Reprinted from JOURNAL OF ECONOMIC THEORY
All Rights Reserved by Academic Press, New York and London

Vol. 44, No. 2, April 1988
Printed in Belgium

On the Robustness of Equilibrium Refinements*

DREW FUDENBERG

*Department of Economics, Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139*

DAVID M. KREPS

Graduate School of Business, Stanford University, Stanford, California 94305

AND

DAVID K. LEVINE

*Department of Economics, University of Minnesota, Minneapolis, Minnesota 55455
and Department of Economics, UCLA, Los Angeles, CA 90024*

Received May 5, 1986; revised March 31, 1987

The philosophy of equilibrium refinements is that the analyst, if he knows things about the structure of the game, can reject some Nash equilibria as unreasonable. The word "know" in the preceding sentence deserves special emphasis. If in a fixed game the analyst can reject a particular equilibrium outcome, but he cannot do so for games arbitrarily "close by," then he may have second thoughts about rejecting the outcome. We consider several notions of distance between games, and we characterize their implications for the robustness of equilibrium refinements. *Journal of Economic Literature* Classification Numbers: 026, 213. © 1988 Academic Press, Inc.

1. INTRODUCTION

Much effort has been devoted recently to refining the notion of a Nash equilibrium. Beginning with Selten's [11, 12] notions of perfection, concepts such as properness (Myerson [8]), sequentiality (Kreps and Wilson [7]), and stability (Kohlberg and Mertens [5]) are often invoked or discussed in the literature. The guiding philosophy of these refinements is that the analyst knows things about the structure of the game that enable him

* We thank Elon Kohlberg, Jean-Francois Mertens, Lloyd Shapley, two referees, and especially Eddie Dekel for helpful discussion and comments. Financial support from the National Science Foundation under Grants SES84-08586, SES85-09484, and SES85-09697 and from the Alfred P. Sloan Foundation is gratefully acknowledged.

0022-0531/88 $3.00

to reject some of the Nash equilibria as unreasonable. The point of this paper is that the word *know* in the preceding sentence deserves emphasis. Specifically, the analyst creates a model of the situation that is a simplification and (he hopes) an approximation. Suppose that, in the model, the analyst can reject a particular equilibrium outcome using the various refinements, but he cannot do so for models that are arbitrary "close" to the one created. Unless the analyst's faith in the model is absolute, he may be wise to have second thoughts about rejecting this outcome.

To study this "robustness" issue, we carry out the following program. Fix a space of games and a notion of "closeness" for the games in the space. For each game in the space, a Nash equilibrium is *strict* if, for each player, the strategy prescribed is a unique best response to the other players' strategies. This is as formidable a refinement criterion as we can think of, implying, for example, Kohlberg and Mertens' [5] *hyperstability*. Now ask: Which Nash equilibria of given game are limit points of strict equilibria for nearby games? We call such equilibria *near strict*. Following the discussion above, we would hesitate to reject any equilibrium that is near strict, insofar as the sense of closeness specified captures our doubts about the exact specification of the game.

In Section 2, we take for the space of games all normal form games over a fixed (finite player and action) normal form, and we measure "closeness" with the Euclidean metric on payoffs. The result is that *every* pure strategy Nash equilibrium is near strict. (We also give a weakening of the strictness requirement that accomodates every equilibrium, pure or mixed.) This is hardly surprising, but then the various refinements of Nash equilibria are motivated for the most part by the analysis of extensive games. In the rest of the paper, we turn to extensive games with the following philosophy: The analyst is certain (in his model) of the "physical" rules of the game: who moves when, with what information about earlier moves, and so on. That is, roughly, a (physical) extensive form is given. But the analyst may entertain doubts about the players' payoffs and/or their knowledge of the payoffs.

The perturbations that arise from such doubts are considered in the work on reputation in repeated games with incomplete information (see Kreps *et al.* [6] and Fudenberg and Maskin [3]). Our theory differs from this work on reputation effects in one important way. The previous work considers the effect of a fixed perturbation on a *family* of repeated games of varying lengths. The typical result there is that if the horizon is sufficiently long, even outcomes that are not Nash equilibria of the unperturbed game are Nash (and even near strict; see below) in the perturbed game. In this paper the game is fixed, and the perturbation is allowed to vary (and made to vanish). Hence only Nash equilibrium outcomes of the original game can be near strict.

Section 3 concerns the program for cases in which the analyst may be unsure of the players' payoffs himself but knows that these payoffs are common knowledge among the players. That is, an extensive form is fixed, the space of games is the space of all payoffs for the extensive form, and "closeness" is measured by the Euclidean metric on payoffs. We do not get very clean results here. There do exist near strict equilibria that are not strict, but the class of near strict equilibria has no apparent simple characterization.

Sections 4 and 5 are the heart of the paper. Here we imagine that the analyst has no doubts about the physical rules of the game, but is not quite certain of payoffs, and is willing to admit the possibility that the players themselves are not quite certain of each others' payoffs. Put another way, close to a given extensive game are games in which players entertain slight doubts about each others' payoffs, and our analyst is not prepared to reject an equilibrium that cannot be rejected in games that are nearby in this sense.

Section 4 deals with a definition of "closeness" under which every (pure) Nash equilibrium of the original game is near strict. We begin with a motivating example in Section 4.1. Technical details and the definition of closeness are given in Section 4.2. This version of closeness allows each player to be uncertain of his own payoffs and to believe that his opponents may (with small probability) have better information about them than he does. In this setting, an unexpected deviation by one player can signal that all players should change their play, as a small ex ante uncertainty can become large if an "unexpected" action is observed. In Section 4.3 we show that, with this notion of closeness, all pure strategy Nash equilibria are near strict. (Once again, mixed equilibria can be accommodated with a minor weakening of the definition.)

In the Section 5 we consider a second notion of closeness, which requires that each player's additional private information relate only to his own payoffs. If each player's additional information must also be independent of the information of the others, then an unexpected deviation by a player signals only that his *own* payoffs are different than had been anticipated. (In the literature on reputation in repeated games referenced above, the perturbations used satisfy this restriction.) With this more restrictive notion of closeness, not all (pure) Nash equilibria are near strict. However, all pure strategy equilibria that are strictly trembling hand perfect in the *normal* form are near strict, including some that are not subgame perfect. We obain a converse by weakening slightly the notion of trembling hand perfection. Finally, we consider relaxing the assumption that players' additional information must be independently distributed. Here we obtain a characterization of near strict equilibria that is like trembling hand perfection, but where the "trembles" of various players may be correlated.

The idea that small uncertainties of the kind we consider can render perfect some otherwise imperfect equilibria is not original to us. Van Damme [13] and Myerson [9], among others, make this point. Our contribution is to define and characterize more precisely the effects of various kinds of uncertainties corresponding to various measures of closeness.

2. Normal Form Games and Payoff Perturbations

Fix a finite player, finite action normal form. (We restrict attention throughout to games with finitely many players, each of whom possesses finitely many strategies.) Let $i = 1, ..., I$ index the players, and let $s_i \in S_i$ index the pure strategies of player i. Denote $\prod_{i=1}^{I} S_i$ by S. Let Γ be the space of games over this normal form: We take $\Gamma = R^{I \times S}$, where for $\gamma \in \Gamma$, $\gamma(i, s)$ is the payoff to player i under strategy s.

The set Γ comes endowed with a natural topology, namely the Euclidean topology. So does Σ_i, the space of mixed strategies for player i, and $\Sigma = \prod_{i=1}^{I} \Sigma_i$. (We will use the term *strategy profile* to refer to elements of Σ).

Two strategies for player i, σ_i and σ_i', are said to be *equivalent* if, no matter what strategies $\sigma_{-i} = (\sigma_1, ..., \sigma_{i-1}, \sigma_{i+1}, ..., \sigma_I)$ are chosen by i's opponents, all players receive the same utility whether i chooses σ_i or σ_i'.[1]

A Nash equilibrium $\sigma = (\sigma_1, ..., \sigma_I)$ for a given game γ is called *strict* if, for each player i, σ_i is i's *unique* best response to the strategies $\sigma_{-i} = (\sigma_1, ..., \sigma_{i-1}, \sigma_{i+1}, ..., \sigma_I)$ of the other players, up to equivalent strategies for i.[2] Note that only an equilibrium in pure strategies can be strict (unless the mixing is over equivalent strategies), according to this definition. Strict equilibria satisfy all the standard refinements; in particular, a strict equilibrium, taken as a singleton set, is *hyperstable* in the sense of Kohlberg and Mertens [5].[3]

Definition. A strategy profile $\sigma \in \Sigma$ is *near strict in the normal form* for a game γ if there is a sequence of games γ^n (over the same normal form

[1] Equivalent strategies arise naturally when the normal form game is constructed from an extensive form game, and one player has multiple information sets in sequence. Then if the player takes an action at an earlier information set that precludes reaching a later information set, the choice of the player at that later information set is irreleant. N.B., it is the player himself who precludes the later information set—equivalence of strategies does not depend on the strategy choices σ_{-i} of other players.

[2] That is, any equally good best response σ_i' must be equivalent to σ_i. We could equally well speak of strictness in the reduced normal form, in which all equivalent strategies are identified.

[3] Recall that Kohlberg and Mertens work with the reduced normal form. Conversely, any pure strategy equilibrium that is, as a singleton set, hyperstable, will be strict. Hence as long as we restrict attention to pure strategy equilibria, we could use "hyperstable as a singleton set" instead of "strict."

as γ) and a sequence of strategies $\{\sigma^n\}$ such that (i) $\lim_n \gamma^n = \gamma$; (ii) for each n, σ^n is a strict equilibrium of γ^n; and (iii) $\lim_n \sigma^n = \sigma$.

PROPOSITION 1. *A strategy combination $\sigma \in \Sigma$ is near strict in the normal form for γ if and only if it is a pure strategy Nash equilibrium for γ.*

The proof is quite simple. To see that any strategy profile that is near strict in the normal form for γ is a Nash equilibrium of γ, use the upper hemi-continuity of the Nash equilibrium correspondence. For the converse, consider the perturbation that adds a small amount to the players' utilities at the outcome prescribed by the strategy in question.

It is easy to deal with mixed strategies, if one permits a small extension to the definition of near strictness. Recall from Kohlberg and Mertens [5] that for any given normal form game, an *equivalent* normal form game is one in which pure strategies that are convex combinations of other pure strategies are added to or deleted from the original game. Corresponding to every strategy profile in an original game are (possibly many) strategy profiles for a given equivalent game. Consider the following modification of the definition of near strictness in the normal form.

DEFINITION. A strategy profile σ for a normal form game γ is *near strict in equivalent normal forms* for γ if it corresponds to some strategy profile σ' of an equivalent game γ' that is *near strict in the normal form* for γ'.

PROPOSITION 2. *A strategy profile σ is near strict in equivalent normal forms for γ if and only if it is a Nash equilibrium of γ.*

Again the proof is easy. A simple example will illustrate the method of proof. Consider the game depicted in Fig. 1 and the particular equilibrium in which player 2 randomizes equally between L and R. In Fig. 2 we have an equivalent game γ', with the particular equilibrium strategy for 2 added as a new strategy for player 2. In this game, this added strategy M corresponds to the mixed strategy of player 2 in the original game; the Nash equilibrium in the original game corresponds to a pure strategy equilibrium in the equivalent game, and we can apply the first proposition.

	2	
	L	R
1	0, 1	2,1

FIGURE 1

		2	
	L	M	R
1	0, 1	1,1	2,1

FIGURE 2

The spirit of this extended definition is that players may derive a little extra utility from playing a particular randomized strategy. This technique can be applied at any point in the development to follow, to extend results from pure to mixed equilibria. However, as both referees emphatically observed, this is an artificial device of little content.[4] Hence we will, in what follows, discuss only the near strictness of pure strategy profiles.

3. Extensive Games and Payoff Perturbations

The result that every pure Nash equilibrium of γ is near strict in the normal form (and, hence, is near hyperstable as a singleton set) is hardly surprising, since we allow for *any* sequence of (vanishing) payoff perturbations to the normal form game. Recall that, for most normal form games, *every* Nash equilibrium satisfies the standard refinements; the refinements were created initially to deal with games that arise from a specified extensive form. Hence we might wish to permit, as perturbations to an initial extensive game model, only perturbations that respect the structure of that extensive form.

We can imagine that our outside analyst is concerned with a game over an extensive form E, and he entertains no doubts about the physical rules of the game so specified. But still our analyst may have doubts about the full specification of the situation. He may, for example, be slightly unsure of the payoffs to the players or their probability assessments concerning nature's moves. We will suppose that there is no question about the probability assessments.[5] Thus we can ask: Fix an extensive form E and

[4] The referees suggested a somewhat more satisfactory treatment for mixed strategy equilibria, in which one considers as a perturbation of a given game the same game in which a player is "replaced" by a continuum of identical copies, as per Harsanyi's [4] classic treatment of purification. We will not follow that route here, except to suggest to the interested reader that, when one comes to general elaborations in Section 4, one will have to be careful with this if one is not to get as near strict all *correlated* equilibria, precisely in the spirit of Myerson [9].

[5] The visual accounting tricks make this without loss of generality, if there is no question about the support of those assessments.

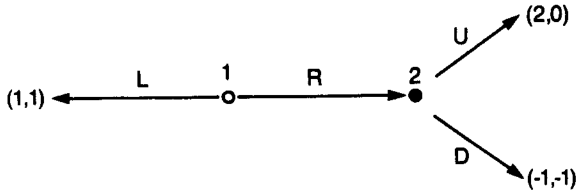

FIGURE 3

probability assessments for nature's moves in E. Let Γ be the space of all payoffs for players in E. For a given $\gamma \in \Gamma$, which strategy profiles are near strict in the following sense?

DEFINITION. Fixing $\gamma \in \Gamma$, a (pure) strategy profile σ of the game γ is *near strict in the extensive form* if there exist a sequence of payoffs $\gamma^n \rightarrow \gamma$ and (normal form) strict equilibria σ^n of γ^n such that $\sigma^n \rightarrow \sigma$.

We know from Section 2 that, for at least some games, there are equilibria that are near strict in the normal form but not strict themselves. Viewing a given normal form as a simple extensive form, we see then that there will be profiles that are near strict in the extensive form but that are not strict. It is easy to show that every strategy profile that is near strict in the extensive form is a pure strategy Nash equilibrium. But not every pure strategy Nash equilibrium over a given extensive form is near strict in the extensive form. Consider the extensive game in Fig. 3. (Our system for diagramming extensive form games is taken from Kreps and Wilson [7].) It should be easy to see that the Nash equilibrium (L, D) is not near strict in the extensive form; only (R, U) is.

Indeed, for some extensive games, *no* equilibrium is near strict in the extensive form. The game depicted in Fig. 4 is an example: While (R, Uu) and (R, Ud) (and, more generally, any randomization between these two strategies) are very nice equilibria, neither is near strict: In fact, for an equilibrium of an extensive game to be near strict in the normal form, it is

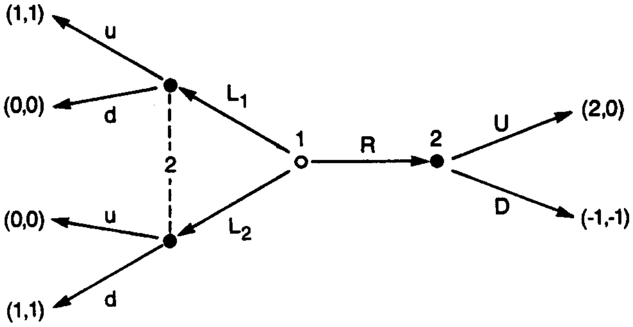

FIGURE 4

necessary that any information set h that is not reached with positive probability either (i) is precluded solely by the action at some earlier information set of the player whose move it is at h or (ii) is a "dummy" information set, where the player's choice is irrelevant to everyone's payoff. This is a very simple manifestation of the type of problem that leads Kohlberg and Mertens to define stable components of equilibria, and we could similarly attempt to define near strict components. However, we are more interested in the implications of the less restrictive notions of "closeness" that correspond to the players not being quite certain of each other's payoffs.

4. General Elaborations

4.1. *An Example*

To motivate this development, we provide a simple example. (A similar example is found in Van Damme [13].) Consider first the game in Fig. 3 and the equilibrium (L, D). Imagine an outside analyst whose thoughts about this situation run as follows:

> I *know* the particular physical rules of this game. Player 1 must first choose between L and R, and player 2 must choose between U and D if R is chosen. And I'm *fairly* sure that the payoffs are as in Fig. 3. But I'm not completely certain of this. I am willing to admit that there is a small chance that the payoffs are quite different from those shown. Moreover, it might not be common knowledge between the players what those payoffs are. That is, the game may have incomplete information, in that players do not know the precise payoffs, although all the players will have priors that give probability close to one for the payoffs in Fig. 3.

To model this, we introduce a game that is more elaborate than that in Fig. 3, as follows. Nature moves first, selecting one of several "versions" of the game. Each version is distinguished by having the same extensive form as in Fig. 3 but may have different payoffs. Nature chooses one version having payoffs close to those in Fig. 3 with probability close to one. Each player has a partition over versions of the game, with each respective player being told in which cell of his partition the true version lies.

One such elaboration is given in Fig. 5. Nature picks one of two versions. The first, which occurs with probability $1-\varepsilon$, has the same payoffs as in Fig. 3; the second, which occurs with probability ε, has very different payoffs. Player 1 knows which version nature picks, while player 2 is not told. Is this *elaboration* of the game in Fig. 3 very different from the original game for small ε? Our outside analyst, plagued by the sort of doubts expressed above, might not be willing to rule out the possibility that the players perceive the situation as being that in Fig. 5.

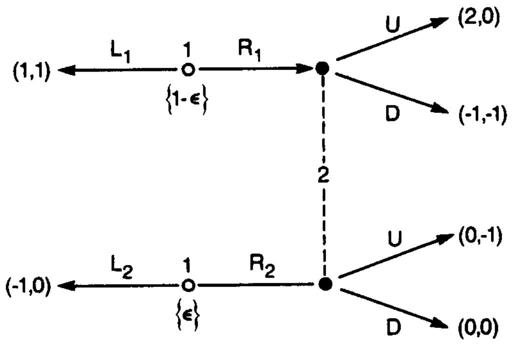

FIGURE 5

The point of this example is now easy to make. For the game in Fig. 5, the equilibrium (L_1R_2, D) is strict in the normal form, as shown in Fig. 6. Moreover, it remains strict as we can decrease ε towards zero. This should not be hard to see: Given that player 2 will play D, player 1's choices at his information sets are both strictly optimal choices. And, given the strategy of player 1, player 2 is given the move only if nature picks version 2, which renders strictly optimal the choice D by 2.

If we were to measure distance so that, as ε goes to zero, this game approaches the game in Fig. 3, then we would conclude, in the spirit of earlier discussion, that (L, D) in Fig. 3 is near strict. This is so even though, in Fig. 3, (L, D) is subgame imperfect. (Note that a nontrivial step is implicit here: In what sense does (L_1R_2, D) approach (L, D) as ε goes to zero? We give a formal criterion below.)

With this as a prelude, we now develop a general treatment.

4.2. *Elaboration Perturbations*

We begin with a precise definition concerning when one game is a small perturbation of another. We will not develop a formal topology on the

	U	D
L_1L_2	1-2ε , 1-ε	1-2ε , 1-ε
L_1R_2	1-ε , 1-2ε	1-ε , 1-ε
R_1L_2	2-3ε ,0	-1,-1+ε
R_1R_2	2-2ε ,-ε	-1+ε ,-1+ε

FIGURE 6

space of games, but instead we give a simple *sufficient* condition for games to be close.

Fix an I-player extensive game of perfect recall, E. This prescribes a game tree T (with nodes denoted by t) which is partitioned into sets of nodes T_i that "belong" to the various players, some of which are initial nodes $w \in W$, and terminal nodes $z \in Z$; an initial assessment ρ over W; information sets $h \in H$, with $H(t)$ denoting the information set containing the (nonterminal) node t; actions $a \in A(h)$ at each information set; and a payoff function $u: I \times Z \to R$ assigning utilities to all players.

The kind of perturbation of E that we have in mind is one in which one of N possible "versions" of the game above is selected by nature at the outset, where each version has the game tree of the game above and, except for initial uncertainty as to nature's choice of version, the same information structure. Versions are distinguished by the players' payoffs. Players are unsure (in a general sense) as to which version prevails. Such perturbations of E will be called *elaborations* of E.

To formalize this, imagine a game $\tilde{E}$ of perfect recall built up as follows. A positive integer N is given, together with a probability distribution μ on $\{1, ..., N\}$. In $\tilde{E}$, the game tree consists of an N-fold copy of T, or $T \times \{1,..., N\}$; we use (t, n) to denote the nth copy of the node t. If player i moves at (nonterminal) node t in E (if $t \in T_i$), then i moves at (t, n) for all n; i.e., $\tilde{T}_i = T_i \times \{1, ..., N\}$. The set of initial nodes consists of the N-fold copy of W, with the probability of initial node (w, n) given by $\rho(w)\,\mu(n)$. The utility to player i at terminal node (z, n) is denoted by $u(i, z, n)$. Finally, information sets are composed as follows. For each player i, a partition P_i of $\{1, 2, ..., N\}$ (with cells denoted by $P_i(n)$) is given, and the information set in $\tilde{E}$ with node $(t, n) \in \tilde{T}_i$ is $H(t) \times P_i(n)$. Actions at information sets are inherited in the obvious fashion.

The information structure bears some scrunity. Player i's exogenously given information concerning which version n is chosen at the outset is described by the partition P_i. That is, if it is player i's turn out to move at node (t, n), his knowledge about t is given by $H(t)$ and his knowledge about n is given by $P_i(n)$. Figures 3 and 5 illustrate the basic construction. The game $\tilde{E}$ in Fig. 5 consists of two version of E in Fig. 3. In the first (upper) version, the payoffs are identical with those in Fig. 3. In the second (bottom) version, they are quite different. Player 1 learns at the outset which version is chosen; i.e., P_1 is the discrete partition. Player 2 learns nothing; i.e., P_2 is the trivial partition.

We consider such elaborations $\tilde{E}$ of a game E as being among the possible perturbations of E. For $\tilde{E}$ to be a "small" perturbation, we take it to be sufficient that the payoffs in $\tilde{E}$ should be, with high probability, from some single version whose payoffs are approximately those in E. Formally, we pose the following criterion.

Convergence Criterion. For a given game E, let $\{\tilde{E}^k\}$ be a sequence of elaborations of E. To say that the sequence approaches E, it is sufficient that

(i) there is a uniform bound on the number of versions of the original game in each elaboration $\tilde{E}^k$ and a uniform bound on the absolute values of the u^k;

(ii) for each i, and z, $\lim_{k \to \infty} u^k(i, z, 1) = u(i, z)$; and

(iii) $\lim_{k \to \infty} \mu^k(1) = 1$.

Conditions (ii) and (iii) state that, along the sequence, the probability that nature picks a version in which payoffs are asymptotically the same as in the original game approaches one. Note that a convergent sequence of extensive form payoff perturbations of E correponds to a (trivial) convergent sequence of elaborations.[6] The first part of condition (i) is probably unnecessary to support a notion of closeness, but since we are giving a sufficient condition for convergence, it cannot hurt. The second part of (i) is quite important, however: Even if the probability of other versions is going to zero, if there is no uniform bound of the payoffs of those versions, then we can inflate the payoffs in other versions so that, in expected utility terms, the players' expected payoffs over all the versions are anything we wish them to be. With the uniform bound on payoffs, however, (iii) implies that in *ex ante* calculations of expected payoffs, only the first version (which gives nearly the payoffs of the original game) will loom large in the limit.

Is this a reasonable sufficient condition for closeness of games in the extensive form? The considerations put into the mouth of our outside analyst at the top of this section would argue that it is, although the reader may already see how broad a class of small perturbations this allows. Insofar as the refinements of Nash equilibrium deal with out-of-equilibrium behavior, a small probability *ex ante* of very different payoffs may loom quite large *ex post*. In what follows, we will see that this is so, and the reader should consider carefully whether the uncertainty of an outside analyst may be so great as this.

4.3. *Near Strictness under General Elaborations*

As in the previous two sections, we wish to identify (equilibrium) strategies for a given game E that are near strict, in this case under the

[6] Strengthening condition (ii) to require that payoffs in the first version exactly equal those in the original game is vacuous: Given a sequence of elaborations satisfying (ii), we can replace version 1 by two versions, called 1′ and 1″, with payoffs in 1′ exactly as in the original game, and specify that no player can distinguish between 1′ and 1″. With appropriate care in picking payoffs in version 1″ and the relative prior probabilities of the two, the new elaborations would be strategically equivalent to the old ones.

notion of closeness developed above. To do so, we must specify a mode of convergence of a sequence of strategies for elaborations to a strategy for the original game. This is not completely trivial, because strategies in an elaborated game are much richer than strategies in the original game. Since we take the point of view that the outside analyst is able to view only the actions that are in the physical game given by E (and not those actions as they are contingent on which version nature selected), one answer is to look at the (marginal) distribution induced by strategies on endpoints of the "physical" game. That is, one could compute the distributions induced on Z, and require that they converge. Note that even if all elaborations are trivial, in the sense that each has a single version of T, this gives us convergence of strategies only at information sets that are reached with positive probability. Hence this is weaker than the usual convergence of entire strategies. We therefore use a slightly sharper convergence criterion: We ask for convergence of behavior prescribed in an elaborated game at all information sets $\tilde{h}$ that contain the nodes $(t, 1)$.

DEFINITION. A strategy σ for the game E is *near strict under general elaborations* if there is a sequence of elaborations of E, $\{\tilde{E}^k\}$, that convergences to E in the sense of the convergence criterion given above, and a sequence of strict (normal form) equilibria $\{\tilde{\sigma}^k\}$ for the respective elaborations, such that the behavior prescribed by the $\tilde{\sigma}^k$ at nodes $(t, 1)$ in their respective games converges to behavior at node t prescribed by σ.

PROPOSITION 3. *A pure strategy profile σ for the game E is near strict under general elaborations if and only if it is a (pure strategy) Nash equilibrium in E.*

Proof. First we show that any strategy profile that is near strict under general elaborations is a Nash equilibrium. Suppose that some σ is near strict but it is not a Nash equilibrium. Then there is some player i, strategy $\hat{\sigma}_i$, and $\varepsilon > 0$ such that $u(i, (\hat{\sigma}_i, \sigma_{-i})) > u(i, \sigma) + \varepsilon$, where $u(i, \sigma)$ denotes the utility to player i in the game if the outcome is determined by the strategy profile σ, and $(\hat{\sigma}_i, \sigma_{-i})$ is the strategy profile composed of $\hat{\sigma}_i$ for i and σ_j for all $j \neq i$.

Fix a sequence of elaborations $\{\tilde{E}_k\}$ and corresponding strict equilibria $\{\tilde{\sigma}^k\}$ which converge to σ in the sense we are using here. Payoffs are uniformly bounded in all versions, payoffs in version 1 of elaboration k approach payoffs in the originally given game, and the prior probability of version 1 approaches 1. Therefore i's overall expected utility in elaboration k approaches $u(i, \sigma)$. Suppose that there were some $\hat{\sigma}_i$ as above. Construct $\tilde{\sigma}_i^k$ so that, in elaboration k, at information sets $(h, P_i(1))$ for $h \in H^i$, i follows the prescription of $\hat{\sigma}_i$, and otherwise i follows the prescription of $\tilde{\sigma}_i^k$.

As k grows, $u(i, (\hat{\sigma}_i^k))$ converges to $u(i, (\hat{\sigma}_i, \sigma_{-i}))$ so that for large enough k, $\hat{\sigma}_i^k$ would be better for i than $\tilde{\sigma}_i^k$ in elaboration k, a contradiction.

For the converse, we fix a (pure strategy) Nash equilibrium σ and construct an elaboration $\tilde{E}^k$ as follows.

For each player i and pure strategy s_i for i, design payoffs for i that make strategy s_i strictly dominant for player i throughout the course of play. To do this, at each terminal node $z \in T$, ask how many times $m_i(z, s_i)$ player i has to deviate from s_i if z is to be reached. (For example, when z lies along a path that passes through no information sets of i, $m_i(z, s_i) = 0$.) To give i the strict incentive to follow s_i at every opportunity, set i's payoffs at $-m_i(z, s_i)$.

Letting $\#S_i$ be the number of pure strategies for player i, in each elaboration there are $\prod_{i=1}^{I} (\#S_i + 1)$ versions of the original game. We label these versions by $(r_1, ..., r_I)$, where each r_i is drawn from the set $S_i \cup \{0\}$. In version $(r_1, ..., r_I)$ of elaboration k, the payoffs to player i are as follows:

(i) If $r_i \in S_i$, then player i is assigned the payoffs described above that make r_i a strictly dominant strategy.

(ii) If $r_i = 0$ and, for some $j \neq i$, $r_j \in S_j$, then player i is assigned payoffs that make playing σ_i a strictly dominant strategy.

(iii) If $r_i = 0$ and, moreover, $r_j = 0$ for all j, then player i is assigned utility equal to $u(i, \cdot) - m_i(\cdot, \sigma_i)/k$, where u is the originally specified utility function and m_i is the number-of-deviations function defined above.

The prior distribution on these versions is set as follows. Version $(r_1, ..., r_I)$ of elaboration k has prior probability $\prod_i \mu_i(r_i)$, where $\mu_i(r_i) = 1/(k \cdot \#S_i)$ if $r_i \in S_i$ and $\mu_i(r_i) = (k-1)/k$ if $r_i = 0$. Finally, for his initial information concerning the version selected, at version $(r_1, ..., r_I)$, player i is told the value of r_i.

This construction may seem quite complex, but it has a simple interpretation. Each player i is either "crazy" of "sane," and there are as many different ways for i to be crazy as i has pure strategies. The overall chance that i is crazy in elaboration k is $1/k$, divided equally among the many different forms of craziness. A player is told whether he is crazy or sane and, if crazy, the form that his craziness takes. Players' "types" are selected independently. (Note that the payoffs of one player will depend on the types of the others, so the payoff perturbations are not independent.) If a player is crazy according to a certain pure strategy, then the player has payoffs that cause him to play that strategy at every available opportunity.[7]

[7] This sort of construction, in which every strategy is played with positive probability, is used by Fudenberg and Maskin [3] to obtain a robust folk theorem for repeated games with long but finite horizons and small levels of incomplete information.

If a player is sane and some other player is crazy (of any type), then the sane player wants to follow the fixed Nash equilibrium strategy σ_i at every available opportunity. Finally, if all players are sane, then payoffs are as in the original game, with a small "kick" in favor of the fixed Nash equilibrium strategies σ_i.

It should be evident that this sequence of elaborations converges to the originally given game, according to the criterion we have given. We claim, moreover, that in each elaboration, the following strategies $\tilde{\sigma}_i$ are strict: Player i follows σ_i if his initial information is $r_i = 0$, and he follows r_i if his initial information is some $r_i \in S_i$. Once this claim is established, we will have the proposition.

Suppose that some (pure) strategy $\hat{\sigma}_i$ is one of i's best responses to the strategies $\{\tilde{\sigma}_j\}$ of his opponents (in some elaboration k). We wish to establish that $\hat{\sigma}_i$ is precisely $\tilde{\sigma}_i$ except possibly at information sets that i's own actions preclude. By the assumption of perfect recall, the information sets of i are ordered by precedence. Take any information set h for i that is earliest in terms of precedence among all of i's information sets in which $\hat{\sigma}_i$ is different from $\tilde{\sigma}_i$ and in which i's previous actions (which would be the same under $\hat{\sigma}_i$ and $\tilde{\sigma}_i$ because this is an earliest information set) do not themselves preclude h. If no information sets of i satisfy these conditions, we are done. There are three possibilities to consider:

(i) The information set h belongs to a crazy variety of i. Because i himself does not preclude h, and because every strategy profile of his opponents is possible under their strategies in $\tilde{\sigma}$ (conditional on whatever type is i), the information set h has positive prior probability. And at that information set, any action other than that prescribed by $\tilde{\sigma}_i$ does strictly worse than $\tilde{\sigma}_i$ (from then on). Hence there is a strategy for i strictly better than $\hat{\sigma}_i$, which contradicts the assumed optimality of $\hat{\sigma}_i$.

(ii) The information set h belongs to the sane variety of i, and it corresponds to an information set that is hit with positive probability under σ in the original game and equilibrium. Then under $\tilde{\sigma}$ it is hit with positive probability in one of two ways: Either we are in the version in which everyone is sane, in which case following $\tilde{\sigma}_i$ is a strict best response for i (recall the "kicker" in defining utilities for the all-sane version), or we are in a version in which someone else is crazy and i is sane, in which case following $\tilde{\sigma}_i$ is a strict best response for the rest of the game. In either case, there is a strategy for i strictly better than $\hat{\sigma}_i$, and again we have a contradiction.

(The kicker perturbation to utility in the all-sane version would be unnecessary as long as there is more than one player, since then we would have that following $\tilde{\sigma}_i$ is at least weakly best in the all-sane version, strictly best in every someone-crazy version, and the latter would have strictly positive probability. For one-player games, however, the latter has zero probability.)

(iii) The information set h belongs to the sane variety of i and corresponds to an information set that is off the equilibrium path in the original game under the strategy σ, and player i does not, by his actions, preclude h. Since there is positive prior probability of every other strategy combination by i's opponents, there is positive prior probability that h will be reached. Moreover, Bayes' rule forces i to conclude that some one or more of his opponents must be crazy at this point, at least insofar as they all follow their parts of $\tilde{\sigma}$. Hence it is strictly better for i to follow $\tilde{\sigma}_i$ at this point and henceforth, and we have the same contradiction as in the previous two steps. Q.E.D.

Proposition 3 shows that any Nash equilibrium can be "justified" if the players entertain the right kinds of doubts about each other's payoffs. Naturally, different forms of payoff uncertainty will justify different equilibria, because the way in which a player responds to an unexpected deviation by his opponents depends on the inferences that he draws from this deviation. However, all of the Nash equilibria are "robust" in the sense that, given a sequence of elaborations that approaches the original game and justifies a given equilibrium, we can consider *any* "small enough" further perturbations of the elaborated games and get the same conclusion. Loosely speaking, every (pure) Nash equilibrium is justified by an "open cone" of elaborations whose vertex is the original game.

5. Elaborations with Types

5.1. *Independent Types*

In terms of our outside analyst story, Section 4 says that unless one is certain that players will not draw unmodelled inferences about their own payoffs from the deviations of their opponents, one should not reject any (pure) Nash equilibrium. Suppose, though, that our outside analyst is prepared to assert that the model as written captures all of the information that players have about each other; he entertains (only) the possibility that each player has unmodelled private information about his own payoffs. Using language from the literature, the analyst entertains the possibility that each player is one of several *types* (specified by payoffs), and each player knows his own type. If it is also true that no player has information (beyond the information possessed by the analyst, which all players have) about the types of the others, then we will say that the types are *independent*. In this case the class of small perturbations that the analyst would

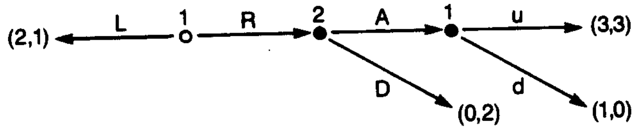

FIGURE 7

consider is smaller than in Section 4.2, and the set of near strict equilibria that results might be smaller as well.

(Note that in the general elaborations considered in Section 4, a player's payoffs were not wholely determined by the player's "type." In particular, in the proof of Proposition 3, the utility function of the "sane" type of player i depends crucially on whether any other player was "crazy." Hence we did not meet the requirement that each player knows his own payoffs; we will now require this. Initially, we also require that the types be independently distributed; that is, that the prior μ will have a product structure.)

To see the force of this restriction, consider the game in Fig. 3 and the Nash equilibrium (L, D). We made this equilibrium near strict by formulating, in Fig. 5, a game in which player 1 received information that was pertinent to player 2's payoffs. Hence, in this elaboration, player 1, by choosing R, was communicating to 2 that 2's best choice lay with D. (This, of course, supposes that 1 chooses L in the top version of the game.) But if we supposed that player 1 could not be given information about 2's payoffs superior to the information that 2 receives, this would be impossible, as player 2, moving in the information set that contains the node from the high probability version, would *know* that U is dominant for him.

This is not to say that "elaborations with independent types" (to be defined formally below) will not render near strict some equilibria that are not themselves strict. Consider the games in Figs. 7 and 8. In Fig. 7, we have an equilibrium (Lu, D) which is subgame imperfect. In Fig. 8 we have an elaboration of this game in which there is some uncertainty as to the payoffs to player 1 *only*. (There are two possible types of player 1). Player 1

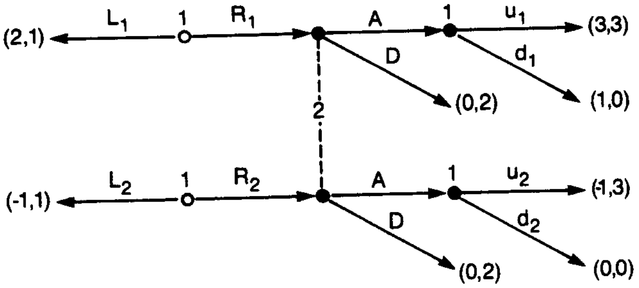

FIGURE 8

is informed about his payoffs; player 2 gets no information. In this game, $(L_1 R_2 u_1 d_2, D)$ is sequential; a more involved elaboration with independent types will make it strict, which renders (Lu, D) near strict under such elaborations.[8] The intuition is that in the alternative, bottom version of the game, player 1 plays R_2 with positive probability and will continue with d_2, because in the bottom version, $R_2 d_2$ is dominant. Player 2, then, given the move, assesses probability one (at the L_1 action equilibrium) that the bottom version is being played. As 1 will therefore continue with d_2 given the chance, 2 chooses D.

With this example to provide motivation, we turn to a formal development. First, we must modify the definition of an allowable elaboration.

DEFINITION. An elaboration is said to be an *elaboration with independent types* if

(i) for each player i there is an integer N_i such that the number of versions in the elaborations is $\prod_i N_i$;

(ii) writing $(j_1, ..., j_I)$ as the index of one of the elaborations, where $j_i \in \{1, ..., N_i\}$, there is for each i a probability distribution μ_i on $\{1, ..., N_i\}$ with $\mu((j_1, ..., j_I)) = \prod_i \mu_i(j_i)$;

(iii) there is for each i and $j \in \{1, ..., N_i\}$ a payoff function $v(i, z; j)$ such that for each i and version $(j_1, ..., j_I)$, $u(i, \cdot, (j_1, ..., j_I)) \equiv v(i, \cdot; j_i)$;

(iv) the information given to i if the true version is $(j_1, ..., j_I)$ concerning which version prevails is simply j_i.

The same convergence criterion as in Section 4.2 is used, although we now restrict to elaborations with independent types for perturbations.

DEFINITION. A strategy σ for an extensive game E is *near strict under elaborations with independent types* if there is a sequence of elaborations with independent types of E, $\{\tilde{E}^k\}$, that converges to E in the sense of the convergence criterion, and a sequence of strict equilibria $\{\tilde{\sigma}^k\}$ for the corresponding normal forms, such that the behavior prescribed by the $\tilde{\sigma}^k$ at nodes $(t, 1)$ converges to behavior at node t prescribed by σ.

To give our result, we need a further definition.

[8] In Fig. 8, $(L_1 R_2 u_1 d_2, D)$ is not strict because player 1's choice between u_2 and d_2 is irrelevant given player 2's choice of D. To make the choice matter, imagine that there is independent uncertainty about 2's payoffs, where, with small prior probability, 2 has A as a dominant strategy. Think of an elaboration with four versions, two (for each type of player 1) by two (for each type of player 2), where each player knows his own type but not the type of the other. If this is too sketchy, the general definition and construction to be given in a moment will clarify matters.

		2		
		A	D	
	Lu=Ld	2,1	2,1	$(\frac{n-2}{n})$
1	Ru	3,3	0,2	$(\frac{1}{n})$
	Rd	1,0	0,2	$(\frac{1}{n})$
		$(\frac{1}{n})$	$(\frac{n-1}{n})$	

FIGURE 9

DEFINITION. A (pure strategy) equilibrium σ of a normal form game γ is *strictly perfect in the normal form* if there is a sequence of totally mixed strategies $\sigma^k \rightarrow \sigma$ such that, for each player i and index k, σ_i is a *strict* best response (up to equivalent strategies for i), to the other players' strategies prescribed by σ^k.[9]

Remark. Figure 9 shows that (Lu, D) is strictly perfect in the normal form. The numbers in parentheses are "trembles" corresponding to a sequence of totally mixed strategies converging to (Lx, D) such that Lx and D are strict best responses for players 1 and 2, respectively, in the reduced normal form. The key is that player 1 trembles onto Rd as often as onto Ru, even though, having played R, player one "should" continue with u.[10]

PROPOSITION 4. *For a given extensive form E, every equilibrium σ that is strictly perfect in the normal form is near strict under elaborations with independent types.*

Proof. Fix a sequence of totally mixed strategies $\sigma^k \rightarrow \sigma$ such that each σ_i is a strict best response to each σ^k. In the kth elaboration, N_i is set equal to $\# S_i$, the number of pure strategies of player i. Payoffs for i in versions corresponding to s_i are as in the original game if $s_i = \sigma_i$, and they make s_i dominant otherwise. The marginal probability that type s_i for i is chosen is exactly that assigned to s_i in σ_i^k. Hence these independent elaborations do converge in the sense of the convergence criterion to the originally specified game.

[9] The term "strictly perfect" is used by Okada [10] to mean something quite different. In the same vein, we later use the term "quasi-perfect" in a sense different from its usage in van Damme [14].

[10] Recall that Selten introduced perfectness in the agent normal form precisely to rule out such correlated trembles.

By construction, it is a strict Nash equilibrium in $\tilde{E}^k$ for each player i, if sane, to play σ_i, and to play s_i when crazy of type s_i. This gives the result. Q.E.D.

Given our previous results, Proposition 4 should not be surprising. Imperfect equilibria that are perfect in the normal form correspond to players inferring that, if an opponent has deviated from equilibrium play once, he is likely to do so again. Such inferences can be justified if each player believes that with small probability his opponents' payoffs may be quite different from payoffs that have high prior probability.

The converse to Proposition 4 is, however, false: Imperfect equilibria can be near strict with respect to elaborations with independent types. The problem is that the perfect equilibrium correspondence is poorly behaved with respect to extensive form payoff perturbations. Consider the following two-person simultaneous move game. If both players choose L, both receive 1. If either choose R, then both receive zero. Only (L, L) is perfect, but (R, R) is near strict under payoff perturbations in the extensive form: Imagine that the payoff to (R, R) was ε for each player, holding the payoffs for (L, R) and (R, L) at zero, and let ε approach zero.

One way to obtain a partial converse to Proposition 4 is to assume that each player's "main type" (the type corresponding to the main version of the game) has exactly the payoffs given in the original game. That is, our outside analyst assesses small probability that he has the payoffs off by (possibly) a large amount, but he assesses high prior probability that he has precisely the correct payoffs, and each player knows this. This class of perturbations is not particularly interesting. Instead, we modify the definition of a perfect equilibrium by taking its closure with respect to payoff perturbations in the extensive form.

DEFINITION. Fix an extensive form E with payoff function u. A (pure strategy) equilibrium σ of the corresponding normal form game is *quasi-perfect* if there is a sequence of payoffs for the extensive form, $\{u^m\}$, with limit u, such that σ is strictly perfect in each of the normal forms corresponding to the u^m.[11]

Remark. Every perfect (pure strategy) equilibrium is quasi-perfect.

PROPOSITION 5. *For a given extensive form game E, a pure strategy equilibrium σ is near strict under elaborations with independent types if and only if it quasi-perfect.*

[11] Eddie Dekel has pointed out that this definition of quasi-perfection is equivalent to a definition in which σ need be only (weakly) perfect in each of the normal forms corresponding to the u^m.

Proof.[12] Suppose that σ is quasi-perfect. Let u be the payoff function for E, and let $u^m \rightarrow u$ be the extensive form payoffs such that σ is strictly perfect in the normal forms corresponding to the u^m. Write E^m for the extensive form game with payoffs u^m. From Proposition 4, there are elaborations with independent types $\tilde{E}^{m,k}$ of E^m and strategy profiles $\tilde{\sigma}^{m,k}$ for those elaborations are such that $\tilde{\sigma}^{m,k}$ are strict equilibria in the corresponding normal forms, and $\tilde{\sigma}^{m,k} \rightarrow \sigma$ in k. Examination of the construction in Proposition 4 shows that we can assume that the number of versions and size of payoffs in the elaborations $\tilde{E}^{m,k}$ are uniformly bounded in m and k. For each positive integer N, take m_M large enough so that u^{m_N} is within $1/2N$ of u, and then take k_N large enough so that (i) $k_N > N$; (ii) the payoffs in version 1 of $\tilde{E}^{m_N,k_N}$ are within $1/2N$ of u^{m_N}; and (iii) the first version has probability $(N-1)/N$ or more in this elaboration. It is immediate that $\{\tilde{E}^{m_N,k_N}\}$ is a sequence of elaborations with independent types approaching (in N) E, with $\tilde{\sigma}^{m_N,k_N}$ a sequence of strict equilibria for those elaborations that approaches σ. Hence σ is near strict.

To establish the converse, we begin with a piece of notation and a preliminary observation. Fix an extensive game E, and let $\tilde{E}$ be an elaboration (with independent types) of E. If $\tilde{\sigma}_i$ is a strategy for i in $\tilde{E}$, we will write $\mathbf{M}\tilde{\sigma}_i$ for the marginal distribution on strategies of i in E induced by $\tilde{\sigma}_i$. That is, averaging over all versions of the original game in $\tilde{E}$ according to μ (or averaging over all of i's personal types according to μ_i, which gives the same result because types are independent) we compute the probability with which i will be "observed" to play various strategies in E.

Now suppose that we have an elaboration $\tilde{E}$ and a pure strategy profile $\tilde{\sigma}$ which is a strict equilibrium for the normal form of $\tilde{E}$. We claim that it is without loss of generality to assume that $\mathbf{M}\tilde{\sigma}_i$ is totally mixed, for each i. This is so because we can append to each player's list of types a further type for each pure strategy of i in E, where in the extra type for strategy s_i, i has s_i as a dominant strategy. Because $\tilde{\sigma}_i$ is a strict best response for i in $\tilde{\sigma}$, we can, in the augmented elaboration (with the extra types) choose the prior probability of those extra types to be sufficiently small (but still strictly positive) so that the composite strategy for each player i, "Follow $\tilde{\sigma}_i$ for types from the original elaboration, and follow the dominant strategy for the appended types," is a strict best response to the composite strategies for the others. (The constraint on the size of the probability of i's extra types depends on the extent to which the *other* players' strategies were strict best responses.)

Suppose then that σ is near strict for E, and fix a sequence of elaborations (with independent types) $\tilde{E}^k \rightarrow E$ and corresponding strict

[12] We thank Eddie Dekel for finding an error in an earlier version of this proof.

equilibria $\tilde{\sigma}^k \to \sigma$. By the observation just above, we can assume that $\mathbf{M}\tilde{\sigma}_i^k$ is totally mixed.

For each player i (holding σ fixed), partition the terminal nodes of the original game tree into two sets: Let A_i be those nodes which, for some combination of strategies by i's opponents, can be reached when i uses σ_i, and let B_i be the complementary set. Using $v^k(i, z; 1)$ to denote the payoffs to player i in the main version of the game in elaboration $\tilde{E}^k$ at node z, consider the payoffs $w^k(i, z)$ for the original extensive form given by

$$w^k(i, z) = \sup_{j \geqslant k} v^j(i, z; 1), \qquad \text{for} \quad z \in A_i,$$

and

$$w^k(i, z) = \inf_{j \geqslant k} v^j(i, z; 1), \qquad \text{for} \quad z \in B_i.$$

We assert that σ is strictly perfect in the game with payoffs given by w^k, against the test sequence $\mathbf{M}\tilde{\sigma}^j$.

To see this, we compare the difference to i between playing σ_i and some other (pure) strategy s_i against $\mathbf{M}\tilde{\sigma}^j_{-i}$ when payoffs are given by w^k with the difference when the payoffs are $v^j(\cdot, \cdot; 1)$, for $j \geqslant k$. Write $\Pi(z)$ for the probability that terminal node z is reached when strategies are $(\sigma_i, \mathbf{M}\tilde{\sigma}^j_{-i})$, and write $\Pi'(z)$ for this probability when the strategies are $(s_i, \mathbf{M}\tilde{\sigma}^j_{-i})$. We wish, then, to sign

$$\Delta = \sum_z (\Pi(z) - \Pi'(z))\, w^k(i, z) - \sum_z (\Pi(z) - \Pi'(z))\, v^j(i, z; 1).$$

Partition the terminal nodes z into four classes: (1) nodes that are reached with positive probability under both σ_i and s_i; (2) nodes such that $\Pi(z) = \Pi'(z) = 0$; (3) nodes with $\Pi(z) > 0 = \Pi'(z)$; and (4) nodes with $\Pi'(z) > 0 = \Pi(z)$. Note that classes (1) and (3) make up A_i, and classes (2) and (4) make up B_i. Now since σ_i and s_i are pure strategies, $\Pi(z) = \Pi'(z)$ in class (1). Thus the terms making up Δ for z in class (1) net zero. For z in class (2), the terms *are* zero. Thus we can rewrite

$$\Delta = \sum_{z \in \text{Class}(3)} \Pi(z)(w^k(i, z) - v^j(i, z; 1)) - \sum_{z \in \text{Class}(4)} \Pi'(z)(w^k(i, z) - v^j(i, z; 1)),$$

where we have used the facts that $\Pi'(z) = 0$ for z in class (3) and $\Pi(z) = 0$ for z in class (4). Since z in class (3) are in A_i, $w^k \geqslant v^j$ there, and conversely for class (4). We conclude, then, that Δ is nonnegative.

But σ_i is a strict best response to $\mathbf{M}\tilde{\sigma}^j_{-i}$ when payoffs are given by v^j. The nonnegativity of Δ shows that the same is true when payoffs are given by w^k. Thus σ is strictly perfect when payoffs are given by w^k. And, since $v^j(\cdot, \cdot; 1)$ converges to u, $w^k \to u$. Thus σ is quasi-perfect. Q.E.D.

5.2. *Personal Types*

The effect of perturbations with independent types is essentially (i) to allow for extensive form payoff perturbations and (ii) to allow correlation between the "trembles" at different information sets of the *same* player. We can, and now will, loosen the definition of elaborations with independent types to allow for the trembles to be correlated across different players.

An elaboration has *personal types* if each player's payoffs depend only on his own information about nature's choice of version; the players' information need not be independent. Formally, an elaboration with personal types must satisfy (i), (iii), and (iv) of the definition of an elaboration with independent types, but the overall measure μ on versions need not have the product form.

Consider the game in Fig. 10.[13] In the subgame where players 1 and 3 have played R and r, players 1, 2, and 3 play a $2 \times 2 \times 2$ simultaneous move game. Against any pair of independent strategies by 1 and 2, player 3 can guarantee himself a payoff of at least 0 in this subgame by playing a 50–50 mixture of A and B, so d is never part of any (normal form) perfect equilibrium of the game. Hence in any normal form perfect equilibrium of the game, 1 and 3 choose R and r, respectively. The subgame must lie along the equilibrium path in any perfect equilibrium, and in the subgame 1 and 2 have unique equilibrium strategies. (They play matching pennies no matter what 3 does.) Hence 3's best choice is A. That is, this game has a unique (normal form) perfect equilibrium, namely $(R, r, A, 0.5H + 0.5T, 0.5h + 0.5t)$. Perturbing the extensive form payoffs will not affect this; hence this is the unique equilibrium that is strict under elaborations with independent types.

But now consider elaborations with personal types. In particular, imagine that there are two other types of players 1 and 2. In the first other types of player 1, RH is a dominant strategy. (We refer to this as type RH.) In the second other type of player 1, RT is dominant. For player 2, there is a type for which h is dominant and a type for which t is dominant. Each player learns only his own type, so, in particular, player 3 learns nothing at all. And, what is crucial to our construction, the probability measure μ is such that there is high probability, say $1 - 2\delta$, that all players are of the "standard" types, probability δ that 1 is type RH and 2 is type h, and probability δ that 1 is type RT and 2 is type t. (If the reader does not like this perfect correlation in the types of 1 and 2, it will suffice to have high but less than perfect correlation.) Now consider the Nash equilibrium outcome $(L, d, A, 0.5H + 0.5T, 0.5h + 0.5t)$ in the original game. It is Nash because d by 3 engenders L by 1, and 3 can do anything he wishes, as L by

[13] This game is a variation on a similar game suggested to us by J.-F. Mertens. Myerson [9] gives another example, making the same point.

1 puts him off the equilibrium path.[14] This equilibrium is made near strict if we adopt the personal type elaborations above, for δ approaching zero. The intuitive argument runs as follows: Since 1 plays L in the main version, the choice of R signals to 3 that some other version has been selected. Given the prior μ we posited, 3 sees it as equally likely that (i) 1 will follow with H and 2 with h, and (ii) 1 will follow with T and 2 with t. When 3 assesses this correlated continuation by 1 and 2, neither A nor B will guarantee him a payoff above -0.5, and so 3 optimally chooses d. (Of course, this choice by 3 validates 1's choice of D in the main version.)

We advance this example and the general definitions to follow in the following spirit. General elaborations may seem to the reader to be too large a class of perturbation, because they allow one player to know more about a second player's payoffs than that second player knows herself. By constraining ourselves to elaborations with the "type" structure, we suppose that each player has all the information about his *own* payoffs that any other player has. The further restiction to independent types implies that no player learns anything about others' types from his own. One might imagine that players have "links" in their past—perhaps they went to the same school, or had similar teachers when they learned game theory, or whatever. Knowing that these links are possible, one suspects that knowing something about the payoffs of one player might tell you something about payoffs of another. This is true, *pari pasu*, of players within a game, which leads to personal, but not independent, types.

Although it does not lead to quite the same thing, we can give another perspective on this sort of elaboration. By analyzing the game noncooperatively and by restricting to Nash equilibria, one is supposing that players cannot access correlating devices, as defined in Aumann [1] (see also Forges [2] and Myerson [9]). Suppose that a more accurate statement of the outside analyst's state of knowledge is that while he is fairly certain that players do not have access to correlating devices, he is not completely certain of this. Then an appropriate perturbations of the game would be an elaboration with personal types, where the utility functions of the players in each version are close to (or perhaps even identical with) their payoffs in the game modelled. (In the example given in Fig. 10, even if players 1 and 2 had access to a correlating device, the fact that they play matching pennies in the subgame means that the device will be of no use to them. But if we change the subgame so that it is a game of coordination, then they might well use a correlating device if one is available. And if player 3 is unable to observe the outcome generated by

[14] Readers may fill in subsequent moves in any way they wish, and we will still have a Nash equilibrium. We have filled them in with the subgame perfect equilibrium strategies for a reason that will become apparent momentarily.

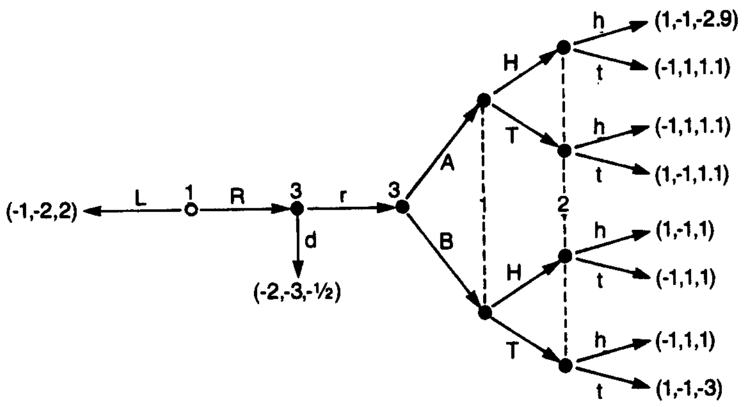

FIGURE 10

the correlating device, then he might choose d at his information set, afraid that 1's choice of R indicates that 1 and 2 have obtained access to the correlating device and will play a correlated equilibrium.) The point is that a low prior probability that players can access correlating devices may become a high posterior probability, if one observes an out-of-equilibrium action from one or more of the players.

To characterize the set of near strict equilibria under elaborations with personal types, we introduce the following definition, which may be of independent interest. The idea is to modify trembling hand perfection, to allow for "correlated trembles."

DEFINITIONS. Fix a pure strategy profile σ for a normal form game. A sequence $\{\{\phi_i^k\}_{i=1}^I\}_{k=1}^\infty$, where for each i and k, ϕ_i^k is a totally mixed distribution over the pure strategy profiles of players other than i, is a *c-perfect test sequence* for σ, if for each i, $\phi_i^k(\sigma) \to 1$. A pure strategy profile σ of a normal form game γ is *c-perfect* if there exists a c-perfect test sequence $\{\{\phi_i^k\}\}$ for σ such that for each player i and index k, σ_i is a best response to ϕ_i^k. The equilibrium is *strictly c-perfect* if each σ_i is a strict best response in the normal form.

Remarks. Note that the distributions ϕ_i^k need not correspond to independently mixed strategies by the players. Also, perfection and c-perfection are identical for the case of two-player games.

PROPOSITION 6. *If an equilibrium σ is strictly c-perfect, then it is near strict with respect to elaborations in personal types.*

Proof. Let ϕ_i^k be the c-perfect test sequence that makes σ strictly c-perfect. We construct elaborations (for $k = 1, 2, \ldots$) that are composed of two kinds of versions of the original game: There will be a "main" version of the game in which players are all sane, having the payoffs given in the original game. And there will be one version of the game for every pair (i, σ'_{-i}), where i runs over all the players and σ'_{-i} over the pure strategy combinations for the other players. In the (i, σ'_{-i}) version, player i will have the payoffs given in the original game, and the other players will have payoffs that make the constituent parts of σ'_{-i} dominant as in our previous constructions. In each elaboration, player i is told whether he is "sane" or not and, if he is not, what his payoffs are. This then meets the requirements for an elaboration in private types.

To specify the kth elaboration, it remains to specify the probabilities of the various versions. We will assign probabilities so that, in elaboration k, the probability of the main version is $1-n\varepsilon_k$ (for some $\varepsilon_k > 0$ to be determined in a bit) and the marginal probability of all versions in which i is sane and the others not will be ε_k. More precisely, we will use $p^k(i, \sigma'_{-i})\, \varepsilon_k$ to denote the probability of the (i, σ'_{-i}) version; that is, $p^k(i, \sigma'_{-i})$ is the conditional probability of crazy types σ'_{-i}, conditional on i being sane and the others crazy.

Suppose that, in elaboration k, players all play according to σ when they are sane and according to their dominant strategies when not. Then conditional on i being sane, the probability that i assess that this opponents will play σ'_{-i} is

$$\frac{(1-n\varepsilon_k)\, \mathbf{I}_{\sigma'_{-i}=\sigma_{-i}} + \varepsilon_k\, p^k(i, \sigma'_{-i})}{1-(n-1)\,\varepsilon_k}$$

where $\mathbf{I}$ denotes the indicator function. The numerator in this expression is the joint probability that everyone is sane and others play σ'_{-i} (which is nonzero only if $\sigma'_{-i} = \sigma_{-i}$), plus the probability that only player i is sane and the others play σ'_{-i}; the denominator is the marginal probability that i is sane. We will want this to equal $\phi_i^k(\sigma'_{-i})$, which, solving for $p^k(i, \sigma'_{-i})$, yields

$$p^k(i, \sigma_{-i}) = \frac{\phi_i^k(\sigma_{-i}) - (1-n\varepsilon_k)/(1-(n-1)\,\varepsilon_k)}{\varepsilon_k/(1-(n-1)\,\varepsilon_k)}$$

and, for $\sigma'_{-i} \neq \sigma_{-i}$,

$$p^k(i, \sigma'_{-i}) = \frac{\phi_i^k(\sigma'_{-i})}{\varepsilon_k/(1-(n-1)\,\varepsilon_k)}.$$

To ensure that $p^k(i, \sigma_i) \geqslant 0$, we must have that

$$\phi_i^k(\sigma_{-i}) \geqslant \frac{1-n\varepsilon_k}{1-(n-1)\,\varepsilon_k}.$$

That is, we must have

$$\varepsilon_k \leqslant \frac{1-\phi_i^k(\sigma_{-i})}{n-(n-1)\,\phi_i^k(\sigma_{-i})}.$$

So if we choose $\varepsilon_k > 0$ to be less than or equal to the minimum (over i) of the terms in the right-hand side of the previous inequality, and use ϕ^k to specify the p^k as above, the elaboration is well defined. By construction, we see immediately that σ is strict in each elaboration k. And since ϕ_i^k is a c-perfect test sequence, $\phi_i^k(\sigma_{-i}) \to 1$, implying that $\varepsilon_k \to 0$, so the elaborations converge to the original game. Q.E.D.

Finally, we have a converse to Proposition 6, exactly parallel to our converse to Proposition 4.

DEFINITION. An equilibrium σ of an extensive form game E with payoffs u is *quasi-c-perfect* if there exists a sequence of payoffs $u^k \to u$ such that σ is strictly c-perfect in the corresponding normal form games.

PROPOSITION 7. *An equilibrium σ of an extensive form game is quasi-c-perfect if and only if it is near strict in elaborations with independent types.*

The proof is just as in the proof of Proposition 5, except that $\mathbf{M}\tilde{\sigma}^k_{-i}$ is replaced, for each i, by the conditional distribution of the others, conditional on i being the first type.

REFERENCES

1. R. AUMANN, Subjectivity and correlation in randomized strategies, *J. Math. Econ.* **1** (1975), 67–96.
2. F. FORGES, An approach to communication equilibrium, *Econometrica* **54** (1986), 1375–1386.
3. D. FUDENBERG AND E. MASKIN, Folk theorems for repeated games with discounting or with incomplete information, *Econometrica* **54** (1986), 533–554.
4. J. HARSANYI, Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points, *Int. J. Game Theory* **2** (1973), 1–23.
5. E. KOHLBERG AND J.-F. MERTENS, On the strategic stability of equilibria, *Econometrica* **54** (1986), 1003–1038.
6. D. KREPS, P. MILGROM, J. ROBERTS, AND R. WILSON, Rational cooperation in the finitely-repeated prisoners' dilemma, *J. Econ. Theory* **27** (1982), 245–252.
7. D. KREPS AND R. WILSON, Sequential equilibria, *Econometrica* **50** (1982), 863–894.
8. R. MYERSON, Refinements of the Nash equilibrium concept, *Int. J. Game Theory* **7** (1978), 73–80.
9. R. MYERSON, Multistage games with communication, *Econometrica* **54** (1986), 323–358.
10. OKADA, On the stability of perfect equilibrium points, *Int. J. Game Theory* **10** (1981), 67–73.

11. R. Selten, Spieltheoretische Behandlung eines Oligopmodells mit nachfragetragheit, *Z. Ges. Staatswiss.* **121** (1965), 301–324.
12. R. Selten, Re-examination of the perfectness concept for equilibrium points in extensive games, *Int. J. Game Theory* **4** (1975), 25–55.
13. E. van Damme, "Refinements of the Nash Equilibrium Concept," Springer-Verlag, Berlin/Heidelberg/New York, 1983.
14. E. van Damme, A relation between perfect equilibria in extensive form games and proper equilibria in normal form games, *Int. J. Game Theory* **13** (1984), 1–13.

Printed by Catherine Press, Ltd., Tempelhof 41, B-8000 Brugge, Belgium