Credible Deviations from Signaling Equilibria*

(Final "Working Paper version")

Péter Eső† James Schummer‡

June 2008

Abstract

In games with costly signaling, some equilibria are vulnerable to deviations which could be "unambiguously" interpreted as coming from a unique set of Sender-types. This occurs when these types are precisely the ones who gain from deviating for any beliefs the Receiver could form over that set. We show that this idea characterizes a unique equilibrium outcome in two classes of games. First, in monotonic signaling games, only the Riley outcome is immune to this sort of deviation. Our result therefore provides a plausible story behind the selection made by Cho and Kreps' (1987) D1 criterion *on this class* of games. Second, we examine a version of Crawford and Sobel's (1982) model with costly signaling, where standard refinements have no effect. We show that only a Riley-like separating equilibrium is immune to these deviations.

Keywords: Signaling games, Sender-Receiver, robust equilibrium, refinements.

JEL classification numbers: C70, C72.

*We thank David Austen-Smith, Drew Fudenberg, Johannes Hörner, Navin Kartik, Marcin Peski, Phil Reny, Marciano Siniscalchi, Jeroen Swinkels, and anonymous reviewers for helpful comments. We are particularly indebted to Sidartha Gordon and Joel Sobel for their insights. We are also grateful to seminar participants at Columbia, Northwestern, Penn State, Rochester, and the 2006 N. American Econometric Society Meetings.

†Kellogg School of Management, MEDS Department, Northwestern University. Email: eso@kellogg.northwestern.edu.

‡Corresponding author. Kellogg School of Management, MEDS Department, Northwestern University. Email: schummer@kellogg.northwestern.edu.

1 Introduction

In games with asymmetric information, the standard notion of equilibrium requires that players' actions be best responses to their beliefs at each information set. On the equilibrium path, these beliefs must be consistent with Bayes' rule; *off* the equilibrium path this requirement does not apply. In Sender-Receiver games, out-of-equilibrium beliefs can be thought of as the Sender's hypothesis of what the Receiver would think upon observing a deviation. This hypothesis rationalizes the Sender's anticipation of what the Receiver would *do*, in turn justifying the Sender's decision not to deviate. In this sense, the concept of out-of-equilibrium *beliefs* should not be taken too literally.

Yet, equilibrium refinements often operate directly on the concept of beliefs. Rather than stopping at the point where these beliefs serve only to rationalize the Sender's actions, much work in this literature interprets beliefs more literally, prescribing exactly what those beliefs should be.

In this paper we show that strong predictions neither require, nor necessarily follow from, such specific impositions on beliefs. Rather than refining the set of admissible beliefs, we ask whether it is possible for the Sender to implicitly *signal a candidate set of deviating types, even if he cannot anticipate exactly which beliefs the Receiver would form over that set.* A Credible Deviation is one that uniquely and unambiguously identifies a set of types that gain from deviating, provided that the Sender anticipates the Receiver to form *some* beliefs over that set. We analyze the extent to which equilibria are immune to such deviations in costly-signaling games.

Our approach is comparatively agnostic in terms of what the Receiver believes when an out-of-equilibrium message is sent. This reflects our less literal interpretation of beliefs. While we do not advocate our concept as a predictive criterion on its own, we show that it describes an equilibrium vulnerability that is not captured by the standard refinements literature.

In Section 1.1 we motivate and explain the concept with an example. Formal definitions are given in Section 2, followed by a comparison with various concepts from the refinements literature in Section 2.2.

Our results concern two classes of signaling games. In Section 3 we show

that on the standard class of *monotonic signaling games*, only the so-called Riley equilibrium is immune to the Credible Deviations we describe. Therefore, on this particular class of games, there is a connection between the predictions made by standard refinements (D1, stability) and immunity to Credible Deviations. That is, our concept provides one behavioral motivation for selecting the stable outcome on this class.

In Section 4 we turn to signaling games whose structure is reminiscent of Crawford and Sobel's (1982) model: The Sender prefers the Receiver to choose an action closer to his actual type, while the Receiver wants to target that type offset by a *bias*. Unlike Crawford and Sobel, we model such a situation with *costly messages*, assuming a single-crossing condition. Because of the lack of monotonic structure in this class of games, the connection between other Credible Deviations and various refinements breaks down. Similar to the previous model, only a "Riley-like" equilibrium is free of Credible Deviations. Despite some similarity between this model and the previous one, however, standard refinements widely used in practice (e.g. D1, D2, Divinity, etc.) can have little predictive power here, as discussed in Section 4.2.

1.1 An example

In a Sender-Receiver game,[1] a Sender privately learns his type θ, then sends a (possibly costly) message m to the Receiver, who finally takes an action a. The two players' payoffs are both functions of (θ, m, a).

In this section we consider the game in Figure 1. The Sender privately knows whether he is a "Quantitative" type or not; both types are equally likely from the Receiver's (prior) perspective. The Sender must decide whether or not to get an MBA.

If the Sender gets an MBA, the employer (Receiver) sees this message and decides whether to promote the Sender to Head of Human Resources (HR), to promote him to Chief Financial Officer (CFO), or to keep him at his current job (Assistant) with a pay raise. If the Sender does not get an MBA, the game ends.

The Receiver would like to promote an MBA in a way corresponding to his

[1] The standard terminology we use in this section is defined formally in Section 2.

	Assist.	HR	CFO	Assist.
Quantitative	2, 2	0, 0	5, 5	3, 3
Non-Quant.	2, 2	1, 5	0, 0	3, 3
	no MBA	get MBA		

FIGURE 1: A Sender-Receiver game. The Sender's payoff is listed first.

type. Neither type wants to be promoted to HR, while only a Quantitative type would like to be CFO. It profits both Sender types (and the Receiver) to get an MBA and a raise with no promotion.

There are three kinds of (pure strategy) equilibria in this game.[2] In one, both Sender types get an MBA, and due to the balanced prior beliefs, the Receiver keeps the employee as an Assistant. In another, only the Quantitative type gets an MBA, which leads to promotion to CFO. We leave it to the reader to see that these two kinds of equilibria are robust, and satisfy all standard refinements in the literature.

There is a third kind of equilibrium where no Sender type gets an MBA. This is supported by the Sender's anticipation that the Receiver would promote an MBA to HR with sufficiently high probability. In turn, this means the Sender thinks the Receiver will believe that only (or with high probability) non-Quantitative types get an MBA.

We now argue that this equilibrium is not robust to the possibility that an out-of-equilibrium message can be interpreted as an implicit statement about the Sender's possible type(s). Before doing so it is worth noting that, perhaps surprisingly, this "pooling on no MBA" equilibrium does satisfy various refinements commonly used in the literature, such as the Intuitive Criterion, D1, and even Kohlberg and Mertens' (1986) stability. Proofs of this are available upon request.

To illustrate the central idea in this paper, suppose the "no MBA" pooling equilibrium is being anticipated by the Receiver, and consider the possibility that if the Receiver sees the out-of-equilibrium choice "get MBA," he interprets it as the following (implicit) statement: *"I am the Quantitative type."* Would this be credible?

[2]We consider only pure strategies throughout the paper. In any case, we make assumptions in both Sections 3 and 4 that imply pure best responses for the Receiver.

If the Receiver were to believe this (implicit) statement, he would choose the CFO action. Therefore the Quantitative type would gain from the Receiver's trust in this statement. The non-Quantitative type however would not. In this sense, this implicit statement is *credible*: the Quantitative type is precisely the only one who would want to "send" it.[3]

On the other hand, "get MBA" cannot credibly convey the statement *"I am the non-Quantitative type."* The Receiver's trust in this statement would cause him to choose HR, under which the non-Quantitative type does not gain. In fact neither type would gain if the Receiver believed such a statement.

Finally, we also consider the possibility that if the Sender gets an MBA, he is trying to convey the less precise statement: *"I am either the Quantitative type or the non-Quantitative type."* Is this a credible statement in the above sense? In order to determine this, we need to predict how the Receiver would respond to such a statement. More precisely, we need to determine what the Sender *anticipates* the Receiver to *believe* about the likelihood of types in order to predict a response.

One could argue that, due to the credibility of the *"I am Quantitative"* speech, it should be less likely for a Quantitative type to send this less precise message.[4] On the other hand, one could admit the possibility that the two types have different abilities to perform forward induction reasoning (which have not been explicitly modeled here). For example, the Quantitative type could be more likely to be able to perform this reasoning, which would make this type *more* likely to have sent the message. A third, more stringent approach would be to explicitly assume that the Receiver simply updates his prior using Bayes' rule when evaluating such a potential implied statement.

Since receiving this out-of-equilibrium message is a counterfactual event, we see little justification for prescribing *any* single, particular belief over the two types when evaluating this speech. In fact, we view the Sender's

[3]This kind of reasoning also appears in Grossman and Perry's (1986) *Perfect Sequential Equilibrium* and in Farrell's (1993) *neologism-proofness*. In fact those concepts would consider the credibility of "I am the Quantitative type" sufficient to rule out this equilibrium. Below we diverge from these two concepts; see also Section 2.2.

[4]Precisely this kind of argument leads Matthews et al. (1991) to require a consistent *set* of "speeches" which may separate different deviant types from each other.

anticipation of the Receiver's posterior beliefs as being ambiguous. Therefore as a first approach, we use a max-min criterion to evaluate preferences when Sender types are deciding whether to deviate. This means that, in this example, we ask whether both types would gain from conveying this less precise message, *regardless of the beliefs* formed by the Receiver.

If the Receiver puts enough weight on the probability that the Quantitative type is trying to make this speech, the Receiver would choose CFO. As argued above, the non-Quantitative type would *not* gain in this scenario. Similarly, with beliefs sufficiently biased toward the non-Quantitative type, the Receiver would choose HR, making both types regret the deviation. Therefore *neither* type would unambiguously gain from conveying this third statement, undermining its credibility.[5]

To summarize, if we interpret the message "get MBA" as an implicit attempt to convey information about a candidate set of types, only one such message is credible: *"I am the Quantitative type."* The uniqueness of this credible message makes this equilibrium vulnerable to a deviation which can be "unambiguously" interpreted to be coming from a unique set of possible Sender-types, namely the singleton "Quantitative type."

In more general games, we say that an equilibrium is *vulnerable to a credible deviation* if there is an out-of-equilibrium message m through which the Sender can convey the following statement. (This "speech" is not really *made* by the Sender; it is implicitly communicated through m.)

> "By sending this out-of-equilibrium message m, I am signaling to you that my type belongs to the set of types C. To see this, observe that if you form *any* belief (distribution) over C and take a corresponding best response with respect to those beliefs (and m), then any type $\theta \in C$ is guaranteed to be better off than he would have been in equilibrium. Conversely, for any remaining type $\theta' \notin C$, there exists a belief over C (and your corresponding

[5] In contrast, Grossman and Perry (*loc. cit.*) would consider this statement credible because they require the Receiver to update his prior off the equilibrium path. This is the crucial difference between our concept and theirs. In a modified version of the example (available upon request), they would eliminate all pure equilibria while we would not. The same can occur in monotonic signaling games as well (Section 3).

best response) that would make θ' worse off than in equilibrium. That is, C is *precisely* the set of types that gains regardless of the beliefs you form, as long as those beliefs are over C. Lastly, C is unique in this respect: Given message m, this speech cannot be made for any other set C'."

The existence of such a message m and set of types C makes the equilibrium in question less plausible than others. Under a mild notion of forward induction, it becomes a self-fulfilling prophecy for the Receiver, upon seeing m, to behave as if the Sender's type is in C.

In this argument, we do not prescribe specific posterior beliefs for the Receiver following the receipt of m. As discussed following the example of Figure 1, this even allows for the possibility that the Sender's type is correlated with the ability to perform this forward-induction reasoning. If the Receiver admits the possibility of such correlation, it is unclear how he would update his beliefs without specifying a more detailed model. It is even less clear how the *Sender* should anticipate the Receiver's understanding of this possibility. Since we think of the Receiver's posterior beliefs simply as a way to rationalize the Sender's equilibrium behavior, a theory with fewer specific assumptions on these posterior beliefs is more appealing.

The ideas outlined above may appear similar to certain concepts used in the literature on equilibrium refinements. We postpone comparisons to this literature to Section 2.2, after we formalize our definitions.

2 Sender-Receiver Games

Our main results concern two different classes of 2-player, Sender-Receiver games with costly signaling. Since those two classes share some structure, we introduce their shared notation here.

The Sender has private information that is summarized by his type $\theta \in \Theta = \{\theta_1, \theta_2, \ldots, \theta_n\} \subset \mathbb{R}$. For notational convenience, we order the types so that $\theta_1 < \theta_2 < \cdots < \theta_n$. The commonly known prior probability that the Sender's type is θ is $\pi(\theta)$. Upon realizing his type, the Sender chooses a message $m \in \mathbb{R}_+$. A strategy for the Sender is a function $M\colon \Theta \to \mathbb{R}_+$. After observing any message m, the Receiver chooses an action $a \in \mathbb{R}$. A

strategy for the Receiver is a function $A\colon \mathbb{R}_+ \to \mathbb{R}$. The Sender and Receiver receive respective payoffs of $u_S(\theta, m, a)$ and $u_R(\theta, m, a)$, which are both continuously differentiable in (m, a). A Sender-Receiver game is given by the tuple (Θ, π, u_S, u_R).

To define a Perfect Bayesian Equilibrium, we introduce the concept of the Receiver's (posterior) beliefs: Upon receiving a message $m \in \mathbb{R}_+$, the Receiver updates his beliefs over Θ regarding the Sender's type. This update-rule is a function $\mu\colon \mathbb{R}_+ \to \triangle(\Theta)$, where $\triangle(\Theta)$ refers to the set of probability distributions on Θ. A *Perfect Bayesian Equilibrium* consists of strategies and an update rule that satisfy the usual incentive compatibility and consistency conditions (Fudenberg and Tirole (1991a)), which we now formalize.

For any message $m \in \mathbb{R}_+$ and any fixed (posterior belief) distribution $\tilde{\pi} \in \triangle(\Theta)$, denote the Receiver's best responses to m (given $\tilde{\pi}$) by $BR(\tilde{\pi}, m) \equiv \arg\max_{a\in\mathbb{R}} \mathbb{E}[u_R(\theta, m, a) \,|\, \tilde{\pi}]$. Assumptions made below guarantee the non-emptiness of this correspondence. In a standard abuse of notation, for any set of types $T \subseteq \Theta$ we write $BR(T, m) \equiv \bigcup_{\tilde{\pi}\in\Delta T} BR(\tilde{\pi}, m)$, which can be thought of as the Receiver's rationalizable actions knowing only that $\theta \in T$.

The triplet (M, A, μ) is a *Perfect Bayesian Equilibrium*[6] when the following conditions hold.

- Sender incentive compatibility:

$$\forall \theta \in \Theta, \quad M(\theta) \in \arg\max_{m\in\mathbb{R}_+} u_S(\theta, m, A(m)).$$

- Receiver incentive compatibility:

$$\forall m \in \mathbb{R}_+, \quad A(m) \in BR(\mu(\cdot \,|\, m), m)$$

- Consistency of beliefs: for any equilibrium message $m \in M(\Theta)$, $\mu(\cdot \,|\, m)$

[6]Based on the results of Fudenberg and Tirole (1991b), Perfect Bayesian Equilibrium is equivalent to Sequential Equilibrium (Kreps and Wilson (1982)) on the classes of games we consider.

is derived from Bayes' rule:

$$\mu(\theta \,|\, m) = \begin{cases} \dfrac{\pi(\theta)}{\sum_{\theta' \in M^{-1}(m)} \pi(\theta')} & \text{if } M(\theta) = m, \\ 0 & \text{otherwise.} \end{cases}$$

There are no restrictions on beliefs following out-of-equilibrium messages m.

Throughout the paper, an *equilibrium* refers to a Perfect Bayesian Equilibrium.

When an equilibrium (M, A, μ) is clearly given in context, we denote the Sender's equilibrium payoff (as a function of his type) as

$$u_S^*(\theta) \equiv u_S(\theta, M(\theta), A(M(\theta))).$$

2.1 Formalizing Credible Deviations

As discussed in Section 1.1, we ask whether the Sender, upon sending an out-of-equilibrium message m, can induce the Receiver to reason that it must have been sent by a type within some set C. Under our definition, this reasoning is justified when C is *precisely* the set of Sender types that would benefit from deviating to m, *whenever the Receiver plays any best response to m with beliefs restricted to C.* Hence, Sender types that do not belong to C are worse off (compared to their equilibrium payoff) under *some* such best response. An equilibrium is Vulnerable to a Credible Deviation if, for some out-of-equilibrium message, there is a *unique* such C.

Definition 1 (Vulnerability to a Credible Deviation) *Given an equilibrium (M, A, μ), we say that an out-of-equilibrium message $m \in \mathbb{R}_+ \backslash M(\Theta)$ is a* Credible Deviation *if the following condition holds for exactly one (non-empty) set of types $C \subseteq \Theta$.*

$$C = \{\theta \in \Theta : u_S^*(\theta) < \min_{a \in BR(C,m)} u_S(\theta, m, a)\} \tag{1}$$

We call C the (unique) Credible Deviators' Club *for message m. If such a message exists, the equilibrium is* Vulnerable to a Credible Deviation.

The fact that (1) is an equality (as opposed to, say, the inclusion relation $C \supseteq$) enforces the precision mentioned above. The uniqueness requirement on C (given m) makes this a minimal attempt to implement the reasoning discussed in Section 1.1. If two such sets, C and C', existed for m then it would be arbitrary for types in C to assume that the Receiver would restrict beliefs to C, and not to C' (or even $C \cup C'$). Hence we specifically require (1) to hold for only one set C, resulting in a weaker invulnerability condition. It is worth noting that the results in this paper would continue to hold under a stronger version in which C is not required to be unique.

We use a max-min criterion to evaluate the Sender's preferences because it is unclear how the Receiver should form beliefs over C. This was illustrated in the example of Section 1.1 for the case in which C consisted of both types. We view our max-min approach as a natural starting point, though alternate definitions could be considered. For example, one could require only that the Receiver possess a single, "worst-case" belief over C that dissuades each $\theta \notin C$ from deviating.[7] It turns out that this weaker condition would yield the same results as our definition for the models in Sections 3 and 4. On the other hand, games exist in which this alternate version has no bite and ours does.

2.2 Relation to the refinements literature

The concept of a Credible Deviation may appear similar to certain equilibrium refinements used in the literature. It turns out, however, that there is no general, logical relation between our vulnerability condition and these equilibrium refinements. For instance there exist games in which *all* equilibria are Vulnerable to Credible Deviations. On the other hand, there are equilibria of other games that are *not* Vulnerable, but still fail certain refinements. In this section we review some of this literature and relate it to our concept.

[7] We thank Johannes Hörner and Jeroen Swinkels for independent comments leading us to these observations.

Perhaps the least controversial concept in this literature is the Intuitive Criterion (see Cho and Kreps (1987)). This refinement deems an equilibrium implausible whenever some Sender type would benefit from deviating to an out-of-equilibrium message, as long as the Receiver makes the minimal assumption that it was sent by types that *could* potentially gain from sending it.

DEFINITION 2 (INTUITIVE CRITERION) *For a given equilibrium* (M, A, μ) *and out-of-equilibrium message* $m \in \mathbb{R}_+ \setminus M(\Theta)$*, denote by* $J(m)$ *the set of types whose equilibrium payoff is higher than any payoff they could get by sending* m*, as long as the Receiver plays a rationalizable action, i.e.*

$$J(m) \equiv \{\theta \in \Theta : u_S^*(\theta) > \max_{a \in BR(\Theta, m)} u_S(\theta, m, a)\}.$$

The equilibrium fails the Intuitive Criterion (via m*) if* $J(m) \neq \Theta$ *and*

$$\{\theta \in \Theta : u_S^*(\theta) < \min_{a \in BR(\Theta \setminus J(m), m)} u_S(\theta, m, a)\} \neq \emptyset. \tag{2}$$

Inequality (2) says that by sending m, at least one type θ gains unambiguously so long as the Receiver restricts his beliefs to $\Theta \setminus J(m)$. This restriction on the Receiver's beliefs is a very minimal requirement, since no type in $J(m)$ could gain by sending m if he anticipates any rational reaction from the Receiver. Given this restriction, the Intuitive Criterion merely checks for the existence of *some* type $\theta \notin J(m)$ who, anticipating such beliefs, would gain unambiguously compared to his equilibrium payoff.

This concept differs from Vulnerability to Credible Deviations in two ways, which can be seen by comparing eqns. (1) and (2). First[8] consider whether *any* type has an incentive to deviate from an equilibrium under some restrictions on the Receiver's response. In eqn. (1) the Receiver's beliefs are restricted more than in eqn. (2). This makes it easier to find deviating types in (1) than in (2), making the Intuitive Criterion a relatively weak concept.

Second, however, consider which types should *not* have an incentive to deviate. While eqn. (2) merely requires non-emptiness of the set of devia-

[8]We thank a referee for making this reasoning precise.

tors, eqn. (1) precludes types (outside C) from wanting to perform certain deviations. This makes it harder to find a deviating set (club) in (1) than in (2), making the Intuitive Criterion a relatively stronger concept.

Therefore it turns out that there is no general logical relation between the Intuitive Criterion and our vulnerability condition. On the classes of games studied in this paper, and for any Sender-Receiver game where the Sender has only two types, the Intuitive Criterion is weak: if an equilibrium fails it, then it is also Vulnerable to Credible Deviations. If, in addition, the Receiver has only two actions following any message, then the conditions are generically equivalent. There are, however, games in which the Intuitive Criterion rules out an equilibrium which is *not* Vulnerable to Credible Deviations.[9]

In certain important classes of Sender-Receiver games with more than two Sender types (e.g. Spence (1973)), the Intuitive Criterion does not reduce the set of equilibrium outcomes. This has led to, among others, a well-known concept that makes specific requirements on posterior beliefs. The D1 Criterion (see Banks and Sobel (1987), Cho and Kreps (1987), Cho and Sobel (1990)) requires the Receiver to disbelieve that a deviating message could be sent by a type θ who weakly gains "less often" (i.e. under fewer $a \in BR(\Theta, m)$) than some other type θ' strictly gains.

Definition 3 (D1 Criterion) *An equilibrium (M, A, μ) fails the D1 Criterion if there exists an out-of-equilibrium message $m \in \mathbb{R}_+ \backslash M(\Theta)$ and types $\theta, \theta' \in \Theta$ such that $\mu(\theta \,|\, m) > 0$ and*

$$\{a \in BR(\Theta, m) : u_S^*(\theta) \leq u_S(\theta, m, a)\} \subsetneq \{a \in BR(\Theta, m) : u_S^*(\theta') < u_S(\theta', m, a)\}.$$

This refinement is stronger than the Intuitive Criterion. For its practical usability, the D1 criterion is one of the most popular refinements used in applied work.

Despite its predictive power, there is little intuitive justification for the Receiver to put infinitely more weight on Sender types that gain from the deviation "more often" (θ') than others (θ). Fudenberg and Tirole (1991a)

[9] Straightforward proofs of these facts are available upon request.

write (p. 460)

> "Since the motivation for D1 is to refine the set of equilibria using 'reasonable' restrictions on beliefs, to the extent that [such] 0-1 restrictions are implausible they may cast doubt on D1 as an equilibrium concept."

While there are arguments against the "speeches approach" as well (e.g. the Stiglitz Critique), one could argue that a missing behavioral motivation is a disadvantage of this practically useful refinement.

This motivates our study of monotonic signaling games in Section 3, where, as a corollary of our results, the D1 Criterion eliminates an equilibrium if and only if it is Vulnerable to Credible Deviations. That is, the Riley outcome can be justified by an intuitive, plausible robustness check: Immunity to Credible Deviations. While we reject (as Vulnerable) the same equilibrium outcomes that D1 eliminates, we do not impose any specific restrictions on out-of-equilibrium beliefs. On the other hand, D1 has little predictive power in a class of non-monotonic signaling games we study in Section 4, while our condition still selects a unique outcome.

A related notion is that of Kohlberg and Mertens' (1986) Stability. In generic Sender-Receiver games all Stable equilibria satisfy the D1 Criterion; furthermore the two concepts are equivalent on the class of discrete monotonic signaling games, resembling the continuous one we study in Section 3 (see Cho and Sobel (1990)). In contrast, on the general class of Sender-Receiver games, the Stability of an equilibrium neither implies nor is implied by its immunity to Credible Deviations.[10]

Our motivation for Credible Deviations has a flavor similar to the motivation behind Grossman and Perry's (1986) Perfect Sequential Equilibrium (PSE). Roughly speaking, under PSE a set of types T breaks an equilibrium with an out-of-equilibrium message m if all types in T improve their payoff by sending that message as long as the Receiver believes that all (and only) the types in T would *always* deviate and send m. The word *always* here implies that the Receiver is specifically assumed to update his priors over T

[10] A less related notion is evolutionary stability in games with pre-play communication, which can select Pareto-efficient outcomes in more general games. See Kim and Sobel (1995) and references therein.

in accordance with Bayes' Rule. This amounts to replacing $BR(C, m)$ with $BR(\pi|_C, m)$ in the right-hand side of eqn. (1).[11]

In Section 1.1, we argued against doing this. When considering the case $C' = \{\text{Quant}, \text{Non-Quant}\}$ in that example, PSE would specify that the Receiver use precisely his prior beliefs, which in turn would cause him to choose the action "Assist." Since both types prefer this outcome, this set of types C' *would* break the pooling equilibrium under PSE. However, $C = \{\text{Quant}\}$ would also break the equilibrium under PSE by inducing the action "CFO". Therefore, *when beliefs over C' are required to coincide with the prior distribution, the Receiver is implicitly forced to ignore the possibility that C is the deviating set.* We find this inconsistent.

More generally, our opinion is that such specific assumptions off the equilibrium path are too prescriptive. While we can think of equilibrium play (and the resulting beliefs) as being self-enforced by, say, repeated interaction, pre-play communication, or even explicit agreement, there is less justification for this reasoning off the equilibrium path.

Even though the definition of PSE may seem to be only marginally different from the ideas which define Credible Deviations, these two concepts can yield strikingly different results. Even on the standard class of monotonic signaling games, Perfect Sequential Equilibria may not exist; see Sec. 10.6 of van Damme (1991). We examine that same class in Section 3, and characterize a single equilibrium outcome as being immune to Credible Deviations.

It is worth noting, however, that if an equilibrium is Vulnerable to a Credible Deviation by a *singleton* C, then it also fails PSE. This is true since there is only one belief the Receiver can form over a singleton set, in which case the distinction between our definitions disappears.

Lastly we mention refinements defined on the class of cheap-talk signaling games, i.e. those in which agents' payoffs are constant with respect to messages m. Farrell's (1993) Neologism-proofness asks whether a set of types can credibly distinguish itself by explicitly sending an out of equilibrium message that identifies a set of potential deviating types, e.g. "We are

[11] To be precise, Grossman and Perry allow the Receiver to put less weight on types in T who are indifferent about deviating, reflecting the idea that such types may randomly choose whether to deviate. Therefore the posterior beliefs may not be exactly $\pi|_T$. PSE also does not require uniqueness of the deviating set of types T, as we do.

the types in T."[12] An equilibrium fails Neologism-proofness if the types in T are precisely the ones who gain when, in response to the message, the Receiver's beliefs are a Bayesian update of his prior beliefs on T. As with PSE, Neologism-proof equilibria need not exist in general. Furthermore, any comparisons between our concept and PSE could be made regarding Neologism-proofness as well.

Matthews et al. (1991) create the more sophisticated concept of *announcement-proofness* by requiring deviating types to (implicitly) announce what could be called a complete deviation strategy. A deviation strategy assigns messages to deviating types; it is credible if each such type has an incentive to send his deviating message, while remaining types have no incentive to send one. This approach avoids the ambiguity discussed in Section 1.1 resulting when the Sender has a choice about how to identify his type. In exchange for the high level of consistency in this approach, the results for this concept are limited.

A similar concept is given by Rabin (1990), who asks whether one can construct profiles of messages that give some types an incentive to "truthfully" communicate their type, and give the rest an incentive to imitate others. Within his class of communication games, there always exist equilibria in which such communication occurs. This idea of the Sender deviating to another profile (or partial profile) of strategies can be traced back to Myerson (1989).

3 Monotonic Signaling Games

A subclass of Sender-Receiver games often used in economic applications is that of monotonic signaling games. These games model situations where a privately informed party chooses a costly action in order to convey payoff-relevant information to another party, e.g. Spence (1973). As in other signaling games there are often multiple Perfect Bayesian Equilibria on this class. Cho and Sobel (1990) show that the D1 Criterion selects a unique

[12] Implicit in the definition is the assumption that such an out of equilibrium message is always available; this escapes the modeling problem that any cheap-talk equilibrium can be rewritten as a "babbling" equilibrium, where all messages are used (perhaps randomly) in equilibrium. See Blume (1994).

equilibrium outcome in the canonical form of this type of game—the Riley outcome—which also happens to be the unique Stable outcome (Kohlberg and Mertens (1986)) in a discrete version of the model. In this section we show that this kind of equilibrium is the only one immune to Credible Deviations.

The class of monotonic signaling games is meant to capture situations in which

- the Sender would prefer the Receiver to take higher actions,
- the Receiver prefers his action to be correlated with the Sender's type,
- higher Sender types are "stronger" in the sense that it is relatively less costly for them to send higher messages than for lower types to do so.

Following Cho and Sobel (1990) and Ramey (1996), we formalize these characteristics with the following assumptions.

For the first, we assume that $u_S(\theta, m, a)$ is strictly increasing in a for all (θ, m). One can think of a as some sort of compensation for the Sender; all Sender types always prefer more. Additionally, in order to avoid solutions involving arbitrarily large messages and actions we assume that $\lim_{m\to\infty} u_S(\theta, m, a) = -\infty$ for all θ and a.

We assume that u_R is such that, for any type θ and message m, the Receiver has a unique best response, i.e. that $BR(\{\theta\}, m)$ is a singleton. Throughout Section 3 we denote this action as

$$\{\beta(\theta, m)\} \equiv BR(\{\theta\}, m).$$

Furthermore we assume that $\beta(\cdot, \cdot)$ is uniformly bounded from above. These two assumptions would be implied by a standard list of assumptions on the primitives; for brevity we assume them directly.

We assume that $\partial u_R/\partial a$ is strictly increasing in θ for all (m, a). As a result, $BR(\tilde{\pi}, m)$ is greater for "more optimistic" beliefs $\tilde{\pi}$ (beliefs that are greater in the first-order stochastic sense), and in particular, $\beta(\theta, m)$ is strictly increasing in θ (Cho and Sobel (1990), p. 392). Together with monotonicity, this captures the idea that the Sender wants to induce the

Receiver to choose larger actions by trying to convince him that the Sender's type is greater.

We make a central assumption in Spencian signaling games, the single-crossing condition: $-(\partial u_S/\partial m)/(\partial u_S/\partial a)$ is strictly decreasing in θ. That is, for a given increase in m, in order to keep the Sender at the same utility level, a higher Sender-type needs less compensation in terms of a (in case m is locally costly for the Sender) or he is willing to give up a larger amount of a (in case m is locally beneficial for him).

Finally, we assume that $u_S(\theta, m, \beta(\theta, m))$ is strictly quasiconcave in m. In many applications, this assumption is implied by stronger assumptions made directly on the primitives of the model.

The above assumptions all are made by Cho and Sobel (1990) and Ramey (1996), and guarantee that the D1 criterion selects the unique "Riley equilibrium outcome."[13] Both cited papers also contain results under other, less restrictive assumptions (e.g. allowing multidimensional signals). However, a unique separating equilibrium is proven to be selected by D1 under exactly the same assumptions that we have here. Our aim is not to reproduce an analog of each result in those two papers; rather it is to demonstrate that in the canonical Spencian signaling model, an equilibrium fails the D1 Criterion if and only if a Credible Deviation exists.

Most applied signaling models have a lot more structure. For example, since m is usually interpreted as a costly action undertaken by the Sender that may be beneficial for the Receiver (e.g. the Sender's education level), it is often assumed that $u_S(\theta, m, a)$ is weakly decreasing in m and $\beta(\theta, m)$ is weakly increasing in m. We need not impose these conditions.

An additional piece of notation simplifies the exposition. Suppose for some θ and m there exists $\hat{a}$ such that $u_S(\theta, m, \hat{a}) = u_S^*(\theta)$. This action by the Receiver would give Sender-type θ his equilibrium payoff after sending m. If such an action exists, it is unique by monotonicity. So for any θ and m,

[13] An unimportant difference is that Cho and Sobel (1990) assume compact message and action sets, while we, following Ramey (1996), assume those sets are unbounded but assume that the best responses are bounded. Ramey (1996) also extends Cho and Sobel's analysis to a continuum of types.

let $\hat{a}(\theta, m)$ be the action to satisfy

$$u_S(\theta, m, \hat{a}(\theta, m)) = u_S^*(\theta) \tag{3}$$

if such an action exists, and denote $\hat{a}(\theta, m) = \infty$ otherwise (for notational convenience).

The single-crossing property suggests that higher types need less compensation for sending higher messages than do lower types. Lemma 1 strengthens that idea, applying it relative to equilibrium payoffs. It states that if a higher type θ^h deviates to a higher message m', he needs less compensation *to keep him at his equilibrium utility* than would a lower type θ^ℓ who sends that same message m'. This is a standard type of result. Proofs of all Lemmas appear in the Appendix.

LEMMA 1 *Fix an equilibrium (M, A, μ) and type $\theta^h \in \mathbb{R}_+$. For all $m' > M(\theta^h)$ and all $\theta^\ell < \theta^h$, $\hat{a}(\theta^h, m') < \infty$ implies $\hat{a}(\theta^h, m') < \hat{a}(\theta^\ell, m')$.*

3.1 Credible Equilibrium

We now turn towards results specific to the existence of Credible Deviations. In order to convey some intuition about this *in the monotonic environment*, it is useful to consider the following lemma. Lemma 2 states that, in searching for a potential deviators' club, it suffices to find a type θ' who would prefer to be self-identified by sending message m', while no lower type would prefer to be perceived as θ' sending that same message. In this sense, "only the lowest type matters" in finding such a club in this class of games.[14]

LEMMA 2 *Fix an equilibrium (M, A, μ), and suppose there exists a type θ' and an out-of-equilibrium message m' such that*

$$\begin{aligned} u_S^*(\theta') &< u_S(\theta', m', \beta(\theta', m')) \quad \textit{and} \\ u_S^*(\theta) &\geq u_S(\theta, m', \beta(\theta', m')) \quad \forall \theta < \theta'. \end{aligned}$$

[14] One of the points we wish to make in this paper is that Credible Deviations may be identified in environments that are less structured than this one. Any intuition for the results in monotonic signaling games may not capture the flavor for them in, for example, the class of games examined in Section 4.

Then there exists a unique credible deviators' club for m'.

As shown in the proof (see Appendix), the unique club is precisely the set of types who prefer (relative to equilibrium payoffs) to be perceived as θ' by sending m', i.e.

$$C' = \{\theta \in \Theta : u_S^*(\theta) < u_S(\theta, m', \beta(\theta', m')).$$

The explanation for this result has two parts. First, in the monotonic setting, the Sender is made worse off as the Receivers beliefs shift towards lower types. Therefore, the minimization in eqn. (1) occurs at $a \in BR(\min C, m)$, for any fixed m; i.e. the "worst" belief is the one putting probability one on the lowest type in C. If C satisfies (1), then the inequalities of the lemma must be satisfied for $\theta' = \min C$. Furthermore these inequalities are sufficient since adding higher types to a set C does not change the set $BR(\min C, m)$. This explains why the inequalities generate *some* credible deviators' club.

The uniqueness result also relies on monotonicity. Since lower types cannot gain by sending m' when being perceived as θ', they also cannot gain by being perceived as themselves, and hence cannot belong to any club C. If only higher types formed a club C by sending m', θ' would want to join this club; hence θ' must belong to any club that exists, and be the minimum member. But then all types in C' would want to join such a club, since θ' is the "worst case" member.

Using this result, we can show that immunity to Credible Deviations is as restrictive as D1 on this class. First, we rule out pooling (or semi-pooling) equilibria.

LEMMA 3 *If an equilibrium (M, A, μ) is not Vulnerable to Credible Deviations, it is a separating equilibrium—no two types send the same message.*

The intuition here is that the highest type θ' in any pooling set would be able to find a sufficiently high message m' with which to satisfy the inequalities of Lemma 2. It is worth noting that the D1 Criterion creates a similar conclusion in this model, but also goes on to require more of beliefs. In this sense one can think of Vulnerability to Credible Deviations as a more minimal requirement with a somewhat more palatable justification.

We have established that if a sequential equilibrium is not Vulnerable to Credible Deviations, then it must be fully separating. Below we show that only the least-distortive separating equilibrium outcome—the Riley outcome—is not Vulnerable.

In a separating equilibrium, each type $\theta_i \in \Theta$ is uniquely identified by his equilibrium message m_i. As a result, $\mu(\theta_i \,|\, m_i) = 1$ and the Receiver's response is $a_i = \beta(\theta_i, m_i)$. Labeling the types in ascending order ($\theta_i < \theta_{i+1}$), the *Riley outcome* is the list of pairs $(m_i^r, a_i^r)_{1 \leq i \leq |\Theta|}$ such that

$$m_1^r = \arg\max_{m \geq 0} u_S(\theta_1, m, \beta(\theta_1, m)) \tag{4}$$

and for all $1 < i \leq n$,

$$\begin{aligned} m_i^r = \arg\max_{m \geq 0} & \, u_S(\theta_i, m, \beta(\theta_i, m)) \\ \text{s.t. } & u_S(\theta_j, m_j^r, \beta(\theta_j, m_j^r)) \geq u_S(\theta_j, m, \beta(\theta_i, m)) \quad \forall j < i, \end{aligned} \tag{5}$$

and $a_i^r = \beta(\theta_i, m_i^r)$ for each i. The uniqueness of such messages is guaranteed by our quasi-concavity assumption. Due to the single-crossing assumption, the Riley messages m_i^r also are increasing in i, This is obvious when messages are always costly, but it also holds on our more-general class of games. In particular, single-crossing implies that if Sender θ_i is indifferent between message-action pairs (m, a) and (m', a') where $m > m'$, then type $\theta_j > \theta_i$ strictly prefers (m, a), which is the essence of Cho and Sobel's (1990) Assumption A4.

To characterize this outcome as the unique one to satisfy our requirement, we first rule out any other separating equilibrium.

Lemma 4 *Any equilibrium whose outcome is different from the Riley outcome is Vulnerable to Credible Deviations.*

The intuition for this result is that, if a separating equilibrium has a "gap" between equilibrium messages beyond that of the Riley outcome, then some type would be able to lower his message and still maintain the inequalities of Lemma 2.

Our main result adds the observation that the Riley outcome is not Vulnerable to Credible Deviations.

THEOREM 1 *The Riley outcome is the unique equilibrium outcome that is not Vulnerable to Credible Deviations.*

PROOF: Lemma 4 makes any other outcome Vulnerable.

To prove that the Riley outcome is not Vulnerable, observe from Cho and Sobel (1990) that the Riley outcome can be supported by *a* sequential equilibrium (in fact, with out-of-equilibrium beliefs that satisfy D1). Fix such an equilibrium, and suppose toward contradiction that C is a deviators' club for some out-of-equilibrium message m', i.e.

$$C = \{\theta \in \Theta : u_S^*(\theta) < \min_{a \in BR(C,m')} u_S(\theta, m', a)\}.$$

Denote the lowest type in C as $\theta_i = \min C$. As in the proof of Lemma 2, due to the monotonicity of the Receiver's best responses with respect to beliefs, and the monotonicity of u_S with respect to a, we have

$$\min_{a \in BR(C,m')} u_S(\theta, m', a) = u_S(\theta, m', \beta(\theta_i, m')) \quad \forall \theta \in \Theta.$$

Therefore for any $j < i$, since $\theta_j \notin C$ we have

$$u_S^*(\theta_j) \geq \min_{a \in BR(C,m')} u_S(\theta_j, m', a) = u_S(\theta_j, m', \beta(\theta_i, m'))$$

These are precisely the constraints in (5), which define m_i^r. Hence (by strict quasi-concavity)

$$u_S(\theta_i, m_i^r, \beta(\theta_i, m_i^r)) > u_S(\theta_i, m', \beta(\theta_i, m')).$$

That is, θ_i prefers *not* to deviate from m_i^r to m' when the Receiver believes the message came from him, which contradicts $\theta_i \in C$. □

Theorem 1 shows that on this class of games, an equilibrium is Vulnerable to Credible Deviations if and only if it fails Cho and Kreps' (1987) D1 refinement. As we show in the next section, however, this similarity breaks

down even on a similar class of games, when we slightly weaken the monotonic structure. It is also worth recalling that monotonic games may contain no equilibria which satisfy Grossman and Perry's (1986) Perfect Sequential Equilibrium concept (van Damme (1991)).

4 Information Transmission and Bias

In this section we consider a class of games which conveys the following type of interaction. The Sender wants the Receiver to take an action that matches his type; messages are costly; and the Receiver wants to take an action that matches the Sender's type offset by some bias. This is a version of Crawford and Sobel's (1982) model, but with discrete types and costly signaling. In fact it is similar to the model of Austen-Smith and Banks (2000), who combine costly signalling with Crawford and Sobel's cheap talk; this extension expands the set of Crawford-Sobel equilibria. More recently, Kartik (2005) examines a model in which miscommunicating one's type is costly, restricting attention to equilibria satisfying a monotonicity condition in the Receiver's beliefs.

While one can imagine a very generalized model for such interactions, it is not our intention to completely analyze every variation on this theme. In particular, we are able to illustrate our main point—that only one equilibrium is immune to Credible Deviations in this non-monotonic environment—by restricting attention to the interesting cases (discussed and defined below) where the bias is not trivially small.

As before, the types are ordered $\theta_1 < \theta_2 < \cdots < \theta_n$, the message space is $\mathbb{R}_+$, and the Receiver chooses an action in $\mathbb{R}$.

The Sender's payoff is of the form

$$u_S(\theta, m, a) = -d(\theta - a) - c(\theta, m)$$

with the following assumptions. The distance function d is convex and symmetric about zero ($d(x) \equiv d(|x|)$); hence increasing on $[0, \infty)$. The cost function c is continuous, strictly increasing in m, satisfies $\lim_{m\to\infty} c(\cdot, m) = \infty$, and satisfies single-crossing: $c(\theta, m') - c(\theta, m) > c(\theta', m') - c(\theta', m)$ for all

$m' > m$, $\theta' > \theta$. In words, the Sender wants the Receiver to choose a as close as possible to θ (with a symmetric convex loss function), while sending larger messages is more costly for him, but relatively less costly if he has a higher type.

The Receiver's payoff is of the form

$$u_R(\theta, m, a) = -(\theta - a - b)^2$$

where $b > 0$ is a commonly known bias. As a consequence, in equilibrium the Receiver chooses action $a = \mathbb{E}[\theta \,|\, \mu, m] - b$, where $\mathbb{E}[\theta \,|\, \mu, m]$ is the Sender's expected type given the observed message m and Receiver's update function μ.[15] The two parties' preferences are misaligned according to the bias b.

By varying the size of b in relation to the values of the θ_i's, one may obtain models with somewhat different flavors. For instance, if b is very small (compared to, say, each $\theta_{i+1} - \theta_i$), there is little conflict of interest between the Sender and Receiver; the situation essentially becomes a coordination game with sequential actions.

A more interesting case is one in which the bias is not "too small" relative to the distance between types, where there is relatively more conflict of interest. We obtain the most striking result by analyzing such cases. Since we wish to impose this restriction without making assumptions on the prior distribution of types, we assume that the bias is not small relative to the distance between *any* two types—specifically that it is greater than half the distance between the highest and lowest types.[16]

ASSUMPTION: The bias is not too small: $b > (\theta_n - \theta_1)/2$.

To get intuition for the role of the bias assumption in our results, see Figure 2. In this model, the no-small-bias assumption—coupled with the

[15] As in related literature, the particular functional form of u_R is not critical; any that implies this objective would yield the same results.

[16] If we merely made an assumption about the relative distance between each consecutive θ_i and θ_{i+1}, it would have little bite in examples where the prior probability of "middle" types goes to zero. An assumption which balances the distance between such types and their relative probabilities would be intractable.

FIGURE 2: If the Receiver's bias is large enough, a low type prefers being perceived as a higher type to being perceived as himself.

fact that $b > 0$—implies that *a low type prefers to be perceived as any higher type.* For instance, θ_1 prefers the action $a = \theta_n - b$ to the action $a = \theta_1 - b$, due to the symmetry of the distance function. In the monotonic games of Section 3, this feature of low types wanting to be perceived as high types is more general in that the Sender *always* prefers *any* higher action to a lower one. Therefore our bias assumption in this section preserves *some* of this incentive in a slightly richer model.

Despite the fact that the two classes of models share this feature, and the fact that Credible Deviations exist precisely when D1 fails in Section 3, we shall see that this equivalence breaks down in this information transmission model.

It is interesting (and perhaps tangential) to note that in the cheap talk literature, *small* bias assumptions are used to obtain separation results. In Crawford and Sobel's (1982) original cheap talk model, more-informative equilibria emerge as the bias becomes smaller. This is intuitive as a small bias implies less conflict of interest.

In models with one Receiver and *two* Senders (who observe the same type but have opposing biases), Morgan and Krishna (2001) and Battaglini (2002) show the existence of a fully type-revealing equilibrium—somewhat analogous to a separating equilibrium in the present model—when biases are sufficiently *small* relative to a bounded, 1-dimensional type space. Ambrus and Takahashi (2006) demonstrate a necessity for small biases (relative to the type space) in order to get full revelation in the multidimensional version of Battaglini's model. The small-bias assumption in these models plays a different role than in ours, however, allowing the Receiver to "punish" out-of-equilibrium (i.e. uncoordinated) behavior of the Senders.

4.1 Credible Equilibrium

The only equilibrium outcome immune to Credible Deviations involves separation. It is the unique outcome that minimizes the Sender's messages subject to the incentive constraints: θ_1 sends $m_1 = 0$, θ_2 sends a different message m_2 low enough to make θ_1 indifferent between sending $m_1 = 0$ and deviating to m_2, and so on. In this sense, this outcome resembles the Riley outcome in Section 3.

To formalize this, observe that in any *separating* equilibrium (M, A, μ), the Receivers equilibrium actions clearly satisfy $A(M(\theta_i)) \equiv \theta_i - b$. We define a *minimal-cost separating equilibrium* to be one where A satisfies that condition, and additionally, $M(\theta_1) = 0$ while for $2 \leq i \leq n$,

$$-d(\theta_{i-1} - A(M(\theta_{i-1}))) - c(\theta_{i-1}, M(\theta_{i-1})) = -d(\theta_{i-1} - A(M(\theta_i))) - c(\theta_{i-1}, M(\theta_i)) \tag{6}$$

which states that θ_{i-1} is indifferent between sending his equilibrium message $M(\theta_{i-1})$ and sending $M(\theta_i)$. Because of the assumption that the bias is not too small, these messages are uniquely defined and strictly monotonic.

To prove that this is the unique surviving equilibrium, we show that in any other equilibrium, a credible deviators' club must exist in one of two ways. First, there could exist a separating type who is greater than any pooling types (if they exist), but for whom eqn. (6) fails to hold. In the proof of Lemma 5 we show that if any such types exist, the highest of them would form a unique deviators' club. (It is unsurprising that this high type could try to deviate; the non-trivial things to show are that no one else would want to, and that he alone is the *unique* deviators' club for some message.) Second, there could exist pooling types. Using the previous case's result, we show (Lemma 6) that the highest one then would form a unique deviators' club. Hence we arrive at Theorem 2: There can be no pooling, and the separating equilibrium must be the one defined above.

Lemma 5 *Suppose an equilibrium (M, A, μ) is immune to Credible Deviations. If for some $s \geq 2$, the types θ_s, $\theta_{s+1}, \ldots, \theta_n$ are all separating (i.e. send unique equilibrium messages), then eqn. (6) holds for all $i \geq s$.*

The second main portion of the proof rules out pooling, given the result

of Lemma 5. The idea of the proof is that, if a (highest) pooling type exists, then he forms a singleton deviators' club with respect to a particular message. That message is one which is sufficiently high so as to make him almost indifferent between forming the singleton club and just playing his equilibrium pooling strategy. (Other messages may also yield a unique deviators' club; we only need to find one such message.) A message this high makes lower types unwilling to join the club. The conclusion of Lemma 5 is then used to show that higher types would not want to deviate to this message either.

LEMMA 6 *Suppose a non-separating equilibrium exists, and let θ_p denote the highest pooling type. If eqn. (6) holds for all $i > p$, then there exists a message for which $\{\theta_p\}$ is a unique credible deviators' club.*

Alternating applications of Lemmas 5 and 6 prove the main result.

THEOREM 2 *If an equilibrium (M, A, μ) is immune to Credible Deviations then it is a minimal-cost separating equilibrium: $A(M(\theta_i)) \equiv \theta_i - b$, $M(\theta_1) = 0$, and for $2 \leq i \leq n$, $M(\theta_i)$ satisfies (6).*

PROOF: If an equilibrium is immune to Credible Deviations, then Lemma 6 implies that the highest type, θ_n, does not pool. Hence Lemma 5 implies that eqn. (6) holds for $i = n$.

In turn, this means (again with Lemma 6) that θ_{n-1} does not pool; hence Lemma 5 implies that eqn. (6) also holds for $i = n - 1$. Continuing this argument for $i = n-2, n-3, \ldots, 2$, θ_i does not pool and eqn. (6) holds.

Therefore θ_1 also does not pool. It remains to be shown that $M(\theta_1) = 0$. This is true in *any* separating equilibrium, though, under our assumption $\theta_n - \theta_1 < 2b$. Indeed, the Receiver's equilibrium response $A(M(\theta_1)) = \theta_1 - b$ is the worst rationalizable action the Receiver could take (from θ_1's perspective), regardless of beliefs. Given this, $M(\theta_1) = 0$ is strictly best for θ_1. □

4.2 D1 AND POOLING

We show that the D1 Criterion does not always restrict the set of equilibria in the class of games examined in this section. Consider a 2-type example

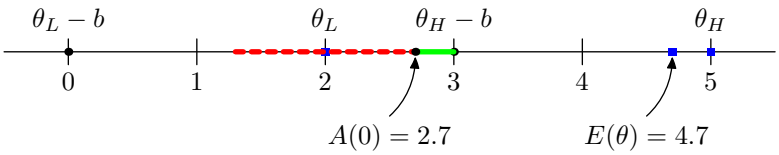

FIGURE 3: Rationalizable actions preferred by θ_L (dashed line) are disjoint from those preferred by θ_H (solid line), so D1 permits pooling.

where $\theta_1 = 2$, $\theta_2 = 5$, $b = 2$, and the prior is $\pi(\theta_2) = 0.9$. These parameters satisfy our previous bias-assumption, namely $\theta_2 - \theta_1 < 2b$. Let $d(x) = |x|$ and $c(\theta_i, m) = m/i$.

There exists a pooling-equilibrium (M, A, μ) such that $M(\theta) \equiv 0$ and, accordingly, $A(0) = E(\theta) - b = 4.7 - 2 = 2.7$. Furthermore μ can be specified so that the equilibrium satisfies D1.[17] Specifically, we show that if $A(m) = \theta_1 - b = 0$ and $\mu(\theta_1 \,|\, m) = 1$ for all $m > 0$, then the equilibrium satisfies D1.

To see this, we examine the potential gains from deviation for both types. Observe that regardless of the Receiver's (posterior) beliefs, he would never choose an action outside the range $[\theta_1 - b, \theta_2 - b] = [0, 3]$; hence we can restrict attention to that interval.

If θ_1 sends an out-of-equilibrium message $m > 0$ and the Receiver responds with $a \in [0, 3]$, then θ_1 gains (relative to his equilibrium payoff) if and only if $m \in (0, .7)$ and $a \in (1.3 + m, 2.7 - m)$. This range of actions is represented by the dashed line in Figure 3. Similarly, θ_2 gains if and only if $m \in (0, 0.6)$ and $a \in (2.7 + m/2, 3]$.[18]

For $m \in (0, 0.6)$, both types could gain from deviation. In those cases, however, $(1.3 + m, 2.7 - m) \cap (2.7 + m/2, 3] = \emptyset$, i.e. the sets of actions which make the two types better-off are not related by inclusion (and in fact do not even overlap). Hence D1 does not restrict out-of-equilibrium beliefs following such a message.

For $m \in [0.6, 0.7)$, only θ_1 could gain from deviation; D1 therefore requires

[17] In fact even the stronger D2 and NWBR conditions can be satisfied. We do not show this, and refer the reader to Cho and Kreps (1987) for definitions.

[18] Type θ_2 could also gain for some values $a > 3$, but we have stated such an action is never a best response for the Receiver.

$\mu(\theta_1|m) = 1$. For $m \geq 0.7$, neither type can gain from deviation and D1 places no restrictions $\mu()$.

Therefore the pooling equilibrium satisfies the D1 Criterion. Since the Receiver respond with action $A(m) = \theta_1 - b = 0$ for $m > 0$, neither type could gain by deviating. On the other hand, this equilibrium is Vulnerable to Credible Deviations since $C = \{\theta_2\}$ is a unique credible deviators' club for various out-of-equilibrium messages.

It is clear that this example is robust to perturbations. More extreme priors would yield the same results, making the out-of-equilibrium beliefs we used (with unit probability on the low type) even less appealing while still satisfying D1. Furthermore, due to the slack in our arguments, it is clear that there even exist D1 equilibria in which all types pool by sending some *positive* message $m > 0$.

5 Conclusion

We have shown that some equilibria of Sender-Receiver games are vulnerable to a particular kind of signaling. Credible signals identify a set of deviating types who gain by deviating as long as the Receiver reacts as if *only* such types could be deviating. Generally, this vulnerability is not captured by standard concepts in the refinements literature. While Credible Deviations are eliminated on the class of Monotonic Signaling Games (Section 3) by, for example, Cho and Kreps' (1987) D1 Criterion, this does not happen in a related class of games (Section 4) where best response sets are not ordered (see also the example in Section 1.1).

On the other hand we wish to emphasize the point that *immunity from Credible Deviations* does not, by itself, serve well as a generally predictive concept. In some games all equilibria may be Vulnerable to Credible Deviations.[19] In other games unappealing equilibria may be immune to Credible Deviations, as in the following example.

It is an equilibrium for all types to send m_1 when the Receiver plans to play a_1 (with beliefs sufficiently biased toward type L). This pooling

[19] For instance, the unique equilibrium of Example 10.5.5 in van Damme (1991) is Vulnerable to Credible Deviations.

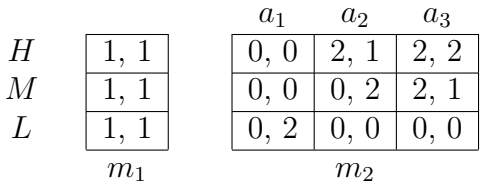

FIGURE 4: Pooling on m_1 is not Vulnerable to Credible Deviations.

equilibrium is ruled out by standard refinements including the Intuitive Criterion. In fact, it is even ruled out by only two rounds of deleting dominated strategies: since L cannot gain by sending m_2, a_1 becomes dominated for the Receiver. This makes H and M strictly prefer playing m_2. Nevertheless, the equilibrium is not Vulnerable to any Credible Deviation.

This observation reinforces the fact that our primary goal is not to propose an equilibrium refinement that selects a unique equilibrium in every game. Instead it is to be aware of a type of non-robustness which some (or all) equilibria may possess in Sender-Receiver games that are used in applications.

The basis for our approach is centered on our view that the Receiver's beliefs (and subsequent action) in response to a deviant message m should be regarded as *ambiguous* to the Sender. While previous work has allowed agents' beliefs to differ off the equilibrium path (e.g. Fudenberg and Levine (1993)), ours is the first formalization (to our knowledge) which explicitly allows ambiguity of the Receiver's beliefs from the Sender's perspective.

REFERENCES

AMBRUS, A. AND S. TAKAHASHI (2006): "Multi-sender Cheap Talk with Restricted State Space," manuscript, Harvard University.

AUSTEN-SMITH, D. AND J. BANKS (2000): "Cheap Talk and Burned Money," *Journal of Economic Theory* **91**, 1–16.

BANKS, J. AND J. SOBEL (1987): "Equilibrium Selection in Signaling Games," *Econometrica* **55**, 647–662.

BATTAGLINI, M. (2002): "Multiple Referrals and Multidimensional Cheap Talk," *Econometrica* **70**, 1379–1401.

Blume, A. (1994): "Equilibrium Refinements in Sender-Receiver Games," *Journal of Economic Theory* **64**, 66–77.

Cho, I. and D. Kreps (1987): "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics* **102**, 179–221.

Cho, I. and J. Sobel (1990): "Strategic Stability and Uniqueness in Signaling Games," *Journal of Economic Theory* **50**, 381–413.

Crawford, V. and J. Sobel (1982): "Strategic Information Transmission," *Econometrica* **50**, 1431–1451.

Farrell, J. (1993): "Meaning and Credibility in Cheap-talk Games," *Games and Economic Behavior* **5**, 514–531.

Fudenberg, D. and D. Levine (1993): "Self-confirming Equilibrium," *Econometrica* **61**, 523–545.

Fudenberg, D. and J. Tirole (1991a): *Game Theory*, MIT Press, Cambridge MA.

Fudenberg, D. and J. Tirole (1991b): "Perfect Bayesian Equilibrium and Sequential Equilibrium," *Journal of Economic Theory* **53**, 236–260.

Gilboa, I. and D. Schmeidler (1989): "Maximin Expected Utility with a Non-unique Prior," *Journal of Mathematical Economics* **18**, 141–153.

Grossman, S. J. and M. Perry (1986): "Perfect Sequential Equilibrium," *Journal of Economic Theory* **39**, 97–119.

Kartik, N. (2005): "Information Transmission with Almost-Cheap Talk," manuscript, UCSD.

Kim, Y.-G. and J. Sobel (1995): "An Evolutionary Approach to Pre-play Communication," *Econometrica* **63**, 1181–1193.

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

Kreps, D. and R. Wilson (1982): "Reputation and Imperfect Information," *Journal of Economic Theory* **27**, 253–279.

Matthews, S., M. Okuno-Fujiwara, and A. Postlewaite (1991): "Refining Cheap-Talk Equilibria," *Journal of Economic Theory* **55**, 247–273.

Morgan, J. and V. Krishna (2001): "A Model of Expertise," *Quarterly Journal of Economics* **116**, 747–775.

Myerson, R. (1989): "Credible Negotiation Statements and Coherent Plans," *Journal of Economic Theory* **48**, 264–303.

Rabin, M. (1990): "Communication between Rational Agents," *Journal of Economic Theory* **51**, 144–170.

Ramey, G. (1996): "D1 Signaling Equilibria with Multiple Signals and a Continuum of Types," *Journal of Economic Theory* **69**, 508–531.

Riley, J. (1979): "Informational Equilibrium," *Econometrica* **47**: 331-359.

Spence, M. (1973): "Job Market Signaling," *Quarterly Journal of Economics* **87**, 355–374.

van Damme, E. (1989): "Stable equilibria and forward induction," *Journal of Economic Theory* **48**, 476–496.

van Damme, E. (1991): *Stability and Perfection of Nash Equilibria*, Springer Verlag, Berlin.

6 Appendix: Proofs of Lemmas

Lemma 1 *Fix an equilibrium (M, A, μ) and type $\theta^h \in \mathbb{R}_+$. For all $m' > M(\theta^h)$ and all $\theta^\ell < \theta^h$, $\hat{a}(\theta^h, m') < \infty$ implies $\hat{a}(\theta^\ell, m') > \hat{a}(\theta^h, m')$.*

Proof: Denote $m^h \equiv M(\theta^h)$. Since θ^h sends m^h in equilibrium, we have $\hat{a}(\theta^h, m^h) = A(m^h)$ (by definition and uniqueness of $\hat{a}()$). Therefore the incentive constraints for θ^ℓ yield

$$u_S(\theta^\ell, m^h, \hat{a}(\theta^\ell, m^h)) \equiv u_S^*(\theta^\ell) \geq u_S(\theta^\ell, m^h, \hat{a}(\theta^h, m^h))$$

implying $\hat{a}(\theta^\ell, m^h) \geq \hat{a}(\theta^h, m^h)$ by monotonicity of u_S.

The derivative of (3) with respect to m is zero, so

$$\frac{\partial \hat{a}}{\partial m}(\theta, m) = -\frac{\partial u_S/\partial m}{\partial u_S/\partial a}(\theta, m, \hat{a}(\theta, m))$$

for any θ and m. The single crossing assumption implies

$$-\frac{\partial u_S/\partial m}{\partial u_S/\partial a}(\theta^\ell, m, a) > -\frac{\partial u_S/\partial m}{\partial u_S/\partial a}(\theta^h, m, a)$$

for any m and a. Combining these observations, for any m we have

$$\left[\hat{a}(\theta^\ell, m) = \hat{a}(\theta^h, m)\right] \quad \Longrightarrow \quad \frac{\partial \hat{a}}{\partial m}(\theta^\ell, m) > \frac{\partial \hat{a}}{\partial m}(\theta^h, m). \tag{7}$$

Recall that $\hat{a}(\theta^\ell, m^h) \geq \hat{a}(\theta^h, m^h)$. If there exists $m' > m^h$ such that $\hat{a}(\theta^\ell, m') \leq \hat{a}(\theta^h, m')$, then there exists $m'' \in (m^h, m']$ such that both $\hat{a}(\theta^\ell, m'') \leq \hat{a}(\theta^h, m'')$ and $\partial \hat{a}/\partial m(\theta^\ell, m'') \leq \partial \hat{a}/\partial m(\theta^h, m'')$, contradicting (7). □

Lemma 2 *Fix an equilibrium (M, A, μ), and suppose there exists a type θ' and an out-of-equilibrium message m' such that*

$$\begin{aligned} u_S^*(\theta') &< u_S(\theta', m', \beta(\theta', m')) \quad \textit{and} \\ u_S^*(\theta) &\geq u_S(\theta, m', \beta(\theta', m')) \quad \forall \theta < \theta'. \end{aligned}$$

Then there exists a unique credible deviators' club for m'.

PROOF: Let θ' and m' satisfy the inequalities in the lemma. We show that

$$C' = \{\theta \in \Theta : u_S^*(\theta) < u_S(\theta, m', \beta(\theta', m')) \tag{8}$$

is the unique set of types to satisfy (1).

The inequalities imply $\theta' \in C'$, and in fact that $\theta' = \min C'$. Perhaps C' contains some $\theta > \theta'$ (if any exist). Regardless, by monotonicity of $u_S()$ in a and by monotonicity of the Receiver's best response with respect to beliefs, respectively, we have the following for any $\theta \in \Theta$.

$$\begin{aligned} \min_{a \in BR(C', m')} u_S(\theta, m', a) &= u_S(\theta, m', \min BR(C', m')) \\ &= u_S(\theta, m', BR(\min C', m')) \\ &= u_S(\theta, m', \beta(\theta', m')) \end{aligned} \tag{9}$$

Hence C' satisfies (1) with respect to m' (showing existence).

To show that C' is the unique such set, let C satisfy (1). For any $\theta < \theta'$, monotonicity of the Receiver's best response implies

$$u_S(\theta, m', \beta(\theta, m')) < u_S(\theta, m', \beta(\theta', m')) \leq u_S^*(\theta)$$

where the last inequality follows from the lemma's assumption. Hence no such type can belong to a deviators' club for m'. Hence $\min C \geq \theta'$.

If $\min C = \theta > \theta'$, then again by monotonicity of the Receiver's best responses (and that any distribution over C stochastically dominates the degenerate one on θ'),

$$u_S^*(\theta') < u_S(\theta', m', \beta(\theta', m')) < u_S(\theta', m', a)$$

for any $a \in BR(C, m')$. But this contradicts the fact $\theta' \notin C$. Hence $\theta' = \min C$.

By (9), a credible deviators' club is uniquely determined by its minimum element; no two distinct clubs can have the same minimum element. Hence $C = C'$ defined by (8). □

LEMMA 3 *If an equilibrium (M, A, μ) is not Vulnerable to Credible Devia-*

tions, it is a separating equilibrium—no two types send the same message.

PROOF: Suppose to the contrary that some equilibrium message m^e is sent by several types, the highest of which is θ'.

Note that $A(m^e) < \beta(\theta', m^e)$ because θ' is the highest of several types that sends m^e (and due to the assumptions on u_R). Therefore

$$u_S^*(\theta') < u_S(\theta', m^e, \beta(\theta', m^e))$$

i.e. θ' would be better off if the Receiver "knew" it was θ' sending m^e and best-responded accordingly.

We claim that there exists $m'' > m^e$ such that both

$$u_S^*(\theta') = u_S(\theta', m'', \beta(\theta', m''))$$

and $u_S(\theta', m'', \beta(\theta', m''))$ is locally decreasing in m. To see this, it is enough to observe that u_S tends to $-\infty$ as $m \to \infty$, and $\beta(\theta', m)$ is bounded from above by assumption.

By choice of m'', $\hat{a}(\theta', m'') = \beta(\theta', m'') < \infty$. By Lemma 1, for all $\theta < \theta'$ we have $\hat{a}(\theta, m'') > \hat{a}(\theta', m'')$, and hence $u_S^*(\theta) > u_S(\theta, m'', \beta(\theta', m''))$.

By continuity, there is an out of equilibrium message $m' < m''$ (sufficiently close to m'') such that

$$\begin{aligned} u_S^*(\theta') &< u_S(\theta', m', \beta(\theta', m')) \quad \text{and} \\ u_S^*(\theta) &> u_S(\theta, m', \beta(\theta', m')) \quad \forall \theta < \theta'. \end{aligned}$$

By Lemma 2, there exists a unique deviators' club with respect to m'. □

LEMMA 4 *Any equilibrium whose outcome is different from the Riley outcome is Vulnerable to Credible Deviations.*

PROOF: Suppose an equilibrium (M, A, μ) is not Vulnerable to Credible Deviations. By Lemma 3 it is separating: $1 \leq i \neq j \leq n$ implies $M(\theta_i) \neq M(\theta_j)$. If the Sender uses Riley messages ($M(\theta_i) \equiv m_i^r$), the Receiver responds accordingly, and we are done.

Otherwise, let θ_i be the lowest type such that $M(\theta_i) \neq m_i^r$. For any $j < i$, we have

$$u_S^*(\theta_j) = u_S(\theta_j, m_j^r, \beta(\theta_j, m_j^r)) \geq u_S(\theta_j, M(\theta_i), \beta(\theta_i, M(\theta_i)))$$

by incentive compatibility. Therefore $M(\theta_i)$ does not maximize $u_S(\theta_i, m, \beta(\theta_i, m))$ subject to the constraints of (5) (since the maximizer m_i^r is unique, by strict quasi-concavity in m).

That is,

$$u_S^*(\theta_i) = u_S(\theta_i, M(\theta_i), \beta(\theta_i, M(\theta_i))) < u_S(\theta_i, m_i^r, \beta(\theta_i, m_i^r))$$

By Lemma 2, there exists a unique deviators' club for message m_i^r. □

LEMMA 5 *Suppose an equilibrium (M, A, μ) is immune to Credible Deviations. If for some $s \geq 2$, the types θ_s, $\theta_{s+1}, \ldots, \theta_n$ are all separating (i.e. send unique equilibrium messages), then eqn. (6) holds for all $i \geq s$.*

PROOF: To derive a contradiction under the hypothesis of the lemma, let $\theta_j \geq \theta_s$ be the highest type for whom eqn. (6) fails; we prove the lemma by showing that $\{\theta_j\}$ forms a unique credible deviators' club. Throughout the proof, denote equilibrium messages $m_i \equiv M(\theta_i)$ and actions $a_i \equiv A(M(\theta_i))$.

(EXISTENCE) Incentive compatibility implies

$$d(\theta_{j-1} - a_{j-1}) - d(\theta_{j-1} - a_j) < c(\theta_{j-1}, m_j) - c(\theta_{j-1}, m_{j-1}) \tag{10}$$

where the strictness follows from the choice of j. By assumption, either θ_{j-1} is a separating type, or pools only with lower types. Therefore the Receiver's response to m_{j-1} satisfies $a_{j-1} \leq \theta_{j-1} - b < \theta_j - b = a_j$. With our assumption that the bias is not too small, this makes the left hand side of (10) positive. The right hand side then implies $m_j > m_{j-1}$.

By the convexity of d, for any $\ell < j - 1$ we have

$$d(\theta_\ell - a_{j-1}) - d(\theta_\ell - (\theta_j - b)) \leq d(\theta_{j-1} - a_{j-1}) - d(\theta_{j-1} - (\theta_j - b)).$$

By the single-crossing property of c,

$$c(\theta_{j-1}, m_j) - c(\theta_{j-1}, m_{j-1}) < c(\theta_\ell, m_j) - c(\theta_\ell, m_{j-1}).$$

Combining the latter two inequalities with (10) we get

$$d(\theta_\ell - a_{j-1}) - d(\theta_\ell - (\theta_j - b)) < c(\theta_\ell, m_j) - c(\theta_\ell, m_{j-1}).$$

The incentive constraint for θ_ℓ not to send m_{j-1} is

$$d(\theta_\ell - a_\ell) - d(\theta_\ell - a_{j-1}) \leq c(\theta_\ell, m_{j-1}) - c(\theta_\ell, m_\ell).$$

Adding it to the previous inequality yields

$$d(\theta_\ell - a_\ell) - d(\theta_\ell - (\theta_j - b)) < c(\theta_\ell, m_j) - c(\theta_\ell, m_\ell) \tag{11}$$

for all $\ell < j - 1$. With (10) this establishes that any $\theta_\ell < \theta_j$ *strictly* prefers his equilibrium payoff to imitating type θ_j.

In the case that $j < n$, types θ_j and θ_{j+1} both separate by assumption, and θ_j is indifferent between sending m_j and m_{j+1}:

$$d(\theta_j - (\theta_j - b)) - d(\theta_j - (\theta_{j+1} - b)) = c(\theta_j, m_{j+1}) - c(\theta_j, m_j).$$

Since $\theta_{j+1} - \theta_j < 2b$, the left hand side of the equality is positive.

By the convexity of d and the single-crossing property of c, for all $h > j$ we have

$$d(\theta_h - (\theta_j - b)) - d(\theta_h - (\theta_{j+1} - b)) > c(\theta_h, m_{j+1}) - c(\theta_h, m_j).$$

The incentive constraint for $\theta_h > \theta_{j+1}$ (if any exist) not to send m_{j+1} is

$$d(\theta_h - (\theta_{j+1} - b)) - d(b) \geq c(\theta_h, m_h) - c(\theta_h, m_{j+1}).$$

Adding the last two inequalities yields

$$d(\theta_h - (\theta_j - b)) - d(b) > c(\theta_h, m_h) - c(\theta_h, m_j).$$

This establishes that any $\theta_h > \theta_j$ *strictly* prefers his equilibrium payoff to imitating type θ_j.

By continuity, this implies that $C = \{\theta_j\}$ satisfies (1) (forms a deviators' club) for any message $m_j - \varepsilon$, as long as $\varepsilon > 0$ is kept sufficiently small so as not to violate the strict inequalities established above. Only type θ_j would gain from sending $m_j - \varepsilon$ if the Receiver would react to it with the action $a = \theta_j - b$.

(Uniqueness) To finish the proof, we need to show that there can exist no other deviators' club C for such a message $m_j - \varepsilon$, whenever ε is sufficiently small.

For $\theta_\ell < \theta_j$ to belong to a credible deviators' club requires that he gain even when the Receiver believes the message came from θ_ℓ, i.e.

$$d(\theta_\ell - (\theta_\ell - b)) - d(\theta_\ell - a_\ell) < c(\theta_\ell, m_\ell) - c(\theta_\ell, m_j - \varepsilon).$$

Adding this to (11) (or (10) for $\ell = j - 1$) yields

$$d(\theta_\ell - (\theta_\ell - b)) - d(\theta_\ell - (\theta_j - b)) < c(\theta_\ell, m_j) - c(\theta_\ell, m_j - \varepsilon).$$

The left hand side of this inequality, which can be written $d(b) - d(b - (\theta_j - \theta_\ell))$, is positive because $0 < \theta_j - \theta_\ell < 2b$. Hence for sufficiently small $\varepsilon > 0$, this inequality is violated; $\theta_\ell < \theta_j$ cannot belong to *any* credible deviators' club C for message $m_j - \varepsilon$, when $\varepsilon > 0$ is sufficiently small.

On the other hand, suppose some deviators' club for $m_j - \varepsilon$ consisted only of types higher than θ_j. Similar reasoning as above implies that θ_j would want to "join that club" since $|\theta_j - (\theta_h - b)| < |\theta_j - (\theta_j - b)|$ when $\theta_h < \theta_j$, i.e. θ_j is even better off when the Receiver believes the message was sent by θ_h than when the Sender believes it was θ_j. This contradicts the fact that such a club C exists without θ_j.

Therefore, any such club C must contain θ_j. But we have already proven that no other type gains by sending $m_j - \varepsilon$ when the Receiver chooses $a =$

$\theta_j - b$. We conclude that $\{\theta_j\}$ is the unique deviators' club for (any out-of-equilibrium) message $m_j - \varepsilon$ when $\varepsilon > 0$ is chosen sufficiently small, making the equilibrium Vulnerable to a Credible Deviation. □

LEMMA 6 *Suppose a non-separating equilibrium exists, and let θ_p denote the highest pooling type. If eqn. (6) holds for all $i > p$, then there exists a message for which $\{\theta_p\}$ is a unique credible deviators' club.*

PROOF: Throughout the proof, denote equilibrium messages $m_i \equiv M(\theta_i)$ and actions $a_i \equiv A(M(\theta_i))$.

Let $\hat{m}_p$ denote the message that would give θ_p his his equilibrium payoff if the Receiver would respond with action $a = \theta_p - b$, i.e.

$$d(\theta_p - a_p) - d(\theta_p - (\theta_p - b)) = c(\theta_p, \hat{m}_p) - c(\theta_p, m_p). \tag{12}$$

We eventually prove that $\{\theta_p\}$ is a unique singleton deviators' club for some message $\hat{m}_p - \varepsilon$.

First, we show that for all $i \neq p$, if θ_i would send $\hat{m}_p$ and the Receiver would respond with $a = \theta_p - b$, then θ_i would be strictly worse off than he is in equilibrium, i.e.

$$d(\theta_i - a_i) - d(\theta_i - (\theta_p - b)) < c(\theta_i, \hat{m}_p) - c(\theta_i, m_i). \tag{13}$$

To prove this claim we separately address types lower and higher than θ_p.

(LOW TYPES) Since θ_p is the highest pooling type, the Receiver's response to his equilibrium message is $a_p < \theta_p - b$. This implies that the left hand side of eqn. (12) is positive, hence $\hat{m}_p > m_p$.

By the convexity of d, for all $\ell < p$ we have

$$d(\theta_\ell - a_p) - d(\theta_\ell - (\theta_p - b)) \leq d(\theta_p - a_p) - d(\theta_p - (\theta_p - b)).$$

Since $\hat{m}_p > m_p$, the single-crossing property of c implies that for all $\ell < p$,

$$c(\theta_p, \hat{m}_p) - c(\theta_p, m_p) < c(\theta_\ell, \hat{m}_p) - c(\theta_\ell, m_p).$$

Combining these two inequalities with eqn. (12) yields

$$d(\theta_\ell - a_p) - d(\theta_\ell - (\theta_p - b)) < c(\theta_\ell, \hat{m}_p) - c(\theta_\ell, m_p)$$

while the incentive constraint for θ_ℓ not to imitate θ_p is

$$-d(\theta_\ell - a_p) - c(\theta_\ell, m_p) \leq -d(\theta_\ell - a_\ell) - c(\theta_\ell, m_\ell).$$

By adding the previous two inequalities, we get

$$d(\theta_\ell - a_\ell) - d(\theta_\ell - (\theta_p - b)) < c(\theta_\ell, \hat{m}_p) - c(\theta_\ell, m_\ell)$$

for all $\ell < p$.

(High Types) Lemma 5 says that θ_{p+1} (if it exists) separates from θ_p at the least cost, that is,

$$-d(\theta_p - (\theta_{p+1} - b)) - c(\theta_p, m_{p+1}) = -d(\theta_p - a_p) - c(\theta_p, m_p). \tag{14}$$

Since $|\theta_p - (\theta_{p+1} - b)| < |\theta_p - (\theta_p - b)|$, this equality with eqn. (12) implies $\hat{m}_p < m_{p+1}$.

Combine eqns. (12) and (14) to get

$$d(\theta_p - (\theta_p - b)) - d(\theta_p - (\theta_{p+1} - b)) = c(\theta_p, m_{p+1}) - c(\theta_p, \hat{m}_p).$$

By the convexity of d and the fact $\theta_{p+1} > \theta_p$, for all $h > p$ we have

$$d(\theta_h - (\theta_p - b)) - d(\theta_h - (\theta_{p+1} - b)) \geq d(\theta_p - (\theta_p - b)) - d(\theta_p - (\theta_{p+1} - b)).$$

By the single-crossing property of c and $m_{p+1} > \hat{m}_p$, for all $h > p$,

$$c(\theta_p, m_{p+1}) - c(\theta_p, \hat{m}_p) > c(\theta_h, m_{p+1}) - c(\theta_h, \hat{m}_p).$$

Therefore, for all $\theta_h > \theta_p$,

$$-d(\theta_h - (\theta_{p+1} - b)) - c(\theta_h, m_{p+1}) > -d(\theta_h - (\theta_p - b)) - c(\theta_h, \hat{m}_p).$$

The incentive constraint for θ_h not to imitate type θ_{p+1} is

$$-d(b) - c(\theta_h, m_h) \geq -d(\theta_h - (\theta_{p+1} - b)) - c(\theta_h, m_{p+1}).$$

The last two inequalities imply that for all $\theta_h > \theta_p$,

$$-d(b) - c(\theta_h, m_h) > -d(\theta_h - (\theta_p - b)) - c(\theta_h, \hat{m}_p).$$

This establishes (13).

We finish the proof by arguing that for any sufficiently small ε, $\{\theta_p\}$ is the unique credible deviators' club with respect to message $\hat{m}_p - \varepsilon$. Since these arguments are mostly the same as those used in the end of the proof of Lemma 5, we keep these arguments brief.

Continuity in eqn. (13) implies that for sufficiently small ε, $C = \{\theta_p\}$ satisfies (1) with respect to message $\hat{m}_p - \varepsilon$.

To show that no *other* deviators' club C can exist, first consider $\theta_\ell < \theta_p$. Since $0 < \theta_p - \theta_\ell < 2b$, any such θ_ℓ would prefer the Receiver to take action $\theta_p - b$ rather than $\theta_\ell - b$. Hence by transitivity and (13), θ_ℓ cannot belong to a deviators' club for message $\hat{m}_p - \varepsilon$, as in the proof of Lemma 5.

Finally, if a deviators' club consisted only of higher types θ_h, θ_p would want to join that club, which is a contradiction. Hence θ_p belongs to any such C, in which case (13) implies θ_h could not belong to the club for message $\hat{m}_p - \varepsilon$, preferring his equilibrium payoff to the one he gets when the Receiver responds with action $a = \theta_p - b$. □