Durable Cheap Talk Equilibria*

Sandeep Baliga
M.E.D.S. Department, Kellogg School of Management

Tomas Sjöström
Department of Economics, Rutgers University

July, 2006

Abstract

We consider a cheap-talk game with one sender and one receiver. If the receiver does not commit to listen to only one message, the equilibrium refinements due to Farrell [5], Grossman and Perry [7] and Matthews, Okuno-Fujiwara and Postlewaite [11] are no longer applicable. We discuss different notions of *durability* and *sequential credibility* when a message can later be followed by more messages, and both parties know this.

1 Introduction

Cheap talk games have been studied by Crawford and Sobel [3], Farrell [5], Grossman and Perry [7], Matthews, Okuno-Fujiwara and Postlewaite [11] and Myerson [12].[1] These authors mainly consider a setting where an informed party, the sender, sends costless messages to an uninformed party, the receiver, who must make some decision. The receiver's optimal decision depends on the sender's information. The focus is on one-shot games where the sender talks just once and then the receiver takes an action immediately. Equilibrium refinements postulate that receivers take the literal meaning of messages seriously. Baliga, Corchon and Sjöström [1] apply this model to an implementation problem where the receiver is a principal who cannot commit to an outcome function (a map from messages to outcomes). However, as the game is one-shot, the principal in effect irrevocably commits to cut off communication after the first message is received. In the present paper, we take the lack of commitment one step further by assuming the principal cannot cut off communication. We define

*We are grateful to Matt Jackson for encouraging us to work on this problem and to Stephen Morris for comments.

[1] See also Maskin and Tirole [10] for a discussion of Farrell [5] and Grossman and Perry [7].

durable neologism proof equilibrium and *durable announcement-proof equilibrium,* inspired by Farrell [5] and Grossman and Perry [7] (henceforth FGP) and Matthews, Okuno-Fujiwara and Postlewaite [11] (henceforth MOP) respectively.

Allowing repeated communication can make partial pooling more difficult. Suppose the cheap talk game takes place on Monday. If after listening to the message on Monday, the receiver becomes convinced that the true type belongs to a subset S of types, then the receiver may initiate further information transmission to try to distinguish a subset of S. One can imagine the receiver returning on Tuesday to give the agent another chance to speak. Of course, the outcome on Tuesday must be consistent with the theory, as there is the possibility to come back again on Wednesday, etc. So our equilibrium concept is essentially recursive. If after having been convinced on Monday that the true state belongs to the set S, the receiver decides to return on Tuesday to sort things out, then the situation on Tuesday is similar to the situation on Monday *except* that he has now ruled out those types that are not in S. The fact that the receiver can continue to "interrogate" the agent can destroy a partially pooling equilibrium and improve information transmission. Moreover, separating equilibria which are *not* neologism- or announcement-proof may be durable if objections or "announcements" that would destroy the separating equilibrium of the one-shot game are not durable objections (i.e. could be broken apart by further interrogation). Therefore, separation of types may be easier with repeated communication (see Example 1 below).

In other cases, repeated communication can lead to *more* pooling of types. This is because credible "objections" [2] can be destroyed if the receiver cannot (credibly) promise not to keep interrogating the sender after he made the objection. As such interrogation can destroy the objection, a lack of commitment to stop talking can lead to more pooling of types (see Example 2). This result is relevant for the interpretation of Baliga, Corchon and Sjöström [1], who (implicitly) assume the receiver commits to listen to one message only. Not being able to commit to stop talking can make the receiver worse off if the repeated communication leads to more pooling. Of course, it is known from other models that inability to commit may make a principal worse off (Dewatripont and Maskin [4]).

In Section 3 we present our solution concept as a *durability* requirement for the one-shot cheap talk model: at no point should the receiver have an incentive to return to get more information. (We do not allow the sender to send several messages in a row unless explicitly permitted by the receiver). Consider the information partition generated by the sender's equilibrium strategy, and consider a set which belongs to this partition. Assume the receiver becomes convinced that the true state belongs to this set. Does he have a (strict) incentive to return for more information? If so, the equilibrium is not durable. Moreover, we test objections the same way. The way we define both *durable neologism proof* and *durable announcement proof* equilibria. The method is similar to Holmström and Myerson [8], except that we have given the

[2] The objections are called *neologisms* by Farrell and *announcements* by MOP.

receiver all of the power to renegotiate. It is also in the spirit of Dewatripont and Maskin [4], Laffont and Martimort [9] and Maskin and Tirole [10] who take the view that a renegotiation or collusion offer made by an uninformed party in the presence of asymmetric information does not suffer from the signalling problem of an offer from an informed party.

Our recursive notion is also in the spirit of Bernheim, Peleg and Whinston's [2] recursive equilibrium concept for games with complete information. In their terminology, a Nash equilibrium is coalition-proof if no coalition of players C can deviate and be better off and there is no sub-coalition $C' \subset C$ that can in turn profitable deviate from the deviation. Also, any potential deviation by the sub-coalition C' must be judged by the same criterion and so on. In our context, a cheap-talk equilibrium concept is durable if no coalition of "types" T has a credible objection and there is no sub-coalition of types $T' \subset T$ that can credibly object to the objection. Moreover, any credible objection by the sub-coalition of types T' must be judged by the same criterion, and so on.

In Section 4, we formally consider a *sequential cheap-talk game* where the sender keeps sending new messages each period until an action is taken. In contrast to the one-shot game, the sender may reveal information slowly over time, and the definition of durability must be modified. If the receiver only listens to *neologisms* of the form "my type is in the set S," then we must allow multi-stage objections of the following form. On day 1 the receiver claims to be in set S, say $S = \{a, b\}$. The receiver anticipates that more precise information will be given on day 2; on day 2 type a says "I am a" and type b says "I am b." (Of course, the relevant incentive compatibility constraints must be satisfied). This leads to a definition of *sequential neologism proof equilibria*. However, the possibility of revealing information slowly over time is spurious if a multi-stage credible neologism could be collapsed to a single more complicated announcement on day 1. Rather than saying "I am in set $S = \{a, b\}$" on day 1, and then "I am a" on day 2, the sender could say all at once: "I am in set $S = \{a, b\}$; if I were type a I would say so, and if I were type b I would say so; actually I am a." The receiver who is sophisticated enough to understand the multi stage speeches should be able to understand this single (but more complex) message. Such complex messages are called *announcements* by Matthews, Okuno-Fujiwara and Postlewaite [11]. Consequently, the definition of sequential neologism proof equilibria is formally equivalent to the definition of durable announcement proof equilibria. One can think of announcements as the collapsing of a sequence of neologisms into one message. Neologisms have the advantage of being simpler speeches. However, sequential neologisms require sophisticated forecasting ability on the part of the receiver, while durable announcements do not.

One final point should be made. Farrell's original argument (see Farrell and Rabin [6]) was that receivers should take the meaning of a message in the natural language as a starting point, and then ask "why would he want me to think that?". In this spirit, we assume that if the sender on Monday claims his type is either a or b,

but when being further interrogated on Tuesday "confesses" that it is actually b, the receiver takes the literal meaning of these *internally consistent* messages as something to be tested for its credibility, "why would he want me to think that his type is b?". On the other hand, no literal meaning can be assigned to *contradictory* statements. Suppose the sender on Monday convinces the receiver that his type is either a or b, but when the receiver returns on Tuesday for more information the sender explains that his type is actually c. As the sender must have been lying either on Monday or on Tuesday, the receiver cannot be *required* to interpret *either* of these statement literally. In this paper we assume that if messages conflict in this way, the receiver disregards the latest message (so in the example he remains convinced that the true type is either a or b).

2 Definitions and Examples

There is one sender and one receiver. Let Θ denote the finite set of feasible states (or types), with generic state $\theta \in \Theta$. Let $\Delta(\Theta)$ denote the set of probability distributions over Θ. Given $T \subseteq \Theta$, let $\mathcal{P}(T)$ be the set of all partitions of T. The sender knows the true θ, but the receiver does not. The sender sends one message $m \in M$ to the receiver. Then the receiver takes an action a in his action space A. Following Farrell [5], we suppose the message space M is sufficiently rich to at least include all the subsets of Θ and "neologisms" to use to deviate from any message profile. A strategy for the sender is denoted $\mu : \Theta \to M$, where $\mu(\theta)$ is the message sent in state θ. Let $\mu^{-1}(m) = \{\theta \in \Theta : \mu(\theta) = m\}$ and $\mu(\Theta) = \{m \in M : \text{there is } \theta \in \Theta \text{ such that } \mu(\theta) = m\}$. A strategy for the receiver is denoted $\alpha : M \to A$, where $\alpha(m)$ is the action taken in response to message m. A strategy profile is denoted $\sigma = (\mu, \alpha)$. The payoff function for the sender is $u(a, \theta)$, for the receiver $v(a, \theta)$. For $q \in \Delta(\Theta)$ define

$$BR(q) \equiv \arg\max_{a \in A} \sum q(\theta) v(a, \theta)$$

Let p be the prior and let p_T be the probability distribution defined by

$$p_T(\theta) \equiv \frac{p(\theta)}{p(T)} \text{ if } \theta \in T$$

and $p_T(\theta) = 0$ if $\theta \notin T$. We write $p_\theta = p_{\{\theta\}}$ for the degenerate probability distribution which puts all probability on state θ, and write $B(T) \equiv BR(p_T)$ for $T \subseteq \Theta$. For simplicity we assume $B(T)$ is single-valued for all $T \subseteq \Theta$.

To motivate our equilibrium concept, consider Example 1. The typical entry in the matrix is the payoff pair $(u(Z, \theta), v(Z, \theta))$ when the true state is $\theta \in \{a, b, c, d\}$ (the row) and the receiver's action is $Z \in \{A, B, C, D, E, F\}$ (the column).

Example 1		Receiver					
		Action A	Action B	Action C	Action D	Action E	Action F
	Type a	2,5	0,0	0,0	0,0	3,1	4,2
Sender	Type b	0,0	2,5	0,0	0,0	3,1	4,2
	Type c	0,0	0,0	2,5	0,0	3,1	0,-8
	Type d	0,0	0,0	0,0	2,5	0,0	8,-8

Suppose $p(\theta) = 1/4$ for all $\theta \in \Theta = \{a, b, c, d\}$. This sender-receiver game has two perfect Bayesian equilibrium (PBE) outcomes. In the first PBE outcome, types a, b and c pool and receive E, while type d receives D. Thus,

$$\mu(a) = \mu(b) = \mu(c) = m \neq m' = \mu(d).$$

$$\alpha(m) = E$$

$$\alpha(m') = D$$

This outcome is neologism-proof.[3] For, a deviation can only be profitable to the sender if it results in the action F. However, this would attract type d, and then F would not be an optimal response for the receiver. On the other hand, if the receiver cannot commit to stop talking, then this equilibrium is not *durable* or *sequentially credible*. For in this equilibrium, the message m convinces the receiver that the true state is in the set $\{a, b, c\}$, each state in this set being equally likely. Suppose rather than taking the action E following message m, the receiver gives the sender one more chance to speak (he cannot commit not to do this). What follows is a cheap talk game where the d-row has been deleted. Complete pooling of the three types a, b and c is implausible here because, using the logic of FGP, if the sender's true type is a or b, then he should say "I am a or b," and following this speech the receiver should assign probability one half each to a and b and take action F. Type c does not want to make this speech because action F is bad for him. Type d likes action F, but this type has been ruled out by hypothesis: message m has *already* convinced the receiver that the type is not d, and (we suppose) the fact that the receiver *himself* returns for more information does not change his belief that the sender is not d. Moreover, after the receiver has ruled out both types c and d, but thinks types a and b are equally likely, there is no way to return yet another time to separate a from b: neither a nor b has any incentive to confess to his true type (because by not confessing he expects F). Therefore, the receiver will be able to obtain more information by further interrogation following message m so this equilibrium is not *durable neologism-proof* or a *sequential neologism-proof*. But, is the sender's speech a credible sequential

[3] In Examples 1 and 2, the set of neologism proof equilibria is identical to the set of equilibria that are announcement-proof in the sense of Matthews, Okuno-Fujiwara and Postlewaite [11]. For simplicity, we couch the discussion in terms of neologisms rather than announcements.

neologism where the subset of types $\{a, b, c\}$ separates out over two periods in to the two subsets $\{a, b\}$ and $\{c\}$, and the final outcome is F for types a and b? This cannot be the case because type d would then be able to get F by imitating a or b.

PBE outcome number two is fully separating: $\mu(\theta) \neq \mu(\theta')$ for all θ, θ'. It is not neologism-proof, because if the sender is type a, b or c he should say "I am a, b or c but I won't tell you which one," and the response should be E. (This is the only credible objection the sender can make at this equilibrium, because any speech which leads the receiver to take the action F attracts type d, but then F is not the best response.) However, the speech "I am a, b or c" is not durable. For, suppose the receiver really does believe this speech and is about to take the action E. As argued before, if the receiver instead gives the senders one more chance to speak, then types a and b should say "I am a or b," the receiver should respond with F, and there is no possibility of further communication. So the initial objection cannot be believable. Since that was the only possible objection against the equilibrium, the separating outcome *is* durable neologism-proof. Nor is there a sequential objection where the subset of types $\{a, b, c\}$ reveal information over two stages and break up into the subsets $\{a, b\}$ and $\{c\}$ - this means type d would want to make the initial objection "I am a, b or c" and then say "I am or a or b". Therefore, the separating outcome is also a sequential neologism-proof equilibrium.

Example 1 shows that repeated interrogation can lead to increased information transmission. On the one hand, it makes it possible for the receiver to destroy partially pooling equilibria, while on the other hand, the objections that would destroy a separating equilibrium in the one-shot game may not survive repeated interrogation and so become irrelevant. In Example 1, both effects combine to give greater information transmission in equilibrium.

A version of the Stiglitz critique applies our analysis. Consider the first of the perfect Bayesian equilibria, where the types $\{a, b, c\}$ pool. We argued that the receiver can press on with a further inquiry to distinguish a and b from c. If the sender predicts that this will occur, type d can pretend to be either a or b, which contradicts the hypothesis of the equilibrium. But what exactly is the cause of the contradiction ? Maybe the conclusion we should draw is that the sender *cannot* break up $\{a, b, c\}$ by further inquiry, so that the equilibrium is reasonable after all ? This critique is not valid. We are trying to show that a particular equilibrium where types a, b and c pool is not plausible. The proof is by contradiction. *If* the receiver becomes convinced that the true type is in $\{a, b, c\}$, *then* he can extract further information. This *if-then* statement follows from application of the logic of FGP to the matrix with the d-row deleted. The if-then statement should therefore be uncontroversial (unless we completely reject the FGP logic and give up the refinements even in one-shot games). Therefore, what must be wrong must be the original hypothesis that pooling of $\{a, b, c\}$ is part of a reasonable equilibrium. This is what we had to prove. This argument would be the same in a sequential game where information can be transmitted slowly. Suppose there is an equilibrium of the sequential game where,

after some sequence of messages, the receiver becomes convinced that the types belong to $\{a, b, c\}$. Consider the time t when he (according to his equilibrium strategy) is about to take the action E. Now suppose he deviates from his equilibrium strategy and asks for another message. After this deviation *by himself* he remains convinced (we suppose) that the sender's type is not d. Then, if the sender says "I am a or b" it should be believed, so pooling of the types a, b and c is not a plausible outcome of the sequential game. A similar argument applies to the analysis of the objections. Against the second (fully separating) equilibrium, "I am a, b or c" should not be a believable objection, because *if* this speech convinces the receiver that the true type is in $\{a, b, c\}$ *then* he can extract further information, which leads to a contradiction.

It is important for the analysis of Example 1 that once the receiver is convinced that the true type is in $\{a, b\}$, he cannot come back yet again to separate a from b. At some point, further interrogation becomes impossible, and this is the point. Durable neologism-proofness is defined recursively, using the stability or instability of smaller sets of pooling types to check the stability or instability of larger sets. The recursion is started off by noticing that a speech which reveals types completely cannot be further destabilized, so durability is trivially satisfied.

Example 1 shows that our equilibrium concept can be neither weaker nor stronger than neologism-proofness (or announcement-proofness, cf. footnote 3). It also shows that the threat to return for more information can make the receiver better off. That this is not always the case is seen in Example 2.

Example 2

		Receiver			
		Action A	Action B	Action C	Action D
	Type a	4,5	0,0	1,3	2,4
Sender	Type b	0,0	4,5	1,3	2,4
	Type c	3,-3	3,-3	1,3	-1,-1

Suppose $p(a) = p(b) = p(c) = 1/3$. This sender receiver game has two PBE outcomes. In the first PBE outcome, $\mu(a) = \mu(b) = m \neq m' = \mu(c)$, $\alpha(m) = D$ and $\alpha(m') = C$. This equilibrium outcome is neologism-proof. Any deviation by the sender that causes the receiver to respond with either A or B would attract type c, but then neither A nor B would be an optimal response. However, this outcome is not durable. For suppose the receiver is about to choose D, convinced that the sender is either a or b (but not c). If he does *not* take any action, but instead tells the sender "now I know you are either a or b, which one is it?", then type a should certainly say a and type b should certainly say b because once the c-row is deleted from the matrix any reasonable criterion predicts separation of a from b. And of course, once a and b have been separated, there is no more information to be obtained. This destroys the original equilibrium. The partially pooling outcome is not durable. Moreover, complete separation is also not a neologism-proof outcome as type c would them imitate type a (or type b).

There also exists a totally uninformative pooling equilibrium where $\mu(a) = \mu(b) = \mu(c) = m$ and the receiver responds with $\alpha(m) = C$. It is not neologism- proof. For

suppose the sender deviates by saying "I am either a or b." The best response to this speech, if believed, is D, which makes types a and b better off, but not type c. Thus, this speech is credible. (It is the only credible speech at the pooling equilibrium, for any objection that causes the receiver to respond with either A or B would be imitated by type c). But the speech "I am either a or b" is not durable, for if it is believed then we obtain the contradiction that the receiver can proceed to find out if it is a or b by further interrogation, as described above. Accordingly, the uninformative pooling outcome is a durable neologism-proof equilibrium.

Example 2 again shows the recursive nature of durability. Here, pooling of a and b is not durable, because the speech "I am a" should be believed if the only possibilities are a and b, and there is no question of the receiver coming back for more information once he is convinced that the type is a. This is used to show that pooling of $\{a, b, c\}$ *is* durable.

In Example 2, complete pooling results when the receiver cannot commit to stop talking after one message is received, while commitment would allow partial separation of types. Here, the fact that the receiver can initiate discussion destroys the partially pooling equilibrium, but the fully pooling equilibrium has no durable objections. In this example *the receiver would like to commit to talk only once*, and in particular he would like to promise not to try to distinguish between types a and b. If he cannot commit, he is made worse off.

3 Durability

Let $T \subseteq \Theta$ and let $\mathcal{T} = \{T_1, ..., T_J\}$ be a partition of T. A *neologism S with respect to $(T, \mathcal{T})$* is message consisting of a subset of types $S \subset T$. It is *credible* if for all T_j :

$$\begin{aligned} \text{(E1) } u(B(S), \theta) &> u(B(T_j), \theta) \text{ for } \theta \in T_j \cap S \text{ and} \\ \text{(E2) } u(B(S), \theta) &\leq u(B(T_j), \theta) \text{ for } \theta \in T_j \backslash S. \end{aligned}$$

If the number of types in S is s, $|S| = s$, then the objection S is of *size* s. A *neologism-proof equilibrium* is a perfect Bayesian equilibrium (μ, α) such that there is no credible neologism with respect to $(\Theta, \mathcal{T}^\mu)$, where $\mathcal{T}^\mu$ is the partitioning of Θ which is induced by the equilibrium messages. (That is, each $T_j \in \mathcal{T}^\mu$ satisfies $T_j = \mu^{-1}(m) \equiv \{\theta \in \Theta : \mu(\theta) = m\}$ for some $m \in \mu(\Theta)$.)

Our definition of *durable neologism-proof equilibrium* differs in two respects. First, suppose the equilibrium is (μ, α) and the receiver receives the message $m \in \mu(\Theta)$. The receiver is convinced that the sender's true type is in $\mu^{-1}(m)$ and is supposed to choose $\alpha(m) = B(\mu^{-1}(m))$. But if the receiver himself tries to obtain more information by allowing the sender to send one more message, this should not change his beliefs about the senders type. If complete pooling is a reasonable outcome of the signalling game with types restricted to the set $\mu^{-1}(m)$, then this consideration does not matter, but if complete pooling is not a reasonable outcome then the receiver can return "the

next day" to obtain more information. Of course, whether or not complete pooling in $\mu^{-1}(m)$ is a reasonable outcome of a cheap talk game is itself something to be tested using our criterion. Hence our definition is recursive.

The second way in which our definition differs is that it will not be enough for an objection of size greater than one to be credible to be believed, if it can be followed by further information transmission. If the receiver becomes convinced by the objection that the sender's true type belongs to S, but complete pooling is not a reasonable outcome of the cheap-talk game when types are restricted to the set S, then the receiver should be able to obtain more information.

These two differences are illustrated in Example 1. It cannot be an equilibrium outcome for $\{a, b, c\}$ to pool, because the receiver will be able to get more information. For the same reason, at the separating equilibrium the receiver cannot believe the objection "I am a, b or c." In both cases, pooling of the types a, b and c is not *durable*, because once the receiver is convinced the state is in $\{a, b, c\}$ he can press on to distinguish a and b from c (but after this there can be no more information transmission because pooling of a and b *is* durable).

We now give the formal recursive definition of durability. A credible neologism of size 1 is always durable. Suppose durability has been defined for credible neologisms of size at most $s-1$. If a credible neologism S is of size s, then it is durable if there is no credible and durable neologism with respect to $(S, \{S\})$. Here $\{S\}$ means the partition of S that has only one element, the whole of S. Notice that a neologism with respect to $(S, \{S\})$ would be of size at most $s-1$. This way durability is defined for neologisms of any size.

Definition 1 *A perfect Bayesian equilibrium $\sigma = (\mu, \alpha)$ is a* durable neologism-proof equilibrium *if: (i) there is no credible and durable neologism with respect to $(\Theta, \mathcal{T}^\mu)$; and (ii) for all $m \in \mu(\Theta)$, there is no credible and durable neologism with respect to $(\mu^{-1}(m), \{\mu^{-1}(m)\})$.*

Definition 2 *A partition $\mathcal{T}$ of Θ is a* durable neologism-proof partition *if there exists a durable neologism-proof equilibrium (μ, α) such that $\mathcal{T}^\mu = \mathcal{T}$.*

Consider Example 2. Observe that the neologism $\{a\}$ is credible with respect to $(S, \{S\})$, where $S = \{a, b\}$. This is because $B(S) = D$ and $B(a) = A$, and $u(A, a) > u(D, a)$ while $u(A, b) < u(D, b)$. It is trivially also durable. Now the first PBE generates the information partition $\mathcal{T}^\mu = \{\mu^{-1}(m), \mu^{-1}(m')\}$, where $\mu^{-1}(m) = \{a, b\}$ and $\mu^{-1}(m') = \{c\}$. There is no credible and durable neologism with respect to $(\Theta, \mathcal{T}^\mu)$, because any deviation that results in A or B attracts c. Thus, condition (i) of Definition 1 is satisfied. But condition (ii) is violated: there exists a credible and durable neologism $\{a\}$ with respect to $(S, \{S\})$ if $S = \mu^{-1}(m) = \{a, b\}$. Thus, this PBE is not durable neologism proof. In the second PBE, $\mu(\Theta) = m$ which generates the completely pooling information partition $\{a, b, c\}$. Here conditions (i) and (ii) both amount to checking for objections with respect to $(\Theta, \{\Theta\})$. There is

only one credible neologism, $S = \{a, b\}$, but this neologism is not durable because $\{a\}$ is credible and durable with respect to $(S, \{S\})$. Thus, both conditions (i) and (ii) are satisfied, so this PBE is durable neologism proof.

Given that the message space M is arbitrarily rich, messages may convey far more information than just a subset of types. This idea was formalized by Matthews, Okuno-Fujiwara and Postlewaite [11] by introducing *announcements*.[4] Fix any $T \subseteq \Theta$ and let $\{T_1, ..., T_J\}$ be a partitioning of T. Let $D \subset T$ be a non-empty set of *deviant types*. An *announcement* (D, δ, m) consists of a set of deviant types D, a *talking strategy* $\delta : D \to M$, and a particular message m, where

$$m \in \delta(D) \equiv \{m \in M : \delta(\theta) = m \text{ for some } \theta \in D\}.$$

The size s of the announcement is

$$s = \max_{m \in \delta(D)} \left|\delta^{-1}(m)\right|$$

For $m \in \delta(D)$ let $\delta^{-1}(m) \equiv \{\theta \in D \mid \delta(\theta) = m\}$ denote the set of types in D that will say m according to δ. If $D \in T$ and $m^* \in \delta(D)$ then we say that (D, δ, m^*) is a *credible announcement with respect to* $(T, \{T_1, ..., T_J\})$ if conditions C1-C3 hold for all $m \in \delta(D)$:

C1. If $\theta \in \delta^{-1}(m) \cap T_j$ then

$$u(B(\delta^{-1}(m)), \theta) > u(B(T_j), \theta);$$

C2. If $\theta \in T_j \backslash D$ then

$$u(B(\delta^{-1}(m)), \theta) \leq u(B(T_j), \theta);$$

C3. If $\theta \in \delta^{-1}(m)$ and $m' \in \delta(D)$ then

$$u(B(\delta^{-1}(m)), \theta) \geq u(B(\delta^{-1}(m')), \theta);$$

MOP define an announcement-proof equilibrium to be a PBE such that there does not exist any credible announcement[5] with respect to $(\Theta, \mathcal{T}^\mu)$. But we must also add a durability requirement.

[4] Myerson's [12] notion of the reliability also contains the idea that a request can be more complicated than a neologism.

[5] Our definition is actually the "strong" announcement-proofness notion of Matthews, Okuno-Fujiwara and Postlewaite [11]. Their notion of "announcement-proofness" in addition to C1,C2,C3 requires:

C4. If (D', δ') also satisfies C1-C3 for all $m' \in \delta'(D')$, then for all $\theta \in D \cap D'$,

$$u\left(B(\delta^{-1}(m)), \theta\right) \geq u\left(B((\delta')^{-1}(m')), \theta\right)$$

where $m = \delta(\theta)$ and $m' = \delta'(\theta)$.

Unfortunately, with this condition durability cannot be defined by induction. For, we would like to argue that the (D', δ') which appears in C4 is only relevant if it itself is durable. But, (D', δ') can be of bigger "size" that (D, δ), so induction does not work. A more abstract definition of durability could be given, but it would seem to be almost impossible to check it in practise. We prefer to focus on the notion of strong announcement-proofness which omits C4.

By definition, a credible announcement of size 1 is always durable. Suppose durability has been defined for credible announcements of size at most $s-1$. If a credible objection (D, δ, m) is of size s, then it is durable if for all $m \in \Delta(D)$, there is no credible and durable announcement with respect to $(\delta^{-1}(m), \{\delta^{-1}(m)\})$. (Any such announcement would be of size at most $s-1$.) This way durability is defined for announcements of any size.

Definition 3 *A perfect Bayesian equilibrium $\sigma = (\mu, \alpha)$ is a* durable announcement-proof equilibrium *if: (i) there is no credible and durable announcement with respect to $(\Theta, \mathcal{T}^\mu)$; and (ii) for all $m \in \mu(\Theta)$, there is no credible and durable announcement with respect to $(\mu^{-1}(m), \{\mu^{-1}(m)\})$.*

4 Sequential Cheap Talk

Consider a sequential cheap talk game of the following form. There are an infinite number of "days," $t = 1, 2, ...$ Each day is divided into two subperiods called morning and evening. In the morning of each day t, the sender sends a message $m_t \in M$ to the receiver. The (time invariant) message space M is again sufficiently rich to at least include all the subsets of Θ and "neologisms" to use to deviate from any message profile. Also, there is a "null" message $\emptyset \in M$ with the interpretation of not saying anything. In the evening of day t, the receiver can either do nothing ($a = \emptyset$) or take an action $a \in A$. When an action $a \in A$ is taken, the game ends, and the payoffs are $u(a, \theta)$ for the sender and $v(a, \theta)$ for the receiver. If instead $a = \emptyset$, the sender sends a new message m_{t+1} in the next period. This continues until the receiver has taken an action in A. There is no discounting.

Let μ be the sender's strategy. The strategy specifies: for each time t, and each string of messages $m_1, ..., m_{t-1}$, if the receiver has not taken an action before time t, which message to send at time t. We assume $v(a, \theta) > 0$ for all a and θ, so the receiver is better off taking an action at some point rather than postponing it indefinitely, which we assume gives him zero. Given the sender's strategy, the receiver's decision problem is simple. The receiver ought to wait until the sender has revealed all the information he will ever reveal, then take the optimal action. Given that there is a finite number of types, if the sender plays according to μ, there will always exist a finite time τ^μ such that after time τ^μ, no more information will be revealed. This generates a partition of Θ denoted $\mathcal{T}^\mu = \{T_j^\mu\}_{j=1}^J$, as follows: θ, θ' belong to the same element of the partition (say T_j^μ) if and only if when the sender plays according to μ then types θ and θ' send the same string of messages from day 1 to day τ^μ. In other words, if the sender plays μ then the receiver will not be able to distinguish between θ and θ', even if he waits until time τ^μ (there is no need to wait longer). Let $T^\mu(\theta)$ be that set in the partition which contains θ, i.e.

$$\theta \in T_j^\mu \quad \leftrightarrow \quad T_j^\mu = T^\mu(\theta)$$

Let α be the receiver's strategy. It specifies, for each time t and each string of messages $m_1, ..., m_t$, if the receiver has not taken an action before time t, whether or not to take an action at time t. In the this model the receiver has a more active role than usual. Suppose the equilibrium is such that after hearing $m_1, ..., m_t$ the receiver is expected to take an action on day t. By *not* taking an action on day t, he initiates further rounds information transmission. We shall assume that if when the receiver is about to act at day t, his beliefs are given by some probability distribution over Θ, then *his own* deviation of not taking any action on day t would not change these beliefs. He does not think *his own* deviation (not taking the action on day t) is correlated with *the sender's* true type.

We can obtain a solution concept for the sequential game by asking whether or not the partition $\mathcal{T}^\mu$ is durable neologism-proof. We will argue that this method is somewhat problematic. An information partition generated by an equilibrium is durable neologism proof if any credible neologism against it is not itself durable. However, non-durable neologisms may be a part of credible multi-period communication: the very speech that could be invoked to show that the neologism was not durable may just be releasing more useful and credible information to the receiver. Consider the following example.

Example 3

		Receiver		
		Action A	Action B	Action C
Sender	Type a	3,4	2,1	1,2
	Type b	2,0	3,4	1,2
	Type c	-1,0	-1,0	1,2

Consider the following PBE σ of the sequential game. The sender's strategy μ is to never say anything. The receiver chooses action C in the evening whatever was said on the morning of the same day or in previous periods. "Off-the-equilibrium-path" beliefs are the prior (all three states are equally likely). The information partition generated by this equilibrium is a durable neologism-proof partition. The only credible neologism at this equilibrium is "I am a or b" (which if believed results in action B), but it is not durable because once type c is ruled out, pooling of a and b can be defeated by the credible neologism "I am a" which causes the receiver to respond with A. Thus, T^μ is a durable neologism-proof partition. (Notice that $\{a\}$ is not a credible neologism at the completely pooling equilibrium, because it attracts type b, and similarly $\{b\}$ attracts type a).

But we claim the completely pooling equilibrium σ is not a reasonable prediction. It can be defeated by a *sequence* of messages, each consisting of a set of types and conforming to the logic of FGP. Suppose the sender's true type is a or b. Suppose in the morning of day 1, the sender deviates from the equilibrium σ by sending the neologism "I am a or b". We claim the rational receiver should do nothing in the evening of day 1. In the morning of day 2, the sender should say "I am a" if the state is a and "I am b" if the state is b. In the evening of day 2, the receiver should

choose action A if the sender said he was type a that morning, and B otherwise. This makes both types a and b better off, as compared to the equilibrium allocation C that would have resulted from σ. Moreover, there is no incentive for type a to pretend to be type b or vice-versa. Finally, type c has no incentive to mimic the deviation as he prefers C to A and B. Thus, if the true state is a or b then the sender can credibly communicate his type by sequentially announcing subsets of states. Notice that communication necessarily is slow (takes two days). If the only believable speeches are announcements of sets of types, then by announcing successively smaller sets of types, the sender can credibly transmit more information than he could in a one-shot game.

Formally, let $T \subseteq \Theta$ and let $\mathcal{T} = \{T_1, ..., T_J\}$ be a partition of T. A *neologism S with respect to $(T, \mathcal{T})$* is a message consisting of a subset of types $S \subset T$. Let $s = |S|$ denote the size of S. If $|S| = 1$ then the neologism is *sequentially credible* if and only if it is credible. Suppose sequentially credible neologisms of size $s-1$ have been defined, and let S be of size s. Then, S is *sequentially credible* if there exists a partition $\{S_1, S_2, .., S_J\} \in \mathcal{P}(S)$ such that the following conditions hold for all T_j :

(1) If $\theta \in S_k \cap T_j$ then

$$u(B(S_k), \theta) > u(B(T_j), \theta);$$

(2) If $\theta \in T_j \backslash S$ then

$$u(B(S_k), \theta) \leq u(B(T_j), \theta);$$

(3) If $\theta \in S_k$ then for all $S_i \neq S_k$,

$$u(B(S_k), \theta) \geq u(B(S_i), \theta);$$

(4) There is no sequentially credible neologism with respect to any $(S_j, \{S_j\})$.

This way, we define sequentially credible neologisms of any size. A *sequential neologism-proof equilibrium* is a perfect Bayesian equilibrium (μ, α) of the sequential cheap talk game such that there is no sequentially credible neologism with respect to $(\Theta, \mathcal{T}^\mu)$, where $\mathcal{T}^\mu$ is the partitioning of Θ which is induced by μ.

A sequentially credible neologism can be thought of as a multi-stage speech where the sender first announces S, and upon hearing S the forward looking receiver does not take any action. Then on the next day the sender announces some subset of S. The eventual outcome of the long conversation satisfies the relevant incentive constraints, and the final information partition is itself sequential credible. This notion imputes a high degree of rationality on the part of the sender and the receiver. The receiver foresees that the subset S that is reported in one period can be split into subsets $\{S_1, S_2, .., S_J\}$, without there being any incentive for a deviating sender of one type to pretend to be another (condition (3)), that the receiver's eventual response will make these deviating types better off (condition (1)), and that the non-deviating types do not want to deviate given his eventual responses (condition (2)).

In the equilibrium σ of Example 3, after the speech "I belong to $S = \{a,b\}$", the receiver should anticipate that different speeches of types a and b will be made the next day. The message "I belong to $S = \{a,b\}$" only serves to alert the receiver that more information is to come, because it is implausible that the sender will stop there and not say anything else. Both type a and type b would go on to separate themselves out: if $S_1 = \{a\}$ and $S_2 = \{b\}$, then the partition $\{S_1, S_2\}$ satisfies the conditions (1) to (4). Therefore, the speech "I belong to $S = \{a,b\}$" is a sequentially credible neologism. Thus, σ is *not* sequential neologism proof. Notice that if the receiver only listens to neologisms of the form "I belong to set S", then it is *necessary* to talk for two periods to transmit the information that breaks the equilibrium. Neologisms do not provide a rich enough language in which to convey a plan for subsequent information revelation and therefore cannot summarize future intentions. But if the receiver is sufficiently sophisticated, upon hearing "I am S", he anticipates that future speeches will be made by types a and b.

However, there should really be no reason to speak for several days: whatever information can be conveyed over several days should be transmittable via a single message on day one. The logical thing is to allow the receiver to understand more complicated messages. This recalls the definition of announcements and, in fact, the two-day speech which we have argued should destroy the equilibrium σ in Example 3 can be compressed into one credible and durable announcement (D, δ, m) as follows. Let $D = \{a, b\}$, $\delta(a) = a$ and $\delta(b) = b$ and $m = a$ (or $m = b$). The sender in effect says, on the first day, "I am a or b; if I were a I would say so, if I were b I would say so; in fact I am a (or b)". This announcement destroys the equilibrium; the information partition generated by the equilibrium σ is not durable announcement-proof. This agrees with the strong intuition that types a and b should be able to separate themselves out. In contrast to the situation with neologisms, there is no need to define "sequentially credible announcements", because with announcements there is no need for drawn-out speeches: if a multi-day speech satisfies the relevant incentive-compatibility conditions, then the final information partition generated by the sequential speech can always be summarized by a once-and-for-all announcement on day one. Moreover, we have:

Proposition 1 *The equilibrium σ is a sequential neologism proof equilibrium if and only if it is a durable announcement proof equilibrium.*

The proof consists of a marshalling of definitions and is omitted. The one-to-one map between sequentially credible neologisms and durable credible announcements is obtained by associating with the partition $\{S_1, S_2, .., S_J\}$ from the definition of sequentially credible neologisms, an announcement (S, δ, m) where for all S_j and $\theta \in S_j$, $\delta(\theta) = S_j$.

5 Conclusion

If the receiver does not commit to listen to only one message, the logic of Farrell [5] and Grossman and Perry [7] can still be used to formalize notions of credible speeches, but their precise definitions no longer apply. Our definition of durable neologism-proof equilibrium is a modification of the static notion of neologism-proofness which gives reasonable predictions, as illustrated in Examples 1 and 2. But it does not sufficiently take into account the possibility sending multi stage speeches. There are two ways of handling this problem. We could explicitly allow objections to be sequential, which leads to the notion of sequential neologism proofness. Or, in the spirit of Matthews, Okuno-Fujiwara and Postlewaite [11], we could allow more complex speeches, where the sender summarizes all information that could be revealed slowly over time in one single message. This leads to the notion of durable announcement proofness. These two ways of arguing lead to the same conclusion, for sequential neologism proofness and durable announcement proofness are just two different ways of expressing the same idea. We think these arguments provide a good foundation for arguing in favor of sequential neologism proofness (or, equivalently, durable announcement proofness) as a solution concept for sequential cheap talk games.

References

[1] Baliga, S. Corchon L. and Sjöström T. (1997) "The Theory of Implementation when the Planner is a Player," *Journal of Economic Theory*, forthcoming.

[2] D. Bernheim, B. Peleg and M. Whinston, "Coalition-Proof Nash Equilibria I: Concepts," *Journal of Economic Theory*, 42 (1987), 1-12.

[3] V. Crawford and J. Sobel, "Strategic Information Transmission," *Econometrica*, 50 (1982), 579-594.

[4] M. Dewatripont and E. Maskin, "Contract Renegotiation in Models with Asymmetric Information," *European Economic Review* 34 (1990), 311-321.

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

[6] J. Farrell and M. Rabin "Cheap Talk," *Journal of Economic Perspectives* (1996) 10:103-118.

[7] S. Grossman and M. Perry, "Perfect Sequential Equilibrium," *Journal of Economic Theory* 39 (1986), 97-119.

[8] B. Holmström and R.B. Myerson, "Efficient and Durable Decision Rules with Incomplete Information," *Econometrica* 51 (1983), 1799-1819.

[9] J-J. Laffont and D. Martimort, "Collusion under Asymmetric Information," *Econometrica* (1997),

[10] Maskin and Tirole, "The Principal-Agent Relationship with an Informed receiver II: Common Values," *Econometrica* 60, (1992), 1-42.

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

[12] R. Myerson, "Credible Negotiation Statements and Coherent Plans," *Journal of Economic Theory* 48 (1989), 264-303.