$.

Now we can formalize the different forms of convergence which will be used. The weakest version, which we call weak convergence, has the interpretation that each player is "almost" sure that the payoffs are "almost" as in the original game. (The latter "almost" requires a definition of close utility functions, and the former is a probabilistic statement -- each player attaches probability almost one to the payoffs being close to those in the original game). Two stronger, and closely related notions of convergence are immediately apparent. One might require that the players are "almost" sure that their payoffs are *precisely* as in the original game; or that they are *absolutely* sure that the payoffs are "almost" as in the original game. These two notions will be called strong convergence and convergence in payoffs respectively. In this section only strong convergence will be examined, since the results are most intuitive and simplest to prove for this case. The other notions, which will be discussed the next section, are important both for clarifying the relationship of this paper with FKL, and to help understand certain issues related to the results in this paper.

In addition to the importance of distinguishing between various notions of convergence, it is important to consider the implications of assuming different restrictions on the information structure of the games of incomplete information. For example, in the context of consistent priors, FKL considered the implications of assuming that the players' beliefs over each others' types are independent (whereas in the "personal types" model p need not be a product measure). With independent types player i's observation of j's play can not effect i's beliefs over k's type. In this paper we examine with care the role of assuming consistent priors ($p^i = p$, for all i). Interestingly, several of our results hold with either consistent or inconsistent priors. This is because the effects of inconsistent priors over the payoffs can be duplicated by appropriately specified inconsistencies in the players' beliefs about each others' strategies. Thus, while the conceptual distinction between strategic uncertainty (beliefs about the strategies) and structural uncertainty (beliefs about the payoffs and other parameters of the game) is clear, assumptions about one of these kinds of uncertainty cannot be separated from assumptions about the other. Another discussion of similar issues can be found in Brandenburger and Dekel (1986) where it is shown that the existence of a 'mediator' is without loss of generality when beliefs over a state space are allowed to be inconsistent, whereas once consistency is required this is no longer the case. In that paper a limited form of consistency in the beliefs over the spaces of strategic uncertainty (namely the existence of a mediator) is achieved by shifting the inconsistency to the beliefs over the state space. In this paper, as the distinction between Propositions 3.1 and 3.2 below clarifies, the consistency in the players beliefs over the type spaces can be achieved (in Proposition 3.1) only by incorporating it into an inconsistency over the beliefs over the strategy spaces. So when the latter is ruled out (as in Proposition 3.2) the consistency of the beliefs over the type spaces can no longer be achieved.

In order to state our main result strong convergence must be defined. The definition is simpler for the case of consistent priors, so we start with that case. Recall that m^i is the support of the limit of μ_n^i (which in the consistent case is equal to the support of the marginal on T^i of the limit of the common prior p_n). The set of possible types of i according to the limit beliefs will often be termed "sane" types and denoted by $\bar{T}^i$. Clearly in the consistent case $\bar{T}^i \equiv m^i$.

DEFINITION 2.1: A sequence $\tilde{E}_n$ of consistent elaborations of E *converges strongly* to E ($\tilde{E}_n \to E$) if:

(i) (a) $|T^i(\tilde{E}_n)| < M$ for all n,

(b) $|\tilde{u}^i_n| < B$ for all i and n;

(ii) For all $\bar{t}^i \in \bar{T}^i$, $\tilde{u}^i_n(\bar{t}^i) = u^i$.

Thus $\tilde{E}_n \to E$ if: (i) the number of types and the absolute value of the payoffs are uniformly bounded in n, and (ii) the set of types with payoffs different than those in E has probability zero in the limit. Note that because of the assumption of consistency the conditional beliefs $\nu_n(\cdot|\bar{t}^i)$ of every "sane" type $\bar{t}^i$ in $\bar{T}^i$ are that the other players are very likely to be "sane," so that $m^i(\bar{t}^i) \subset \bar{T}^{-i}$. With this notion of convergence we are treating as identical a game E and an elaboration $\tilde{E}$ where all versions in $\tilde{E}$ have the same payoffs as in E. So the two games in Figure 1 are identical (with the obvious mapping of strategies of player 2). This way, each type plays a pure strategy, but a player can have a nondegenerate belief over the strategies of the other players (because beliefs over their types is nondegenerate) which is equivalent to them playing a mixed strategy.

DEFINITION 2.2: A sequence of strategies $\tilde{\sigma}^i_n$ will be said to converge to σ^i (written $\tilde{\sigma}^i_n \to \sigma^i$) if

$$\lim \sum_{t^i \in T^i} \mu^i_n(t^i)\tilde{\sigma}^i_n(t^i) = \sigma^i.$$

This notion of convergence requires that player i's play converge to σ^i at every information set (even those which are not reached by σ^i regardless of the other players' strategies).

3. Iterated Weak Dominance and 1*WIZ*

A strategy s^i is weakly dominated if there is another strategy $\hat{s}^i$ such that $u^i(\hat{s}^i, \sigma^{-i}) \geq u^i(s^i, \sigma^{-i})$ for all $\sigma^{-i} \in \Delta(S^{-i})$, and if the inequality is strict for some σ^{-i}. Any strategy s^i which is not weakly dominated, is said to be admissible, and is a best reply to some *full support* belief σ^{-i} (i.e the support of σ^{-i} is S^{-i}) over i's opponents strategies (Pearce (1984, Appendix B), Van Damme (1983), Gale and Sherman(1950)). Kohlberg and Mertens (1986) provide the following argument in support of iterated weak dominance (that is, iteratively deleting strategies which are weakly dominated) as a minimal solution concept. First note that if the players are uncertain of their environment, they should never play a weakly dominated strategy. (Kohlberg and Mertens (1986) argue in fact that admissibility itself is a basic postulate of decision theory -- and not only a consequence of uncertainty about the environment. Axiomatic characterizations of preferences under uncertainty which lead to this postulate, in a way which is analogous to the relationship between strong dominance and expected utility rationality, are provided in Blume (1986), Brandenburger and Dekel (1986) and Luce and Raiffa (1957).) In any case if the players know their opponents payoffs, they should not expect them to play a weakly dominated strategy, and thus each player should only play strategies which survive two rounds of deletion of weakly dominated strategies. If the payoffs are common knowledge this argument leads to iterated weak dominance, denoted *IW*. More generally *kWIZ* is used to denote the set of strategies remaining after k rounds of simultaneous deletion of weakly dominated strategies, followed by l rounds of deletion of strongly dominated strategies. When k or l is infinity it is convenient to use I (for iterated), for example 1*WIZ*. Each of these sets is a Cartesian product of strategies for each player, so $kWIZ^i$ denotes the projection of *kWIZ* on i's strategy space.

Proposition 3.1 below says that any strategy in 1*WIZ* is close to a strategy in *IW* for some sequence of nearby games, and any strategy in *IW* for nearby games is close to a strategy in 1*WIZ*. Thus if there is "small" payoff uncertainty in the sense described by strong convergence (as, we would argue, is typically the case) then ruling out any strategy in 1*WIZ* is questionable, even if we

agree to rule out all strategies not in IW when payoffs are common knowledge.

PROPOSITION 3.1: $s^i \in 1WIZ^i(E)$ if and only if there is a sequence of consistent elaborations $\tilde{E}_n \to E$, and strategies $\tilde{s}^i_n \in IW^i(\tilde{E}_n)$ such that $\tilde{s}^i_n \to s^i$.

PROOF: **Only If:** In this direction of the proof the sequence of elaborations $\tilde{E}_n$ is constructed. Let $T^i_n = \{\bar{t}^i, \hat{t}^i\}$, where $u^i(\bar{t}) = u^i$ and $u^i(\hat{t}^i) \equiv 0$. So i can either be a "sane" type (with payoffs as in E), or "crazy" and completely indifferent among all his/her strategy choices. The common prior p assigns probability $1 - n$ to all the players i being of type $\bar{t}^i$, and for each player i probability $1/nI$ (I is the number of players) to the event that only i is sane and all the other players are crazy. Thus, when i is sane, the conditional probability $\nu^i_n(\cdot|\bar{t}^i)$ that (s)he assigns to the event that all the players are sane is $I(n-1)/(I(n-1)+1)$ and the conditional probability that all the others are crazy is $1/(I(n-1))$. Player i's strategy in $\tilde{E}_n$ is written as an ordered pair $(\bar{\sigma}^i_n, \hat{\sigma}^i_n)$ where the first element is i's play when sane, and the second is his/her play when crazy. Since, when i is sane, his/her opponents are either all sane or all crazy, we can consider his/her beliefs over $\tilde{S}^{-i}_n$ (the opponents strategies) as elements of $\Delta(S^{-i}) \times \Delta(S^{-i})$. Such beliefs are denoted by ordered pairs $(\bar{\sigma}^{-i}_n, \hat{\sigma}^{-i}_n)$ We claim that: $IW(\tilde{E}_n) = \{(\bar{s}^i, s^i) \mid \bar{s}^i \in 1WIZ^i(E), s^i \in S^i\}$ Obviously this implies the only if part of Proposition 1. Proving the claim involves two steps.

Step 1: $(\bar{s}^i, s^i) \in W(\tilde{E}_n)$. Since $\bar{s}^i \in 1WIZ^i(E) \subseteq W^i(E)$, there exists a full support belief $\bar{\sigma}^{-i} \in \Delta(S^{-i})$ such that $\bar{s}^i$ is a best reply to $\bar{\sigma}^{-i}$. So $(\bar{s}^i, s^i)$ is a best reply to $(\bar{\sigma}^{-i}, \bar{\sigma}^{-i})$ which is equivalent to a full support belief over $\tilde{S}^{-i}_n$. For future reference let α be the smallest weight assigned to any pure strategy s^{-i} by $\bar{\sigma}^{-i}$.

Step 2: If $(\bar{s}^i, s^i) \in 1W1Z^i(E) \times S^i$ then $(\bar{s}^i, s^i) \in 2W(\tilde{E}_n)$. This can be seen as follows. We need to show that $(\bar{s}^i, s^i)$ is a best reply to a full support belief over $W^{-i}(\tilde{E}_n)$,which by step 1 is a superset of $1WIZ^{-i}(E) \times S^{-i}$. Since $\bar{s}^i \in 1W1Z^i(E)$ there is a $\sigma^{-i} \in 1W^{-i}(E)$ to which $\bar{s}^i$ is a

best reply. Specify that the sane types of the opponents play σ^{-i} with probability $1-\beta$ (where β is small and is specified below), and with complementary probability β the sane types play any full support distribution $\hat{\sigma}^{-i}$ over all the strategies in $1W^{-i}(E)$. The crazy types of the opponents play a strategy σ'^{-i} such that the weighted average (weighted by the probabilities of the crazy and sane opponents, and of the sane opponents playing $\hat{\sigma}^{-i}$) of σ'^{-i} and $\hat{\sigma}^{-i}$) is $\bar{\sigma}^{-i}$. (For convenience set $N \equiv I(n-1)$. Then $I(n-1)/(I(n-1)+1) = 1-1/N$, and $1/(I(n-1)) = 1/N$.) The induced strategy for i's opponents is $(1-1/N)(1-\beta)\sigma^{-i} + (1-1/N)\beta\hat{\sigma}^{-i} + (1/N)\sigma'^{-i}$ which we want to be equal to $(1-1/N)(1-\beta)\sigma^{-i} + [1-(1-1/N)(1-\beta)]\bar{\sigma}^{-i}$. This is achieved by setting $\sigma'^{-i} = \bar{\sigma}^{-i} + \beta(N-1)[\bar{\sigma}^{-i} - \hat{\sigma}^{-i}]$ which will be a probability measure as long as $\beta < \alpha/(N-1)$.

Step 2 can now be iterated to show that if $(\bar{s}^i, s^i) \in 1WIZ^i(E) \times S^i$ then $(\bar{s}^i, s^i) \in IW^i(\tilde{E}_n)$.

Remark: Note that in step 2 the fact that $\bar{s}^i \in 1W1Z^i(E)$ was used in finding $\sigma^{-i} \in 1W^{-i}(E)$ to which $\bar{s}^i$ is a best reply. This suggests that we could not have found an elaboration to "justify" s^i if that strategy could be deleted by strong dominance. Intuitively, "small" uncertainties about payoffs should not be able to undo the iteration of strict dominance, so that the IZ step in $1WIZ$ should be necessary for a characterization of the "closure" of IW. This is verified in the proof of the "if" direction below.

If: This direction also involves two steps.

Step 1: $\tilde{s}^i_n \in 1W^i(\tilde{E}_n)$ implies $\tilde{s}^i(\bar{t}^i) \in 1W^i(E)$ for all $\bar{t}^i \in \bar{T}^i$. This follows from the fact that $\tilde{s}^i_n$ is a best reply to some full support belief $\tilde{\sigma}^{-i}_n$ over $\tilde{S}^{-i}_n$. Hence $\tilde{s}^i_n(\bar{t}^i)$ is a best reply to $\hat{\sigma}^{-i}_n \equiv \sum_{t^{-i}} \nu^i_n(t^{-i} \mid \bar{t}^i)\tilde{\sigma}^{-i}_n(t^{-i})$ which is a full support belief over S^{-i}. Since player i's utility function when (s)he is of type $\bar{t}^i$ is the same as his/her utility function in E, clearly $\tilde{s}^i_n$ is not weakly dominated in E.

Step 2: $\tilde{s}^i_n \in 1W1Z(\tilde{E}_n)$ implies $\tilde{s}^i_n(\bar{t}^i) \in 2W(E)$. We know that $\tilde{s}^i_n(\bar{t}^i)$ is a best reply to some $\hat{\sigma}^{-i} \equiv \sum_{t^{-i}} \tilde{\sigma}^{-i}_n(t^{-i})\nu_n(t^{-i} \mid \bar{t}^i)$ for some $\tilde{\sigma}^{-i}_n$ which is supported by strategies in $W^{-i}(\tilde{E}_n)$ since $\tilde{s}^i_n \in$

$2W^i(\tilde{E}_n)$. As noted earlier, by condition (ii) of the definition of convergence in types $\nu_n(t^{-i}|\bar{t}^i)$ converges to a measure supported by $\bar{T}^{-i}$, i.e. player i is almost certain that the others have the same payoffs as in E. Further, by step 1, for those types t^{-i} in $\bar{T}^{-i}$ we know that $\tilde{\sigma}_n(t^{-i})$ is a belief over $1W^{-i}(E)$. Taking limits now in the definition of $\hat{\sigma}_n^{-i}$ (in the second sentence of this step) it has been shown that $\tilde{s}_n^i(\bar{t}^i)$ is a best reply to $\lim\hat{\sigma}_n^{-i}$ which is supported by $1W^{-i}(E)$, hence $\tilde{s}_n^i(\bar{t}^i)$ is not strongly dominated within $1W^i(E)$.

Step 2 can now be iterated to show that $\tilde{s}_n^i(\bar{t}^i)$ is an element of $1WIZ^i(E)$. □

The reason that after one round of deletion of weakly dominated strategies only strongly dominated strategies could be deleted follows from the difference between steps 1 and 2 in the *if* part of the proof. In step 1 s^i is a best reply to a strategy $\hat{\sigma}_n^{-i}$ which has full support. In step 2 a similar $\hat{\sigma}_n^{-i}$ was found, but it does not have full support within $1W^{-i}(E)$: It's support is larger because of the possibility of crazy types of $j \neq i$, and of course its limit may have smaller support than $1W^{-i}(E)$.

Since the solution concept used here involves iterated deletion procedures it inherently allows for inconsistencies in the strategic beliefs of the players. In the proof of the "only if" part of Proposition 3.1 a players' beliefs in steps 1, 2 and in the iteration of step 2 need not be the same. In particular, in the first step the crazy types were expected to play $\bar{\sigma}^{-i}$, in the second step σ'^{-i}, and in iterating the second step the beliefs over the opponents would be different each time. This is very similar to the hierarchies of beliefs in Bernheim's definition of rationalizability, where i may think that j is playing a certain strategy, but i thinks j thinks i thinks that j is playing something else. The *strategic* beliefs are not consistent. This suggests that allowing in addition for inconsistent beliefs over the *types* will not change the result, as Corollary 1 below confirms.

In order to formalize the inconsistent case the definition of convergence of elaborations must be extended accordingly. Recall that in Definition 2.1 the common prior p was used in defining the

set $\bar{T}^i \equiv m^i$ of possible types of player i in the limit. For a sequence of elaborations to converge we then required that for any player i all types t^i in $\bar{T}^i$ had the same payoffs as in the original game. When the player's priors differ this convergence requirement is too weak. We still want all the possible types of player i in the limit to have the same payoffs as in E. However we use an expanded definition of $\bar{T}^i$. Of course we want $m^i \subset \bar{T}^i$ (so i's prior must asymptotically assign probability 1 to having the same payoffs as in E). Furthermore, for any t^j in m^j, if t^j thinks a type t^i has positive probability in the limit, then t^i should also be in $\bar{T}^i$. Continuing iteratively, we need to ask whether (in the limit) player h thinks i thinks j thinks ... k thinks l may be of type t^l. If so then we will want to require that t^l has payoffs as in the original game E. Formally, for any player l, we define $\bar{T}^l$ as follows. If $t^h \in m^h$, $t^i \in m^h(t^h)$, $t^j \in m^i(t^i)$,..., $t^k \in m^l(t^l)$ for some permutation of players h, i, j,..., k, l, then $t^l \in \bar{T}^l$. A general sequence of elaborations then converges strongly when conditions (i) and (ii) of Definition 2.1 are satisfied with respect to the extended definition of $\bar{T}^i$. Since the two definitions of $\bar{T}^i$ coincide when $p^i = p$ for all$i$, the extended definition of convergence agrees with the previous one when beliefs are consistent.

COROLLARY 3.1: $s^i \in 1WIZ^i(E)$ if and only if there is a sequence of elaborations $\tilde{E}_n \to E$, and strategies $\tilde{s}^i_n \in IW^i(\tilde{E}_n)$ such that $\tilde{s}^i_n \to s^i$.

PROOF: The proof of Proposition 1 proves the corollary also, when $\bar{T}^i$ is redefined as discussed above. The "only if" direction is exactly the same. The iterative definition of $\bar{T}^i$ in the inconsistent case corresponds to the iteration applied in the proof of the "if" direction. □

4. Payoff Perturbations and Strict Best Replies.

This section discusses the implications of using weak convergence, instead of strong convergence, to characterize "small" doubts. The difference is that in weak convergence the types in $\bar{T}^i$ may have payoffs u_n^i which *converge* to the payoffs u^i in E, instead of $u_n^i = u^i$ for all n. As one would expect, the consequence of allowing more convergent sequences of elaborations is that more strategies in E survive IW in nearby games. In fact, the resulting set is the closure of $1WIZ$ with respect to extensive form payoff perturbations, which we denote $\overline{1WIZ}$. Moreover (again because more sequences of elaborations converge to a given game E) we can show that any strategy in $1WIZ$ is close to a strategy which satisfies a stronger requirement than IW in nearby games, namely the iterated deletion of strategies which are never strict best replies. A strategy which is weakly dominated is never a strict best reply, but the converse is in general false. In considering weak convergence and strict best replies we are also able to clarify the relationship between our results and those of FKL.

To understand the results of this section it is helpful to review briefly a result on rationalizability. In Brandenburger and Dekel (1987) it is shown that correlated rationalizability is the same as *a posteriori* equilibrium (Aumann 1974) which is roughly the same as a Nash equilibrium with a subjective correlating device (about which the players may have inconsistent beliefs) explicitly introduced. So, an alternative to IW as a refinement of IZ is to look at strict Nash equilibrium with subjective correlating devices.

DEFINITION 4.1: $\tilde{E}_n$ converges in payoffs to E ($\tilde{E}_n \xrightarrow{p} E$) if condition (i) of Definition 2.1 holds, and:

(ii) For all $t^i \in T^i$, $\tilde{u}_n^i(t^i) \to u^i$.

DEFINITION 4.2: Two strategies for player i are *equivalent* if they lead to the same probability distribution over endpoints for all strategies of the opponents. A Nash equilibrium $(s^1, \ldots, s^I)$ is

strict if each players' strategy s^i does strictly better against s^{-i} than any other strategy $\hat{s}^i$ which is not equivalent to s^i.

LEMMA 4.1: If s^i is not weakly dominated then there exists a consistent sequence $\tilde{E}_n \xrightarrow{P} E$ where s^i is a strict best reply (up to equivalent strategies) to some $\sigma^{-i} \in \Delta(S_n^{-i})$.

PROOF: If s^i is not weakly dominated then it is a best reply to some s^{-i} with full support. Let T be a singleton in each elaboration $\tilde{E}_n$ so that the utility functions (defined next) are common knowledge. Let $u_n^i(z) = u^i(z)+1/n$ on all endpoints z reached by s^i and σ^{-i}, and $u_n^i(z) = u^i(z)$ otherwise. □

Lemma 4.1 provides the intuition for Proposition 4.1 below. It shows that by allowing for small extensive form payoff perturbations, strategies which are not weakly dominated can be made strict best replies. Proposition 4.1 below is an analog to Proposition 3.1, where the notion of "not weakly dominated" is strengthened to "is a strict best reply" and convergence is weakened to allow for extensive form payoff perturbations.

DEFINITION 4.3: $\tilde{E}_n$ converges weakly to E ($\tilde{E}_n \overset{w}{\to} E$) if condition (i) of Definition 2.1 holds, and:

(ii) For all $t^i \in \bar{T}^i$, $\tilde{u}_n^i \to u^i$.

PROPOSITION 4.1: If $s^i \in 1WIZ^i(E)$ then there is a sequence of elaborations $\tilde{E}_n \overset{w}{\to} E$, and strategies $\tilde{s}_n^i \overset{w}{\to} s^i$, such that $\tilde{s}_n^i$ is a strategy in a strict Nash equilibrium of $\tilde{E}_n$.

Remark: Proposition 4.1 relies on inconsistent elaborations in an essential way to obtain as a Nash equilibrium strategies that may not be played in any *objective correlated* equilibrium of the original game E. Any *subjective correlated* equilibrium is a Nash equilibrium of the game where the appropriate subjective correlating device is explicitly incorporated into the strategy spaces. The

point is that nature's move at the beginning of the game, which determines the types of the players, serves also as a subjective correlating device. (The difference between subjective and objective correlating devices corresponds to the cases of consistent and inconsistent priors.)

PROOF: The elaborations $\tilde{E}^n$ are constructed as follows. In each elaboration each player's set of possible types T^i is partitioned into two sets, the "sane" types $\bar{T}^i$ and the "crazy" types $\hat{T}^i$. $\bar{T}^i$ is isomorphic to $1WIZ^i(E)$ and $\hat{T}^i$ is isomorphic to the set S^i of i's pure strategies in E. (Using these isomorphisms we will write $\bar{t}^i = \bar{s}^i$, and $\hat{t}^i = \hat{s}^i$.) The priors p_n^i will be chosen so that only types in $\bar{T}^i$ are possible in the limit, which explains the abuse of notation. If i's type is $\hat{s}_k^i = \hat{t}_k^i \in \hat{T}^i$, we say that i was "told" to play $\hat{s}_k^i$, and if i's type is $\bar{t}_k^i = \bar{s}_k^i \in 1WIZ^i$ we say that i was told to play $\bar{s}_k^i$. The payoffs and beliefs will be chosen so that in each elaboration playing as told will be a strict best reply for each possible type of player i, and so that the elaborations converge weakly to E.

For each crazy type $\hat{t}_k^i$, simply set the payoffs $\tilde{u}_n^i(\hat{t}_k^i)$ so that $\hat{s}_k^i$ is a strict best reply (up to equivalent strategies) to *any* belief σ^{-i} over the other players. (See FKL for an explicit construction.) Note that since these payoffs may be very different than those in E, the types in $\hat{T}$ must have probability zero in the limit.

To make $\bar{s}_k^i$ in $1WIZ^i$ a strict best reply for type $\bar{t}_k^i$ we proceed as follows. First fix a sequence $\varepsilon_n \downarrow 0$. Since $\bar{s}_k^i \in 1WIZ^i$ there exists a $\sigma_k^{-i} \in \Delta(\prod_{j \neq i} S^j)$ with full support, such that $\bar{s}_k^i$ is a best reply to σ_k^{-i}. Also there exists a $\hat{\sigma}_k^{-i} \in \Delta(\prod_{j \neq i} 1WIZ^j)$, such that $\bar{s}_k^i$ is a best reply to $\hat{\sigma}_k^{-i}$. Since σ_k^{-i} has full support we can increase the payoffs at all endpoints reached under σ_k^{-i} and $\bar{s}_k^i$ by ε_n and thus make $\bar{s}_k^i$ a strict best reply against σ_k^{-i}. Furthermore this change in payoffs will not change the fact that $\bar{s}_k^i$ is a best reply against $\hat{\sigma}_k^{-i}$. This is because no other pure strategy of i can

increase the probability of reaching the endpoints for which payoffs were increased.

Next we specify the beliefs in an elaboration. Let i's beliefs over the others' types, conditional on his/her type be as follows. For "sane" types $\bar{t}_k^i$ the beliefs $\nu_n(\cdot|\bar{t}_k^i)$: (i) assign probability ε_n to all the others being crazy, with the distribution of crazy types corresponding to σ_k^{-i}; and (ii) assign probability $1-\varepsilon_n$ to all the others being sane. For crazy types $\hat{t}_k^i$, the beliefs are arbitrary. For each i choose a sequence of marginals μ_n^i over T^i which has full support on $\bar{T}^i \cup \hat{T}^i$, and which converges with probability one to the sane type of player i which was in the hypothesis of the Proposition (say $\bar{s}_1^i = \bar{t}_1^i$). The priors p_n^i which are generated by the ν_n^i and μ_n^i are such that the sets of types which, in the limit, players think that others think that ... have positive probability are exactly the sane types $\bar{T}^i$. Thus $\tilde{E}_n$ converges weakly to E.

Finally we observe that by construction, for each n and each player i, each type playing as told is a strict best reply to the others playing as told, hence playing as told is a strict Nash equilibrium. □

Now we turn to the question of finding a converse to Proposition 4.1, i.e. we ask which strategies can be justified using elaborations that converge weakly to the original game. The problem is that the converse to Proposition 4.1 is not precisely correct. There are strategies in E which are the limit of strategies that survive iterated deletion of weakly dominated strategies in a sequence of elaborations that converge to E, but which are not elements of $1WIZ(E)$. This is because $1WIZ$ is a normal form solution concept, whereas in the sequence of elaborations converging to E, in addition to incomplete information on the payoffs, we allow for perturbations of the extensive form payoffs of E. Hence, roughly speaking, since the 'closure' of IW allows for extensive form payoff perturbations it can only be equal to a solution concept which is closed with respect to such perturbations. Since weak dominance is not closed in this sense, neither is $1WIZ$. Although this suggests that we could achieve a generic converse to Proposition 2.1, we believe it is more interesting to provide a complete characterization. In order to clarify the nature of the converse direction we

begin with a partial converse that is easier to prove. The converse is partial in that the 'closure' with respect to payoff perturbations is taken in a consistent manner (that is, the players agree about the perturbation) and only the closure of the first round of deletion is taken. Formally, we say that a strategy $s^i \in S^i$ is not weakly* dominated if there exists a sequence of extensive form games E_n identical to E except for the payoffs which are $u_n^i \to u^i$ such that s^i is not weakly dominated in E_n. If there is no such sequence, then s^i is weakly* dominated. We denote strategies which survive deletion of weakly* dominated strategies by $1W^*$ instead of $1W$. Proposition 4.2 is an analog to the "if" part of Proposition 3.1 when strong convergence is replaced by weak convergence.

PROPOSITION 4.2: If $\tilde{s}_n^i \in IW^i(\tilde{E}_n)$, $\tilde{E}_n \twoheadrightarrow E$, $\tilde{s}_n^i \twoheadrightarrow s^i \in S^i(E)$, then $s^i \in 1W^*IZ^i(E)$.

PROOF: Step 1: Consider the versions of $\tilde{E}_n$ where i is of any type t^i that satisfies $\tilde{u}_n^i(t^i) \to u^i$. Any $\tilde{s}_n^i$ which is not weakly dominated in $\tilde{E}_n$ is a best reply to $\hat{\sigma}_n^{-i} \equiv \sum_{t^{-i}} \tilde{\sigma}_n^{-i}(t^{-i})\nu(t^{-i} \mid t^i)$ for some $\tilde{\sigma}_n^{-i}$ with full support. Let E_n be an extensive form identical to E with payoffs $\tilde{u}_n^i(t^i)$. Hence any $s^i \in S^i$ such that $\tilde{s}_n^i \to s^i$ is not weak* dominated, i.e. $s^i \in 1W^*(E)$.

Step 2: Since $\tilde{s}_n^i \in 2W^i(\tilde{E}_n)$, $\tilde{s}_n^i(t^i)$ is a best reply to some $\hat{\sigma}_n^{-i} \equiv \sum_{t^{-i}} \tilde{\sigma}_n^{-i}(t^{-i})\nu_n(t^{-i} \mid t^i)$ for some $\tilde{\sigma}_n^{-i}$ which is supported by strategies in $1W^{-i}(\tilde{E}_n)$. The sequence of beliefs $\nu_n(t^{-i} \mid t^i)$ converges to a measure which is supported by types t^{-i} for which $\tilde{u}_n^{-i}(t^{-i}) \to u^{-i}$ (this follows from $\tilde{E}_n \twoheadrightarrow E$). Hence by step 1, for those types t^{-i} in the support of $\lim_n \nu_n(\cdot \mid t^i)$ we know that $\tilde{\sigma}_n^{-i}(t^{-i}) \in \Delta(1W^{*-i}(E))$. Taking limits in the first sentence of step 2 we then have that s^i is a best reply to some weighted average of $\lim_n \tilde{\sigma}_n^{-i}(t^{-i})$ where $\lim_n \tilde{\sigma}_n^{-i}(t^{-i})$ is supported by $1W^{*-i}(E)$ (and where the weights are given by $\nu_n(\cdot \mid t^i)$). Hence s^i is a best reply to the strategy $\lim_n \hat{\sigma}_n^{-i}$ which is supported by $1W^{*-i}(E)$, i.e. $s^i \in 1W^*1Z^i(E)$. Step 2 can be iterated to show that $s^i \in 1W^*IZ^i(E)$. □

Proposition 4.2 is not a converse to Proposition 4.1, so we have not yet characterized the 'closure' of the set of iteratively admissible strategies with respect to weak convergence of elaborations. In order to do so both Propositions 4.1 and 4.2 need to be strengthened. The sets $1WIZ^i$ and $1W^*IZ^i$ must be replaced by the same set, and this set will be the 'closure' of $1WIZ^i$ with respect to convergence in payoffs. This set is denoted by $\overline{1WIZ}$.

DEFINITION 4.4: $s^i \in \overline{1WIZ}^i(E)$ if $\tilde{s}^i_n \to s^i$ and $\tilde{s}^i_n \in 1WIZ\ \tilde{E}_n$ for some sequence of elaborations $\tilde{E}_n \xrightarrow{p} E$.

PROPOSITION 4.3: $\tilde{s}^i_n \in IW(\tilde{E}_n)$, $\tilde{E}_n \twoheadrightarrow E$, $\tilde{s}^i_n \twoheadrightarrow s^i \in E$, if and only if $s^i \in \overline{1WIZ}^i(E)$.

PROOF: *If:* This follows from a simple diagonal argument and Proposition 3.1. If $s^i \in \overline{1WIZ}^i(E)$ then there exists a sequence $\tilde{E}_n \xrightarrow{p}$ with $\tilde{s}^i_n \to s^i$ and $\tilde{s}^i_n \in 1WIZ^i(\tilde{E}_n)$. By Proposition 3.1 there exists $\tilde{\tilde{E}}_{k,n} \to \tilde{E}_n$ and $\tilde{\tilde{s}}_{k,n} \to \tilde{s}^i_n$ with $\tilde{\tilde{s}}_{k,n} \in IW^i(\tilde{\tilde{E}}^i_n)$. Clearly $\tilde{\tilde{s}}^i_{m,m} \to s^i$ and $\tilde{\tilde{E}}^i_{m,m} \twoheadrightarrow E$ as required.

Only if: We are given a sequence $\tilde{E}_n \twoheadrightarrow E$. Let R^i denote the strategies played by sane types, that is $R^i \succ\equiv \text{limsup} R^i_n$ where $R^i_n \equiv \{s^i \in S^i \mid$ for some $\tilde{s}^i_n \in IW^i(\tilde{E}_n)$and some $\bar{t}^i \in \bar{T}^i$, $\tilde{s}^i_n(\bar{t}^i) = s^i\}$. Construct the following elaborations $\bar{E}_n$ which will converge in payoffs to E. The set of possible types for each player i is isomorphic to M^i copies of R^i, where $M^i = |\prod_{j \neq i} R^j|$. The types in $\bar{E}^n$ will be denoted by $s^i_k(m)$ where $s^i_k \in R^i$ for $k = 1,\ldots, K$, and $m = 1,\ldots, M^i$. For a given i and k all types $s^i_k(m)$ have the same payoffs (independent of m), and these payoffs are determined as follows. Since $s^i_k \in R^i$, then taking a subsequence if necessary, there exists $\tilde{s}^i_n$ and $\bar{t}^i$ with $\tilde{s}^i_n(\bar{t}^i) = s^i_k$ and $\tilde{s}^i_n \in IW^i(\tilde{E}_n)$. Hence there exists $\tilde{\sigma}^{-i}_n \in \Delta(IW^{-i}(\tilde{E}_n))$ such that $\tilde{s}^i_n$ is a best reply to σ^{-i}_n with payoffs as in E_n. That means in particular that $\tilde{s}^i_n(\bar{t}^i) = s^i_k$ is a best reply to $\hat{\sigma}^{-i}_n \equiv$

$\sum \tilde{\sigma}_n^{-i}(t^{-i})\nu(t^{-i}|\bar{t}^i)$, with payoffs $\tilde{u}_n^i(\bar{t}^i)$. Although $\tilde{s}_n^i(\bar{t}^i)$ is not necessarily a best reply to $\hat{\sigma}^{-i} \equiv \lim_n \hat{\sigma}_n^{-i}$, it is a best reply to $\hat{\sigma}^{-i}$ if the payoffs at all the endpoints reached by the strategies $s_n^i(\bar{t}^i)$ and $\hat{\sigma}^{-i}$ are increased by a sufficiently large 'bonus' of ε_n. Furthermore, the bonus required converges to zero since $\lim \tilde{s}_n^i(\bar{t}^i)$ *is* a best reply to $\hat{\sigma}^{-i}$ with payoffs $\lim \tilde{u}_n^i(\bar{t}^i)$. Let the payoffs of type $s_k^i(m)$ be equal to $\tilde{u}_n^i(\bar{t}^i)$ with the ε_n bonus. Since $\varepsilon_n \to 0$, $\bar{E}_n \xrightarrow{p} E$. We now claim that there exist beliefs $\bar{\nu}_n(\cdot|\cdot)$ for the elaboration $\bar{E}_n$ such that the strategy I-tuple where each type $s_k^i(m)$ of each player i plays s_k^i is a Nash equilibrium in undominated strategies, hence this strategy I-tuple is in $1WIZ(\bar{E}_n)$. Recall that i's opponents will be an $I-1$ tuple of types in $R^{-i} \times \{1,...,K\}$. Let i's beliefs be such that if (s)he is of type $s_k^i(m)$, then (s)he believes that the opponents can only be of type $R^{-i} \times \{k\}$, and the distribution over R^{-i} is determined by $\hat{\sigma}^{-i}$ (see above). Then $s_k^i(m)$ is a best reply to the opponents all playing the strategy to which their type is matched. This shows that each type $s_k^i(m)$ playing s_k^i is a Nash equilibrium. Now we show that it is not weakly dominated. Since $\tilde{s}_n^i(\bar{t}^i) \in IW^i(\bar{E}_n)$, there is a $\tau_k^{-i} \in \Delta(S^{-i})$ such that $\tilde{s}_n^i(\bar{t}^i) = s_k^i$ is a best reply to τ^{-i} with full support when the payoffs are $\tilde{u}_n^i(\bar{t}^i)$. The strategy s_k^i is still a best reply to τ^{-i} when the payoffs are changed to include the bonus ε_n described above. So each type $s_k^i(m)$ playing s_k^i is a best reply to the full support strategy of the opponents where each type in $R^{-i} \times \{k\}$ plays τ_k^{-i}. □

Remark: The type spaces need to be expanded to M^i copies of R^i, rather than using only R^i, because of the second stage in the proof above. With the payoffs as described above if we aggregated all the types $s_k^i(m)$ into one type $s_k^i(1)$ then each type $s_k^i(1)$ playing s_k^i would still be a Nash equilibrium. But these strategies are not necessarily admissible. This is because to show admissibility we need one particular full support strategy of the opponents to which "each type $s_k^i(1)$ playing s_k^i" is a best reply. And for each k the type $s_k^i(1)$ is a best reply against a *different* full support

strategy τ_k^{-i}. Hence we need to allow different types for i's opponents as a function of k. □

The above results suggest an additional interpretation of the following result in FKL (Section 5.2). Any quasi-c-perfect equilibrium in E is equal to the the limit of a sequence of strict Nash equilibrium of a sequence of consistent elaborations $\tilde{E}_n$ which converges weakly to E. The observations below provide a simpler statement of this result. First we recall the definition of a quasi-c-perfect equilibrium. A c-perfect equilibrium is a perfect equilibrium where the test sequences for each player may be correlated and inconsistent (that is i and j may assign a third player k different trembles). A strict c-perfect equilibrium is a c-perfect equilibrium where the limit strategies are strict best replies to the test sequence. A quasi-c-perfect equilibrium is the limit of a sequence of strictly c-perfect equilibria in a sequence of consistent elaborations E_n which converge in payoffs to E. The first observation is that when considering weakly convergent elaborations, the set of limits of strictly c-perfect equilibria of games $\tilde{E}_n$ is the same as the set of limits of c-perfect equilibria: any c-perfect equilibrium can be made strict by the small payoff perturbations allowed by weak convergence. This follows from Lemma 4.1. The second observation is that c-perfect equilibria are Nash equilibria in strategies that are not weakly dominated. So, the theorem cited above says that the closure with respect to convergence in payoffs of undominated Nash equilibria is equal to the closure with respect to (consistent) weak convergence of the set of strict Nash equilibria. This result is closest in spirit to Proposition 4.3, where we showed that the closure with respect to convergence in payoffs of $1WIZ$ (which is the same as correlated rationalizable strategies which are undominated) is equal to the closure with respect to (inconsistent) weak convergence of the set ΠW.

5. An Alternative Interpretation

We motivated the consideration of payoff uncertainty by asking what players should infer when they observe play that is not consistent with their understanding of the game. This section sketches a different way to formalize those inferences. To begin, note that the state space for each player i is $\Omega^i \equiv \prod_{j \neq i}(S^j \times T^j)$ and specify i's beliefs over Ω^i by $q^i \in \Delta\,\Omega^i$. (Recall that $T^j \equiv U$, the set of all possible utility functions for i.) The assumption that players only update their beliefs about the payoffs if surprised is formalized by Supp $\text{marg}_{T^j} q^i = u^j$. Each player i has partitions on Ω^i determined by the extensive form. The traditional assumption (implicit in the refinements literature) which is questioned in this paper is that even when observing an unexpected strategy choice the player does not update his/her beliefs on T^j. Here we allow for Supp $\text{marg}_{T^j} q^i(\cdot \mid H^{-i} \subset S^{-i}) \neq u^j$ if $\text{marg}_{S^{-i}} q^i = 0$. This approach is related to the formalization in this paper in essentially the same way that beliefs at all information sets in sequential equilibrium are determined by a sequence of beliefs (generated by completely mixed strategies). Here a conditional probability $q^i(\cdot \mid H^{-i})$ is determined by a sequence of elaborations and the strategies in the elaborations. Our purpose in this section is to show how our results and the different notions of convergence used are related to the idea of updating beliefs on payoffs when observing unexpected strategy choices. This is best seen in a simple example which mimics the construction in the proof of Proposition 3.1. In Figure 2, player 1 believes at node a that player 2 will play L and that the payoffs are as in E, so 1 will play L; player 2 believes that 1 will play L and that the payoffs are as in E, and so player 2 does not expect to play. At node b 2 has been surprised and updates his beliefs to assigns probability one to 1 playing Rl and the payoffs being as in E', so 2 will play L as 1 expected at node a. There is no need to specify the beliefs at the third node since it is already clear that 1 playing L satisfies a natural form of backwards induction rationality when the players' beliefs over payoffs can be updated.

The above argument shows how strong convergence corresponds precisely to the ideas of updating beliefs. That is, it satisfies Supp $\text{marg}_{T^j} q^i(\cdot \mid H^{-i}) = u^j$ if H^j was assigned positive prior

probability by q^i. This interpretation does not allow for the payoffs to be "almost" equal to u^j when the player is not surprised, which points out an interesting distinction between strong and weak convergence, analogous to the difference between sequential and perfect equilibrium. In sequential equilibrium each player's beliefs at information sets along the equilibrium path are *precisely* that the equilibrium strategies are being played. Similarly the definition of strong convergence requires that (in the limit game) any types which receive positive probability are believed to have *precisely* the payoffs of the limit game. So at player 2's information set in Figure 3a, 2's beliefs are that the game is as in Figure 3b (so 1 can not play R because it is weakly dominated). Therefore $1WIZ = \{L\} \times \{l\}$. On the other hand in perfect equilibrium, even at information sets along the equilibrium path the players allow for "trembles" in the opponents strategies. The fact that "trembles" are allowed for even along the equilibrium path can be formally understood using the approach of lexicographic beliefs in Blume (1986) and Brandenburger and Dekel (1986). These papers suggest that the limit game which corresponds to weak convergence is as in Figure 3c below (where ε is an infinitessimal). In this case R is no longer weakly dominated, hence in fact $\overline{1WIZ} = \{R, L\} \times \{r, l\}$.

To conclude this comparison between weak and strong convergence, we note that the latter seems more appropriate for modeling the idea that a player i may update his/her beliefs about an opponent's payoffs if i observes an unexpected strategy choice by the opponent. On the other hand for modeling the question of robustness of a refinement it seems more natural to allow for the wider class of perturbed games which is formalized by weak convergence. The similarity between the closure of iteratively admissible strategies with respect to either notion of convergence emphasizes the close relationship between these two objectives.

One more point regarding this interpretation of the model is worth clarifying. Our approach allows a player to update his/her beliefs about the opponents' payoffs *whenever* surprised -- even if there is a "rational" explanation which does not require changing beliefs about the payoffs. One interesting extension of this model involves imposing the restriction that a player who is surprised by an opponent's strategy first tries to explain the observation without violating the assumptions

that payoffs and rationality are common knowledge. Instead the player assumes that his/her beliefs about the opponents strategy choice (or the opponents beliefs about other players' strategies, etc.) were wrong. Only if the "deviation" can not be explained by questioning the players' beliefs over the elements of strategic uncertainty is the more basic assumption regarding common knowledge of payoffs doubted.

6. Examples and Conclusion

We conclude with two examples. The first is meant both to motivate the simultaneous deletion of weakly dominated strategies in the first stage of 1*WIZ*, and to further clarify the relationship between our results and the idea of updating beliefs about payoffs after observing unexpected strategies. In the example of Figure 4 the order in which strategies are deleted matters. Deleting both players' dominated strategies simultaneously, and then iterating, yields $\{U\} \times \{L\}$. The argument against simultaneous deletion in the first round follows the intuition of backwards induction: If the payoffs are certain to be as specified, player 1 will never play M, and knowing this 2 should find both L and R reasonable. This argument yields $\{U, D\} \times \{L, R\}$ as the set of reasonable strategies. However the only reason for 2 to be willing to play R is if (s)he entertains no doubts about 1. But precisely these doubts are needed in the intuition for ruling out a weakly dominated strategy for either player.

Our second example helps explain why we do not feel comfortable with a prediction based on *IW*. In this example the question we would like to ask is whether 1 will play U. It is true that 2 should be sure that (s)he is at node b. So 2 will not play L or M. But confronted with playing at c what is the appropriate thought process for 1? We might argue that 1 should play C because U is only best if 2 plays L which is strictly dominated by M, and hence 2 shouldn't be expected to play L. But since 2 also shouldn't have played M what explanation is there for 1 being at c? Perhaps 1 will doubt his beliefs about 2's payoffs -- and then both U and C can be justified.

In conclusion we would like to review the main points of this paper. First, as we argued in the Introduction, the questions of robustness and "what to believe when surprised," are particularly relevant in models which assume only common knowledge of rationality and payoffs. Including payoff uncertainty in the model and using weak convergence yields a sharp and intuitive characterization of the "closure" of iterated weak dominance (Section 3). Also, the distinction between weak and strong convergence is helpful in understanding the relationships between strict best replies and weak dominance; and between the two objectives of this line of research, namely robustness and the updating of beliefs on null events.

References

R. Aumann (1974): "Subjectivity and Correlation in Randomized Strategies," *Journal of Mathematical Economics, 1, 67-96.*

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

K. Basu (1985): "Strategic Irrationality in Extensive Games," mimeo, Princeton University.

D. Bernheim (1984): "Rationalizable Strategic Behavior," *Econometrica, 52,* 1007-1028.

L. Blume (1986): "Lexicographic Refinements of Nash Equilibrium," mimeo, University of Michigan.

A. Brandenburger and E. Dekel (1987): "Rationalizability and Correlated Equilibria," forthcoming, *Econometrica.*

----- (1986): "Bayesian Rationality in Games," mimeo, Graduate School of Business, Stanford University.

----- (1986): "On An Axiomatic Approach to Refinements of Nash Equilibrium," mimeo, University of California at Berkeley.

D. Fudenberg, D. Kreps and D. Levine (1987): "On The Robustness of Equilibrium Refinements," forthcoming, *Journal of Economic Theory*

D. Gale and S. Sherman (1950): "Solutions of Finite Two-Person Games," in *Contributions to the Theory of Games,* Vol. 1, ed. by H. Kuhn and A. Tucker, Princeton University Press.

J. Harsanyi (1967-68): "Games of Incomplete Information Played By Bayesian Players," I, II, and III, *Management Science, 14,* 159-182, 320-334, 486,502.

E. Kohlberg and J. F. Mertens (1986): "On the Strategic Stability of Equilibria," *Econometrica 54,* 1003-1038.

D. Kreps and R. Wilson (1982): Sequential Equilibria, Econometrica 50, 863-894.

R. Luce and H. Raiffa (1957): *Games and Decisions,* Wiley: New York.

D. Pearce (1984): "Rationalizable Strategic Behavior and The Problem of Perfection," *Econometrica, 52,* 1029-1050.

R. Rosenthal (1981): "Games of perfect information, predatory pricing and the chain store paradox," *Journal of Economic Theory, 25,* 92-100.

P. Reny (1985): "Rationality, Common Knowledge, and the Theory of Games," mimeo, Princeton University.

R. Selten (1975): "Re-examination of the perfectness concept for equilibrium points in extensive games," *International Journal of Game Theory, 4* 25-55.

T. Tan and S. Werlang (1984): "The Bayesian Foundations of Rationalizable Strategic Behavior and Nash Equilibrium Behavior," mimeo, Princeton.

E. van Damme (1983): "Refinements of the Nash Equilibrium Concept," Springer-Verlag: Berlin, Hedelberg, New York.

Figure 1:

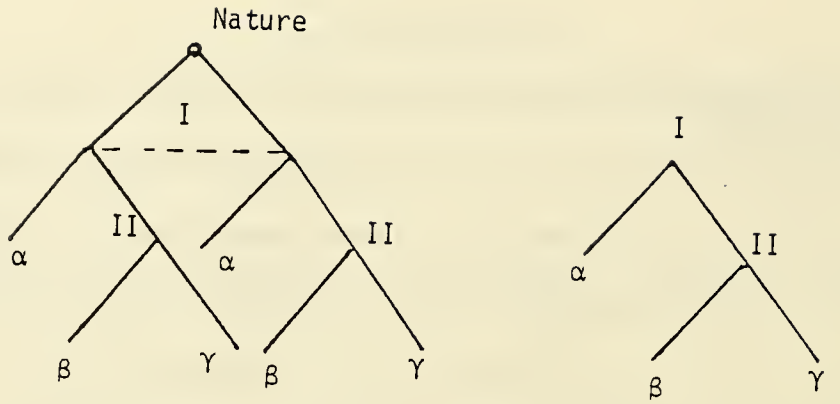


Figure 2:

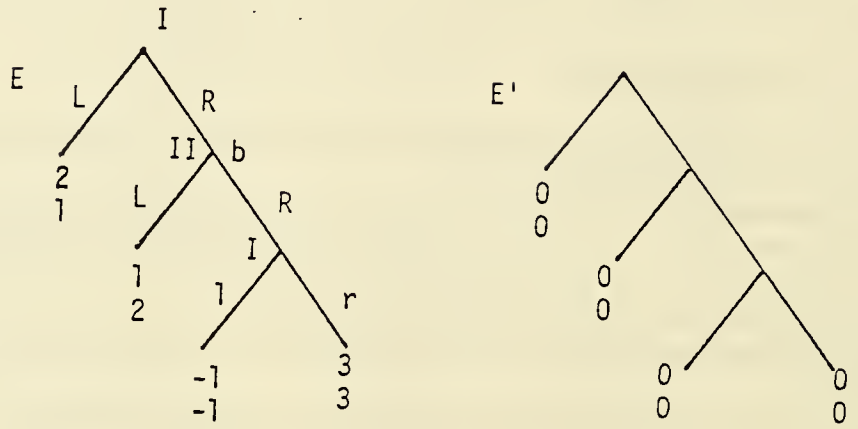
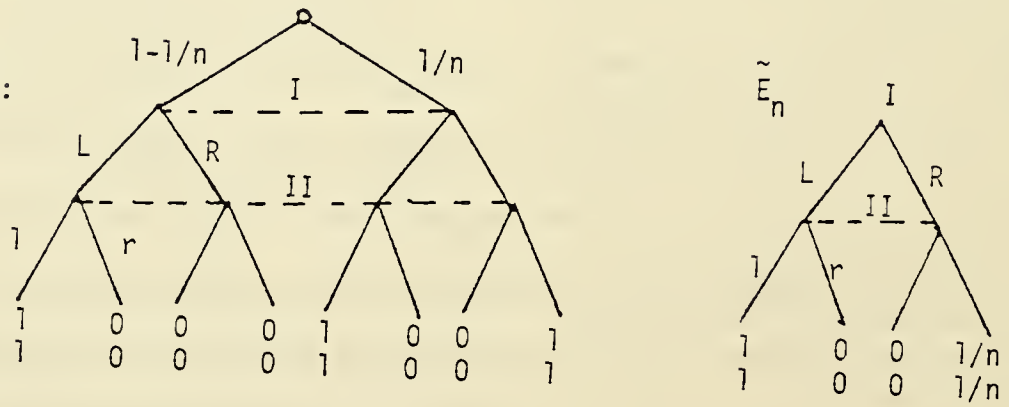
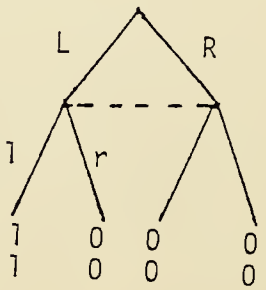


Figure 3:



b:



c:

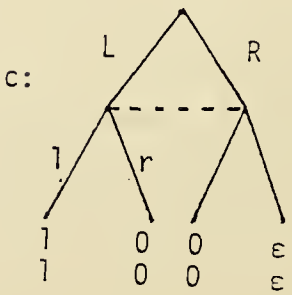
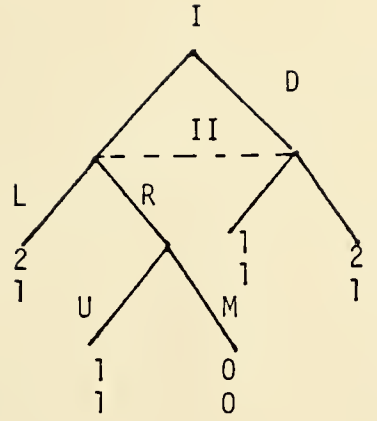
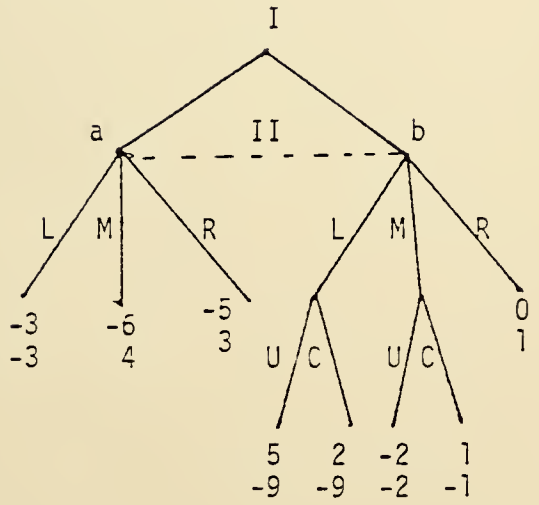


Figure 4:



	L	R
U	2,1	1,1
M	2,1	0,0
D	1,1	2,1

Figure 5:



	L	M	R
U	5,-9	-2,-2	0,1
C	2,-9	1,-1	0,1
D	-3,-3	-6, 4	-5,3

3611 102

Date Due

DEC 20 1991

Lib-26-67

MIT LIBRARIES
3 9080 005 130 247