THE

QUARTERLY JOURNAL OF ECONOMICS

Vol. CII May 1987 Issue 2

SIGNALING GAMES AND STABLE EQUILIBRIA*

IN-KOO CHO AND DAVID M. KREPS

Games in which one party conveys private information to a second through messages typically admit large numbers of sequential equilibria, as the second party may entertain a wealth of beliefs in response to out-of-equilibrium messages. By restricting those out-of-equilibrium beliefs, one can sometimes eliminate many unintuitive equilibria. We present a number of formal restrictions of this sort, investigate their behavior in specific examples, and relate these restrictions to Kohlberg and Mertens' notion of *stability*.

I. INTRODUCTION

Much of information economics has been concerned with situations in which the following simple *signaling game* is embeded: one party, hereafter called party A, possesses private information. On the basis of this information, A sends a signal to a second party B, who thereupon takes an action. Examples abound: Spence's [1974] model of job market signaling is one example, if we modify things slightly so there are two or more parties B. (We shall develop a simple case of the Spence model in this format in Section V.) Grossman [1981] examines the role of warranties and product quality using this sort of model. Models of bargaining with incomplete information (see, for example, Grossman and Perry [1986a] or

*We are grateful to Anat Admati, Drew Fudenberg, Elon Kohlberg, Paul Milgrom, Richard McKelvey, Jean-Fra^cois Mertens, Motty Perry, John Roberts, Joel Sobel, Gyu Ho Wang, and especially Hugo Sonnenschein for helpful discussion, and to three referees and an editor for helpful suggestions. The financial support of Harvard University, the Korea Foundation for Advanced Studies, the National Science Foundation (Grants SES80-06407 and SES84-05865), the Sloan Foundation, and the Institute for Advanced Studies at the Hebrew University, are all gratefully acknowledged. The material in this paper originally appeared in two separate papers, one with the above title, and a second entitled "More Signaling Games and Stable Equilibria." We hope that anachronistic references to the earlier incarnations of these ideas will not prove too troublesome to the reader.

Rubinstein [1985]) constitute another class of examples. In the literature of industrial organization, there is the entry-deterrence limit pricing model of Milgrom and Roberts [1982a], the analyses of the chain-store game of Kreps and Wilson [1982b] and Milgrom and Roberts [1982b], and recent work on the role of advertising by Milgrom and Roberts [1986]. Theoretical accountants often employ this sort of model (see, for example, Demski and Sappington [1986]). And, on a slightly higher plane, there is the general analysis of a game of this sort due to Crawford and Sobel [1982], and related work on mechanism design by an informed principal [Myerson, 1983]. This is only a partial list (and apologies are tendered to those left off), and in most cases the models are variations on the general theme outlined above. But this theme, together with variations, has been played a lot recently.

In most of these recitals, one finds a plethora of equilibria. This paper takes a noncooperative game-theoretic approach, and we mean here a plethora of Nash equilibria. One can cut back on the number of equilibria by invoking notions of perfection (or sequentiality), but this is only of minor help—in many games the wealth of off-the-equilibrium path beliefs that can be imposed gives rise to a wealth of equilibria. That is, what constitutes an equilibrium is powerfully affected by the "interpretations" that would be given by B to messages that A *might* have sent, but in equilibrium *does not* send. In a sequential equilibrium, B is required to frame some hypothesis (probability assessment) over what is A's private information and respond accordingly. As one varies those hypotheses, one varies the optimal responses of B, and hence the incentives of A to send the various messages.

At this point, in many of these contextually based analyses, the analyst(s) resorts to various intuitive criteria based on the conclusions that B "ought" to draw from sundry out-of-equilibrium messages. If one can restrict the out-of-equilibrium beliefs (or hypotheses) of B, one can sometimes eliminate many of the equilibria. An example of this that is particularly prevalent runs as follows: suppose, for simplicity, that A's private information must be one of two things, called t and t'. Suppose that in equilibrium A sends message m with probability one. Suppose that there is a second possible message m' with the following properties: if A knows t, then A would strictly prefer (in comparison with the equilibrium outcome) not to send m', no matter how B interprets this. And if A knows t', then A would prefer to send m' to what A gets in the equilibrium *if by sending m' A could convince B that A knew t'.*

The former condition, it is argued, implies that B should not entertain the hypothesis that the message did come from an A who knows t. B should infer from the message that A knows t'. And, therefore, if A knows t', he should send the message (thus upsetting the given equilibrium). It is as if A, if he knows t', is (by sending m') implicitly making the speech:

> I am sending the message m', which ought to convince you that I know t'. For I would never wish to send m' if I know t, while if I know t', and if sending this message so convinces you, then, as you can see, it is in my interest to send it.

This particular criterion, and minor extensions to it, have appeared in several of the applications described above. It has been applied directly in Grossman [1981], Milgrom and Roberts [1982a], and Kreps and Wilson [1982b], and it is applied indirectly in Rubinstein [1985]. Its powers can be considerable: in a simple (two type) Spence signaling model, there is a single equilibrium outcome that survives this criterion (see Section V).

While analyses of particular examples have been based on intuitive criteria for out-of-equilibrium beliefs such as the one just given, there have been, at the same time, further attempts to refine generally the notion of a Nash equilibrium. An important recent example of this is the Kohlberg-Mertens [1986] theory of *stability* and *stable equilibrium outcomes.*

Because the Kohlberg-Mertens development takes place in a very abstract context, it is hard to see what stability entails for concrete examples. One point of this paper is to see what stability does entail for (generic) signaling games. Roughly put, we find that stability implies a number of (progressively stronger) restrictions on out-of-equilibrium beliefs in this simple class of games. Some of the restrictions we find quite intuitive; for example, stability implies the intuitive restriction given above. As the restrictions mount, however, our intuition becomes progressively weaker; until we come to implications of stability that we (at least) are unable to motivate so nicely. In the end, we have mixed feelings about stability, at least insofar as it applies to signaling games: it captures quite beautifully some restrictions that we find very satisfactory; but in other cases it seems very strong.

We have two objectives in this study. Our first concern is with signaling games alone. These games, and elaborations of these games, have proved to be very important to recent work in theoretical microeconomics. By developing a sequence of (progressively stronger) general criteria for the equilibria in these games, we hope to provide analysts with a general language for the discussion of what level of restrictions they must impose in order to obtain a particular equilibrium outcome. Second, the theory of stability, either as it stands or as it develops, will certainly prove to be an important idea in noncooperative game theory. We hope that our examples and characterization of stability for signaling games will help in the further general development of these ideas.

The paper is organized as follows. We begin, in Section II, with an example that illustrates the basic program that we are following. Then, in Section III we introduce a general framework for our analysis. In subsection III.1 we define the general signaling game, and we recall in subsection III.2 some propositions concerning equilibria of extensive games from the literature. In subsection III.3, we recall basic concepts and definitions from Kohlberg and Mertens [1986], with emphasis on a particular result that they give.

Section IV is the heart of the paper. The general program that we follow for restricting beliefs in testing equilibrium outcomes of signaling games and the connection of this program to stability are given in subsection IV.1. The rest of Section IV develops some specifications of the general program: subsection IV.2 concerns the well-known and much used criterion of domination. Subsection IV.3 takes up what we call *equilibrium domination*; included here is the criterion that follows from the "speech" given above, which we refer to as *the Intuitive Criterion* (the uppercase letters signifying this particular criterion). Subsection IV.4 briefly discusses (variations on) the Banks and Sobel [forthcoming] criteria of divinity and universal divinity. And subsection IV.5 discusses the "never a weak best response" criterion of Kohlberg and Mertens [1986].

In Section V we apply these various criteria to a simple version of Spence's signaling model, showing how they work to rule out all but a single equilibrium outcome in the game (namely the separating equilibrium identified by Riley [1979]). We conclude in Section VI with a discussion of the full implications of stability for signaling games, and with a summary of what (we think) we have learned.

McLennan [1985] pioneers the approach of refining Nash equilibria by formal restrictions on out-of-equilibrium beliefs. His approach differs from our own in one important respect, and we shall offer a few remarks on this in subsection IV.3. He should be credited, however, with initiating the general program we follow here.

Contemporaneously, Banks and Sobel [forthcoming] have analyzed the same basic problem as do we, arriving at very similar answers. We have benefited from seeing their results, and we have, for completeness, related their criteria of divinity and universal divinity to our approach. They should be given all the credit for those two criteria, and (at least) equal credit for other results that appear in both papers. The reader will benefit from reading their treatment of these issues. Also, we are greatly indebted to Kohlberg and Mertens [1986] for many of the ideas here. In particular, they are responsible for the "never a weak best response" criterion that dominates our mathematical analysis.

Because our focus here is on simple signaling games, many interesting questions that arise in games with a richer dynamic structure are moot. Cho [1986, forthcoming] presents an analysis of some of our ideas, adapted to more interesting games. McLennan [1985] also deals with general extensive games.

II. A Simple Example

The basic ideas in this paper are illustrated by the following simple game. The reader should refer to Figure I throughout.

We tell the story of a two-player game with incomplete information concerning one of the two. The first player is called A, and A either is a *wimp* or is *surly*. Nature has selected the disposition of A, with probability 0.9 that the A selected is surly. In terms of Figure I, nature has chosen to start the game at one of the two open dots labeled t_w (for the wimp) and t_s (for the surly type of A). The prior

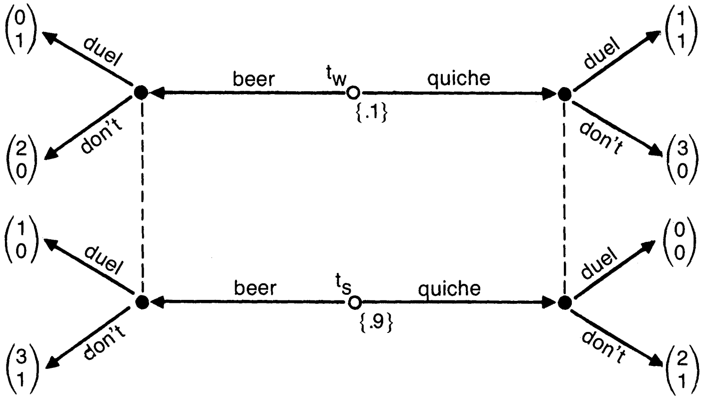

FIGURE I

probability of nature's choice is indicated by the numbers in curly brackets.

At the start of the game, A knows his own disposition or type, and A is faced with the choice of what breakfast to have, before setting off for the day. The choices available are *quiche* and *beer.* These choices are denoted by the pairs of arrows pointing out from the open dots. A's preferences concerning breakfast depends on his type: if A is a wimp, he derives incremental payoff 1 for having quiche and 0 from beer; if A is surly, beer is worth 1, and quiche is worth 0.

After breakfast, A meets with B. There are four conceivable circumstances under which the meeting could take place, corresponding to the two types of A and the two types of breakfast; these are depicted by the four filled dots. B, at this meeting, chooses whether to *duel A*. B's choices are represented by the arrows emanating from the four filled dots. *When B chooses whether to duel, he does so knowing what A had for breakfast, but not knowing for sure what is A's type.* This is depicted in the picture by the dashed lines connecting pairs of solid dots, representing the *information sets* of B—to the right is the information set of B if he knows that A had quiche for breakfast, and to the left is the information set of B if he has observed A have beer.

B's choice whether to duel effectively ends the game. A, whether surly or not, wishes that B choose not to duel. We imagine that A gets (incremental) payoff 2 if B chooses not to duel, and 0 if B does duel. A's total payoff (the sum of the two increments) is the first number in each column vector, at the end of each sequence of choices (by nature, then A, and then B). B wishes to duel with A if and only if A is a wimp—B's payoffs, reflecting this, are the second number in each column vector.

Note well that it is more important (in terms of payoff) to A that he deter B from dueling than that he have his preferred breakfast. And B's prior on what type is A would, absent any further information, induce B to avoid the duel.

This extensive game has two Nash equilibrium outcomes. In the first, A, regardless of type, has beer for breakfast. B, having seen a breakfast of beer, will not duel—this makes sense as (anticipating A's strategy) B's posterior, given a breakfast of beer, is that A is surly with (prior) probability 0.9. Now to make this a Nash equilibrium, we must keep the wimpish A from having a breakfast of quiche—this will happen if, upon seeing a breakfast of quiche, A chooses to duel with probability 0.5 or greater. We can, moreover, "make sense" of such a reaction by B to a quiche breakfast as follows: quiche is taken as a sign that A is a wimp; B revises his probability assessment that A is a wimp to 0.5 (or more). If this "posterior probability" is 0.5, then B is indifferent between dueling or not; if it exceeds 0.5, then B strictly prefers to duel. Note that, at the equilibria we have described, B's "posterior beliefs" at the quiche information set are not computable using Bayes' rule, since there is zero prior probability that B will observe the event (quiche for breakfast) that he is meant to condition upon. But there do exist beliefs at this out-of-equilibrium information set that rationalize B's dueling (with sufficiently high probability). That is to say, the equilibrium outcome is *sequential.*

The second equilibrium outcome is much like the first, except that the breakfast changes. A, regardless of type, has quiche for breakfast. Seeing quiche for breakfast, then, B learns nothing; his posterior on A's type is the prior, and B chooses to avoid the duel. To keep the surly A from having a breakfast of beer, B, in the event of a beer breakfast, duels with probability 0.5 or more. And this response by B to the out-of-equilibrium meal of beer can be rationalized; B's out of equilibrium "posterior beliefs" are that, if A has beer, A is a wimp with probability 0.5 or more.

This multiplicity of equilibria, and their source, is typical of this type of signaling game. (A general definition will be given shortly.) Even if we insist that B rationalize responses to out-of-equilibrium messages (such as A's choice of breakfast) with some beliefs as to A's type, the wealth of possible out-of-equilibrium beliefs gives us a wealth of out-of-equilibrium responses by B. And, therefore, many equilibrium choices of message by A can be supported.

But, in the second equilibrium, are B's out-of-equilibrium beliefs sensible? Here is an argument to say that they are not. If this is the equilibrium, then a wimpish A will net utility 3 at the equilibrium. (He gets his preferred breakfast of quiche and no duel in the bargain.) By having beer for breakfast, the very best he could do is a payoff of 2. Sending the out-of-equilibrium message "beer for breakfast" makes no sense for him. But it *might* make sense for the surly A to have this breakfast, in that, in equilibrium, the surly A receives 2 in equilibrium, and he can *conceivably* get 3 from a breakfast of beer. Suppose then that B is restricted to beliefs which put no weight on the wimpish A having beer for breakfast. (Think in terms of *removing* from the game the possibility that a wimpish A could have beer.) In this case, the only beliefs B could hold are that

A must be surly, and *B* would not duel. If the surly *A* realizes all this, he knows that he can have his beer and safely anticipate no duel. This breaks the second equilibrium.

The first equilibrium is unbroken by such considerations. There it is quiche that is the out-of-equilibrium meal. The surly *A* has no reason to defect (getting 3 in equilibrium, and getting a maximum of 2 if he has quiche), whereas the wimp could conceivably gain by a defection. So, in the spirit of the previous paragraph, we would say that *B* should rule out the possibility that it is the surly *A* that is sending the out-of-equilibrium message. This causes *B*, in the event of receiving that message, to hold a "posterior" assessment that he faces a wimp, which in turn causes him to choose to duel. But this *supports* the equilibrium outcome we have described.

To take the argument a level further, consider the variation in Figure II. Here we have given *B* three options: to duel; to walk away; to give *A* $1,000,000. This third option does not affect the set of equilibria, since giving away the million is always a dominated strategy for *B*. But this third option does ruin the specific argument we gave against the "quiche for breakfast" equilibria. We said before that *B* should discount the possibility that an out-of-equilibrium breakfast of beer comes from the wimpish *A*, since this type of *A* could conceivably benefit from this defection, relative to

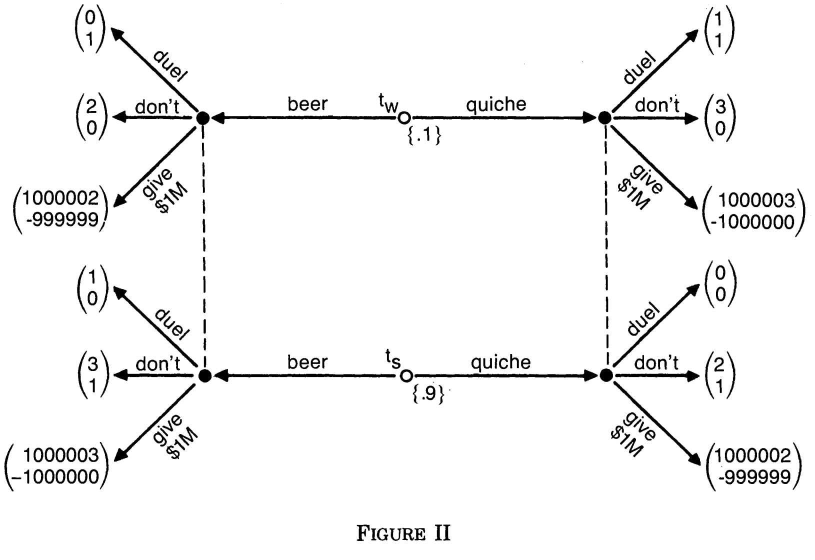

FIGURE II

what he gets in equilibrium. When we add the third response for B, we can no longer say that the wimpish A cannot conceivably benefit from a defection—he does benefit if this induces B to give him the million. We must modify our test to read: could the wimpish A benefit from the out-of-equilibrium breakfast, relative to his in-equilibrium expected payoff, *for any response by B to this out-of-equilibrium breakfast that B might conceivably take?* Define "responses that B might conceivably take" as those that are undominated by other available responses. Then with this modification, we again dispose of the quiche for breakfast outcome.

It is this type of argument that we formalize here. We give answers to the following questions: what is the precise criterion being applied in this example, and what are other, similar criteria? Can we be assured that some equilibrium outcome will always survive a given criterion? And we shall seek to connect these criteria to formal refinements of Nash equilibrium, and especially to stability. To do so requires some formal setup and a review of stability.

III. Formulation and Preliminaries

1. Signaling Games

We focus in this paper on what we call the general *signaling game* with two players. The first, player A, receives private information. Following standard practice, we shall say that this player learns his *type* t, drawn from a finite set T. The player's type is drawn according to some probability distribution π over T that is common knowledge. Player A, having learned his type, sends a message m to player B chosen out of some finite set M. We allow the set of messages available to A to depend upon A's type; we write $M(t)$ for the set of messages available to type t, and $T(m)$ for the set of types that have available the message m. Player B, having heard this message, chooses a response r from a finite set of responses R. We allow the available responses to depend on the message received, writing $R(m)$. The game ends with this response, and payoffs are made to the two players, depending on the type of player A, the message A sent, and the response B took. The utility payoff to player A is denoted $u(t,m,r)$, and the utility to player B is denoted $v(t,m,r)$.

There are very simple games in extensive form, and one can describe their (sequential) equilibria very easily. Some notation will be helpful: we shall write behavior strategies for player A as $\rho(m;t)$, where, for each t, $\rho(\cdot;t)$ is a probability distribution over $M(t)$. The interpretation is that t sends m with probability $\rho(m;t)$. For behavior strategies of player B, we shall write $\phi(r;m)$, where, for each m, $\phi(\cdot;m)$ is a probability distribution on $R(m)$; the interpretation is that B, observing m, chooses response r with probability $\phi(r;m)$.

When player B chooses a response r in response to some message m, he does so (in any sequential equilibrium, at least) on the basis of some posterior probability assessment μ over the set $T(m)$ of types of player A who might have sent that message. Write $BR(\mu,m)$ for that subset of R of best responses (for player B) to m if player B has posterior assessment μ. That is,

$$BR(\mu,m) = \arg\max_{r\in R(m)} \sum_{t\in T(m)} v(t,m,r)\ \mu(t).$$

For subsets I of $T(m)$, let $BR(I,m)$ denote the set of best responses by B to probability assessments concentrated on the set I. That is,

$$BR(I,m) = \bigcup_{\{\mu:\mu(I)=1\}} BR(\mu,m).$$

Write $MBR(\mu,m)$ and $MBR(I,m)$ for the mixed best responses by B to, respectively, beliefs μ and any beliefs whose support is I.[1]

A Nash equilibrium for a signaling game is described by the obvious conditions: given B's strategy ϕ, each type t evaluates the utility from sending message m as $\Sigma_r u(t,m,r)\phi(r,m)$, and $\rho(\cdot;t)$ puts weight on m only if it is among the maximizing m's in this expected utility. And given A's strategy ρ, B proceeds in two steps: first, for any message m that is sent with positive probability by some t, B uses Bayes' rule to compute the posterior assessment that m comes from each type $t \in T(m)$ as $\mu(t|m) = [\pi(t)\rho(m;t)]/[\Sigma_{t'\in T(m)}\ \pi(t')\rho(m;t')]$. And then the Nash condition is that for all m that are sent by some type t with positive probability, every response r in the support of B's response must be a best response to m given beliefs $\mu(\cdot|m)$ that are computed using Bayes' rule; or, in symbols,

$$\phi(\cdot;m) \in MBR(\mu(\cdot|m),m). \tag{1}$$

To require that the Nash equilibrium is sequential is to add the requirement that, for every message m that is sent with zero probability by A (for all m such that $\Sigma_t\pi(t)\rho(m;t) = 0$), there must

[1] Note well that while $MBR(\mu,m)$ is the set of probability distributions over $BR(\mu,m)$, $MBR(I,m)$ may be smaller than all probability distributions over $BR(I,m)$—for each $\phi \in MBR(I,m)$, we must produce a single μ with support I such that $\phi \in MBR(\mu,m)$. The example in subsection IV.5 demonstrates this.

be *some* probability distribution over types $T(m)$, which we shall write $\mu(\cdot|m)$, such that (1) holds. That is, B's responses to out-of-equilibrium messages must be rationalized by *some* beliefs on the part of B. The general program we shall follow is to restrict the set of sequential equilibria by posing restrictions on these out-of-equilibrium beliefs.

2. *Three Facts about the Equilibria of Games*

We next wish to record three useful facts about the sets of Nash and sequential equilibria of various classes of games. The first fact holds for *all* noncooperative games with finitely many players, each of whom has finitely many pure strategies. It is given by Kohlberg and Mertens [1985].

FACT 1. The set of Nash equilibria (and also sequential equilibria) of any finite player, finite pure strategy noncooperative game, viewed as a set of probability distributions over the product space of pure strategy profiles, consists of a finite number of connected sets.

The second fact pertains to games that are *generic for a given extensive form.* By this is meant: fix any (finite player, finite action) extensive form. Fix as well the probability distributions for any moves by nature. Let Z denote the set of terminal endpoints for the extensive form, and let I denote the set of players. Then the specification of the game is completed by an assignment of payoffs to the players, one for each player at each point in Z. That is, the *space of games* over the given extensive form is $\mathcal{R}^{Z\times I}$ (where $\mathcal{R}$ denotes the real line). A statement is said to be *true for generic extensive games* if, for every fixed (finite) extensive form and probability distribution for nature's moves, the set of payoffs (for that form) for which the statement is false has closure whose Lebesgue measure in $\mathcal{R}^{Z\times I}$ is zero. Put another way, if a statement is generically true in this sense, and if the payoffs for a given extensive form are chosen from $\mathcal{R}^{Z\times I}$ *at random,* according to some probability distribution that is absolutely continuous with respect to Lebesgue measure, then there is probability 1 that the statement is true for all games in an open neighborhood of the payoffs chosen.

We require one further piece of terminology. For a given extensive form with terminal endpoints Z, each strategy profile (an assignment of one strategy for each player) induces a probability distribution over which endpoint is reached. Fixing the strategy profile, call this probability distribution the *outcome* of the game associated with the strategy profile. If a particular outcome is the outcome for some Nash equilibrium, call it a *Nash equilibrium outcome.* If it is the outcome of a sequential equilibrium, call it a *sequential equilibrium outcome,* and so forth. With all this buildup, the following is shown by Kreps and Wilson [1982a] and Kohlberg and Mertens [1986].

FACT 2. For generic extensive games, the set of Nash equilibrium outcomes consists of a finite number of points.

This means that, while the set of Nash equilibria for an extensive form game may be infinite, the infinite variety (generically) concerns *out-of-equilibrium* actions and reactions.

Because the map from strategy profiles to outcomes is continuous, facts 1 and 2 combine to establish the following simple corollary:

FACT 3. For generic extensive games, a single equilibrium outcome is associated with each individual connected set of Nash equilibria (cf. Fact 1).

The structure that is implied for generic extensive games by these three facts is illustrated by the beer-quiche example. In this example there are two connected sets of equilibria, each one resembling a line segment. The first set of equilibria corresponds to the outcome where both types have beer for breakfast, followed by no dueling. The *set* of equilibria associated with this outcome arises from the many possible equilibrium responses by B to an out-of-equilibrium breakfast of quiche, namely any (mixed strategy) response that puts weight 0.5 or more on dueling. Similarly, the equilibrium outcome of quiche-no duel is associated with a line segment of equilibria, arising from many possible responses by B to the out-of-equilibrium breakfast of beer.

Note that *genericity* for our signaling games is defined in the space of all payoff assignments to all endpoints. Thus, in applying Fact 2, we cannot be confident that there are a finite number of outcomes for a signaling game that is randomly selected from the subspace of signaling games in which the message sent by A has no impact on B's utility, or where A's message selection has no impact on either player's payoffs. There may be a finite number of equilibrium outcomes for such games—witness the beer-quiche game. But we cannot appeal to Fact 2 to justify an assumption that such a game, selected at random, has a finite number of equilibrium outcomes. In the analysis to follow, we almost always fix a signaling game and assume that it has a finite number of equilibrium outcomes, justifying this assumption by appeal to Fact 2. Thus, our justification is not valid for the games analyzed by Crawford and Sobel [1982] or Farrell [1985]. (In any event, the criteria that we subsequently develop would not have force in games in which signals are costless to the sender.)

3. *Stability—A Review*

Kohlberg and Mertens [1986] develop a number of criteria for Nash equilibria and sets of Nash equilibria that fit under the general rubric of *stability*. It is not our intention to repeat their development in detail; the reader is urged to read their paper to get a complete feel for their intentions and accomplishments. (In particular, it is very important to understand why they have moved from Selten's [1965, 1975] basic program of perfecting individual equilibria to a consideration of "perfect" (strategically stable) *sets* of equilibria). But to make this paper close to self-contained, salient features of their development will be summarized in this subsection.

Suppose that we have a class of games Γ on which some metric is defined. Let Σ be the space of strategy profiles for games from Γ, also endowed with some metric. Let $N : \Gamma \Rightarrow \Sigma$ be the Nash correspondence. For a particular game $\gamma \in \Gamma$, a subset $M \subset N(\gamma)$ of the Nash equilibria of γ is said to be *stable* if for every $\epsilon > 0$ there is a $\delta > 0$ such that every game γ' that is within δ of γ has some Nash equilibrium that is less than ϵ distant from the set M.

This may seem rather a strong condition, but if we allow much latitude in selecting the set M, it is rather weak. For most sensible metrics on games, the Nash correspondence is upper hemi-continuous. In such cases, one stable set of equilibria for the game γ is $N(\gamma)$, the set of *all* equilibria of γ. But $N(\gamma)$ is "large"; one is interested in showing that a set of equilibria smaller than $N(\gamma)$ is stable.

This is the basic mathematical structure in Kohlberg and Mertens [1986]. (The definitions that we use come from earlier versions of their paper and vary somewhat from their most recent treatment. We shall alert the reader to the single important distinction near the end of this section). For the space of games Γ, they fix a Normal form—a positive integer number I of players, and, for each player $i = 1, \ldots, I$, a positive integer number s_i of pure strategies, and they consider for Γ all assignments of utilities to (pure) strategy profiles. That is, $\Gamma = (\mathcal{R}^{s_1 s_2 \dots s_n})^I$. And the space of strategy profiles Σ is then the n-tuple of simplices of mixed strategies for the fixed game form (always with the usual Euclidean metric). As for the metric on Γ, they work with three, corresponding to their notions of hyperstability, full stability, and (plain) stability. Hyperstability, for example, corresponds to the standard Euclidean metric on Γ. Stability (which we shall deal with in the sequel) is defined as follows. A neighborhood base for the game γ is generated by taking, for each $\delta > 0$, the set of all games γ' whose payoffs can be realized as follows: for each player i there is a (mixed) strategy σ_i^* and a real number $\delta_i \in [0,\delta]$ such that the outcome to player j in γ' if the players chose strategies σ_i, respectively, is the outcome in game γ if they choose the strategies $\delta_i\sigma_i^* + (1 - \delta_i)\sigma_i$. (Kohlberg and Mertens have σ^* completely mixed and $\delta_i \in (0,\delta)$; we have closed these so that we can rephrase their construction in terms of a topology on the space of games.)

In what follows, we shall use stability only. But the reader should note that the basic existence result, given below as Fact 4, is true for hyperstability, and so much of our analysis can be carried over to that stronger criterion.

Since the Nash correspondence is upper hemi-continuous in the metric of stability, the set $N(\gamma)$ is always stable. The idea, as indicated above, is to find smaller stable sets. For games that are *generic in the normal form,* that is rather too easy to do; it turns out that stability per se adds no restrictions to the original definition of Nash. Let us explain: *genericity in the normal form* is defined in the obvious way, where the space of payoffs is taken to be all payoff assignments over the normal form. Now if we think of a game in normal form as a trivial game in extensive form (where there is one information set per player, at which the player chooses among his pure strategies), Fact 2 is seen to imply that, for generic normal form games, there are a finite number of Nash equilibria. Moreover, Kohlberg and Mertens [1986] establish that, for generic normal form games, each of these Nash equilibria, taken as a singleton set, is stable (and even hyperstable).

Stability acquires cutting power (for generic games) when applied to normal form games that arise from and are generic for a given extensive form. Consider, for example, the signaling game of Section III. The space of payoffs for the extensive game has dimension $(TMR)^2$ (where we use the uppercase letters to denote both the sets and the cardinality of the sets, and we assume, for purposes of this paragraph, that $M(t) \equiv M$ and $R(m) \equiv R$). Viewed as a normal form game, though, A has M^T pure strategies, and B has R^M, so the space of normal form playoffs for the same has dimension

$(M^T R^M)^2$. Clearly, a generic set of payoffs for the extensive form will typically be a very small set in the set of all payoffs for the normal form, so knowing that a statement is true for generic normal form games tells us nothing about the truth of statement for payoffs that arise generically in the underlying extensive form.

Indeed, fixing an extensive form, for games that arise from payoffs generic in the extensive form, there are sometimes infinitely many Nash equilibria (e.g., in the beer-quiche example). And it is sometimes the case that no single equilibrium, taken as a singleton set, will be stable. Kohlberg and Mertens [1986], however, establish the following basic existence result: recall that the set of equilibria (for any finite game) consists of a finite number of connected sets.

FACT 4. For any finite game, some one (or more) of those connected sets of equilibria, taken by itself, is stable.

Since, by Fact 3, in generic extensive games each connected set of equilibria is associated with a single equilibrium outcome, we can enlist Kohlberg and Mertens' [1986] basic existence result to say: *generic extensive games possess at least one stable equilibrium outcome,* where an equilibrium outcome is stable if the set of equilibria that give rises to it is a stable set. That is, for generic extensive games, while we cannot guarantee the existence of an equilibrium that is stable as a singleton set, we can guarantee that some single equilibrium outcome is stable.

Moreover, while for games generic in their normal form *every* Nash equilibrium is stable, it is not the case that, for generic extensive games, each equilibrium outcome is stable. That is, stability as a criterion for equilibrium outcomes does have cutting power in generic extensive games. It is this cutting power, in the context of signaling games, that we investigate.

In doing so, we make use of a result from Kohlberg and Mertens [1986]. This requires a piece of terminology: take any game and any set of equilibria for the game. A strategy for one of the players is said to be *never a weak best response* for the player, relative to the set of equilibria, if in no equilibrium from the set is the strategy in question as good for the player as the strategy prescribed by the equilibrium.

FACT 5. Suppose that we are given a set of equilibria for a game and a particular pure strategy for a given player that is never a weak best response for the player relative to the set. Consider the game with this pure strategy removed (from the normal form) entirely, and consider the subset of the original stable set of equilibria that consists of all strategy profiles from the set in which the given player puts zero weight on the particular pure strategy. This subset is stable in the game that results after the particular strategy for the player is "pruned." The same is true if the strategy for the player that is pruned is weakly dominated by some other strategy for the player.

The reader should be warned that the terminology we are using is not quite consistent with Kohlberg and Mertens [1986]. The important difference is that they reserve the term *stable set* for a minimal (by set inclusion) closed set of Nash equilibria having the stability property we have given. It is easier for us to to use the term *stable set* for any set of equilibria which has the stability property, and we shall do so.[2] It makes no difference to the results, as long as one defines a stable equilibrium outcome as a single outcome that is the projection of some (minimal) stable set of equilibria. This warning is particularly apt here, as the reader searching for Fact 5 in Kohlberg and Mertens will find a slightly different formulation; namely, the subset *contains* a stable set of equilibria.

We require a couple of preliminary results, concerning the connection between signaling games and stable equilibria. We state them in a single lemma.

LEMMA 1. Fix a signaling game. Suppose that an equilibrium outcome is stable, in the sense that the set of *all* the equilibria that give rise to the outcome is a stable set. Then the set of all *sequential* equilibria that give rise to the outcome is also a stable set. Also, every stable set of equilibria contains at least one sequential equilibrium.

The condition that the game is a signaling game is needed here: Kohlberg and Mertens [1986] present an example (which they attribute to Gul) of a game having a stable set of equilibria, each member of which gives an outcome different from the unique sequential equilibrium of the game. (That is, not only does the stable set not contain any sequential equilibrium; every equilibrium in the stable set gives an equilibrium outcome different from the unique sequential equilibrium outcome.) What saves us from such an unhappy situation here is that, for signaling games, equilibria that are perfect in the normal form are also sequential. From the

[2] Our usage is consistent with the 1982 and 1983 versions of Kohlberg and Mertens.

definition of stability, it is easy to see that every stable set of equilibria contains some equilibrium that is perfect in the normal form. (See Kohlberg and Mertens [1986, p. 1028].) And it is likewise easy to show that, for a given stable set of equilibria, the subset of all normal form perfect equilibria in that set will be stable.

IV. Selection Guide

1. *The General Program*

We have finally gotten through the preliminaries, and we can outline the general program of the paper.

Recall the beer-quiche example. We fixed a particular equilibrium outcome, and we restricted B's out-of-equilibrium beliefs by giving and applying a criterion for saying that a particular out-of-equilibrium message was "unreasonable" for a particular type. The criterion used there was that type t would not reasonably be expected by B to send out-of-equilibrium message m if the best t could do from m was less than t got at the equilibrium outcome. This is but one criterion we might think of, and, later in this section, we shall formalize it and others. In the variation on beer-quiche, we found that this criterion might be sharpened by first discarding "unreasonable" responses by B to certain out-of-equilibrium messages. The criterion used there was that response r to message m is unreasonable if it is dominated by some other response. We shall use a criterion that is equivalent to this for ruling out responses to messages in what follows.

After restricting B's out-of-equilibrium beliefs, we asked whether the originally fixed equilibrium outcome could be "supported" by out-of-equilibrium beliefs that obey the restrictions. In the case of the equilibrium outcome where both types ate quiche, the answer was no, because B's out-of-equilibrium beliefs would cause the surly type of A to defect.

We formalize all this as follows: for a given signaling game with a finite number of equilibrium outcomes, fix some one of those outcomes. Let $u^*(t)$ denote the expected utility of type t in the fixed equilibrium outcome. Construct a test of that outcome in two steps.

STEP 1. Pose some criterion for saying that a particular out-of-equilibrium message cannot "reasonably" be expected to be sent by a particular type. In each of the four subsections to follow, a different criterion of this sort will be investigated.

Also, say that message m will never "reasonably" be expected to be met with response r if $r \notin BR(T(m),m)$.

Now specify some order and number of times for applying these two criteria. Apply the first by deleting from $M(t)$, for each type t, any out-of-equilibrium message m such that (t,m) is "unreasonable." Apply the second by deleting from $R(m)$, for each out of equilibrium message m, any response $r \notin BR\ (T(m),m)$. Apply the two iteratively, in an order, and for a number of times that is part of the specification of the test.[3] That is, the order and number of applications, together with the criterion used to strike type-message pairs, is the specification of step 1 of the test.

When step 1 is completed, we shall have, for each out-of-equilibrium message m, a set of types that have not been ruled out for that message. Call this set of types $T^s(m)$.

STEP 2. For each out-of-equilibrium message m, consider all sequential equilibrium responses of B to m in the original game. Are any of these sequentially rational for B, given that B's beliefs are restricted to $T^s(m)$? If not, then the equilibrium outcome has failed the test. (If so, it has passed.)

(It may happen that $T^s(m)$ is empty. In this case, to pass the test, it is necessary that, for every probability distribution μ on $T(m)$, there is some $\phi \in MBR(\mu,m)$ that supports the equilibrium outcome in the original game.)

We shall be attempting in Step 1 (as we delete type-message and message-response pairs) to make restrictions in B's beliefs that are intuitive in the sense that no one playing the game would reasonably expect B to put positive weight, given an out-of-equilibrium message m, on a type t that has been excluded for m. If one were trying to envisage the thought process of the players, we might imagine two stages to Step 1. First, there is the asserted *fact* that, given the equilibrium outcome, a given out of equilibrium message is not sensible for certain of the types of A. Second, insofar as this asserted *fact* is held by the players to be correct, *introspection* on their part will tell them that B will attach zero weight to this out-of-equilibrium message coming from those types, which will cause B to consider only certain responses to the message. This restriction on B's responses, if believed by all to be valid, may lead

[3] In the variation on beer-quiche, we saw that applying the second criterion could sharpen the first. Since the first will "reduce" $T(m)$ for a given m, it will sharpen the second. Examples are easy to concoct to show that iterating back and forth can lead to continued reductions in the set of types that are reasonable for a given out-of-equilibrium message, which is the point of this exercise.

to more asserted *facts* about which types of A might conceivably send the given out-of-equilibrium messages; introspection by the players concerning this will lead to further restrictions on B's responses, and so on. Since this cycle (and its consequences in Step 2) depends on both players coming to these conclusions through introspection, our success in this endeavor will depend upon the extent to which the asserted *facts* (the criterion used to delete type-message pairs) are judged to be intuitive, and on the number of iterations that are required. (Restrictions made in one or a few iterations are presumably easier to swallow than restrictions that require many iterations, since each level of introspection requires faith that the other side has made it to the previous level.)

Having made those restrictions, in Step 2 we suppose that everyone expects B to respond to an out-of-equilibrium message m that is based on beliefs which give no weight to types that have been excluded for that message. With this supposition can we continue to sustain the equilibrium outcome? Again envisaging the thought process of the players, we have in mind something like Kohlberg and Merten's process of *forward induction*: will some type of player A, having arrived introspectively at restrictions in B's beliefs (hence B's conceivable actions), see that deviation will lead to a higher payoff than will following the equilibrium?

Note that we shall fail Step 2 if and only if there is some out-of-equilibrium message m such that, for every $\phi \in MBR(T^s(m),m)$, there is some type $t \in T(m)$ with $u^*(t) < \Sigma_r u(t,m,r)\phi(r)$. We can pose a slightly weaker test, by asking whether there is some *single* type t (for the given m) that would send m no matter what response B picks out of $BR(T^s(m),m)$. That is, to fail the weaker test, the restrictions from Step 1 must suffice to give us a single type who will then certainly wish to break the equilibrium; in Step 2 as formulated, we rely on the (usual) assumption that B's out-of-equilibrium response to m is commonly known to all types of A. The weaker test, when failed, may be slightly more convincing as evidence against the fixed equilibrium outcome, so we shall consider, in the sequel, tests that are composed of some specification of Step 1, followed by this variation in Step 2, which we refer to as Step 2A.[4]

[4] The reader may wonder whether a type that has been excluded for message m in Step 1 might subsequently prove to be the undoing of the equilibrium in Step 2 or 2A, in the sense that, with B's beliefs restricted by Step 1 to exclude type t, B will respond to m in a manner that causes t to send m. If we could be sure that this would not happen, then we could pose Step 2 a bit more simply, as: is the original equilibrium outcome still a sequential equilibrium outcome in the game where only types $t \in T^s(m)$ can send m? We shall return to this issue in subsection IV.5, at which point we shall see a particular test for which the answer is no.

Besides posing "intuitive tests" of equilibrium outcomes in signaling games, we follow the program above in order to relate our tests to Kohlberg-Mertens stability. We seek, in general, to show that any equilibrium outcome that fails any of the tests we construct fails as well to be a stable equilibrium outcome. To do this, we apply Fact 5, Lemma 1, and the following lemma.

LEMMA 2. For signaling games, if a response r is not sequentially rational for B in response to message m' for some beliefs over $T(m')$, then it is dominated by some convex combination of B's other available responses.

This is a simple application of the separating hyperplane theorem and is left to the reader.

This lemma shows that deletion of message-response pairs according to the "never sequentially rational" criterion is a special case of deletion of weakly dominated strategies. Fact 5, therefore, implies that any time we begin with a stable set of equilibria in a given signaling game, what remains of the set is stable in the signaling game that results from the pruning of any such message-response pairs.[5] Suppose, then, that the criterion used in Step 1 to delete type-message pairs also falls under the "permitted categories" of Fact 5. The iterated application of Fact 5 is clearly legitimate, so if the set of equilibria is stable to begin with, what remains of it must remain so after as many of these deletions as we care to make. Now invoke Lemma 1. In a signaling game with a finite number of equilibrium outcomes, suppose that some outcome is stable. Then the set of all sequential equilibria giving rise to that outcome is stable, according to the first half of Lemma 1. The iterated deletion from this set of type-message and message-response pairs according to any criteria that are permitted by Fact 5 will leave us at each stage with a stable set of equilibria that are sequential for the original game. When this is completed, apply the second half of Lemma 1 to extract an equilibrium from this set that is sequential for the reduced game. Necessarily, B's response in this equilibrium must be a best response to beliefs on the types in $T^s(m)$. Hence the test in Step 2 would be passed. We summarize this discussion as

PROPOSITION 1. Insofar as the deletion of type-message pairs falls under either category permitted by Fact 5, any equilibrium outcome that fails our test fails to be stable.

[5] Fact 5 permits domination in mixed strategies.

Moreover, the Kohlberg-Mertens basic existence result gives us an instant corollary.

COROLLARY. Insofar as the deletion of type-message pairs falls under either category permitted by Fact 5, some one or more of the equilibrium outcomes of the fixed signaling game must pass the test that has been posed.

2. *Dominance*

We can now pose specific tests that follow the general scheme above. An obvious test, and one that is well-known to practitioners of signaling games, involves dominance. Pose the following criterion for eliminating type-message pairs:

Elimination of type-message pairs by dominance. For out of equilibrium message m, type t may be eliminated for this message if there is some other message m' with

$$\min_{r \in R(m')} u(t,m',r) > \max_{r \in R(m)} u(t,m,r). \tag{2}$$

(We could get by with a weak inequality here, but we use the strict inequality to facilitate comparison with later tests.)

From this criterion for the elimination of type-message pairs, we can construct many tests. We might allow for a single application of this criterion, which corresponds to one round of elimination of strategies dominated for A. We might allow for the deletion of message-response pairs as in IV.1, followed by one round of elimination of type-message pairs using this criterion. We might allow iterated application of the two, for as long as it is profitable. (Such an iteration must terminate eventually, as there are only finitely many message-response and type-message pairs to delete.) This corresponds to the iterated application of (sometimes weak) dominance. And we could follow any of these with either Step 2 or 2A. It is evident that the criterion above is "permitted" under Fact 5, so that Proposition 1 and the corollary apply.

Are any of these tests subsumed by less stringent refinements of Nash equilibrium? The answer is no for both trembling-hand perfection [Selten, 1975] and properness [Myerson, 1978]. Consider, for example, the game in Figure III. (The pictures are as before, except that in this case, as B is given no choice of response to the message m, we do not bother to put in an information set for him.) Consider the equilibrium outcome in which both types of A send the message m, and B responds to m' by choosing r_1 with probability 0.5 or greater. To support this equilibrium, B's beliefs at

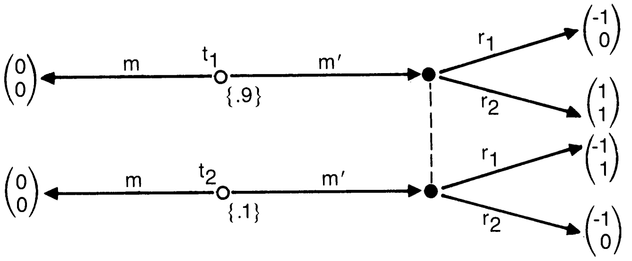

FIGURE III

the m' information set must put weight 0.5 or more on A being type t_2. But for type t_2, m dominates m'. So, by any of the tests constructed from the dominance criterion above, we can prune the type-message pair (t_2,m') from the game. In the game that is left, B must respond to m' with r_2. This causes the equilibrium outcome to fail the test, using either Step 2 or 2A, since this response causes t_1 to defect.

The game in normal form is given in Table I. (Note that the prior enters into the expected payoff calculations.) We leave to the reader the simple task of verifying that the equilibrium in which A chooses m regardless of type and B responds to m' with r_1 is indeed proper. (Moreover, it is easily shown to be perfect in the agent normal form.)

This example can be used to make another point, concerning properness for signaling games. (The material in this paragraph is a bit esoteric, and it may be skipped without loss of comprehension of most of the rest of the paper.) Consider changing the prior on A's type, from 0.9 that A is t_1 to 0.9 that A is t_2. Since, to support the m equilibrium outcome, it is necessary that B "assess" high posterior

TABLE I
GAME OF FIGURE III IN NORMAL FORM

Message if		Response	
		r_1	r_2
t_1	t_2		
m	m	0, 0	0, 0
m	m'	−0.1, 0.1	−0.1, 0
m'	m	−0.9, 0	0.9, 0.9
m'	m'	−1, 0.1	0.8, 0.9

probability that m' comes from t_2, this change would seem to make it *harder* to disqualify the equilibrium outcome under scrutiny. In any event, this change should make it no easier. But if the reader constructs the associated normal form, he or she will find that this change renders the m,r_1 equilibrium improper. Properness now does act to disqualify this equilibrium. This seems rather counterintuitive, but it is not hard to see why this is happening. If we view this game as a two-player game (A and B), then we must make some *intertypal* comparisons of payoffs. That is, we have to aggregate the payoffs of type t_1 and type t_2 of A. The prior, in this case, serves to scale these payoffs; when the prior is high that A is type t_1, then it is a "worse" mistake for t_1 to send m' than it is for t_2 (if B is responding with r_1) simply because it is a mistake that happens with higher probability. We would see the same thing if we rescaled the utility of one type (but not the other) of A; if, say, we changed t_1's payoffs to 0 if m, -100 if m',r_1, and 100 if m',r_2, then the range of priors for which m,r_1 is proper expands. Especially if we regard these as games of incomplete (as opposed to imperfect) information, this intertypal comparison of utility seems nearly as suspect as would be an inter*personal* comparison. We shall try to avoid intertypal comparisons of utility for the rest of the paper, which means, among other things, that we abandon properness. The most expedient means of being sure that we are not making intertypal comparisons of utility is to regard signaling games not as two-player but rather as $T+1$-player games, where T stands here for the number of types of A; each type of A is regarded as a separate player. We shall on occasion, use language appropriate to this interpretation in what follows. One further word on this: while the properness of the m,r_1 equilibrium depends on the prior, there is another equilibrium with the m outcome, namely where B randomizes evenly between the two responses, which is not proper for any prior. Moreover, the improperness of this equilibrium is unaffected by rescaling of one type's utility. The reader will be better able to see why this is happening when we move on in our development to divinity.

The power of iterated dominance in signaling games has long been noted. See, for example, the development in Milgrom and Roberts [1986].

3. *Equilibrium Dominance and the Intuitive Criterion*

Fix a particular equilibrium outcome, and use, as before, $u^*(t)$ to denote the expected payoff at this outcome to type t of A.

In subsection IV.2, the criterion used to eliminate type-message pairs was domination: the pair (t,m) is eliminated if there is a message m' available to t that does better, no matter what was the response of B to that message, than the best that t can obtain if he sends m. Consider weakening this to read: the pair (t,m) may be eliminated if

$$u^*(t) > \max_r u(t,m,r). \tag{3}$$

Comparing with (2), we see that the difference is that the value against which the consequences of sending m are tested is the expected value that t obtains at the given equilibrium, rather than the worst that can be gotten by sending some other message. Clearly, this new criterion will allow us to eliminate more type-message pairs in the application of the first sort of substep. In the sequel, it is called *equilibrium domination,* or *domination by the equilibrium value.*

Construct from equilibrium domination the following test. First, throw out all message-response pairs (m,r) such that $r \notin BR(T(m),m)$. Then use equilibrium domination to dispose of type-message pairs. Then apply Step 2A. In aggregate, this test amounts to the following.

THE INTUITIVE CRITERION. For each out of equilibrium message m, form the set $S(m)$ consisting of all types t such that

$$u^*(t) > \max_{r \in BR(T(m),m)} u(t,m,r).$$

If for any one message m there is some type $t' \in T$ (necessarily not in $S(m)$) such that

$$u^*(t') < \min_{r \in BR(T(m) \setminus S(m),m)} u(t',m',r),$$

then the equilibrium outcome is said to fail the Intuitive Criterion.

This is the criterion used in the beer-quiche example. (Since we begin with a round of elimination of message-response pairs, it serves both for the original example and the variation.) It has been used in a number of applications: by Grossman [1981] directly; by Kreps and Wilson [1982b] and Milgrom and Roberts [1982a] almost directly (those analyses involve a richer dynamic structure than we have here); Rubinstein's [1985] assumption B-1 is closely related (albeit again with a richer dynamic structure).

Despite the name we have given it, the Intuitive Criterion is not completely intuitive. (It is certainly less intuitive than applications of dominance.) Equilibrium dominance accords a crucial role to the particular equilibrium outcome under discussion, yet, when it works, it proceeds to discredit that equilibrium outcome.[6] Consider in this regard the example of Section II, *beer-quiche.* We argued that, at the equilibrium outcome in which both types have quiche for breakfast, the wimp would never willingly defect to a breakfast of beer, because the best he could do with this breakfast gave him a lower utility than what he got at the equilibrium. If player B regards this as logical, then *introspection* would cause B to respond to beer without a duel, based on beliefs that this breakfast was a sure sign that A is surly. The surly A, capable of replicating this *introspection,* then applies *forward induction* to conclude that a breakfast of beer is worthwhile.

But now take this a step further. If B can, through introspection, come to this conclusion, and if B believes that A can come to it as well, then B will expect A, if surly, to have chosen beer. Hence quiche is a sure sign of a wimp, and should be met with a duel. And, therefore, having quiche will not net the wimpish A a utility of 3, but rather 1, which means that a breakfast of beer cannot be taken as a sure sign that A is surly, which breaks the chain of the previous argument just where it started.

We respond to this counterargument from the following general perspective. An equilibrium is meant to be a candidate for a mode of self-enforcing behavior that is common knowledge among the players. (Most justifications for Nash equilibria come down to something like this. See, for example, Aumann [1987] or Kreps [forthcoming]). In testing a particular equilibrium (or equilibrium outcome), one holds to the hypothesis that the equilibrium (outcome) is common knowledge among the players, and one looks for "contradictions." Thus, to argue, in our example, that beer might conceivably be better for the wimp than is quiche, because quiche might engender a duel, is to *accept* the contention that the quiche outcome equilibrium is not a good candidate for self-enforcing behavior.

A comparison with McLennan's [1985] *justifiable beliefs* is appropriate here. Note that McLennan's concept derives from an inequality that is very similar to (3). The major difference, roughly, is that on the left-hand side he puts the minimum sequential equilibrium outcome to the player, for any sequential equilibrium, whereas we use the payoff from the equilibrium payoff under

[6] The argument to follow was first given in our hearing by Joe Stiglitz.

consideration. Thus, we give a much greater role to the particular equilibrium payoff under consideration. We hold very strictly, in this, to the notion that the particular equilibrium is common knowledge among the players, and we have in mind a "story" that says that out-of-equilibrium messages should be construed as conscious defections from *the* equilibrium. If one thought that out-of-equilibrium messages were (probably) the manifestation of some player or other being unaware of the equilibrium, and if one further thought that this "defecting" player (who is unaware that he is defecting) believes that some other sequential equilibrium prevails in this instance, then McLennan's weaker criterion is the more sensible. We wish to stress here that the Intuitive Criterion relies heavily on the common knowledge of the fixed candidate equilibrium outcome and, in particular, attaches a very specific meaning (a conscious attempt to break that equilibrium) to defections from the supposed equilibrium.

Whatever its intuitive merits, the Intuitive Criterion is based on a criterion for striking type-message pairs that fits into the general program we are following.

PROPOSITION 2. If message m is equilibrium dominated for type t at some equilibrium outcome, then it is never a weak best response at any equilibrium that gives the outcome.

This requires no proof; it is almost a matter of definition. Accordingly, we can post a test (stronger than the intuitive criterion), which we call the equilibrium domination test:

EQUILIBRIUM DOMINATION TEST. Fixing an equilibrium outcome, strike type-message pairs using equilibrium domination and message-response pairs using the "never a best response" criterion, iterating between the two for as long as either has effect. (Finite termination is assured by the usual argument.) Then apply Step 2.

COROLLARY. For a generic class of signaling games, a stable equilibrium outcome will pass the equilibrium domination test. Every signaling game from this generic class (therefore) has at least one equilibrium outcome that will pass this test.

4. Divinity

Banks and Sobel [forthcoming] propose tests of equilibrium outcomes in signaling games that they call *divinity* and *universal divinity*. (In addition, they develop independently many of the other results here.) The reader is urged to read their paper to obtain a detailed analysis of these tests; but, for completeness, we briefly adopt their tests to the framework we are using.

Consider the following two criteria for disposing of type-message pairs. Fix the equilibrium outcome. For a given out-of-equilibrium message m and for each type t, find all (mixed) responses $\phi \in MBR(T(m),m)$ by B that would cause t to defect from the equilibrium. That is, for each t, form the set,

$$D_t = \{\phi \in MBR(T(m),m) : u^*(t) < \sum_r u(t,m,r)\phi(r)\}.$$

And define

$$D_t^0 = \{\phi \in MBR(T(m),m) : u^*(t) = \sum_r u(t,m,r)\phi(r)\}.$$

CRITERION D1. If for some type t there exists a second type t' with $D_t \cup D_t^0 \subseteq D_{t'}$, then (t,m) may be pruned from the game.

CRITERION D2. If for some type t, $D_t \cup D_t^0 \subseteq \bigcup_{t' \neq t} D_{t'}$, then (t,m) may be pruned from the game.

The intuition that is meant to be conveyed by these criteria is that whenever type t either wishes to defect and send m or is indifferent, some other type t' strictly wishes to defect. Hence it should be accorded (by B) more likely that the defection came from some other t' than that it comes from t. In D1, we require that there is a single type t' that always strictly wishes to defect whenever t does. In D2, we pose the weaker requirement that for any response that causes t to defect, there is some type (which may change with the particular response) that wishes strictly to do so. Note well that D2 will permit, in general, more type-message pairs to be struck than will D1.

By striking the pair (t,m), B is assumed to believe that it is "infinitely more likely" that m has come from this other t'. One might therefore seek a milder restriction on B's beliefs—for t such that $D_t \cup D_t^0$ is nonempty, require only that B's beliefs given m do not raise the probability that A is t relative to the probability that A is some other t'. (That is, require that $\mu(t;m)/\mu(t';m) \leq \pi(m)/\pi(m')$.) (When $D_t \cup D_t^0$ is empty, strike (t,m) as before.) The intuition is that t should be no more likely to send m than is t'.

Divinity is, roughly, a test formed from iterated application of the milder sort of restriction just described, combined with D2. It does not correspond to the striking of type-message pair and thus does not fit into our general scheme, but it does have considerable intuitive appeal. *Universal divinity,* on the other hand, fits into our general scheme: it corresponds (again in spirit, as a precise comparison is a bit subtle) to iterated application of the strong restrictions that arise from pruning type-message pairs on the basis of D2.

One could build as well weakened forms of divinity and university divinity, based on D1 instead of D2. We shall see an important difference between the two in the next subsection.

Note that both divinity and universal divinity subsume equilibrium domination. For if message m is equilibrium dominated for a type t, then for this message, D_t and D_t^0 are both empty. As long as there is any type that would gain from defecting to m, we prune (t,m). (If no type can ever gain from sending m, then the equilibrium outcome will not fail the equilibrium dominance test or any other that we can think of on account of this message.)

The connection with stability is established in the usual fashion.

PROPOSITION 3. Any type-message pair disposed of by either criterion D1 or D2 given above is never a weak best response at the given equilibrium outcome. Hence a stable equilibrium outcome will pass the Banks-Sobel test of universal divinity, and a universally divine equilibrium outcome exists, for generic signaling games.

The proof is simple. Since D2 strikes more type-message pairs, we shall work with it. If m were a weak best response for type t at some equilibrium giving this outcome, then, at that equilibrium, the response $\phi(\cdot;m)$ would have to lie in D_t^0. By assumption, this $\phi(\cdot;m)$ lies in $D_{t'}$ for some other type t', which immediately implies that it cannot be an equilibrium response to the out-of-equilibrium message m; it would cause t' to defect.

5. *Never a Weak Best Response*

The proof of Proposition 3 makes clear that we would not run afoul of stability if we modified the criterion to read as follows:

Fix the equilibrium outcome and an out-of-equilibrium message m, and define D_t^0 and D_t as above. Then the pair (t,m) may be pruned if $D_t^0 \subseteq \bigcup_{t' \neq t} D_{t'}$.

In other words, prune (t,m) precisely when there is no sequential equilibrium response to m at which t is indifferent between the equilibrium and sending message m; when m is never a weak best

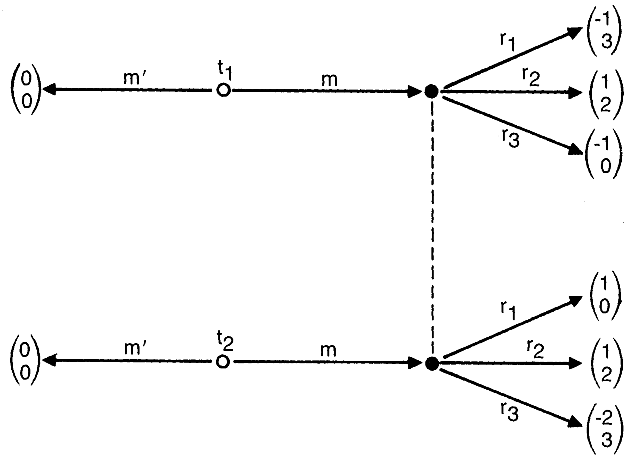

FIGURE IV

response relative to the set of sequential equilibria giving this outcome.

Tests built up out of this criterion will be stronger than those built up out of the criteria of the previous section, as the game depicted in Figure IV shows.[7] We consider the equilibrium outcome in which both types send message m', and B responds to m with high weight on response r_3. In this game, B's mixed best responses to m include all three pure strategies, mixtures of r_1 and r_2, and mixtures of r_2 and r_3. Simple algebra shows that D_{t_1} consists of mixtures of r_1 and r_2 and mixtures of r_2 and r_3, where in each case r_2 has weight more than ½; $D^0_{t_1}$ the frontier of this set. And D_{t_2} consists of all mixtures of r_1 and r_2, plus mixtures of r_2 and r_3 that put weight greater than ⅔ on r_2; $D^0_{t_2}$ consists solely of the mixture ⅔ on r_2 and ⅓ on r_3. By the criterion of the previous section, no pruning is possible. But by the never a best weak response criterion, we can prune type t_2 for the message m. Doing so causes B to play the pure strategy r_1, which is not an equilibrium response in the original game (type t_2 would defect).

With reference to footnote 5, note that in this example we prune type t_2. This restricts the beliefs of B, who then takes a response that causes the *pruned* player to defect from the original equilibrium. If there is some implicit "speech" to go with the "never a weak best response" criterion, similar to the speech that goes with the Intuitive Criterion, it would have to run something like:

[7] This example is based on another, similar example, from Banks and Sobel.

> At this equilibrium outcome, you should realize that I am not t_2, because. . . . Of course, you should avoid drawing the accompanying inference, which is that only if I am t_2 would I wish to be making this speech.

We do not wish to make too much of these "speeches." But we cannot suggest an intuitive inferential process for B (of the type we have been considering) that accompanies a defection from this equilibrium outcome and that leads B to conclude that the defection cannot be from the only type that would benefit from the defection if B makes that inference. On intuitive grounds, one might wish to insist that a defection that breaks an equilibrium is accompanied by a process of inference that leads B to put weight on those types that would break the equilibrium. (This philosophy finds favor in the related work of Farrell [1985] and Grossman and Perry [1986b].) Certainly, such a restriction is obeyed by dominance and by equilibrium dominance. In other words, we could rewrite Step 2 of our tests to read: with beliefs restricted in Step 1, the equilibrium should founder or not based on a defection from a type that has not been pruned for the out-of-equilibrium message under consideration. And the tests based on criteria up through equilibrium dominance would not be affected. But the never a weak best response criterion would be changed. We do *not* find this intuitive.

What of universal divinity in this regard? If one insisted on criterion D1 in order to strike a type-message pair, then the resulting test would be safe; an equilibrium outcome that failed the test would fail because of (at least) one uneliminated type. But if criterion D2 is used, the resulting test is not safe. One can construct examples in which, at a given stage, one type is eliminated by virtue of several others, one of which is simultaneously eliminated because the first type is not yet eliminated. That is, each helps to eliminate the other. We are, in consequence, happier with tests built up out of D1 than with those built up out of D2.[8]

V. The Spence Signaling Model

As an example of the various tests posed above, we consider a simple case of the Spence [1974] signaling model. In so doing, we shall be a bit casual, leaving it to the reader to fill in the gaps.[9]

We imaging that a given worker is one of T types, indexed 1, 2,

[8] We are grateful to Gyu Ho Wang for this observation, and for correcting an earlier version of this paper on this point.

[9] The reader, wishing to see our basic argument made both exact and more general, should consult Cho [1986].

$\ldots, T$. (The usual story in the Spence model is that there are many workers, divided by types. We could use this formulation equally easily.) The prior probability that the worker is of type n is π_n. The worker moves first, choosing an education level e from the set $[0,\infty)$. Then two risk-neutral firms, having observed the education choice but not the type of the worker, bid for the services of the worker. The bidding is in the style of Bertrand, with each naming a wage $w \in [0,\infty)$ that it is willing to pay the worker; the worker chooses whichever firm bids the most; if the firms offer the same wage, the worker chooses by means of a coin flip. The worker is worth ne to the firm if n is the worker's type and e is the level of education he obtained. If the worker is of type n, is paid w, and obtains education level e, then the worker's utility is $w - k_n e^2$, for strictly positive constants k_n that satisfy $k_1 > k_2 > \cdots > k_T$. (This particular parametric family of indifference curves is irrelevant to the analysis—any family with the property that the marginal disutility of education strictly decreases with type will do.)

We should note immediately that while we have assumed only a finite number of types of players, we are allowing infinitely many actions for each type, and infinitely many responses by the firms. Thus, stability cannot properly be applied to this analysis; indeed, the definition of a sequential equilibrium must be specially adapted to this context. We shall therefore continue in the spirit of sequential equilibrium and of the restrictions on beliefs posed formally above.

A sequential equilibrium is defined as follows: type n selects education levels according to some probability distribution $\rho(\cdot;n)$; we shall restrict attention to equilibria in which these distributions are discrete and have finite support, so that $\rho(e;n)$ will denote the probability with which type n selects education level e.[10] Firms respond to education level e according to commonly held beliefs $\mu(\cdot\,|\,e)$ as to the type of worker that has selected e; $\phi_i(\cdot;e)$ gives the wage offered by firm i if e is selected. In equilibrium:

(i) Education levels must be chosen by the worker in a way that maximizes his expected utility, taking as given the offers of the two firms.

(ii) At each education level e, firms must bid optimally, given the bidding function of the other firm, and holding to beliefs about the quality of the worker given by μ.

(iii) At every education level e that is selected by the worker with

[10] The interested reader can verify that if any Borel distributions are allowed, all equilibria will have the character of finite support.

positive probability, the firms' beliefs should be generated in the usual fashion by Bayes' rule.

Note that in (ii) we require optimality of the firms' bids at every education level, given their beliefs. So it is (iii) that ties the firms' equilibrium strategies to the worker's.

Bertrand competition among the firms ensures that in any equilibrium, they bid precisely the expected value (to them) of the worker. That is, $\phi(\cdot;e)$ is a point mass at the value $\Sigma_{n=1}^{T} \mu(n|e)ne$. We abbreviate this sum by $W(e;\mu)$ in what follows.

The equilibrium is said to be a *screening equilibrium* if the supports of the various $\rho(\cdot;n)$ do not intersect. A particular candidate for a screening equilibrium outcome, which we call the *Riley outcome* (after Riley [1979]), is constructed as follows. The least able type $n = 1$ chooses the (unique) education level e that maximizes $e - k_1e^2$. That is, type 1 picks the best education level for himself, given that the wage commanded will be the wage appropriate for him. Call this e_1^*. Then type 2 chooses the (unique) level of education that maximizes $2e - k_2e^2$, subject to the constraint that $2e - k_1e^2 \leq e_1^* - k_1e_1^*$. In words, type 2 chooses the best education level for himself, assuming that he will be paid at a wage appropriate to his type, subject to the constraint that 1 would not strictly prefer to choose this education level (and wage) over e_1^*. And so on: type n chooses the best level given that assuming that he will be paid appropriate to his type, subject to the constraint that no lower index type would strictly prefer to pretend to be type n, in preference to staying at the (previously determined) education level (and appropriate wage) for the lower type. Denote the education levels so derived by e_n^*.

We leave to the reader the task of showing that this does indeed give a screening equilibrium outcome. (One must fill in the out-of-equilibrium beliefs and wage offers, and there are many choices for this.) We wish instead to demonstrate that the Riley outcome is the *only* equilibrium outcome that survives the sorts of tests we have posed previously. More precisely, if there are only two types of worker, then the Riley outcome is the only outcome that survives the Intuitive Criterion. With more than two types, it is the only equilibrium outcome to survive Banks-Sobel universal divinity, as defined in subsection IV.4. (With two or more types, dominance arguments as in subsection IV.2 can be applied to restrict somewhat the equilibrium outcomes; and with more than two types, equilibrium dominance can provide somewhat more restrictions. We do not chase down all these partial implications, in the interest of brevity.)

Case A. Two Types

When there are only two types of worker, the Intuitive Criterion suffices to make the argument. Consider first the possibility that the two types *pool*; they each pick some education level e_p with positive probability. Refer to Figure V. Since each picks e_p with positive probability, $\mu(2;e_p) < 1$, and $W(e_p,\mu)$ must be less than $2e_p$. In Figure V we draw the indifference curves of the two types through the equilibrium education and wage pair $(e_p, W(e_p))$. By assumption, the indifference curve of type 2 is less steeply sloped than that of type 1, so points such as those shaded along the line $w = 2e$ exist. Since wages can never exceed twice education level in any sequential equilibrium, type 1 would be strictly worse off picking education level e^* (as shown) than he is at the equilibrium. Hence by the Intuitive Criterion, we must be able to support the equilibrium with beliefs by the firm that e^* is chosen by a worker of type 2 with probability one. But this would lead to a wage $W(e^*;\mu) = 2e^*$, which causes type 2 to defect from the equilibrium.

Hence only screening equilibrium can survive the Intuitive Criteria. Since in any sequential equilibrium wages at level e must be at least e, it is easy to see that, in any screening equilibrium, type 1 will select his Riley level e_1^*. And then the Intuitive Criterion again tells us that the equilibrium must be supportable by beliefs which put weight one on type 2 for any education level e such that

$$2e - k_1 e^2 < e_1^* - k_1(e_2^*)^2; \tag{4}$$

type 1 is certain to do worse (given a sequential equilibrium

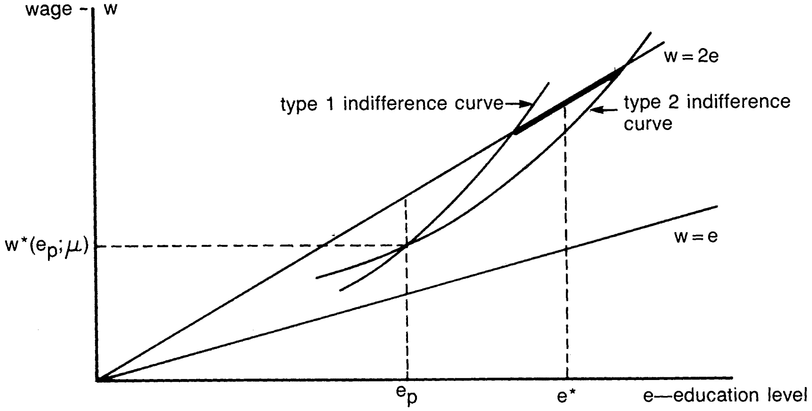

FIGURE V

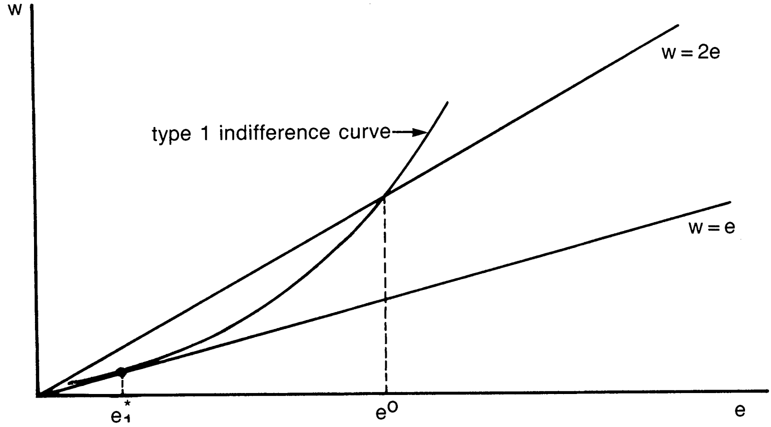

FIGURE VI

response from the firms) at such levels e than at his equilibrium value. As in Figure VI, let e^0 be the level of e where we get equality in (4). It will not be a screening equilibrium for type 2 to select an education level less than e^0, and the argument just given tells us that we cannot have a screening equilibrium that survives the Intuitive Criterion if we force 2 to some education level above e^0 other than his "constrained first best" from that set. Hence the Riley outcome is the only outcome that can survive.

Case B. More Than Two Types[11]

The Intuitive Criterion does not suffice to get us to the Riley outcome, if there are more than two types. Consider Figure VII, for three types. Here we have drawn an equilibrium outcome at which types 1 and 2 pool at education level e_p, and type 3 is screened at education level e_3. To break the pool, we would want (in the spirit of the previous arguments) to have type 2 offer an education level so high that type 1 would never do so in preference to the equilibrium. But since the firms could *conceivably* respond as if this out-of-equilibrium signal came from type 3, the education level needed to do this is at least e^0 (as shown). If type 2 picks a level a bit higher than this, he can be sure to get a wage of $2e$ or more. But this will not guarantee that he gets more from the defection than he gets from the equilibrium. Indeed, this equilibrium does survive the Intuitive Criterion (and equilibrium dominance).

[11] We are especially grateful to John Roberts, who pointed out an error in our earlier analysis of this case, and who helped us to find the correct argument.

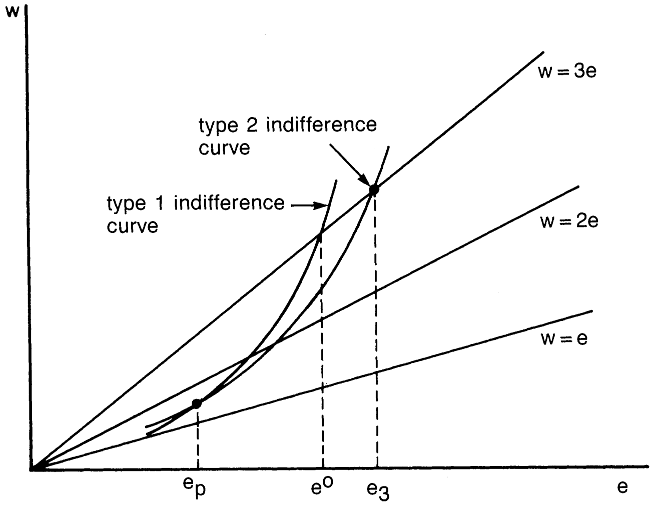

FIGURE VII

But it falls when the test constructed out of criterion D1 is applied. Indeed, all pooling equilibria fail this test. Suppose in some equilibrium that we had pooling of two types or more at an education level *e*. Let *n* index the highest type in the pool. Then at any education level above *e*, any response (wage) that a lower index type would prefer to the equilibrium, the higher index type would strictly prefer. Hence the equilibrium outcome would have to be supportable by beliefs in which the highest index type in any pool can, by choosing a slightly higher education level than the pooling level, be assured of a wage appropriate to (at least) his type. No pooling equilibrium can survive this. (The test built out of criterion D1 is powerful stuff indeed!)

Hence only screening equilibria can survive this test. And among those, only the Riley outcome will do so. We leave to the reader the task of showing that the only way that the Riley outcome can be missed in a screening equilibrium is if, for some two successive types *n* and *n* + 1, the picture is as in Figure VIII. But then for education levels above the point marked *e′*, the "better than equilibrium" set for *n* (and lower types) is strictly included in the same set for type *n* + 1. The D1 test requires that the outcome be supported by beliefs that put no weight on types *n* or less, which is manifestly impossible.

Does the Riley outcome survive the D1 (and even, the D2) test? We leave it to the reader to show that the answer is yes.

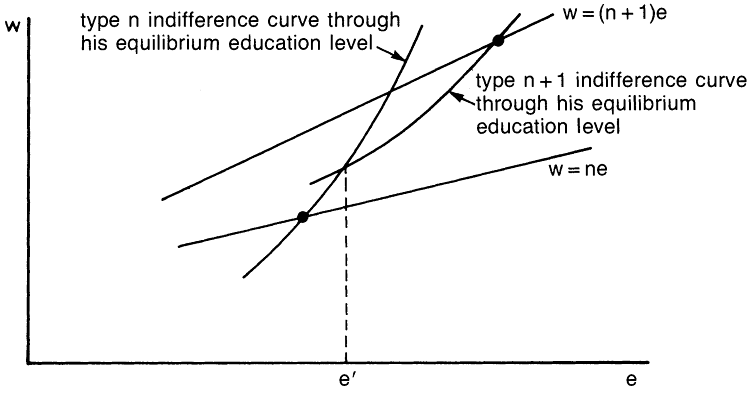
w
type n indifference curve through his equilibrium education level
w = (n + 1)e
type n + 1 indifference curve through his equilibrium education level
w = ne
e′
e

FIGURE VIII

Finally, we observe that the Intuitive Criterion would suffice if we modified the game form.[12] We have supposed that the worker obtains an education level, and then the firms bid for his services. Suppose that we modified things slightly, so that the worker obtains education, and then the worker proposes the wage that he wishes; the firms then (simultaneously and independently) signify whether they are willing to hire the worker at that wage. In this game, there are sequential equilibria in which the worker, in equilibrium, is paid *less* than his expected value to the firm. (The worker cannot ask for more because this would change beliefs.) But when the Intuitive Criterion is applied to this game, one can show that the unique equilibrium outcome that survives is the Riley outcome, no matter how many types there are (as long as the number is finite). This observation is especially pertinent when one thinks of applying this sort of criterion to alternating move bargaining games, as there the party who is "on the move" is allowed to propose an entire deal, which the other party must accept or reject. This gives the intuitive criterion (and similarly based tests) more bite in those games; cf. Admati and Perry [forthcoming].

VI. CONCLUDING REMARKS—THE FULL IMPLICATIONS OF STABILITY

The arguments just given are *not* meant to justify restriction to the Riley outcome in the Spence signaling model. In the first place,

[12] We thank Anat Admati for this observation.

the specific game form used is crucial.[13] But more importantly to this paper is that we do not mean to advocate all the tests we have described. The demonstration above shows the tests we have devised are very powerful in applications; perhaps too powerful. We have posed these tests in a general framework in order to provide a somewhat general typology of such tests, and to see them at work in examples. The reader must be the judge of which, if any, of them provide reasonable tests of equilibrium outcomes in particular manifestations in signaling games.

We also have sought to relate these tests to stability as it applies to signaling games. Since stability entails them all, if any of them is thought to be unintuitive, then the implications of stability cannot be accepted without some further thought. We ourselves find the D1 test very strong in the context of the Spence model, and we find "never a weak best response," at least as it applies to the example in subsection IV.5, to be downright unintuitive. But the reader should be warned that stability does not end with "never a weak best response." A general characterization of stability for the outcomes of (a generic class of) signaling games runs as follows.

Fix a signaling game and some equilibrium outcome. For each unsent message m, let Ψ_m denote the set of all pairs (μ,S), where μ is a probability distribution on $T(m)$ and S is a subset of $T(m)$ such that, at *some* sequential equilibrium with the given outcome, B's response $\phi(\cdot;m)$ to m satisfies

(i) $\phi(\cdot;m) \in MBR(\mu,m)$, and

(ii) $u^*(t) = \Sigma_r u(t,m,r)\phi(r;m)$ for all $t \in S$.

That is, at beliefs μ, B has an equilibrium response that makes m a weak best response for all the types in S simultaneously.

PROPOSITION 4. For a generic class of signaling games, an equilibrium outcome is stable if and only if: for every unsent message m and probability distribution θ over $T(m)$, there is some $(\mu,S) \in \Psi_m$ such that μ is in the convex hull of θ and the space of all probability distributions on S.

We do not attempt to prove this proposition here. It is far from trivial, and the reader should be warned that the generic class for

[13] Martin Hellwig [1985] has shown that, with a different game form, C. Wilson's reactive equilibrium outcome is the only outcome to satisfy the stability-like restrictions.

which it is true is smaller than the class of signaling games that have only a finite number of equilibrium outcomes.[14]

To see this proposition in action (and to see the strength of stability), consider a signaling game with three types, t_1,t_2,t_3, two messages, m' and m, and three responses to m, r_1,r_2,r_3. We shall examine the outcome in which all three types send m' with probability one, assuming that B has a unique best response which he chooses. The out-of-equilibrium data of the game are depicted in Figures IXa and IXb.

Figure IXa depicts the best responses of B to m as a function of his "posterior" assessment on the type of A. So, for example, if B is certain that the type is t_1, he chooses response r_1. If B is certain that the type is t_3, he chooses r_2. If he has an assessment that puts probability ½ on each of t_2 and t_3, he chooses r_3. Note the point ω; at these beliefs, B is indifferent between all three responses, and any mixed strategy is a best response.

Figure IXb depicts, as a function of B's response to m, which types (if any) of A would prefer m to m' (thus breaking the equilibrium outcome we are examining). So, for each n, if B responds to m with more than weight ⅔ on r_n, type T_n would prefer m to m'. At precisely weight ⅔ on r_n, type t_n is indifferent.

These data are consistent with the following assignment of payoffs: let all equilibrium payoffs if m' is sent be 0. In case the message is m, payoffs to A and B, depending on the type of A and the response by B, are given in Table II, with A's payoff first.

The payoffs are in "general position" as concerns the arguments we shall make—the same arguments could be made for all payoffs in some open neighborhood of the payoffs that give these data.

Note that it is indeed an equilibrium outcome for each type to send m'. This can be seen in Figure IXb, where we find a region (shaded) of responses by B at which no type strictly prefers to send m. Moreover, the outcome is sequential, since beliefs ω justify any response by B. And all the tests posed above are passed: every response to m is justified by some beliefs, and sending m is a weak best response for each type, at some equilibrium.

Yet the equilibrium outcome is not stable. Consider a perturbation of the game in which type t_1 sends m by "trembling" much more than do types t_2 and t_3. Unless something is done to change the beliefs of B, B will respond with r_1, which will break the equilibrium.

[14] Banks and Sobel [forthcoming], who arrived at this proposition independently, give a sketch of the proof.

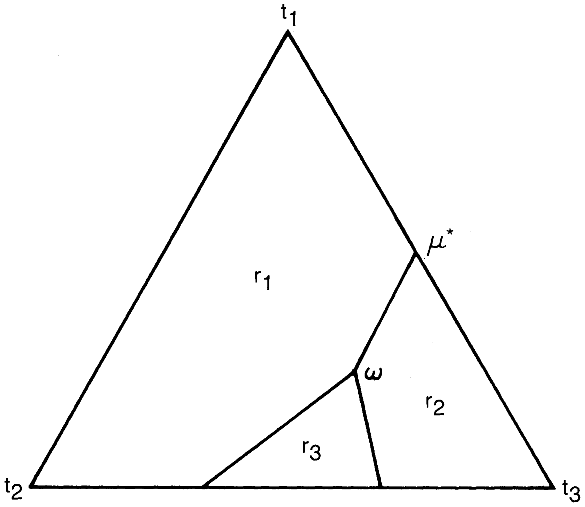

FIGURE IXA

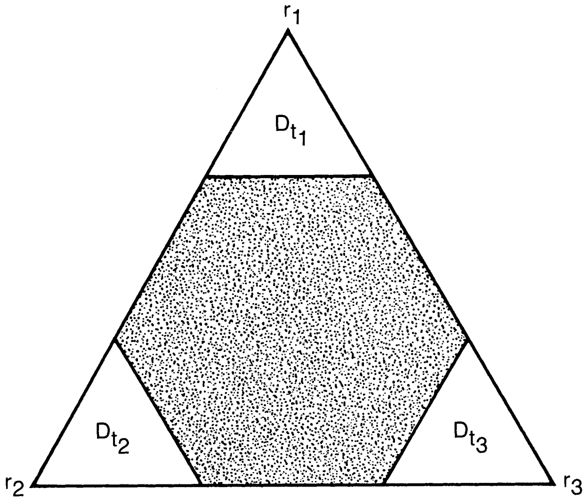

FIGURE IXB

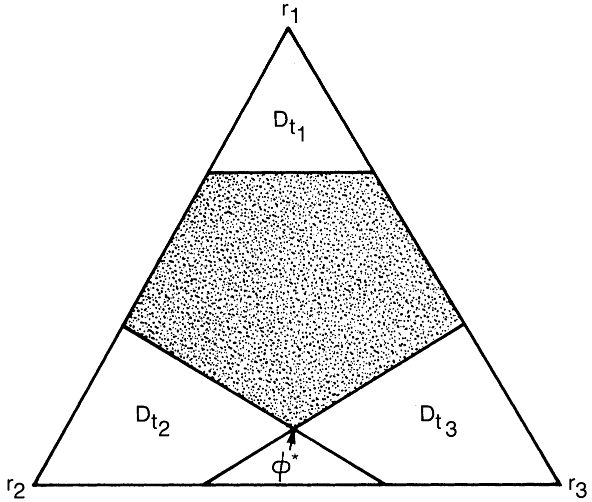

FIGURE IXc

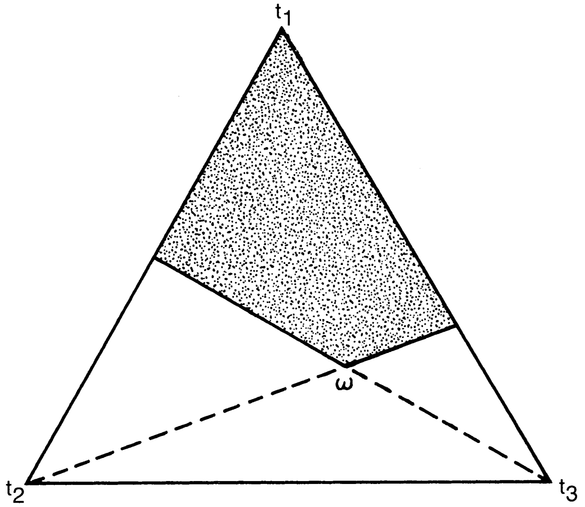

FIGURE IXd

TABLE II

PAYOFFS TO A AND B IF MESSAGE m IS SENT

	Type of A		
	t_1	t_2	t_3
Response of B			
r_1	1, 3	−2, 3	−2, 0
r_2	−2, 0	1, 0	−2, 3
r_3	−2, 0	−2, 2	1, 2

Now we can, by looking at equilibria where r_2 is played with probability ⅔, have t_2 indifferent between m and m', and so we could at such an equilibrium increase the posterior belief that m comes from t_2. But to get B to play r_2 with any probability at all, B must have posterior beliefs that put substantial weight (at least ½) on m coming from t_3. (Refer to Figure IXa and the region in which r_2 is a best response.) Alternatively, we can look at equilibria where r_3 is played with probability ⅔, and thus make t_3 indifferent. But to have B respond with positive probability on r_3, the posterior weight on t_2 must be at least ¼. And this would require a response that puts weight at least ⅔ on r_2. Because *there is no equilibrium* (at the given outcome) *at which both t_2 and t_3 are simultaneously indifferent between m* and *m'*, *it is impossible to increase simultaneously B*'s *posterior assessment that m comes from each. And to raise the probability of either, we need B* to *hold beliefs that put substantial weight on the other.* The equilibrium outcome is not stable.

In contrast, if the data were consistent with Figure IXc instead of IXb, there would be an equilibrium response (namely, the point marked ϕ^* in IXc) at which both t_2 and t_3 are indifferent between m and m', which would allow us to move from trembles that put most of the weight on t_1 to the point ω, which in turn supports the response ϕ^*. (Payoffs for A that are consistent with IXc are easy to compute.)

In terms of the proposition let $\theta = (1,0,0)$ (where the vector refers to the probability of the types in order). With the data of IXb, the candidates for sets S in Ψ_m are $\{t_1\}$, $\{t_2\}$, and $\{t_3\}$. Hence we can only "pull" θ along the two faces of the simplex. Pulling in the direction of t_2 is clearly useless, as this will not change B's response at all. It is possible, moving from θ in the direction of t_3, to reach beliefs that are equilibrium beliefs—namely those that are labeled μ^*. But the set of equilibrium responses that go with the beliefs μ^*

includes none that make t_3 indifferent. If, on the other hand, the data were as in IXc, then the pair $(\omega,\{t_2,t_3\})$ is in Ψ_m. Thus, with the equilibrium response ϕ^*, we can "stabilize" any initial perturbation θ such that ω is in the convex hull of θ and the face of types t_2 and t_3. That is, all perturbations θ in the shaded region of Figure IXd are stabilized by ω and ϕ^*, including (1,0,0)—the reader can verify that every other initial perturbation can be stabilized at some other equilibrium with the m' outcome.

The characterization given in Proposition 4 shows that stability (for generic signaling games) entails two considerations that our earlier criteria did not. First, one must consider for which subsets of types it is possible to find an equilibrium at which all types in the subset are indifferent between the equilibrium and some out-of-equilibrium message. Second (and less apparent from our example) is that perturbations that can be stabilized at a particular equilibrium depend on "direction"—one projects from the face of indifferent types, past beliefs that support the equilibrium response, to find what perturbations are stabilized at the given equilibrium. (If this second consideration is not clear, it should provide the reader with sufficient motivation to consult Banks and Sobel [forthcoming], whose example of an unstable outcome that survives "never a weak best response" trades on this second consideration.)

We do not mean to say that the m' equilibrium outcome in the examples of Figure IX is not breakable by intuitive agreements. For example, the criterion proposed by Grossman and Perry [1986b] does break this equilibrium.[15] But their criterion works equally well if the data are given by IXa and IXc as if they are given by IXa and IXb, so stability makes a distinction here that we cannot motivate intuitively.[16] We conclude that, if there is an intuitive story to go with the full strength of stability, it is beyond our powers to offer it here.

UNIVERSITY OF CHICAGO
STANFORD UNIVERSITY

REFERENCES

Admati, Anat, and Motty Perry, "Strategic Delay in Bargaining," Stanford University, *Review of Economic Studies,* forthcoming.

[15] See also Farrell [1985], although his analysis takes place in a setting in which messages are free.

[16] It is also worth noting that the example of subsection IV.5 is an unstable equilibrium that Grossman and Perry accept.

Aumann, Robert J., "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica,* LV (1987), 1–18.

Banks, Jeffrey S., and Joel Sobel, "Equilibrium Selection in Signaling Games," mimeo, U.C. San Diego, *Econometrica,* forthcoming.

Cho, In-koo, "A Refinement of Sequential Equilibrium," Princeton University, *Econometrica,* forthcoming.

——, "Refinement of Sequential Equilibrium: Theory and Application," Ph.D. thesis, Princeton University, 1986.

Crawford, Vince, and Joel Sobel, "Strategic Information Transmission," *Econometrica,* L (1982), 1431–51.

Demski, Joel, and David Sappington, "Delegated Expertise," mimeo, Yale University, 1986.

Farrell, Joseph, "Credible Neologisms in Games of Communication," mimeo, MIT, 1985.

Grossman, Sanford, "The Informational Role of Warranties and Private Disclosure about Product Quality," *Journal of Law and Economics,* (1981), 461–83.

——, and Motty Perry, "Sequential Bargaining under Asymmetric Information," *Journal of Economic Theory,* XXXIX (1986a), 120–54.

——, and ——, "Perfect Sequential Equilibrium," *Journal of Economic Theory,* XXXIX (1986b), 97–119.

Hellwig, Martin, private communication, 1985.

Kohlberg, Elon, and Jean-Francois Mertens, "On the Strategic Stability of Equilibria," *Econometrica,* LIV (1986), 1003–38.

Kreps, David M., "Nash Equilibrium," mimeo, Stanford University, *The New Palgrave,* forthcoming.

——, and Robert Wilson, "Sequential Equilibria," *Econometrica,* L (1982a), 863–94.

——, and ——, "Reputation and Imperfect Information," *Journal of Economic Theory,* XXVII (1982b), 253–79.

McLennan, Andrew, "Justifiable Beliefs in Sequential Equilibrium," *Econometrica,* LIII (1985), 889–904.

Milgrom, Paul, and John Roberts, "Limit Pricing and Entry under Incomplete Information: An Equilibrium Analysis," *Econometrica,* L (1982a), 443–59.

——, and ——, "Predation, Reputation and Entry Deterrence," *Journal of Economic Theory,* XXVII (1982b), 280–312.

——, and ——, "Price and Advertising Signals of Product Quality," *Journal of Political Economy,* XCIV (1986), 796–821.

Myerson, Roger, "Refinements of the Nash Equilibrium Concept," *International Journal of Game Theory,* VII (1978), 73–80.

——, "Mechanism Design by an Informed Principal," *Econometrica,* LI (1983), 1767–98.

Riley, John, "Informational Equilibrium," *Econometrica,* XLVII (1979), 331–59.

Rothschild, Michael, and Joseph E. Stiglitz, "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," this *Journal,* LXXX (1976), 629–49.

Rubinstein, Ariel, "A Bargaining Model with Incomplete Information about Preferences," *Econometrica,* LIII (1985), 1151–72.

Selten, Reinhard, "Spieltheoretische Behandlung Eines Oligopolmodells mit Nachfragetragheit," *Zeitschrift fur die Gesamte Staatswissenschaft,* CXXI (1965), 301–24.

——, "A Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory,* IV (1975), 25–55.

Spence, A. Michael, *Market Signaling* (Cambridge, MA: Harvard University Press, 1974).