JOURNAL OF ECONOMIC THEORY 52, 243–267 (1990)

Rational Behavior with Payoff Uncertainty

EDDIE DEKEL

University of California, Berkeley, California 94720

AND

DREW FUDENBERG*

Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Received February 27, 1989; revised January 31, 1990

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 strategies that are strongly dominated. This extends the Fudenberg, Kreps, and Levine (*J. Econ. Theory* **12** (1988), 354–380) study of the robustness of Nash equilibrium refinements to the robustness of solution concepts based only on rationality. Our results also clarify the relationship between various notions of what it means for payoff uncertainty to be "small." *Journal of Economic Literature* Classification Number: 026.

1. 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 when the players' strategic beliefs may be inconsistent, that is, using only the assumption that it is common knowledge that the players are rational. Bernheim [4] and Pearce [19] 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

* This work was begun while the second author was at the University of California, Berkeley. We acknowledge many helpful conversations with Adam Brandenburger and Matthew Rabin. Financial support from the Miller Institute and NSF grants SES 88-08133 and 88-08204 is gratefully acknowledged.

0022-0531/90 $3.00

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 [4, Sect. 6(b)], Pearce [19, Sect. 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 [12], (henceforth referred to as FKL) have argued that standard Nash equilibrium refinements such as trembling hand perfection [23] and sequential equilibrium [16] 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 s/he expects the opponents to play subsequently. If the observed play to date is not consistent with the player' initial understanding of the game, one plausible inference is that the reason for the deviation is that the deviators' 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.

The question of what players can infer from behavior they did not expect to occur is not restricted to equilibrium analysis: Rosenthal [22], Reny [21], Basu [3], Binmore [5, 6], and Bonanno [8] 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 characterize the implications of introducing small uncertainties about the payoffs for predictions 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 inconsistent beliefs (inconsistent in the sense that they may disagree) about each other's 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

[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 by Aumann [2], Brandenburger and Dekel [10], and Tan and Werlang [25].

achieve our characterizations. First, a sharp notion for the implications of rational behavior must be given for games with small doubts. We chose the notion of iterated deletion of weakly dominated strategies[2] since it clearly incorporates certain 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 other's strategies. Hence it seems natural here to consider the case where they have inconsistent doubts about each other's payoffs as well.

The latter modeling issue emphasizes the fact that the key question in evaluating the robustness of various solution concepts is which sequences 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 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 $\mathbf{S}^{\infty}\mathbf{W}$. This set is computed by first 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 uncertainty about the payoffs, then iterated weak dominance goes too far. Recently Börgers [9] generalized the argument for $\mathbf{S}^{\infty}\mathbf{W}$. He shows that it captures *any* situation where it is "almost common knowledge" (in the

[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 [15].

[4] In two person games this coincides with Bernheim's [4] 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.

sense of Monderer and Samet [18] and Stinchcombe [24]) that players do not use weakly dominated strategies, and not only the case where players' doubts result from payoff uncertainty.[5]

Section 4 shows that weak convergence yields the set $\overline{\mathbf{S}^{\infty}\mathbf{W}}$ which is the closure of $\mathbf{S}^{\infty}\mathbf{W}$ 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 explains why the concept $\mathbf{S}^{\infty}\mathbf{W}$ allows cooperation in the finitely repeated prisoner's dilemma. This is of particular interest since the Nash equilibria of this game do not exhibit cooperation, and FKL have shown that even under payoff uncertainty Nash equilibria are robust to (consistent) payoff perturbations. We use these contrasting predictions of $\mathbf{S}^{\infty}\mathbf{W}$ and Nash equilibrium to illustrate the distinction between consistent and inconsistent strategic beliefs in the context of examining the robustness of solution concepts. Section 6 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, Brandenburger, and Dekel [7], it is argued that the distinction between the two notions of convergence is analogous to the difference between perfect and sequential equilibrium.

To summarize, this paper shows that the FKL critique of refinements of equilibrium can be extended to iterated deletion of weakly dominated strategies, which allows for strategic uncertainty. Payoff uncertainty is shown to directly cast doubts on the deletion of weakly dominated strategies. Thus, we pinpoint weak dominance arguments, which underlie many refinements, as the feature which led to the failure of robustness demonstrated by FKL. Finally, we believe that the robustness of $\mathbf{S}^{\infty}\mathbf{W}$ suggests it as a useful concept in its own right.

2. Perturbations, Elaborations, and Convergence

Since this paper examines some implications of "small" amounts of payoff uncertainty, a crucial issue to consider is 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

[5] Börgers' almost common knowledge assumption is satisfied in one version of our model since strong convergence implies that in nearby games it is almost common knowledge (at states of the world which are possible in the limit) that the limit game is actually being played. Hence our results imply that Börger's assumption cannot imply more restrictions than $\mathbf{S}^{\infty}\mathbf{W}$.

that the physical extensive form (who moves when, and the players' information regarding their opponents' actions) is common knowledge, and the only doubts the players entertain (other than those explicitly specified in the given extensive form) are about each other's 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 $z \in Z$, information sets H, and a utility function $u^i \in U \equiv \{f \mid f\colon Z \to R\}$ for each player i.

Following Harsanyi [14], 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., are summarized by the player's 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.[6] Under this assumption 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 T 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 $v^i(\cdot \mid t^i)$ 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 v^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 $\sigma^i \in \Delta(S^i)$, and beliefs over S^{-i} are denoted by $\sigma^{-i} \in \Delta(S^{-i})$, where $\Delta(X)$ is the set of probability 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 "$\tilde{\ }_n$" to the appropriate symbol; e.g., $\tilde{S}^i_n$ denotes the pure strategies of i in $\tilde{E}_n$. When a particular elaboration $\tilde{E}_n$ is discussed 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; e.g., $\tilde{u}^i_n(t^i, t^{-i})$ 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.

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

[6] FKL call this assumption "personal types" in distinction from the "general types" case in which i may receive better information than j about j's payoffs. We restrict attention to personal types because we feel that this is most often the relevant kind of uncertainty.

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 of 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 be "almost" sure that their payoffs are *precisely* as in the original game; or that they be *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 the next section only strong convergence will be examined, since the results are most intuitive and simplest to prove for this case. Furthermore, as we argue in Section 6, strong convergence is most appropriate for modeling players' inferences when "surprised." The other notions, which will be discussed in Section 4, are important both for clarifying the relationship of this paper with FKL, and to verify that our results do not depend in an essential way on whether weak or strong convergence is used.

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 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.[7] 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, so long as players are almost certain that the payoffs are as in the original game, the effects of inconsistent priors over the payoffs can be duplicated by appropriately specified inconsistencies in the players' beliefs about each others' strategies. Brandenburger and Dekel [11] prove a version of the converse result: some of the effects of inconsistent strategic beliefs can be achieved in a model where strategic beliefs are consistent, but players have inconsistent beliefs about the structure of the game. 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. However, once the structural beliefs are required to be consistent, the assumption of a mediator does entail a loss of generality. Similarly, in the present paper, the consistency in the players' beliefs over the type spaces can be achieved (in Proposition 3.1) only by incorporating the inconsistency into the beliefs over the strategy spaces. So when the latter is ruled out (as in Proposition 4.1) the consistency of the beliefs over the type spaces can no longer be achieved.

[7] With independent types player i's observation of j's play cannot affect i's beliefs over k's type, whereas in the "personal types" model p need not be a product measure.

The two results here, combined with their converses in Brandenburger and Dekel [11], show that 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 always be separated from assumptions about the other.

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.

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) $|\text{support } p_n| < M$ for all n;

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

(ii) For all i there is a subset $\bar{T}^i$ of T^i such that

(a) $\lim_n p_n(\bar{T}^i) = 1$;

(b) for all $\bar{t}^i \in \bar{T}^i$, $\tilde{u}_n^i(\bar{t}^i) = u^i$.

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 $\nu_n(\cdot \mid \bar{t}^i)$ of every "sane" type $\bar{t}^i$ in $\bar{T}^i$ are that the other players are very likely to be "sane." 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 Fig. 2.1 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 the belief over their types may be nondegenerate) which is equivalent to their playing a mixed strategy.

DEFINITION 2.2. A sequence of strategies $\tilde{\sigma}_n^i$ will be said to converge to σ^i (written $\tilde{\sigma}_n^i \to \sigma^i$) if $\lim_n \sum_{t^i \in T^i} \mu_n^i(t^i)\, \tilde{\sigma}_n^i(t^i) = \sigma^i$. A sequence of strategy profiles $\tilde{\sigma}_n = (\tilde{\sigma}_n^1, \ldots, \tilde{\sigma}_n^I)$ converges to σ if $\tilde{\sigma}_n^i$ converges to σ^i for each i.

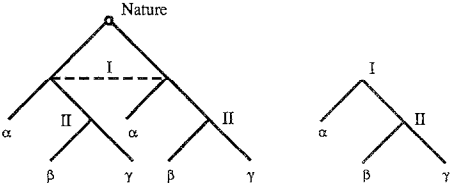


FIGURE 2.1

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 player's strategies.

3. Iterated Weak Dominance and $\mathbf{S}^\infty\mathbf{W}$

A pure 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 the inequality is strict for some σ^{-i}. Any strategy s^i which is not weakly dominated is said to be admissible (Luce and Raiffa [17]), 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 [19, Appendix B], van Damme [26], Gale and Sherman [13]). Any mixture over admissible actions is admissible. Kohlberg and Mertens [15] have argued in favor of iterated weak dominance, denoted $\mathbf{W}^\infty$ (that is, iteratively deleting strategies which are weakly dominated), as a minimal requirement of a solution concept. More generally, $\mathbf{S}^l\mathbf{W}^k$ 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. Each of these sets is a Cartesian product of strategies for each player, so $(\mathbf{S}^l\mathbf{W}^k)^i$ denotes the projection of $\mathbf{S}^l\mathbf{W}^k$ on i's strategy space.

Proposition 3.1 below says that any strategy profile[8] in $\mathbf{S}^\infty\mathbf{W}$ is close to a strategy profile in $\mathbf{W}^\infty$ for some sequence of nearby games, and any strategy profile in $\mathbf{W}^\infty$ for nearby games is close to a strategy profile in $\mathbf{S}^\infty\mathbf{W}$. 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 $\mathbf{S}^\infty\mathbf{W}$ is questionable, even if we agree to rule out all strategies not in $\mathbf{W}^\infty$ when payoffs are common knowledge.[9]

Proposition 3.1. *$s \in \mathbf{S}^\infty\mathbf{W}(E)$ if and only if there are a sequence of consistent elaborations $\tilde{E}_n \to E$, and strategies $\tilde{s}_n \in \mathbf{W}^\infty(\tilde{E}_n)$ such that $\tilde{s}_n \to s$.*

Proof. (Only if) In this direction of the proof the sequence of elaborations $\tilde{E}_n$ is constructed. Let $T_n^i = \{\bar{t}^i, \hat{t}^i\}$, 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 $1 - 1/n$ 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

[8] We thank Matthew Rabin for pointing out that Proposition 3.1 can be stated in terms of strategy profiles as well as individual strategies.

[9] Under the assumption of general types, instead of personal types, the "closure" of $\mathbf{W}^\infty$ would simply be $\mathbf{S}^\infty$.

i is sane and all the other players are crazy. Thus, when i is sane, the conditional probability $\nu_n^i(\cdot \mid \bar{t}^i)$ 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 $1/(I(n-1)+1)$. Player i's strategy in $\tilde{E}_n$ is written as an ordered pair $(\bar{\sigma}_n^i, \hat{\sigma}_n^i)$ 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/here beliefs over $\tilde{S}_n^{-i}$ (the opponents' strategies) as elements of $\Delta(S^{-i}) \times \Delta(S^{-i})$. Such beliefs are denoted by ordered pairs $(\bar{\sigma}_n^{-i}, \hat{\sigma}_n^{-i})$. We claim that $(\mathbf{W}^\infty(\tilde{E}_n))^i = \{(\bar{s}^i, s^i) \mid \bar{s}^i \in (\mathbf{S}^\infty \mathbf{W}(E))^i, \quad s^i \in S^i\}$. Since $\mathbf{W}^\infty(\tilde{E}_n) = \prod_i (\mathbf{W}^\infty(\tilde{E}_n))^i$, this implies the only if part of Proposition 1. Proving the claim involves two steps.

Step 1. $(\bar{s}^i, s^i) \in \mathbf{W}(\tilde{E}_n)$. Since $\bar{s}^i \in (\mathbf{S}^\infty \mathbf{W}(E))^i \subseteq (\mathbf{W}(E))^i$, 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 $\tilde{S}_n^{-i}$. For future reference let α be the smallest weight assigned to any pure strategy s^{-i} by $\bar{\sigma}^{-i}$.

Step 2. If $(\bar{s}^i, s^i) \in (\mathbf{SW}(E))^i \times S^i$ then $(\bar{s}^i, s^i) \in \mathbf{W}^2(\tilde{E}_n)$. 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 $(\mathbf{W}(\tilde{E}_n))^{-i}$, which by step 1 is a superset of $(\mathbf{S}^\infty \mathbf{W}(E))^{-i} \times S^{-i}$. Since $\bar{s}^i \in (\mathbf{SW}(E))^i$ there is a $\sigma^{-i} \in (\mathbf{W}(E))^{-i}$ 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 $(\mathbf{W}(E))^{-i}$. 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, with the sane opponents playing $\hat{\sigma}^{-i}$) of σ'^{-i} and $\hat{\sigma}^{-i}$ is $\bar{\sigma}^{-i}$. To make this precise 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 $(\bar{s}^i, s^i) \in (\mathbf{S}^\infty \mathbf{W}(E))^i \times S^i$ then $(\bar{s}^i, s^i) \in (\mathbf{W}^\infty(\tilde{E}_n))^i$.

Remark 3.1. Note that in step 2 the fact that $\bar{s}^i \in (\mathbf{SW}(E))^i$ was used in finding $\sigma^{-i} \in (\mathbf{W}(E))^{-i}$ 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 $\mathbf{S}^\infty$ step in $\mathbf{S}^\infty\mathbf{W}$ should be necessary for a characterization of the "closure" of $\mathbf{W}^\infty$. This is verified in the proof of the "if" direction below.

(If) To prove that if $\tilde{s}_n \in \mathbf{W}^\infty(\tilde{E}_n)$ converges to s, then s is in $\mathbf{S}^\infty\mathbf{W}(E)$, we establish the slightly stronger fact that this inclusion holds player-by-player.

Step 1. $\tilde{s}_n^i \in (\mathbf{W}(\tilde{E}_n))^i$ implies $\bar{s}^i(\bar{t}^i) \in (\mathbf{W}(E))^i$ for all $\bar{t}^i \in \bar{T}^i$. This follows from the fact that $\tilde{s}_n^i$ is a best reply to some full support belief $\tilde{\sigma}_n^{-i}$ over $\tilde{S}_n^{-i}$. Hence $\tilde{s}_n^i(\bar{t}^i)$ is a best reply to $\hat{\sigma}_n^{-i} \equiv \sum_{t^{-i}} \nu_n^i(t^{-i} \mid \bar{t}^i)\, \tilde{\sigma}_n^{-i}(t^{-i})$ 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 $\tilde{s}_n^i$ is not weakly dominated in E.

Step 2. $\tilde{s}_n^i \in \mathbf{W}^2(\tilde{E}_n)$ implies $\tilde{s}_n^i(\bar{t}^i) \in \mathbf{SW}(E)$. We know that $\tilde{s}_n^i(\bar{t}^i)$ is a best reply to some $\hat{\sigma}^{-i} \equiv \sum_{t^{-i}} \tilde{\sigma}_n^{-i}(t^{-i})\, \nu_n(t^{-i} \mid \bar{t}^i)$ for some $\tilde{\sigma}_n^{-i}$ which is suppported by strategies in $(\mathbf{W}(\tilde{E}_n))^{-i}$ since $\tilde{s}_n^i \in (\mathbf{W}^2(\tilde{E}_n))^i$. As noted earlier, by condition (ii) of the definition of convergence in types $\nu_n(t^{-i} \mid \bar{t}^i)$ 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 $(\mathbf{W}(E))^{-i}$. 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 $(\mathbf{W}(E))^{-i}$; hence $\tilde{s}_n^i(\bar{t}^i)$ is not strongly dominated within $(\mathbf{W}(E))^i$.

Step 2 can now be iterated to show that $\tilde{s}_n^i(\bar{t}^i)$ is an element of $(\mathbf{S}^\infty\mathbf{W}(E))^i$. ∎

Remark 3.2. 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 $(\mathbf{W}(E))^{-i}$: Its support is larger because of the possibility of crazy types of $j \neq i$, and of course its limit may have smaller support than $(\mathbf{W}(E))^{-i}$.

The sequence of elaborations $\tilde{E}_n$ we constructed in the proof of the "only if" direction has the property that, when a player is sane, s/he assigns probability $\nu_n(\bar{T}^{-i} \mid \bar{t}^i) = I(n-1)/(I(n-1)+1)$ to all his/her opponents being sane as well. This means that, when all players are sane, the event "all players are sane" is evident $\nu_n(\bar{T}^{-i} \mid \bar{t}^i)$-belief in the sense of Monderer and Samet [18]. Since both the marginal probability that all players are sane and these conditional probabilities converge to one, the event "all players are sane" is "almost common knowledge" (at states which are

possible in the limit). (More generally, we believe that strong convergence implies that the payoffs are "almost common knowledge.") Since it is common knowledge that players conform to $\mathbf{W}^\infty$ in any elaboration, we can conclude that for n large it is almost common knowledge that players do not use strategies that are weakly dominated with respect to the sane (i.e., limit) payoffs. Thus our results show that the hypothesis of almost common knowledge of admissibility cannot imply more restrictions than $\mathbf{S}^\infty\mathbf{W}$. This observation was prompted by Börgers' [9] study of the implications of almost common knowledge of admissibility in a more general setting.

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 player's beliefs in steps 1 and 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 the iteration of the second step the beliefs over the opponents would be different each time. Since the beliefs about the crazy types' *strategies* are allowed to be inconsistent, one suspects that allowing for inconsistent beliefs over the types will not change this result. Corollary 3.1 below confirms this intuition. We should point out that inconsistent beliefs about types do, in general, matter.[10] The reason the inconsistency is innocuous here is that we consider distributions that put probability close to one on all players being sane.

In order to formalize the inconsistent case the definition of convergence of elaborations must be extended accordingly. Recall that the common prior p was used in condition (ii) of Definition 2.1 to verify that $\bar{T}^i$ was the set of types with positive limit probability. For a sequence of elaborations to converge we then required that for any player i all types t^i in $\bar{T}^i$ have the same payoffs as in the original game. When the player's priors can differ, the definition of convergence—i.e., the analog of condition (ii)—is more delicate. We certainly want $\lim_n p_n^i(\bar{T}^i) = 1$, so that each player believes that s/he almost certainly will be sane. However, we also need each player i to be almost certain that i's opponents are sane as well, and so on iteratively. To formalize this, let $m^i \equiv \lim_n \operatorname{support}(\mu_n^i)$, and let $m^i(t^i) \equiv \lim_n \operatorname{support}(\nu_n^i(\cdot \mid t^i))$ be player i's beliefs about his/her opponents conditional on his/her own type. We will still require $\bar{T}^i$ as in condition (ii)(b) of Definition 2.1 to be a set of types whose payoffs are the same as in E,

[10] For example let player two have two types: t_2' has s_2' as a strongly dominating strategy, and s_2'' is strongly dominant for t_2''. Assume further that s_1' is a strict best response to s_2' and that s_1'' is best against s_2''. Then the set of rationalizable strategies depends on 1's beliefs about 2. If we allow inconsistent beliefs then we can obtain rationalizable outcomes which are not rationalizable for any consistent beliefs about the types.

but now we replace condition (ii)(a) of the definition by the following iterative test:

(ii) (a′) $m^i \subseteq T^i$, and 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 $t^l \in \bar{T}^l$.

Thus, if t^h has positive probability in the limit, and t^h thinks t^l has positive limit probability, then $t^l \in \bar{T}^l$; and the same is true for arbitrary chains of players. 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 $\bar{T}^i$, i.e., replacement of (ii)(a) with (ii)(a′). Since (ii)(a) and (ii)(a′) 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. *$s \in \mathbf{S}^\infty\mathbf{W}(E)$ if and only if there is a sequence of elaborations $\tilde{E}_n \to E$, and strategies $\tilde{s}_n \in \mathbf{W}^\infty(\tilde{E}_n)$ such that $\tilde{s}_n \to s$.*

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 $\tilde{t}^i$ in $\bar{T}^i$ may have payoffs $\tilde{u}^i_n(\tilde{t}^i)$ which *converge* to the payoffs u^i in E, instead of $\tilde{u}^i_n(\tilde{t}^i) = u^i$ for all n and $\tilde{t}^i \in \bar{T}^i$. As one would expect, the consequence of allowing more convergent sequences of elaborations is that more strategies in E survive $\mathbf{W}^\infty$ in nearby games. In fact, the resulting set is the closure of $\mathbf{S}^\infty W$ with respect to extensive form payoff perturbations, which we denote $\overline{\mathbf{S}^\infty\mathbf{W}}$. Moreover (again because more sequences of elaborations converge to a given game E) we can show that any strategy in $\mathbf{S}^\infty\mathbf{W}$ is close to a strategy which satisfies a stronger requirement than $\mathbf{W}^\infty$ 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. Brandenburger and Dekel [10] show that correlated rationalizability is the same as *a posteriori* equilibrium (Aumann

[1]). This is the same as a Nash equilibrium where a subjective correlating device is explicitly introduced, and the players' strategies are required to be optimal conditional on all observations of the device, including those assigned prior probability zero. So, an alternative to $\mathbf{W}^\infty$ as a refinement of $\mathbf{S}^\infty$ is to look at strict Nash equilibrium with subjective correlating devices.

DEFINITION 4.1. $\tilde{E}_n$ converges in payoffs to E ($\tilde{E}_n \to^p E$) 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 $(s^1, ..., s^I)$ is *strict* if each player's 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 $\tilde{E}_n \to^p E$, 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 allowing for small extensive form payoff perturbations permits strategies which are not weakly dominated to be made strict best replies. Proposition 4.1 below is an analog to Proposition 3.1, where the notion of "not weakly dominated" is strenghtened 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 ($\tilde{E}_n \succsim E$) if conditions (i) and (ii)(a) of Definition 2.1 hold, and:

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

PROPOSITION 4.1. *If $s^i \in (\mathbf{S}^\infty \mathbf{W}(E))^i$ then there is a sequence of elaborations $\tilde{E}_n \succsim E$ and strategies $\tilde{s}_n^i \succsim s^i$ such that $\tilde{s}_n^i$ is a strategy in a strict Nash equilibrium of $\tilde{E}_n$.*

Remark 4.1. 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 $(\mathbf{S}^\infty\mathbf{W}(E))^i$ 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 $\bar{t}_k^i = \bar{s}_k^i \in (\mathbf{S}^\infty\mathbf{W})^i$ 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 $\hat{t}_k^i$, 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 from those in E, the types in $\hat{T}$ must have probability zero in the limit.

To make $\bar{s}_k^i$ in $(\mathbf{S}^\infty\mathbf{W})^i$ a strict best reply for type $\bar{t}_k^i$ we proceed as follows. First fix a sequence $\varepsilon_n \downarrow 0$. Since $\bar{s}_k^i \in (\mathbf{S}^\infty\mathbf{W})^i$ 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 $\hat{\sigma}_k^{-i} \in \Delta(\prod_{j \neq i} (\mathbf{S}^\infty\mathbf{W})^j)$, 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 $\nu_n(\cdot \mid \bar{t}_k^i)$ (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's playing as told is a strict best reply to the other playing as told, hence playing as told is a strict Nash equilibrium. ∎

Remark 4.2. In our construction we used several different sane types for each player. This is because we require each strategy $s_k^i \in (\mathbf{S}^\infty\mathbf{W})^i$ to be a *strict* best reply to some strategies of the opponents, and to ensure that this preference is strict we may need to use a different small payoff perturbation for each k.

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 limits 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 $\mathbf{S}^\infty\mathbf{W}(E)$. This is because $\mathbf{S}^\infty\mathbf{W}$ 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 $\mathbf{W}^\infty$ 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 $\mathbf{S}^\infty\mathbf{W}$. 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. For this purpose we replace $\mathbf{S}^\infty\mathbf{W}$ by its "closure," denoted by $\overline{\mathbf{S}^\infty\mathbf{W}}$.

DEFINITION 4.4. $s^i \in (\overline{\mathbf{S}^\infty\mathbf{W}}(E))^i$ if $\tilde{s}_n^i \to s^i$ and $\tilde{s}_n^i \in (\mathbf{S}^\infty\mathbf{W}(\tilde{E}_n))^i$ for some sequence of elaborations $\tilde{E}_n \to^p E$.

PROPOSITION 4.2. $\tilde{s}_n^i \in (\mathbf{W}^\infty(\tilde{E}_n))^i$, $\tilde{E}_n \simeq E$, $\tilde{s}_n^i \simeq s^i \in E$, *if and only if* $s^i \in (\overline{\mathbf{S}^\infty\mathbf{W}}(E))^i$.

Proof. (If) This follows from a simple diagonal argument and Proposition 3.1. If $s^i \in (\overline{\mathbf{S}^\infty\mathbf{W}}(E))^i$ then there exists a sequence $\tilde{E}_n \to^p$ with $\tilde{s}_n^i \to s^i$ and $\tilde{s}_n^i \in (\mathbf{S}^\infty\mathbf{W}(\tilde{E}_n))^i$. By Proposition 3.1 there exist $\tilde{\tilde{E}}_{k,n} \to \tilde{E}_n$ and $\tilde{\tilde{s}}_{k,n} \to \tilde{s}_n^i$ with $\tilde{\tilde{s}}_{k,n} \in (\mathbf{W}^\infty(\tilde{E}_n))^i$. Clearly $\tilde{\tilde{s}}_{m,n}^i \to s^i$ and $\tilde{\tilde{E}}_{m,m}^i \simeq E$ as required.

(Only if) We are given a sequence $\tilde{E}_n \rightrightarrows E$. Let R^i denote the strategies played by sane types; that is, $R^i \equiv \limsup R_n^i$, where $R_n^i \equiv \{s^i \in S^i \mid \text{for some } \tilde{s}_n^i \in (\mathbf{W}^\infty(\tilde{E}_n))^i \text{ and some } \bar{t}^i \in \bar{T}^i,\ \tilde{s}_n^i(\bar{t}^i) = s^i\}$. 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 $R^i \times R^{-i}$. The types in $\bar{E}^n$ will be denoted by $s_k^i(m)$, where $s_k^i \in R^i$ for $k = 1, ..., |R^i|$, and $m = 1, ..., |R^{-i}|$. For a given i and k all types $s_k^i(m)$ have the same payoffs (independent of m), and these payoffs are determined as follows. Since $s_k^i \in R^i$, then taking a subsequence if necessary, there exist $\tilde{s}_n^i$ and $\bar{t}^i$ with $\tilde{s}_n^i(\bar{t}^i) = s_k^i$ and $\tilde{s}_n^i \in (\mathbf{W}^\infty(\tilde{E}_n))^i$. Hence there exists $\tilde{\sigma}_n^{-i} \in \Delta((\mathbf{W}^\infty(\tilde{E}_n))^{-i})$ such that $\tilde{s}_n^i$ is a best reply to σ_n^{-i} with payoffs as in E_n. That means in particular that $\tilde{s}_n^i(\bar{t}^i) = s_k^i$ is a best reply to $\hat{\sigma}_n^{-i} \equiv \sum \tilde{\sigma}_n^{-i}(t^{-i})\, \nu(t^{-i} \mid \bar{t}^i)$, 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 $\varepsilon_n \to 0$, $\bar{E}_n \to^p E$. We now claim that there exist beliefs $\bar{\nu}_n(\cdot \mid \cdot)$ 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 $\mathbf{S}^\infty\mathbf{W}(\bar{E}_n)$. Recall that i's opponents will be an $(I-1)$-tuple of types in $R^{-i} \times R^i$. 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 $\{k\} \times R^{-i}$, 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.

Finally we show that it is not weakly dominated. Since $\tilde{s}_n^i(\bar{t}^i) \in (\mathbf{W}^\infty(\tilde{E}_n))^i$, 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 4.3. The reader may wonder why we did not simply take each player's type space to be R^i, instead of $R^i \times R^{-i}$, since all types $s_k^i(m)$ with the same m have the same payoffs. Indeed, a single type $s_k^i(1)$ for each $k \in R^i$ would suffice for us to have a Nash equilibrium with each type $s_k^i(1)$ playing s_k^i. However, to make the strategy "play s_k^i whenever told to" *admissible* for player i in a two person game, we will need as many distributions over the types of $j \neq i$ as $|R^j|$, even if this is larger than $|R^i|$. The easiest way to guarantee enough types for the strategy profile to be a Nash equilibrium and to be admissible is to use the product structure.

Remark 4.4. The above results enable us to provide two clarifying observations regarding the relation of this work to FKL.

1. Our results show that in order to be robust an equilibrium refinement can, at most, apply one round of deletion of weakly dominated strategies. To see this recall Proposition 7 of FKL which states that a strategy profile is robust if and only if it is quasi-c-perfect. Using Lemma 4.1 and the characterization of weakly dominated strategies discussed at the beginning of Section 3, this proposition is equivalent to one stating that a refinement is robust if and only if it is contained in the closure (with respect to convergence in payoffs) of the set of Nash equilibria in the game remaining after weakly dominated strategies are deleted. This reinterpretation of Proposition 7 is similar to our Proposition 4.3, which states that the closure with respect to weak convergence of the set $\mathbf{W}^\infty$ is equal to the closure with respect to convergence in payoffs of $\mathbf{S}^\infty\mathbf{W}$ (which is the same as correlated rationalizable strategies in the game remaining after weakly dominated strategies are deleted.

2. By adopting admissibility as our basic notion of rational behavior we obtain similar characterizations of the closure of $\mathbf{W}^\infty$ when weak or strong convergence is used (Propositions 3.1 and 4.3). Since FKL use strict best replies as their tight notion of rationality, their results hold only for weak convergence. As Section 6 argues, strong convergence may be more appropriate for some purposes.

5. $\mathbf{S}^\infty\mathbf{W}$ AND THE FINITELY REPEATED PRISONER'S DILEMMA

The concept $\mathbf{S}^\infty\mathbf{W}$ allows there to be cooperation in the finitely repeated prisoner's dilemma, even though in a Nash equilibrium the players always fink. By expanding on this observation we can illustrate the intuition for our approach and the substantive way in which it differs from that of FKL. The stage game for the prisoner's dilemma is shown in Fig. 5.1; this game is to be repeated T times.

First note that for any horizon T the unique outcome consistent with $\mathbf{W}^\infty$ is for both players to always fink. (Any strategy s of the repeated game

	Coop	Fink
Coop	1,1	-1,2
Fink	2,-1	0,0

FIGURE 5.1

that specifies cooperation in the last period for some sequence of play through period $T-1$ is weakly dominated by the strategy that agrees with s through period $T-1$ and then finks at T. Proceeding with the familiar induction, we argue that if the opponent is sure to fink for the last k periods then cooperating at $T-k$ is weakly dominated.) However, with $\mathbf{S}^{\infty}\mathbf{W}$ the induction stops after one round: After all strategies that cooperate in the last period (which are weakly dominated) are deleted, cooperating in the next-to-last period cannot be ruled out by *strict* dominance.

Next we examine the strategies that survive $\mathbf{W}^{\infty}$ once we allow for a small amount of payoff uncertainty. When $T=2$, $\mathbf{W}^{\infty}$ applied to the perturbed game predicts that both players will fink in both periods. This is because every sane type of each player will find it weakly dominating to fink in the last period. Since the likelihood of a sane opponent is almost one, each sane type will judge that cooperating in the first period is unlikely to induce the opponent to cooperate in the second, so (in the perturbed game) cooperating in the first period can be ruled out by $\mathbf{W}^{\infty}$. However, when $T=3$, applying $\mathbf{W}^{\infty}$ to the perturbed game no longer yields strong conclusions. We demonstrate below that strategies for player 2 which specify cooperation in the second period if player 1 cooperated in the first period are not removed at the second round of deletion. The failure of "cooperate at period 2" to be weakly dominated in the perturbed game suggests that the arguments for a unique outcome (under payoff uncertainty!) must rely on more than backwards induction: Information about period 1 play must be "moved forward" to generate constraints in period 2. It will soon be seen how the hypothesis that play corresponds to a Nash equilibrium can provide the necessary forward link between periods 1 and 2.

To show that "cooperate in period 2" is not weakly dominated in a perturbed game, and to illustrate how the consistency of Nash equilibrium rules out cooperation, a particular elaboration is constructed. In this elaboration $\mathbf{W}^{\infty}$ is consistent with cooperation while Nash equilibrium rules out cooperation. Player 1 has only two types, "sane" and "crazy," where the latter occurs with probability ε, and player 2 is sane with probability one. The sane players have the same payoffs as in the original game; the crazy player 1 plays σ: Cooperate in the first two periods, and cooperate in the third if and only if player 2 cooperated in the second.

Weak dominance implies that the sane types fink in the last period. The second round of weak dominance then implies that the sane player 1 should fink in period 2, and that player 2 should fink in period 2 if player 1 finked in period 1. On the third roud of deletion, we conclude that since player 1's play in independent of player 2's first period action, player 2 should fink in period 1. These arguments reduce the normal form to the

one shown in Fig. 5.2. Strategy U for player 1 is: "if sane always fink, if crazy play σ;" and strategy D is: "if sane cooperate in period 1 and fink in periods 2 and 3, if crazy play σ." Strategy L is "always fink," and R is "cheat in periods 1 and 3, cooperate in period 2 iff player 1 cooperates in period 1." The payoffs for player 1 are expected values for player 1's two types, where we have set the payoff of the crazy type to equal zero if s/he plays σ.

The $\mathbf{W}^\infty$ algorithm terminates at this stage as each remaining strategy is a strict best response to an opponent's strategy. Cooperation in this construction relies upon inconsistent expectations: Player 2 plays his/her cooperative strategy R when s/he thinks player 1's play reveals 1's type; player 1 plays the cooperative strategy D to exploit player 2's expectations, and R is *not* a best response to D. In fact this inconsistency is necessary since FKL showed that Nash equilibria are robust to the kind of uncertainty we consider, and clearly the Nash equilibria of the prisoner's dilemma involve finking along the equilibrium path.

To see in more detail how the hypothesis of Nash equilibrium yields finking in the perturbed game consider Fig. 5.2. While all four strategy combinations are rationalizable, the unique Nash equilibrium has player 1 playing U ("always fink if sane") with probability $(1-4\varepsilon)/(1+\varepsilon)$. Thus in equilibrium player 1 is almost certain to fink in period 1, and so player 2 is almost certain to fink in period 2. For a fixed $\varepsilon > 0$, Nash equilibrium *is* consistent with the sane types *sometimes* acting "crazy," but the total probability of crazy play is of order ε. In contrast, iterated weak dominance accepts and rejects pure strategies *without* describing their likelihood. Indeed, a key difference with equilibrium analysis is that these likelihoods are not objective quantities, and can differ for the two players. In an *a posteriori* equilibrium, when player 2 is told to play R s/he must believe that player 1 is unlikely to play D. Playing R is only optimal if the sane player 1 is unlikely to cooperate. Thus player 2 will always assess a small probability of actually cooperating in the second period. However, whenever player 1 plays D s/he assesses a high probability of player 2 cooperating in period 2. This inconsistency of beliefs *combined* with payoff uncertainty is what makes cooperation possible under $\mathbf{W}^\infty$.

	L	R
U	0,0	0,5ε
D	-1,2	1,1+4ε

FIGURE 5.2

The discussion above studied one particular "elaboration" of the prisoner's dilemma. Additional insight into the distinction between equilibrium and non-equilibrium analysis is provided by the general argument that the only Nash equilibrium path of *any* sufficiently close elaboration is "always cheat."

The proof proceeds as follows. Let the prior probability that each player is sane be $1-\varepsilon$ and let θ_i^c be the event player i is a crazy type who cooperates in period 3 iff j cooperates in period 2. Fix a Nash equilibrium, and use it to calculate a joint probability distribution π over types and over the outcome x_i in period i for $i=1, 2, 3$. Since "sane" players fink in period 3 along the equilibrium path, player j will fink in period 2 following any first period outcome x_1 with $\pi(x_1)>0$ *except* for x_1 which leads to a sufficiently large posterior probability $\pi(\theta_i^c \mid x_1)$ that player i is a crazy type who cooperates in period 3 iff j cooperates in period 2. We denote this critical probability by $\bar{\pi}$; in our example $\bar{\pi}=\frac{1}{2}$.

Now consider the probability of the event that player j's second period beliefs that player i is such a crazy type exceed $\bar{\pi}$. That is, determine $Q_1 \equiv \pi(\{x_1 : \pi(\theta_i^c \mid x_1) \geqslant \bar{\pi}\})$. Bayes' rule implies that $Q_1 \leqslant \varepsilon/\pi$, so $Q_1 \to 0$ as $\varepsilon \to 0$. This step would also be valid in an *a posteriori* equilibrium: Neither player thinks it is likely that s/he will cooperate in the second period. The distinction between Nash equilibrium and *a posteriori* equilibrium, i.e., the distinction between consistent and inconsistent strategic beliefs, comes about as we work backward to the first period. Player i will only cooperate in period 1 if this is sufficiently likely to induce cooperation in periods 2 and 3. Since period 1 is the first period, player i's beliefs are that player j is probably sane, so player i can only cooperate if s/he believes that by his/her doing so the sane type of player j is likely to cooperate. We have just seen that in a Nash equilibrium, the sane type of player j is unlikely (*ex ante*) to cooperate in period 2, so that when ε is sufficiently small the sane player i finks in the first period. On the other hand *ex post*—after an x_1 such that $\pi(\theta_i^c \mid x_1) > \bar{\pi}$—player j is likely to cooperate in period two (as in the elaboration above). Hence this cooperative strategy may be undominated. Furthermore, when the player's *strategic beliefs* are not consistent, player i can believe j will cooperate more often than j expects to. Indeed, player i can attach probability one to player j using any strategy not ruled out by dominance. Therefore in the case of inconsistent strategic beliefs player i can cooperate in period 1 with the intent to lead j to cooperate in period 2. Thus a crucial distinction between consistent and inconsistent beliefs for robustness arguments is one of how likely each player can believe the other is to change his play in response to the possibility of crazy types.

6. 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 an alternative, direct, formalization of 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 $q^i \in \Delta(\Omega^i)$. 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. By retaining this assumption extensive form refinements are led to an internal inconsistency: they impose restrictions based on a particular hypothesis of play even after that hypothesis has been contradicted (see Reny [21], Basu [3], Binmore [5, 6], and Bonanno [8] among others). By relaxing it in our model, players can always have an inference which is consistent with the theory. Formally, here we allow for Support $\text{marg}_{T^j} q^i(\cdot \mid H^i) \neq u^j$ if $q^i(H^i) = 0$ (where H^i is a cell in i's information partition). 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 Fig. 6.1, 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/her beliefs to assign 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

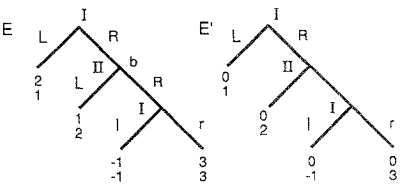


Figure 6.1

precisely to the ideas of updating beliefs. That is, it satisfies Support $\text{marg}_{T^j} q^i(\cdot \mid H^j) = u^j$ if H^j was assigned positive prior probability by q^i. Thus strong convergence seems more appropriate for modeling the idea that a player i may update his/her beliefs about an opponent's payoffs if and only if i observes an unexpected strategy choice by the opponent. This interpretation does not allow for the payoffs to be "almost" equal to u^j when the player is not surprised. 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.

In terms of the formal mathematical properties of the definitions the previous discussion points to an interesting distinction between strong and weak convergence, which is roughly 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. Analogously the definition of strong convergence specifies that an elaboration $\tilde{E}_n$ is close to a game E only if all types which receive positive probability according to E have *precisely* the same payoffs in $\tilde{E}_n$ as in E. On the other hand in perfect equilibrium, even at information sets along the equilibrium path the players allow for "trembles" in the opponents' strategies. This is analogous to weak convergence which allows for (small) payoff perturbations even for those types which receive positive probability in the limit. The notion of lexiographic beliefs[11] is useful for modelling such perturbations *within* the limit game. To see this consider Fig. 6.2. The elaborations $\tilde{E}_n$ in Fig. 6.2a converge weakly, but not strongly, to the game E in Fig. 6.2b. (Therefore $\mathbf{S}^\infty\mathbf{W}(E)$, whcih equals $\{L\} \times \{l\}$, does not include the limit of $\mathbf{W}^\infty(\tilde{E}_n)$, which equal $\{R, L\} \times \{r, l\}$. Of course $\overline{\mathbf{S}^\infty\mathbf{W}}$ does include the latter limit.) However, using lexicographic beliefs, the limit of $\tilde{E}_n$ is naturally defined to be $\hat{E}$ in Fig. 6.2c, where ε is a positive infinitesimal. Now $\mathbf{S}^\infty\mathbf{W}(\hat{E})$ *does* include the limit of $\mathbf{W}^\infty(\tilde{E}_n)$. The payoff perturbations in the elaborations are incorporated in $\hat{E}$, so that the closure of $\mathbf{S}^\infty\mathbf{W}(\hat{E})$ with respect to payoff perturbations is not required.

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. An interesting extension of this model involves imposing the restriction that a

[11] Refinements of Nash equilibrium such as perfect equilibrium are characterized using lexicographic beliefs in [7].

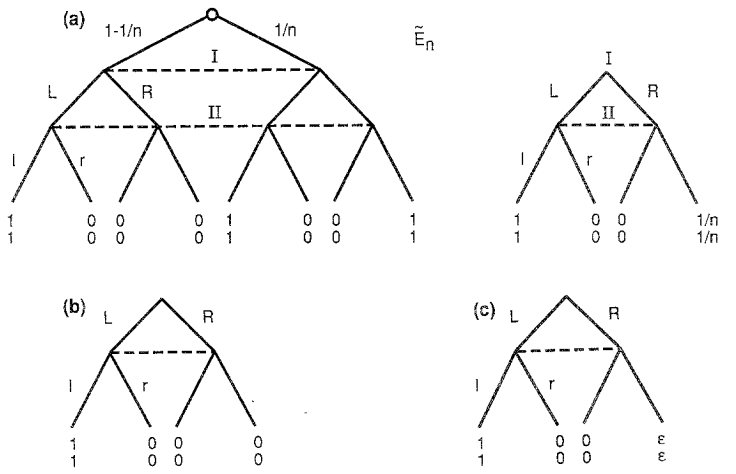
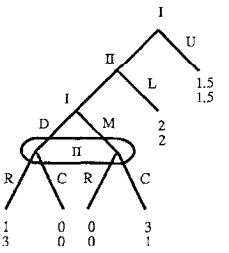


FIGURE 6.2

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. (One might even consider more precise orderings on what assumptions players revise when surprised, and thus attempt to characterize different refinements.) Thus the player may assume that his/her beliefs about the opponents' strategy choice (or the opponents' beliefs about other players' strategies, etc.) were wrong. Only if the "deviation" cannot be explained by questioning the players' beliefs over the elements of strategic uncertainty is the more basic assumption regarding common knowledge of payoffs doubted. Rabin's [20] idea of "focal rationalizability" can be interpreted as implementing this idea.

7. Concluding Example and Remarks

We conclude with an example, described in Fig. 7.1, that helps explain why we do not feel comfortable with a prediction based on $\mathbf{W}^\infty$. This game is similar to a modification in van Damme [27] of an example from Kohlberg and Mertens [15]. The unique strategy pair surviving $\mathbf{W}^\infty$ in this game is (U, R). However, we feel that (D, L) is reasonable. This is because, if player 2 accepts the $\mathbf{W}^\infty$ solution (which is based on the intuition of forwards induction) and then is given the opportunity to play, 2 must conclude that "something basic has changed," and 2 might conclude that 1's payoffs will lead 1 to violate the unique $\mathbf{W}^\infty$ outcome in the subgame.



		II L	II C	II R
	U	1.5,1.5	1.5,1.5	1.5,1.5
I	M	2,2	3,1	0,0
	D	2,2	0,0	1,3

FIGURE 7.1

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). This approach also yields a model which is more restrictive than rationalizability yet provides an explanation for being at *any* information set; moreover, the explanation does not contradict the model. In fact, it suggests how more restrictive theories can be developed while this form of internal consistency is retained. Finally, the distinction between weak and strong convergence is helpful in understanding the relationships between strict best replies and weak dominance, and between robustness and the updating of beliefs on null events.

REFERENCES

1. R. AUMANN, Subjectivity and correlation in randomized strategies, *J. Math. Econ.* **1** (1974), 67–96.
2. R. AUMANN, Correlated equilibrium as an expression of Bayesian rationality, *Econometrica* **55** (1987), 1–18.
3. K. BASU, Strategic irrationality in extensive games, mimeo, Princeton University, 1985.
4. D. BERNHEIM, Rationalizable strategic behavior, *Econometrica* **52** (1984), 1007–1028.
5. K. BINMORE, Modeling rational players, I, *J. Econ. Philosophy* **3** (1987), 179–214.
6. K. BINMORE, Modeling rational players, II, *J. Econ. Philosophy* **4** (1988), 9–55.

7. L. BLUME, A. BRANDENBURGER, AND E. DEKEL, Equilibrium refinements and lexicographic probabilities, *Econometrica*, forthcoming.
8. G. BONANNO, The logic of rational play in extensive games of perfect information, mimeo, University of California, Davis, 1988.
9. T. BÖRGERS, Bayesian optimization and Dominance in normal form games, mimeo, Institut für Volkswirtschaft, Universität Basel, 1989.
10. A. BRANDENBURGER AND E. DEKEL, Rationalizability and correlated equilibria, *Econometrica* **55** (1987), 1391–1402.
11. A. BRANDENBURGER AND E. DEKEL, Bayesian rationality in games, mimeo, Graduate School of Business, Stanford University, 1986.
12. D. FUDENBERG, D. KREPS, AND D. LEVINE, On the robustness of equilibrium refinements, *J. Econ. Theory* **44** (1988), 354–380.
13. D. GALE AND S. SHERMAN, Solutions of finite two-person games, *in* "Contributions to the Theory of Games" (H. Kuhn and A. Tucker, Eds.), Vol. I, Princeton Univ. Press, Princeton, NJ, 1950.
14. J. HARSANYI, Games of incomplete information played by Bayesian players, I, II, III, *Managem. Sci.* **14** (1967–1968), 159–182, 320–334, 486–502.
15. E. KOHLBERG AND J. F. MERTENS, On the strategic stability of equilibria, *Econometrica* **54** (1986), 1003–1038.
16. D. KREPS AND R. WILSON, Sequential equilibria, *Econometrica* **50** (1982), 863–894.
17. R. LUCE AND H. RAIFFA, "Games and Decisions," Dover, 1989.
18. D. MONDERER AND P. SAMET, Approximating common knowledge with common beliefs, *Games Econ. Behav.* **1** (1989), 170–190.
19. D. PEARCE, Rationalizable strategic behavior and the problem of perfection, *Econometrica* **52** (1984), 1029–1050.
20. M. RABIN, Consistency and robustness criteria for game theory, mimeo, Massachusetts Institute of Technology, 1988.
21. P. RENY, Rationality, common knowledge, and the Theory of games, mimeo, Princeton University, 1985.
22. R. W. ROSENTHAL, Games of perfect information, predatory pricing and the chain store paradox, *J. Econ. Theory* **25** (1981), 92–100.
23. R. SELTEN, Re-examination of the perfectness concept for equilibrium points in extensive games, *Int. J. Game Theory* **4** (1975), 25–55.
24. M. STINCHCOMBE, Approximate common knowledge, mimeo, Department of Economics, University of California, San Diego, 1988.
25. T. TAN AND S. WERLANG, The Bayesian foundations of solution concepts of games, *J. Econ. Theory* **45** (1988), 370–391.
26. E. VAN DAMME, "Refinements of the Nash Equilibrium Concept," Springer-Verlag, Berlin/New York, 1983.
27. E. VAN DAMME, Stable equilibria and forward induction, *J. Econ. Theory* **48** (1989), 476–498.