Equilibrium Selection in Signaling Games

Jeffrey S. Banks; Joel Sobel

Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28198705%2955%3A3%3C647%3AESISG%3E2.0.CO%3B2-T

Econometrica is currently published by The Econometric Society.

EQUILIBRIUM SELECTION IN SIGNALING GAMES

By Jeffrey S. Banks and Joel Sobel[1]

This paper studies the sequential equilibria of signaling games. It introduces a new solution concept, divine equilibrium, that refines the set of sequential equilibria by requiring that off-the-equilibrium-path beliefs satisfy an additional restriction. This restriction rules out implausible sequential equilibria in many examples. We show that divine equilibria exist by demonstrating that a sequential equilibrium that fails to be divine cannot be in a stable component. However, the stable component of signaling games is typically smaller than the set of divine equilibria. We demonstrate this fact through examples. We also present a characterization of the stable equilibria in generic signaling games.

KEYWORDS: Strategic stability, equilibrium selection, signaling, game theory.

1. INTRODUCTION

THIS PAPER INVESTIGATES the relationship between Kreps and Wilson's (1982) concept of sequential equilibria and Kohlberg and Mertens's (1986) concept of stability. It introduces a restriction on off-the-equilibrium-path beliefs that refines the set of sequential equilibria in signaling games. We call all sequential equilibria that satisfy our restriction on beliefs divine. For generic signaling games, every equilibrium contained in a stable component is divine. Moreover, the solution concept is restrictive enough to rule out all of the equilibria that Kreps (1985)[2] and others dismiss on intuitive grounds. Thus, divinity provides an independent theoretical foundation for discarding nonintuitive equilibria in signaling games.

We provide a generic example to show that divine equilibria may not be contained in any stable component. However, the paper presents an explicit characterization of stability in terms of off-the-equilibrium-path beliefs. That is, an equilibrium of a generic signaling game is in a stable component if and only if it can be supported as a sequential equilibrium with restricted off-the-equilibrium-path beliefs. Just as Kreps and Wilson (1982) characterize perfect equilibria for generic extensive-form games in terms of sequential equilibrium strategies and beliefs, our result characterizes stable outcomes for generic signaling games in terms of sequential equilibrium strategies and restriction on beliefs. The characterization may be a useful way to compute stable equilibrium outcomes and to evaluate the consequences of using stability to select equilibria in extensive-form games.

Independent of our work, Cho and Kreps (1987) analyze the power of stability to select equilibria in signaling games. Their results closely parallel our own. They identify restrictions on equilibria similar to those embodied by divinity. In

addition, they also state our characterization result (Theorem 3). Cho (1987) extends a restriction identified in Cho and Kreps to obtain a solution concept that refines the set of sequential equilibria in general extensive-form games.

Our debt to the existing literature on solution concepts for noncooperative games is obvious. Recent work on this topic includes papers by Selten (1975), Kreps and Wilson (1982), and McLennan (1985), who present refinement concepts for extensive-form games; and Myerson (1978), Kalai and Samet (1984), and Kohlberg and Mertens (1986), who present refinement concepts for normal-form games.

2. THE MODEL

In this paper we analyze the equilibria of signaling games with finite action sets. There are two players, a Sender (S) and a Receiver (R). The Sender has private information, summarized by his type, t, an element of a finite set T. There is a strictly positive probability distribution $p(t)$ on T; $p(t)$, which is common knowledge, is the ex ante probability that S's type is t. After S learns his type he sends a message, m, to R; m is an element of a finite set M. In response to m, R selects an action, a, from a finite set $A(m)$; $k(m)$ is the cardinality of $A(m)$. S and R have von Neumann–Morgenstern utility functions $u(t, m, a)$ and $v(t, m, a)$, respectively.

For fixed T, M, and $A(m)$ for $m \in M$, the utility functions $u(t, m, a)$ and $v(t, m, a)$ completely determine the game. Therefore, if $L = [\bar{T} \times \sum_{i=1}^{\bar{M}} k(i)]^2$, where $\bar{T}$ is the cardinality of T and $\bar{M}$ is the cardinality of M, then every element of $\mathbb{R}^L$ determines a signaling game. We call a property of a signaling game *generic* if there exists $D \subset \mathbb{R}^L$ such that the property holds for all signaling games determined by $d \in D$ and a closed set of Lebesgue measure zero contains $\mathbb{R}^L \backslash D$. If a property of a signaling game is generic, then we say it holds for generic signaling games.

For any positive integer k, let $\Delta_k = \{\delta = (\delta(1), \ldots, \delta(k)): \delta(i) \geq 0 \; \forall i$ and $\sum_{i=1}^k \delta(i) = 1\}$ be the $(k-1)$-dimensional simplex. We refer to the $(\bar{T}-1)$-dimensional simplex most often; to simplify notation, we write Δ instead of $\Delta_{\bar{T}}$. A *signaling rule* for S is a function $q: T \to \Delta_{\bar{M}}$; $q(m|t)$ is the probability that S sends the message m, given that his type is t. An *action rule* for R is an element of $\prod_{m \in M} \Delta_{k(m)}$; $r(a|m)$ is the probability that R uses the pure strategy a when he receives the message m.

We extend the utility functions u and v to the strategy spaces $\Delta_{k(m)}$ by taking expected values; for all $t \in T$, let

$$u(t, m, r) = \sum_{a \in A(m)} u(t, m, a) r(a \mid m),$$

$$v(t, m, r) = \sum_{a \in A(m)} v(t, m, a) r(a \mid m).$$

Also, for each $\lambda \in \Delta$ and $m \in M$ let

$$BR(\lambda, m) \equiv \arg \max_{r(m) \in \Delta_{k(m)}} \sum_{t \in T} v(t, m, r(m)) \lambda(t)$$

be the best-response correspondence for R and for $\Lambda \subset \Delta_{k(m)}$, let $BR(\Lambda, m) \equiv \bigcup_{\lambda \in \Lambda} BR(\lambda, m)$.

DEFINITION: A sequential equilibrium for a signaling game consists of signaling rules $q(t)$ for S, action rules $r(m)$ for R, and beliefs $\mu(\cdot \mid m) \in \Delta$ for R, such that (i) $\forall t \in T$, $q(m^* \mid t) > 0$ only if

$$u(t, m^*, r(m^*)) = \max_{m \in M} u(t, m, r(m));$$

(ii) $\forall m \in M$, $r(a^* \mid m) > 0$ only if

$$\sum_{t \in T} v(t, m, a^*)\mu(t \mid m) = \max_{a \in A(m)} \sum_{t \in T} v(t, m, a)\mu(t \mid m);$$

(iii) if $\sum_{t \in T} q(m \mid t)p(t) > 0$, then

$$\mu(t^* \mid m) = \frac{q(m \mid t^*)p(t^*)}{\sum_{t \in T} q(m \mid t)p(t)}.$$

In words, (i) states that $q(\cdot)$ maximizes S's expected utility, given R's strategy; (ii) states that $r(\cdot)$ maximizes R's expected utility, given beliefs $\mu(\cdot)$; and (iii) states that R's beliefs given S's strategy are rational in the sense that Bayes' Rule determines $\mu(t \mid m)$ whenever the probability that S sends m in equilibrium is positive. If $q(m \mid t) = 0$, for all $t \in T$, then sequential rationality does not determine $\mu(t \mid m)$. However, the refinement concept introduced in Section 3 restricts the values that these beliefs may take.

Next, we describe stable equilibria. Our introduction follows Cho and Kreps (1987). Fix a signaling game; let $\tilde{\rho} = (\tilde{\rho}_R, \tilde{\rho}_S)$ satisfy $0 < \tilde{\rho}_i < 1$, $i = R, S$, and let $\tilde{q}$ and $\tilde{r}$ be strategies for S and R respectively that satisfy $\tilde{q}(m \mid t) > 0$, $\forall m \in M$, $\forall t \in T$ and $\tilde{r}(a \mid m) > 0$, $\forall a \in A(m)$, $\forall m \in M$. A $(\tilde{\rho}, \tilde{q}, \tilde{r})$-perturbation of the original game is the signaling game in which, if the players choose strategies q and r from the original game, then the outcome is the outcome of the original game if the strategy chosen by S is $(1 - \tilde{\rho}_S)q + \tilde{\rho}_S\tilde{q}$ and the strategy chosen by R is $(1 - \tilde{\rho}_R)r + \tilde{\rho}_R\tilde{r}$. We refer to $(\tilde{\rho}, \tilde{q}, \tilde{r})$ as trembles. Let (q, r) be Nash equilibrium strategies for a perturbed game. If $q(m \mid t) > 0$, we say that a type t Sender *voluntarily* sends m and we say that R voluntarily uses the mixed strategy $r(m)$.

For a given signaling game, we call a subset C of the set of Nash equilibria *stable* if, for every $\varepsilon > 0$, there exists $\delta > 0$ such that every $(\tilde{\rho}, \tilde{q}, \tilde{r})$-perturbation of the original game with $0 < \tilde{\rho}_i < \delta$, $i = R, S$ has an equilibrium no more than ε from the set C.

DEFINITION: A *stable component* is a minimal (by set inclusion) stable set of equilibria.

Our analysis depends on several facts about extensive-form games and stable sets.[3] To state these facts, we need one more definition. Given an extensive-form

[3] Kreps and Wilson (1982) prove Proposition 1. Kohlberg and Mertens (1986) prove Propositions 1-3.

game, the strategy choices for the players induce a probability distribution over the endpoints of the game. We call this probability distribution the *outcome* of the game (associated with particular strategies).

PROPOSITION 1: *For generic extensive-form games, the set of Nash equilibrium outcomes is finite and all Nash equilibria within a given connected component induce the same outcome.*

PROPOSITION 2: *Every game has at least one stable component.*

PROPOSITION 3: *A stable set of equilibria for a given game remains a stable set for the game obtained by deleting a strategy that is not a weak best response against any equilibrium in the set.*

Therefore, in generic signaling games, there exists a stable set of equilibria with the property that every equilibrium in the set agrees along the equilibrium path; the equilibria may vary off the equilibrium path. A variety of off-the-equilibrium-path responses may be needed to guarantee that any perturbation of the game has an equilibrium path close to a particular equilibrium path. Therefore, a single equilibrium need not be a stable set. However, we use Proposition 1 to justify an abuse of terminology. We call an equilibrium *stable* if it agrees with an element of a stable component along the equilibrium path. In particular, in generic signaling games, if an equilibrium is stable, then every perturbation has an equilibrium with payoffs close to the original equilibrium payoffs.

3. DIVINE EQUILIBRIA

Previous refinements of the Nash equilibrium concept place rationality restrictions on zero-probability events. In particular, sequential rationality requires that players respond optimally to some consistent assessment of how the game has been played. These equilibrium concepts do not *require* a player to draw any conclusion when a zero-probability event takes place. That is, although the refinements concepts embodied in sequential rationality and perfectness require that equilibria of games induce equilibria on any continuation of the game, these concepts do not require that a player systematically draw an inference from an opponent's unexpected move. Nevertheless, in order to decide how to respond to an unexpected signal, R should evaluate the willingness of S-types to deviate from equilibrium, and then incorporate into his beliefs the information that deviations from equilibrium might reveal.

This section presents an equilibrium concept that refines the set of sequential equilibria in signaling games by placing restrictions on off-the-equilibrium-path beliefs. We begin by describing two restrictions on beliefs along with the intuition behind them, and then proceed to define an equilibrium concept that incorporates these restrictions.

The first intuitive restriction on beliefs that we discuss requires R's off-the-equilibrium-path beliefs to place zero probability on those Sender types who certainly lose from a defection. Formally, this condition requires that if a type t Sender receives utility $u^*(t)$ in equilibrium and $J = \{t: u^*(t) > u(t, m, r(m))$ for all $r(m) \in BR(\Delta, m)\}$, then $r^*(m) \in BR(\Delta_{T \setminus J}, m)$.[4] Cho and Kreps (1987) also identify this condition and show that if an equilibrium is stable, then the condition must hold.[5] Our refinement notion includes this type of restriction on beliefs.

Figure 1[6] describes a special case of a sequential settlement game (see Salant (1984) or Sobel (1986)). There are two types of S (the "defendant"): type t_2 defendants are negligent; type t_1 defendants are not negligent. S offers a low settlement, m_1, or a high settlement, m_2. R (the "plaintiff") either accepts (a_1) or rejects (a_2) the offer. If R accepts S's offer, S pays R an amount that depends only on the offer. If R rejects the offer, S must pay court costs and a transfer depending only on his type (e.g. the court finds out with certainty whether or not S was negligent). If $p(t_1) = p(t_2) = \frac{1}{2}$, then the game depicted in Figure 1 has two types of equilibria. In one type of equilibrium, both types of S offer m_1, and R accepts any offer; $q(m_1 \mid t_i) = 1$, $i = 1, 2$, $r(a_1 \mid m_j) = 1$, $j = 1, 2$. In the other type of equilibrium, both types of S offer m_2 and R accepts m_2 and rejects m_1; $q(m_1 \mid t_i) = 0$, $i = 1, 2$, $r(a_1 \mid m_1) = 0$, $r(a_1 \mid m_2) = 1$. In order to support this behavior, we need $\mu(t_1 \mid m_1) \leq \frac{2}{5}$. We claim that the second equilibrium is not plausible because, in order to support it, R must believe that t_2 is more likely than t_1 to offer m_1. However, t_1 prefers to defect whenever t_2 does (and not conversely: consider an equal mixture of a_1 and a_2 given m_1). Thus, a reasonable restriction on beliefs would require that the relative probability of t_1 should increase if R observes m_1. Our refinement notion captures this argument as well.

Fix an equilibrium in which a Sender of type t obtains utility $u^*(t)$, and, for all $t \in T$, the probability that t sends m is zero. We intend to restrict the beliefs that R can have given the message m. Since we deal with only one unsent message at a time, for notational convenience we drop the argument m from R's response function.

Recall that $\Delta_{k(m)}$ consists of all actions, r, available to R given m. Let

$$A_G = \{r \in \Delta_{k(m)}: u(t, m, r) \geq u^*(t), \text{ for some } t \in T\}$$

m_1	a_1	a_2		m_2	a_1	a_2
t_1	$-3, 3$	$-6, 0$		t_1	$-5, 5$	$-6, 0$
t_2	$-3, 3$	$-11, 5$		t_2	$-5, 5$	$-11, 5$

FIGURE 1.

[4] If $J = T$, then no action R can take in response to the signal m induces S to send m. In this case, any beliefs are permissible.

[5] Kreps (1985) suggests a less restrictive version of this condition. Kreps discards an equilibrium in which there exists a Sender type who would like to defect for every action in $BR(\Delta_{T \setminus J}, m)$.

[6] We represent examples with a bi-matrix $B(m)$ for each $m \in M$. There is one column in $B(m)$ for each strategy in $A(m)$ and one row for each type. The entry in the tth row and the ath column is $(u(t, m, a), v(t, m, a))$. In each of these examples, the qualitative properties that we discuss in the text remain valid if we perturb the entries in $B(m)$.

be the set of actions that some S-type weakly prefers to equilibrium actions, conditional on sending m. Our initial restriction is that R should believe that any type who sends m instead of the equilibrium signal does not expect to lose by doing so.[7] Thus, if R receives the signal m (as a defection from equilibrium), he should believe that S expects him to take an action in A_G.

For all $r \in \Delta_{k(m)}$, let

$$\bar{\mu}(t, r) = \begin{cases} 1 & \text{if } u(t, m, r) > u^*(t), \\ [0, 1] & \text{if } u(t, m, r) = u^*(t), \\ 0 & \text{if } u(t, m, r) < u^*(t), \end{cases}$$

be the frequency that $t \in T$ would send m if he believes that m would induce the action r and t had a choice between sending m or obtaining $u^*(t)$. Next, let

$$\Gamma(r) = \{\gamma \in \Delta : \exists \mu(t) \in \bar{\mu}(t, r) \text{ and } c > 0 \text{ such that}$$

$$\gamma(t) = c\mu(t)p(t), \forall t \in T\}.$$

Notice that $\Gamma(r)$ is nonempty if and only if $r \in A_G$. If it is common knowledge that m induces r, then the posterior probability distribution over T must be an element of $\Gamma(r)$. Thus, $\Gamma(r)$ is the set of beliefs consistent with R taking the action r in response to m (and t earning $u^*(t)$ otherwise).

Finally, let

$$\bar{\Gamma}(A) = \text{convex hull} \left[\bigcup_{r \in A} \Gamma(r) \right].$$

Thus, if A is closed, then $\bar{\Gamma}(A)$ is a closed, convex subset of the simplex Δ, and is empty if and only if $A_G \cap A$ is empty. Since $\bar{\Gamma}(\Delta_{k(m)})$ is empty only if $u^*(t) > u(t, m, r)$, $\forall t \in T$, $\forall r \in \Delta_{k(m)}$, R truly would be surprised by a defection from equilibrium, and there seems to be no reason to select one inference over another in response to m. Indeed, in this case, *any* conjecture supports the equilibrium. When $A_G \neq \varnothing$, and hence $\bar{\Gamma}(\Delta_{k(m)}) \neq \varnothing$, we think that it is not plausible for R to hold beliefs outside of $\bar{\Gamma}(\Delta_{k(m)})$ given the signal m. If R observes a defection from the equilibrium path, then he must form a conjecture over T *based on that defection*.

Notice that any equilibrium in which beliefs lie in $\bar{\Gamma}(\Delta_{k(m)})$ satisfies the intuitive restrictions that we described earlier. All conjectures in $\bar{\Gamma}(\Delta_{k(m)})$ assign zero probability to any $t \in T$ with $u(t, m, r) < u^*(t)$, $\forall r \in \Delta_{k(m)}$. Furthermore, if there exists $t, t' \in T$ such that $\bar{\mu}(t, r) = 1$ implies $\bar{\mu}(t', r) = 1$, $\forall r \in \Delta_{k(m)}$, then for all beliefs in $\bar{\Gamma}(\Delta_{k(m)})$, the ratio of the probability of t' given m to the probability of t given m is at least as great as $p(t')/p(t)$. That is, R believes that t' is at least as likely to defect as t.

Beliefs must lie in $\bar{\Gamma}(\Delta_{k(m)})$ provided two conditions hold. First, R believes that no type t would use m if t expected R to take an action that resulted in

---

[7] It does not change our results to require that R believes that any type who sends m instead of the equilibrium signal expects to benefit strictly by doing so. Thus, we can use a strong inequality in the definition of A_G.

utility less than $u^*(t)$. This means that S expects R to take actions in A_G given the signal m. Second, S-types have a common conjecture over the distribution of actions that R would take as a response to a defection. This second condition may seem odd, since there is only one Sender. However, a "type" is a specification of the information S has concerning decision parameters that are not common knowledge. Thus, it is possible for two S-types to have different conjectures over R's actions in equilibrium. If it is common knowledge that R holds beliefs in $\bar{\Gamma}(\Delta_{k(m)})$, then S should expect m to induce an action in $BR(\bar{\Gamma}(\Delta_{k(m)}), m)$. This observation suggests the following iterative procedure. Let

$$\Gamma_0 = \Delta, \quad A_0 = \Delta_{k(m)}, \quad \text{and for } n > 0,$$

$$\Gamma_n = \begin{cases} \bar{\Gamma}(A_{n-1}) & \text{if } \bar{\Gamma}(A_{n-1}) \neq \varnothing, \\ \Gamma_{n-1} & \text{if } \bar{\Gamma}(A_{n-1}) = \varnothing, \end{cases}$$

$$A_n \equiv BR(\Gamma_n, m), \quad \Gamma^* = \bigcap_n \Gamma_n, \quad A^* = \bigcap_n A_n.$$

Others use iterative procedures in the definition of equilibrium concepts. Specifically, given the assumptions that S expects R to take actions in A_G given an unexpected signal m and that S-types have a common conjecture over the actions that R would take in response to m, our iterative procedure coincides with that used by Bernheim (1984) and Pearce (1984) to define the set of rationalizable equilibria.

THEOREM 1: *In generic signaling games, if an equilibrium in which* $q(m|t) = 0 \; \forall t \in T$ *is stable, then there exists* $r^* \in A^*$ *such that* $u(t, m, r^*) \leqslant u^*(t), \; \forall t \in T$.

Theorem 1 is a direct consequence of Proposition 3. It states that if an equilibrium is stable, then there exist beliefs in Γ^* that support it. We discuss the proof later in this section.

DEFINITION: A sequential equilibrium in a signaling game is *divine* if it is supported by beliefs in Γ^*.

Thus, by Theorem 1, every stable component contains a divine equilibrium. Therefore, Proposition 2 implies our next result.[8]

THEOREM 2: *Every signaling game has a divine equilibrium.*

We believe that divinity captures a minimal restriction on off-the-equilibrium path beliefs. Stability implies much more, but we are not convinced that these restrictions are plausible.

The set of beliefs in Γ^* depend on the prior distribution of Sender types. To check this property, one need only note that in the game that Figure 1 describes,

$$\Gamma^* = \{\lambda \in \Delta : \lambda(t_1) \geqslant p(t_1)\}$$

for the equilibrium in which both t_1 and t_2 send m_2 with probability one. Let

---

[8] Strictly speaking, Theorem 1 and Proposition 3 imply the existence of divine equilibria in generic signaling games. A limiting argument, based on the upper hemi-continuity of divine equilibrium paths, establishes Theorem 2. Cho (1987) gives the details of a related argument.

Γ^{**} be the intersection of the Γ^* taken over all nondegenerate priors on Sender types. We can show that in generic signaling games, if an equilibrium is stable, then it can be supported by beliefs in Γ^{**}. Call an equilibrium supported by beliefs in Γ^{**} *universally divine*. To see that universal divinity is more restrictive than divinity alone, note that in Figure 1, the sequential equilibrium in which S sends m_2 with probability one is divine provided that $p(t_1) \leqslant \frac{2}{3}$, but it is never universally divine since, regardless of the prior probability that S is t_1, R must believe that the unexpected signal m_1 comes from t_1.[9]

Cho and Kreps use Proposition 3 to further refine the equilibrium set. Fix an equilibrium outcome and an unsent signal m. Proposition 3 guarantees that a stable outcome passes the never a weak best response criterion. That is, any stable outcome can be supported by beliefs that give no weight to any type t who for every sequential equilibrium giving rise to this outcome strictly prefers the equilibrium outcome to sending m[10] (if m is never a weak best response for all types, then the equilibrium payoffs strictly dominate any payoff S can obtain from a best response to m). This condition is more restrictive than universal divinity because for generic signaling games if m is ever weak best response for t, then $e(t)$, the element of Δ with tth component equal to one, is an element of Γ^{**}.[11] Thus, Proposition 3 also implies that in generic signaling games, if an equilibrium is stable, then there exist beliefs in Γ^{**} that support it. Since $\Gamma^{**} \subset \Gamma^*$, Theorem 1 follows from Proposition 3.

4. A CHARACTERIZATION OF STABLE EQUILIBRIA

This section gives necessary and sufficient conditions for a sequential equilibrium in a generic signaling game to be stable. First, we present an example of

---

[9] Harris and Raviv (1985) study a game in which there is a divine equilibrium that is not universally divine, hence not stable. Their comparative-statics analysis concentrates on the stable path.

[10] McLennan (1985) defines a refinement concept that is similar in spirit to this requirement. Specifically, call an action *useless* if it has a suboptimal payoff in every sequential equilibrium of a game (not just those equilibria in a stable component). McLennan shows that there exist sequential equilibria with beliefs restricted so that, at each information set, they assign positive probability only to nodes reached by the fewest useless actions. From this, McLennan recursively defines higher-order uselessness and arrives at a set of *justifiable* equilibria. In generic signaling games, only strongly dominated actions are useless; thus any divine equilibrium is justifiable.

[11] This condition is strictly more restrictive than universal divinity. In the game described in Figure 2, there is a sequential equilibrium in which both S types send m_1 with probability one and R takes a_3 given m_2. It is straightforward to check that $\Gamma^{**} = \Delta$. However, the never a weak best response criterion demands that R believe only t_1 would send m_2. Hence R must respond to m_2 with a_1. As $u(t_2, m_2, a_1) = 1 > 0$, the outcome is not stable. Cho and Kreps observe that in this example, the never a weak best response criterion demands that R believe that t_1 sends the unexpected signal m_2, but the outcome in which both S types send m_1 fails to be stable because t_2 wishes to defect given $\mu(t_1 | m_2) = 1$.

m_1	a		m_2	a_1	a_2	a_3	a_4
t_1	0, 0		t_1	$-1, 3$	$1, 2$	$-1, 0$	$1, -2$
t_2	0, 0		t_2	$1, -2$	$1, 0$	$-2, 2$	$-1, 3$

FIGURE 2.

m_1	a		m_2	a_1	a_2	a_3	a_4
t_1	0, 0		t_1	$-1, 3$	$-1, 2$	1, 0	$-1, -2$
t_2	0, 0		t_2	$-1, -2$	1, 0	1, 2	$-2, 3$

FIGURE 3.

a signaling game that has an unstable, divine equilibrium. The example motivates the notion of stable beliefs that we need to prove our equivalence theorem.

Consider the signaling game in Figure 3. Let $p(t_1) = .4$. There exists a sequential equilibrium to this game in which $q(m_1 | t_i) = 1$, $i = 1, 2$, $r(a_1 | m_2) = 1$ supported by beliefs $\mu(t_1 | m_2) \geq \frac{2}{3}$. This equilibrium is universally divine since

$$\Gamma^* = \Gamma^{**} = \Delta$$

and $a_1 \in BR(\Gamma^*, m_2)$. However, this equilibrium is not stable.

The stable equilibrium for this example involves both t_1 and t_2 sending m_2 with probability one and R responding to m_2 with action a_3.

In this game R's best responses to m_2 consist of all four pure strategies and mixtures between a_i and a_{i+1} for $i = 1, 2, 3$. Figure 4 plots the expected utility for both types of Sender given signal m_2. The horizontal axis represents R's response; points between two pure strategies represent mixed-strategy responses. (Figure 4 shows that if S voluntarily sends m_2 in an equilibrium to the perturbed game in which S expects to receive 0, then R must respond to m_2 with an equal mixture of a_1 and a_2 or an equal mixture of a_3 and a_4.)

If the equilibrium in which S sends m_1 with probability one were part of a stable component, then any perturbed game would have an equilibrium that yields both S types utility zero. Therefore, there must exist an equilibrium to the
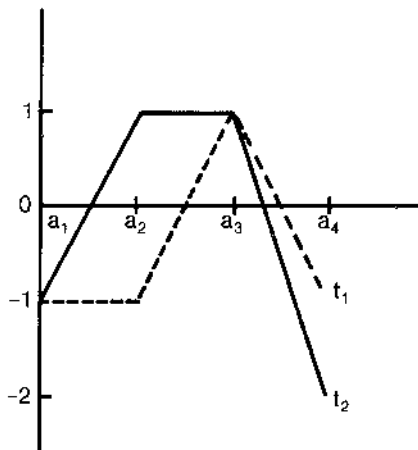

FIGURE 4.

perturbed game in which R believes that the probability that t_1 sent m_2 is either less than or equal to one-third (so that R weakly prefers a_4) or greater than or equal to two-thirds (so that R weakly prefers a_1). Moreover, t_1 voluntarily sends m_2 only if R responds with a mixture of a_3 and a_4 that provides utility 0 to t_1. Thus, t_1 voluntarily sends m_2 only if $\mu(t_1 | m_2) = \frac{1}{3}$ in equilibrium. Similarly, t_2 voluntarily sends m_2 only if $\mu(t_1 | m_2) = \frac{2}{3}$ in equilibrium. Consequently, if we select a perturbation that induces R to select a_3 given m_2 if S does not voluntarily send m_2, then this game cannot have equilibrium that yields both S types 0 payoffs. With no voluntary use of m_2, R believes that the probability of t_1 given m_2 is between $\frac{1}{2}$ and $\frac{2}{3}$. Thus, R would take a_3 given m_2 and destroy the equilibrium. R changes his belief only if S voluntarily sends m_2. However, if t_1 voluntarily sends m_2, then $\mu(t_1 | m_2)$ would exceed $\frac{1}{2}$. Since $\mu(t_1 | m_2) = \frac{1}{3}$ in any sequential equilibrium in which S receives 0 and t_1 voluntarily sends m_2, t_1 does not voluntarily send m_2. Similarly, t_2 does not voluntarily send m_2. This argument establishes that the sequential equilibrium in which S sends m_1 with probability one is not stable.

We next present a characterization of stable outcomes in generic signaling games. Fix an equilibrium that leads to utility levels $u^*(t)$ for all $t \in T$ and in which $q(m | t) = 0$ for all $t \in T$. We identify the set of actions that R could take given m in an equilibrium to a perturbed game in which S's payoffs are $u^*(t)$. In the previous example, this set consists of a_1, a_4, mixtures of a_1 and a_2 that place more weight on a_1 than on a_2, and mixtures of a_3 and a_4 that place more weight on a_4 than on a_3. It is useful to divide these actions into different sets, $I(J)$, $J \subset T$.

$$I(J) \equiv \{r \in \Delta_{k(m)} : u^*(t) \geq u(t, m, r) \ \forall t \in T, \text{ and}$$

$$u^*(t) = u(t, m, r) \text{ if } t \in J\}.$$

$I(J)$ contains those actions in which S types in J are indifferent between sending m or following the equilibrium path.

Fix a perturbation that leads to a belief μ given m provided that S does not voluntarily send m. We wish to find conditions under which there exists an equilibrium to the perturbed game close to the original equilibrium. Doing this is easy if $BR(\mu, m) \cap I(J) \neq \emptyset$ for some J; in this case S need not voluntarily send m. Otherwise some non-empty set of S types J must voluntarily send m and therefore R must respond to m with an action $r \in I(J)$. If S types in J voluntarily send m, then R's equilibrium belief given m will be a convex combination of μ and $\{e(t)\}_{t \in J}$ where $e(t) \in \Delta$ is the vector with tth component equal to one and all other components equal to zero. Thus,

$$\hat{A}(J, r) \equiv \left\{ \lambda \in \text{int } \Delta : \exists \lambda^* \in \Delta \text{ with } r \in BR(\lambda^*, m) \text{ such that} \right.$$

$$\left. \lambda^* = \sum_{t \in J} \alpha(t)e(t) + \beta\lambda, \text{ for } \alpha(t) \geq 0, 1 - \sum_{t \in J} \alpha(t) \equiv \beta > 0 \right\},$$

is the set of beliefs λ that cannot be "stabilized" through voluntary action by

types in J if R takes the action r. Finally, if

$$A(J) = \begin{cases} \bigcap_{r \in I(J)} \hat{A}(J, r) & \text{if } I(J) \neq \varnothing, \\ \Delta & \text{if } I(J) = \varnothing, \end{cases}$$

and $A^* = \bigcap_{J \subset T} A(J)$, then A^* consists of all of the perturbations that can be stabilized. This discussion motivates our characterization theorem.

THEOREM 3: *In generic signaling games, an equilibrium is stable if and only if, for all unused signals m, $A^* = \varnothing$.*

Cho and Kreps (1987) obtain the same result. The Appendix contains a proof of Theorem 3.

5. EXTENSIONS

While we confine our discussion in this paper to signaling games, Propositions 1–3 hold for generic extensive-form games. Since these results combine to imply Theorems 1 and 2, we can use our techniques to rule out implausible sequential equilibria in more general extensive-form games. We suspect that divinity is easier to verify than stability and may be simpler to generalize to games with infinite strategy spaces. On the other hand, Theorem 3 and possible generalizations appear to be valuable only as a characterization of stable equilibria.

We conclude by noting that our techniques do not refine the set of sequential equilibria in signaling games in which signals are costless. Specifically, let $A(m)$, $u(t, m, a)$, and $v(t, m, a)$ be independent of m. These games are not generic, so we cannot apply our results directly. However, it is easy to verify that $\Gamma^* = \Delta$ for any unused signal. This is because if t induces the action $a \in \bar{A}$ with signal m', then there exist beliefs for which a is a best response to the (unused) signal m. When signaling is costless, t is indifferent between sending m and m' and no other agent strictly prefers m to his equilibrium payoff. In addition, straightforward arguments show that stability does not restrict the set of equilibria, although this kind of game always has an equilibrium in which all types of S send the same signal and typically has other, more appealing, equilibria. Farrell[12] (1984) and Myerson (1983) present ideas that apply to costless signaling games. Myerson presents an axiomatic solution that limits the outcomes in a mechanism-design problem that usually has a large number of sequential equilibria, but it is not clear that his ideas extend in a sensible way to a noncooperative framework. Farrell argues that an equilibrium outcome is not plausible if there exists an unused signal m, a nonempty set J, and an action $r \in BR(\lambda, m)$ such that

$$J = \{t: u^*(t) < u(t, m, r)\}, \quad \text{where}$$

---

[12] Grossman and Perry's (1986) concept of perfect sequential equilibrium is similar to Farrell's concept. However, the perfect sequential equilibrium concept does not refine the equilibrium set in games with costless signaling.

$$\lambda(t) = \begin{cases} p(t) \Big/ \sum_{t' \in J} p(t') & \text{if } t \in J, \\ 0 & \text{if } t \notin J, \end{cases}$$

is the conditional probability of t given $t \in J$.[13] That is, Farrell argues that R should interpret a defection that benefits exactly the set J as evidence that exactly those t in J use m. Farrell calls an equilibrium in which this type of defection does not exist neologism proof. Neologism-proof equilibria do not exist in general, and, in games with costly signaling, need not be divine.

Department of Economics, University of Rochester, Rochester, NY 14627, U.S.A.

and

Department of Economics, University of California—San Diego, La Jolla, CA 92093, U.S.A.

APPENDIX

We analyze an equilibrium in which a Sender of type t receives expected utility $u^*(t)$ and in which no Sender uses signal m with positive probability.

First we discuss our genericity assumptions. We deal only with signaling games for which Propositions 1 and 2 apply.

PROPERTY 1: In generic signaling games, an equilibrium in which all signals are sent with positive probability, taken by itself, is a stable component.

We have not seen a proof of Property 1. Nevertheless, Property 1 follows from standard dimension-counting arguments similar to those found in Kreps and Wilson (1982). Property 1 allows us to concentrate on unused signals when we characterize stable equilibria.

In our proofs we use several other properties. To state these properties precisely, we must state explicitly a regularity condition. Let $U(t, r; \xi(t)) \equiv u(t, m, r) - (u^*(t) - \xi(t))$ be the gain over $u^*(t) - \xi(t)$ to a Sender of type t if he sends m and R takes the action r in response to m. Think of $\xi(\cdot)$ as a perturbation in payoffs due to small trembles. For fixed t and $\xi(\cdot)$, viewed as a function of its second argument, $U(\cdot)$ is a linear function defined on $\Delta_{k(m)}$, the simplex of (mixed) strategies available to R given m. Let J be a nonempty subset of T with cardinality j; let F be an arbitrary j-dimensional face of $\Delta_{k(m)}$; and, for $\xi(\cdot)$ fixed, let $\hat{U}_J^F : F \to R^J$ be the restriction of $U(\cdot)$ to F for $t \in J$. Thus,

$$\hat{U}_J^F(r; \xi(t)) \equiv (U(t, r; \xi(t))), \quad t \in J.$$

$\hat{U}_J^F(\cdot)$ is linear in $r \in F$. The regularity condition that we need is:

(RC)   the derivative of $\hat{U}_J^F(r; 0)$ is nonsingular.

(RC) is true generically for fixed $u^*(t)$. Even though the signaling game itself determines $u^*(t)$, (RC) holds for almost every signaling game as well. This result follows because the set of equilibrium payoffs to S when no one uses m is generically finite and does not depend on $u(t, m, r)$.

We need (RC) in order to establish the next property. Property 2 states that small perturbations do not change Λ^* at a stable equilibrium in generic signaling games. To state this result, we need to introduce some notation.

Fix a $(\bar{p}, \bar{q}, \bar{r})$-perturbation and let $\hat{u}(t, m, r)$ be the utility t receives if he sends the signal m and R voluntarily takes the action r; let $\bar{u}^*(t)$ be an equilibrium utility to a t-Sender; and let $\hat{\Phi}$ be derived from the perturbed game as Φ is derived from the original game. $\hat{\Phi}$ depends on $(\bar{p}, \bar{q}, \bar{r})$ and $\bar{u}^*(t)$.

---

[13] We have modified Farrell's definition to accommodate situations in which $BR(\lambda, m)$ is not single valued.

PROPERTY 2: In generic signaling games, if an equilibrium is stable, then for any $\varepsilon > 0$ there exists a $\delta > 0$ such that there exists an equilibrium to any $(\tilde{p}, \tilde{q}, \tilde{r})$-perturbation with $0 < \tilde{p}_i < \delta$ for $i = R$ and S in which $\hat{\Lambda}^* = \Lambda^*$ and $|u^*(t) - \hat{u}^*(t)| < \varepsilon$.

We omit the straightforward proof of this fact.

For $q = (q(1), \ldots, q(T))$, $q > 0$, define $\mu(\cdot)$ by $\mu(q) \equiv (p(t)q(t)/\sum_{t' \in T} p(t')q(t'))$, $t \in T$. When $q(t)$ is the probability that t sends m, then $\mu(q)$ is R's posterior belief on t given m.

LEMMA 1: *In a generic signaling game, if there exists an equilibrium in which $\Lambda^* \neq \varnothing$, then the equilibrium is not stable.*

PROOF: Assume that the equilibrium is stable. We argue to a contradiction. Fix $\lambda^* \in \Lambda^*$. If the original equilibrium is stable, then one can find $\varepsilon > 0$ and $\delta > 0$ such that if any $(\tilde{p}, \tilde{q}, \tilde{r})$-perturbation with $0 < \tilde{p}_i < \delta$ for $i = R$ and S has an equilibrium such that

(1) $\qquad \Lambda^* = \tilde{\Lambda}^*$

and for every $r \in BR(\lambda^*, m)$, there exists $t(r) \in T$ such that

(2) $\qquad \hat{u}(t(r), m, r) > \hat{u}^*(t(r))$.

Property 2 implies (1) directly; since $\lambda^* \in \Lambda^*$ implies that for each $r \in BR(\lambda^*, m)$ there exists $t(r)$ such that $u(t(r), m, r) > u^*(t(r))$ we obtain (2) from Property 2 by taking ε and δ small enough. In the perturbation, pick $\hat{q}$ such that $\mu(\hat{q}) = \lambda^*$. We can find $\hat{q}$ because $\lambda^* \in \Lambda^* \subset \text{int } \Delta$. Consider an equilibrium to the $(\tilde{p}, \tilde{q}, \tilde{r})$-perturbation for which (1) and (2) hold. If $\tilde{r}$ is R's voluntary action given m, then it follows from (2) that $\tilde{r} \notin BR(\lambda^*, m)$. Therefore, S must voluntarily send m in the equilibrium to the perturbed game. In order for S to be optimizing, there must be a nonempty subset J of T such that

$$\hat{u}^*(t) = \hat{u}(t, m, \tilde{r}) \quad \text{if } t \in J \quad \text{and}$$

$$\hat{u}^*(t) > \hat{u}(t, m, \tilde{r}) \quad \text{if } t \notin J.$$

In the equilibrium to the perturbed game, a Sender of type t voluntarily sends m only if $t \in J$. Therefore, if $\hat{\lambda}(t)$ is R's posterior probability of t given m, then

$$\hat{\lambda}(t) = c(\tilde{\rho}_S p(t)\hat{q}(t \mid m) + (1 - \tilde{\rho}_S)p(t)q(t))$$

where $0 \leq q(t) \leq 1$, $q(t) = 0$ for $t \notin J$, and the condition $\hat{\lambda} \in \Delta$ determines the value of the positive constant c. If we let

$$\beta = c\tilde{\rho}_S \sum_{t \in J} p(t)\hat{q}(t) \quad \text{and} \quad \alpha(t) = c(1 - \tilde{\rho}_S)p(t)q(t) \quad \text{for } t \in J,$$

then $\alpha(t) \geq 0$, $\beta = 1 - \sum_{t \in J} \alpha(t) > 0$, and $\hat{\lambda} = \beta\lambda^* + \sum_{t \in J} \alpha(t)e(t)$. Hence, $\lambda^* \notin \bar{\bar{A}}(J, \tilde{r})$. This is a contradiction since by (1), $\lambda^* \in \Lambda^* = \tilde{\Lambda}^* \subset \bar{\bar{A}}(J, \tilde{r})$. Therefore, if the equilibrium is stable, then $\Lambda^* = \varnothing$.
$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ Q.E.D.

Now we show that if $\Lambda^* = \varnothing$ for all unused signals, then the equilibrium is stable. We establish this in two steps. First, we show that if $\Lambda^* = \varnothing$, then with an arbitrarily small amount of voluntary action, we can fix the off-the-equilibrium-path strategies so that they do not disrupt an equilibrium with payoffs close to $u^*(t)$. The second step of the argument shows that a tremble does not disrupt the equilibrium path.

LEMMA 2: *In a generic signaling game, if $\Lambda^* = \varnothing$, then there exists $\xi > 0$ such that for any $\eta > 0$, there exists $\rho > 0$ such that if $|\hat{u}^*(t) - u^*(t)| < \xi$ for all t and $0 < \tilde{\rho}_i < \rho$ for $i = R$ and S, then for any $(\tilde{p}, \hat{q}, \tilde{r})$-perturbation there exists $q(t)$, r, and $J \subset T$ such that*

(3) $\qquad \hat{u}(t, m, r) < \hat{u}^*(t) \quad \text{if } t \notin J,$

(4) $\qquad \hat{u}(t, m, r) = \hat{u}^*(t) \quad \text{if } t \in J,$

(5) $\qquad q(t) = 0 \quad \text{if } t \notin J,$

(6) $\qquad 0 \leq q(t) \leq \eta \quad \text{if } t \in J, \quad \text{and}$

(7) $\qquad r \in BR(\hat{\lambda}, m), \quad \text{where} \quad \hat{\lambda} \in \Delta \quad \text{and} \quad \hat{\lambda} = \mu[\tilde{\rho}_S \hat{q} + (1 - \tilde{\rho}_S)q] \quad \text{for some } c > 0.$

Lemma 2 states that if $\hat{u}^*(t)$ represent equilibrium payoffs to a perturbed game, then we can find voluntary strategies $q(t)$ and r such that r is a best response for R given m in the perturbed game (condition (7)), $q(t)$ is an equilibrium strategy for S in the sense that $q(t) > 0$ only if the Sender of type t expects a payoff of $\hat{u}^*(t)$ if he uses m (conditions (3), (4), and (5)), and that the probability that S voluntarily uses m is arbitrarily small.

PROOF: Pick $\xi > 0$ and $\rho_R > 0$ so that if $|\hat{u}^*(t) - u^*(t)| < \xi$ for all t and $0 < \hat{\rho}_R < \rho_R$, then for all $\hat{q}$ and $\hat{r}$, $\hat{A}^* = A^* = \varnothing$. This is possible by Property 2. Since $A^* = \varnothing$, there exists a nonempty set J, an action $\hat{r}$, and $\bar{\lambda} \in \Delta$ such that (3) and (4) hold, $\hat{r} \in BR(\bar{\lambda}, m)$, and

$$\bar{\lambda}(t) = \beta p(t)\hat{q}(t) + \alpha(t)$$

for some $\alpha(t)$ with $\alpha(t) \geqslant 0$, $\alpha(t) = 0$ for $t \notin J$, and $1 - \sum_{t \in J} \alpha(t) = \beta > 0$. Let $q(t) = 0$ for $t \notin J$, and, for $t \in J$, let

$$q(t) = \left[(\alpha(t)\hat{\rho}_S) \sum_{t' \in T} p(t')\hat{q}(t'|m) \right] \Big/ [\beta(1 - \hat{\rho}_S)p(t)].$$

This choice of $q(t)$ satisfies (5) and, since $\beta > 0$, $q(t) \geqslant 0$. A simple computation confirms that $\bar{\lambda} = \mu(t; \hat{\rho}_S\hat{q} + (1 - \hat{\rho}_S)q)$. Therefore, there is a $\rho_S > 0$ such that if $0 < \hat{\rho}_S < \rho_S$, then (6) holds. This establishes the lemma for $\rho = \min(\rho_R, \rho_S)$.   Q.E.D.

To complete the proof of Theorem 3, we construct an equilibrium to a perturbed game by piecing together strategies identified in Lemma 2, which we designed to be equilibrium behavior in the perturbed game off the path of the original equilibrium, with strategies that support an equilibrium to the perturbed game on the original equilibrium path. Therefore, we must restrict attention to only those signals used in the original equilibrium. For sufficiently small perturbations of this game, we can find an equilibrium close to the original equilibrium. Moreover, we can do this even if we save a small amount of S's strategy to control the off-the-equilibrium path trembles. This is a consequence of Property 1, but we must be a bit careful because the construction in Lemma 2 does not guarantee that each type of S voluntarily sends m with the same probability. We need to introduce another concept. For $0 \leqslant \nu(t) < 1$, a ν-pseudo-equilibrium to a signaling game consists of strategies q and r for S and R, respectively, such that

> $r(m) \in \Delta_{k(m)}$ is a best response to q given m,
>
> q is a best response to r,
>
> $q(m|t) \geqslant 0$ and $\sum_{m \in M} q(m|t) = 1 - \nu(t)$.

Thus, if $\nu(t) = 0$ for all t, then a ν-pseudo-equilibrium is an equilibrium. If $u^*(t)$ is the payoff to a Sender of type t in a pseudo-equilibrium, then

$$u^*(t) = \max_{m \in M} \sum_{a \in A(m)} u(t, m, a)r(a|m).$$

For our purposes, we obtain an equivalent concept if we allow the tremble $\hat{\rho}_S$ in the definition of stability to depend on the type of Sender. The next lemma follows from Property 1.

LEMMA 3: *In generic signaling games, for any equilibrium in which every signal is sent with positive probability and for every $\varepsilon > 0$, there exist $\rho > 0$ and $\eta > 0$ such that if $0 < \hat{\rho}_i < \rho$ for $i = R$ and S and if $0 \leqslant \nu(t) \leqslant \eta$ for any t, then any $(\hat{\rho}, \hat{q}, \hat{r})$-perturbation has a ν-pseudo-equilibrium within ε of the original equilibrium. Moreover, S's utility in the pseudo-equilibrium does not depend on ν.*

PROOF OF THEOREM 3: Lemma 1 proves one implication. To complete the proof, we must show that if $A^* = \varnothing$ for all unused signals m, then the equilibrium is stable. We assume that there is only one unused signal. This does not change the substance of the proof. We may take ε in the definition of stability to be less than ξ of Lemma 2. Pick η in Lemma 2 equal to the smaller of ε and η (corresponding to the given ε) in Lemma 3 and the δ in the definition of stability equal to the minimum of the ρ's from Lemma 2 and Lemma 3. For any $(\hat{\rho}, \hat{q}, \hat{r})$-perturbation with $0 < \hat{\rho}_i < \delta$ for $i = R$ and S, first compute the payoffs $\hat{u}^*(t)$ from Lemma 2. By construction, next we can apply Lemma 2 to obtain the perturbed-equilibrium voluntary strategies $q(m|t)$. The manner in which we selected δ and η guarantee that $0 \leqslant q(m|t) \leqslant \eta$. Consequently, we may apply Lemma 3 with $\nu(t) = q(m|t)$ in order to obtain strategies for S for signals used in the original equilibrium, that describe equilibrium behavior for S in the perturbed game. This construction provides an equilibrium to the perturbed game within ε of the original equilibrium.   Q.E.D.

REFERENCES

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

CHO, I. (1987): "A Refinement of the Sequential Equilibrium Concept," *Econometrica*, 55, forthcoming.

CHO, I., AND D. KREPS (1987): "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics*, 102, forthcoming.

FARRELL, J. (1984): "Credible Neologisms in Games of Communication," mimeo, MIT.

GROSSMAN, S., AND M. PERRY (1986): "Perfect Sequential Equilibria," *Journal of Economic Theory*, 39, 97-119.

HARRIS, M., AND A. RAVIV (1985): "A Sequential Signaling Model of Convertible Debt Call Policy," *Journal of Finance*, 40, 1263-1281.

KALAI, E., AND D. SAMET (1984): "Persistent Equilibria," *International Journal of Game Theory*, 13, 129-144.

KOHLBERG, E., AND J.-F. MERTENS (1986): "On the Strategic Stability of Equilibria," *Econometrica*, 54, 1003-1038.

KREPS, D. (1985): "Signaling Games and Stable Equilibria," mimeo, Stanford University.

KREPS, D., AND R. WILSON (1982): "Sequential Equilibria," *Econometrica*, 50, 863-894.

McLENNAN, A. (1985): "Justifiable Beliefs in Sequential Equilibrium," *Econometrica*, 53, 889-904.

MYERSON, R. (1978): "Refinement of the Nash Equilibrium Concept," *International Journal of Game Theory*, 7, 73-80.

——— (1983): "Mechanism Design by an Informed Principal," *Econometrica*, 51, 1767-1798.

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

SALANT, S. (1984): "Litigation of Settlements Demands Questioned by Bayesian Defendants," Caltech Social Science Working Paper 516.

SELTEN, R. (1975): "A Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 4, 25-55.

SOBEL, J. (1986): "An Analysis of Discovery Rules," UCSD Discussion Paper.