Interim Bayesian Persuasion: First Steps

Eduardo Perez

Interim Bayesian Persuasion: First Steps*

Eduardo Perez-Richet†

January 23, 2014

Abstract

This paper makes a first attempt at building a theory of interim Bayesian persuasion. I work in a minimalist model where a low or high type sender seeks validation from a receiver who is willing to validate high types exclusively. After learning her type, the sender chooses a complete conditional information structure for the receiver from a possibly restricted feasible set. I suggest a solution to this game that takes into account the signaling potential of the sender's choice.

1 Introduction

Some recent information transmission models (Kamenica and Gentzkow, 2011; Rayo and Segal, 2010) view the sender as an *information structure designer*. This view emphasizes a connection between design and information transmission. It departs from traditional models of information transmission (Crawford and Sobel, 1982; Grossman, 1981; Milgrom, 1981; Spence, 1973) since it assumes that the sender commits to an information system before learning her type so that her choice is by itself uninformative. By contrast, there is a small but active tradition in

*I thank Pierre Fleckinger, Jeanne Hagenbach, Emir Kamenica, Navin Kartik, Frederic Koessler, Delphine Prady and Joel Sobel for discussions and ideas related to this topic. I am especially grateful to Jeanne Hagenbach, Navin Kartik and Frederic Koessler for specific comments on this paper.

†École Polytechnique, e-mail: eduardo.perez@polytechnique.edu

mechanism design (Myerson, 1983; Maskin and Tirole, 1990, 1992; Mylovanov and Troeger, 2012) that considers the possibility that a choice of mechanism by an *informed principal* may be used as a signal by the participants in the mechanism. Through the study of an example, this short paper proposes a modest first step towards building a theory of information structure design by an informed principal.

I work in a simple framework where a sender, whose type is either high or low, seeks validation from a receiver who wants to validate high types exclusively. The sender chooses an *information system*, *i.e.* a conditional distribution of signals for each of her types, from a feasible set. I allow for potential restrictions on the set of feasible information systems, which is important to make Bayesian persuasion models more flexible. I consider perfect Bayesian equilibria of the game in which the sender chooses an information system after learning her type, and the receiver chooses whether to validate after observing the signal generated by the information system. I show that there is no loss of generality in considering only pooling equilibria by an argument partly reminiscent of the *inscrutability principle* of Myerson (1983).[1] In general, perfect Bayesian equilibrium has little predictive power and it is unsatisfying to stop the analysis at this point. My main result is to show that three different refinement concepts lead to the selection of the high type optimal equilibria. The first refinement is a version of the notion of *undefeated equilibria* (Mailath et al., 1993; Umbhauer, 1994) which was developed in the context of signaling games. The second refinement is an adaptation of the notion of *core mechanism* of Myerson (1983), developed for the analysis of informed principal problems, to my framework. Finally, the third refinement builds on a discussion in Farrell (1993) to define a refinement notion that is stronger than the related notions of neologism proofness in Farrell (1993) and perfect sequential equilibrium in Grossman and Perry (1986). I also characterize selection by the D1 refinement (Cho and Kreps, 1987; Banks and Sobel, 1987). I then analyze the prediction of this method for several common restrictions on feasible information systems. All results are proved in the appendix.

[1] When all information systems are feasible, Myerson's argument that any information revealed by the strategy of the *designer* can be replicated by pooling on a different information system holds. When this set is restricted, however, the focus on pooling equilibria relies on the particular structure of the model which implies that the low type always wants to pool with the high type.

The idea of this paper is inspired from Perez-Richet and Prady (2012), which shares a similar structure, and where the information system is actually chosen by the receiver in a restricted set, but effectively controlled by the sender insofar as she can choose a complexity parameter that determines the cost of information systems for the receiver. This complexity parameter can then be used as a signal about the type of the receiver. A few other papers analyze the signaling effect of a choice of information transmission technology. In Gill and Sgroi (2012), a monopolist chooses the toughness (a one dimensional parameter) of a pass or fail test for her product, and analyzes the signaling effect of this choice.[2] As in this paper and Perez-Richet and Prady (2012), all equilibria are pooling, and they select undefeated equilibria. In Miyamoto (2013), the information of the sender is two-dimensional, and her choice of an information system on the first dimension is potentially informative about the second dimension. All these papers parameterize the set of feasible information systems along one dimension, and also make the signaling action of the sender one-dimensional. By contrast, this paper shows how to handle a much richer set of feasible information systems, and by the same token a much richer set of signaling actions.

2 The Model

The Players. A sender of type $t \in \{L, H\}$ seeks validation from a receiver. The receiver favors validation if and only if $t = H$. The prior is that type H occurs with probability $p \in (0, 1)$. The gain from validation is set 1 for both types of the sender. I normalize the payoff of the receiver to be 0 when she makes the right decision, while undue validation entails a loss $\omega_v > 0$, and undue rejection a loss $\omega_r > 0$. If $(1-p)\omega_v < p\omega_r$, the receiver is (ex ante) *pro-validation* and incentives are (ex ante) aligned, and if $(1-p)\omega_v > p\omega_r$, she is (ex ante) *pro-rejection* and there is a conflict of interests.

Strategies and Information. The sender can selectively reveal information to the receiver. She does so by choosing an *information system*, which consists of a pair of distribution measures

[2] See also Gill and Sgroi (2008).

$\pi_L(\cdot)$ and $\pi_H(\cdot)$ on $\mathbb{R}$.[3] This choice may be constrained, so I denote by $\mathcal{S}$ the set of feasible information systems. For example, in the usual persuasion framework à la Milgrom (1981), the players either reveal their type or reveal nothing, so $\mathcal{S}$ can be described as a pair of information systems (π_L, π_H) and $(\tilde{\pi}_L, \tilde{\pi}_H)$ supported on $\{0,1\}$ with $\pi_L(0) = \pi_H(1) = 1$, and $\tilde{\pi}_L(0) = \tilde{\pi}_H(0)$.[4] In the Bayesian persuasion model of Kamenica and Gentzkow (2011), all information systems are available.[5] Let $\bar{\mathcal{S}}$ denote the corresponding feasible set. I say that an information system π is *fully revealing* whenever π_L and π_H have disjoint support. When π_L and π_H have overlapping but unequal supports, π is partially revealing.

I consider only pure strategies on the sender's side but I allow the receiver to mix. For an information system π, I let Σ^π be the reunion of the supports of π_L and π_H. The receiver uses a behavioral validation strategy which is a measurable function $\lambda : \Sigma^\pi \to [0,1]$, and I denote by Λ^π the set of such functions.

Timing. I use the *interim timing* in which nature first draws the type of the sender which she observes, second the sender chooses a feasible information system which is observed by the receiver, third a signal is generated according to the information system and observed by the players, and finally the receiver decides whether to validate. For comparison purposes, I sometimes refer to the ex ante timing, which is the timing under which the sender chooses an information system before learning her type.

Equilibrium. I consider perfect Bayesian equilibria. The receiver must update her information consistently with equilibrium strategies and with the informational content of the signal generated by the chosen information system. Perfect Bayesian equilibrium implies that the signal generated by the information system has a sort of preeminence off the equilibrium path in the following sense. Suppose for example that an off path information system π' is chosen

[3] This description includes any pair of distributions on a finite set as in Kamenica and Gentzkow (2011).

[4] Note that this model also subsumes cheap talk as I could limit $\mathcal{S}$ to any number of uninformative pairs of distribution.

[5] In fact they constrain the support of the distributions to be finite but this is without loss of generality as far as feasible outcomes are concerned. As they show, one could even constrain the support to be of cardinality 2.

and that it generates a signal that could only have come from one of the two types, say t. Then, following this signal, the receiver must believe that she is facing type t with probability 1, even though her belief may have put probability 0 on t after observing the off path choice of π'. This treatment is consistent with the way evidence is treated in models with hard information.

3 Analysis

3.1 General Results under the Interim Timing

For the players, the relevant properties of an equilibrium are perfectly described by the probability of justified rejection ρ and the probability of justified validation ν. Suppose that a certain information system π is chosen by the sender, and that the receiver uses the behavioral validation strategy $\lambda \in \Lambda^\pi$. Then the outcome is summarized by $\nu(\lambda) = \int \lambda(\sigma) d\pi_H$ and $\rho(\lambda) = \int \big(1 - \lambda(\sigma)\big) d\pi_L$. So for an information system π, the set of feasible outcomes is $\Phi^\pi = \big\{\big(\rho(\lambda), \nu(\lambda)\big) : \lambda \in \Lambda^\pi\big\}$, which is a compact and convex set. [6]

An outcome is possible under belief β of the receiver if it corresponds to a best response of the receiver so the set of possible outcomes is [7] $\mathcal{P}(\beta, \pi) = \arg\max_{(\rho,\nu)\in\Phi^\pi} \ \nu\beta\omega_v + \rho(1-\beta)\omega_r$. Then the set of outcomes that can be attained by the sender if her actions lead to a belief β is described by $\mathcal{P}(\beta, \mathcal{S}) = \bigcup_{\pi\in\mathcal{S}} \mathcal{P}(\beta, \pi)$. Clearly, all equilibrium outcomes must lie in $\bigcup_{\beta\in[0,1]} \mathcal{P}(\beta, \mathcal{S})$. But in fact one can restrain attention to a smaller set.

Proposition 1. *All equilibrium outcomes lie in* $\mathcal{P}(p, \mathcal{S})$.

This is due to the fact that one can restrain attention to pooling equilibria. It is reminiscent of the *inscrutability principle* of Myerson (1983) for informed principal problems. When all information systems are available, the logic is exactly the same, and can be generalized to other setups: any information revealed in a separating equilibrium can be revealed in a pooling

[6] It is bounded because it is a subset of $[0,1]^2$, it is closed because it is the image of the set of measurable functions Λ^π, which is closed under pointwise convergence, by a function which is continuous by the dominated convergence theorem. Convexity can be proved directly without difficulty.

[7] This is well defined since Φ^π is compact and the objective function is continuous.

equilibrium. When $\mathcal{S}$ is constrained, the result is dependent on the particular incentive structure of the game I consider, which is such that the low type always wants to imitate the high type

If perfect revelation is available, then perfect Bayesian equilibria have some predictive power. Indeed, any pooling equilibrium must achieve validation with probability one for the high type, for otherwise a high type sender would deviate to perfect revelation. Therefore an information system π is an equilibrium if and only if it maximizes the utility of the high type under the constraint that the receiver chooses an optimal validation policy given an initial belief equal to the prior, and the choice of information system. Let $\mathcal{H}(p,\mathcal{S}) = \{(\rho,\nu) \in \mathcal{P}(p,\mathcal{S}) : \nu \geq \nu',\ \forall(\rho',\nu') \in \mathcal{P}(p,\mathcal{S})\}$ denote the set of high type optimal possible outcomes, which I assume to be non empty.[8]

Proposition 2 (Perfect Bayesian Equilibria)**.**

1. *If a perfectly revealing information system is available, then the set of equilibrium outcomes is exactly the set of high type optimal outcomes $\mathcal{H}(p,\mathcal{S})$, and hence $\nu = 1$ in any such outcome.*

2. *If there exists an outcome $(\hat{\rho},\hat{\nu}) \in \mathcal{P}(p,\mathcal{S})$ such that $\hat{\rho} = 1$, and for every $\nu' > \hat{\nu}$, $(1,\nu') \notin \mathcal{P}(p,\mathcal{S})$, then the set of equilibrium outcomes is the set of outcomes $(\rho,\nu) \in \mathcal{P}(p,\mathcal{S})$ such that $\nu \geq \hat{\nu}$.*

3. *Otherwise, all outcomes in $\mathcal{P}(p,\mathcal{S})$ can be supported as an equilibrium.*

Hence, the predictive power of perfect Bayesian equilibrium is weak when perfect revelation is not available. But any feasible information system, and therefore any outcome in $\mathcal{P}(p,\mathcal{S})$, can be supported as a pooling equilibrium when no information system allows the high type to be proved with positive probability. To support any information system π, it is then sufficient to let the receiver attribute any alternative choice π' to the low type.[9] This lack of predictivepower is unsatisfying. Furthermore, some of these equilibria seem unreasonable. In this case, equilibrium

[8]This set may be empty if, for example, $\mathcal{P}(p,\mathcal{S})$ is an open set of $[0,1]^2$. It is closed when $\mathcal{S} = \bar{\mathcal{S}}$, and otherwise it is always possible to ensure that it is closed by making reasonable restrictions on $\mathcal{S}$.

[9]The intuitive criterion, formalized by Cho and Kreps (1987), does not refine prediction since all information systems perform equally from the point of view of the sender when the receiver forms the belief that she faces the high type with probability one.

refinements seem indispensable. My main result is to show that three different but related belief based refinements lead to the selection of the high type optimal outcomes $\mathcal{H}(p, \mathcal{S})$. All of these refinements, in spite of their unavoidable limitations, capture the following logic, which seems convincing for this game. First, a successful deviation from an equilibrium path should be attributed to $\{H\}$ or $\{L, H\}$. Second, if a deviation is attributed to $\{H\}$, then both types will want to use it, which brings the receiver back to $\{L, H\}$. Third, at this stage there are three possibilities: (i) if attributing the deviation to $\{L, H\}$ makes it profitable for both L and H, then $\{L, H\}$ is a self-signaling set; (ii) if attributing the deviation to $\{L, H\}$ is only profitable to $\{L\}$, then the receiver should put more weight on L but that makes the deviation unattractive to any type of the sender; (iii) if attributing the deviation to $\{L, H\}$ is only profitable to $\{H\}$, then the receiver could reasonably believe anything between p and 1, but all these beliefs would make the deviation attractive to either the high type or both types.

Next, I describe these refinements. To gain space, I define them more formally and prove the result in the appendix. The first of these refinements, $\mathcal{R}_1$, is a variant on the notion of undefeated equilibrium (Mailath et al., 1993; Umbhauer, 1994). The concept of undefeated equilibrium relies on the idea that a deviation of the sender from her equilibrium action π to an information system π' should be interpreted as an attempt to indicate that she would prefer to play the equilibrium outcome associated to π' in $\mathcal{P}(p, \mathcal{S})$. The beliefs of the receiver associated to such a deviation should therefore anticipate the fact that the sender is of a type that benefits from the new equilibrium. If they do not, the original equilibrium is said to be defeated by the new one.[10] The second notion, $\mathcal{R}_2$, is an adaptation of the notion of core mechanisms of Myerson (1983) to my framework so as to take into account the fact that the designer chooses an information system rather than a full mechanism.The third notion, $\mathcal{R}_3$ refines the notions of neologism proofness of Farrell (1993) and of perfect sequential equilibrium of Grossman and Perry (1986).[11] For exposition purposes, suppose as in Farrell (1993) that a deviating player

[10]Contrary to the original definitions, I do not assume that all types using π' in the new equilibrium must prefer either the new equilibrium (Mailath et al., 1993), or some best response of the receiver to the belief generated by π' in the new equilibrium, to the original equilibrium (Umbhauer, 1994).

[11]It is easy to show that in this example all the perfect Bayesian equilibria are selected by the notions of Farrell (1993) and Grossman and Perry (1986).

is allowed to make a suggestion to the receiver as to how her deviation should be interpreted. The idea is to use the suggestion to start a chain of subsets of types that could be making the deviation by asking which types benefits if the receiver best-responds to the suggestion (a set S_1), and then which types benefit if the receiver best-responds to the belief generated by S_1 etc.

Proposition 3. *All three refinements select exactly the equilibrium outcomes in* $\mathcal{H}(p, \mathcal{S})$.

In the next section I analyze the consequences of selecting the high type optimal outcomes under several common restrictions on the set of available information systems. Before doing that, I remark that the D1 criterion[12] of Cho and Kreps (1987) may in some cases lead to a different selection. In fact, D1 and the NWBR criterion of Kohlberg and Mertens (1986) make the same selection, which is characterized as follows.

Proposition 4. *An outcome* $(\rho, \nu) \in \mathcal{P}(p, \mathcal{S})$ *is selected by D1 or NWBR if and only if either of the following holds:*

(i) $\nu = 1$.

(ii) $\rho = 1$ *and for every* ν' *such that* $(1, \nu') \in \mathcal{P}(p, \mathcal{S})$, $\nu' \leq \nu$.

(iii) For every $(\rho', \nu') \in \mathcal{P}(p, \mathcal{S})$, $\rho' \leq \rho$ *or* $\nu' \leq \nu$.

Instead of a complete proof, I provide an example of how D1 operates in the online appendix, together with a graphic intuition. Applying this intuition repeatedly leads to the result of the proposition.

4 Applications

First, I consider the case where all information systems are available. It is represented in panels (a) and (b) of Figure 1 for, respectively, the pro rejection and pro validation cases. The set $\mathcal{P}(p, \bar{\mathcal{S}})$ is the set of policies that lie above the blue dotted line, whose equation is given by

[12] See also Banks and Sobel (1987) and Cho and Sobel (1990).

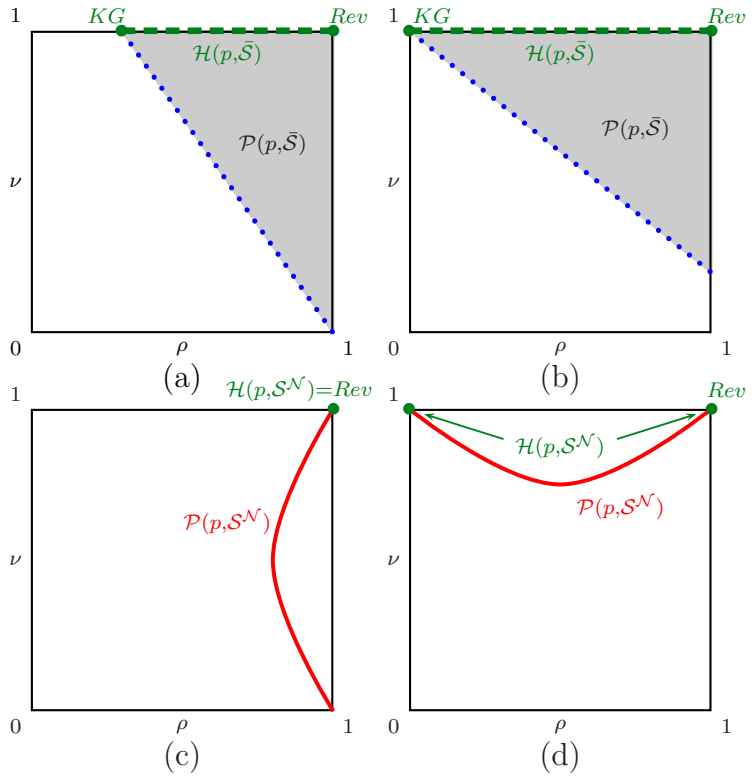

Figure 1: Applications – (a) all information systems available; pro rejection receiver (b) all information systems available; pro validation receiver (c) information systems: one of the normal families; pro rejection receiver (d) information systems: one of the normal families; pro validation receiver.

$\nu p\omega_r + \rho(1-p)\omega_v = \chi$, where the left hand-side is the objective that the receiver tries to maximize, and the right hand-side is the value of this objective function if she always rejects for the pro-rejection case and if she always validates for the pro-validation case. By Proposition 2, the set of equilibrium outcomes is exactly the set of policies such that the high type is validated with probability 1, which is represented by the green dashed line in both panels. To the left of this line, *KG* denotes the ex ante (Kamenica-Gentzkow) solution. To the right, *Rev* is the receiver-optimal policy, which can only be attained under full revelation. By refining the solution concept even more, it may be possible to select either of these outcomes in the interim case. If one requires the selected outcome to be Pareto optimal across types, for example, then the ex ante outcome is the unique selection. To select the full revelation outcome, one can modify the preferences of the sender as follows. Suppose that the sender cares lexicographically: first, about the probability of validation; second about the belief of the receiver. The prediction

under the ex ante timing remains the same as before. Under the interim timing, however, the unique high type optimal outcome now corresponds to full revelation.

Second, I consider information systems such that π_H is a normal distribution with mean $\ell > 0$ and variance V, while π_L is a normal distribution with mean $-\ell$ and variance V. Then, there are two natural ways of generating a family of feasible information systems: first, by considering variances from $V = 0$ to ∞, and adding a completely uninformative information system; second by letting ℓ go from 0 to ∞, and adding a perfectly revealing information system. In the discussion that follows, both families are denoted by $\mathcal{S}^{\mathcal{N}}$ as they have the same qualitative properties. Panels (c) and (d) in Figure 1 illustrate the properties of these families for, respectively, the pro rejection and the pro validation cases. The red curve represents the set $\mathcal{P}(p, \mathcal{S}^{\mathcal{N}})$ of feasible outcomes, and $\mathcal{H}(p, \mathcal{S}^{\mathcal{N}})$ corresponds to the green dots. In the pro rejection case, full revelation is the unique equilibrium outcome, whereas in the pro validation case, both the full revelation outcome and the completely uninformative outcome leading to certain validation are possible. Under the ex ante timing, the unique prediction would lie somewhere on the upper half of the red curve for the pro rejection case, and at the uninformative outcome for the pro validation case.

In the pro-validation case, it is interesting to constrain the precision of available information systems (taken to be ℓ or the inverse V) to a limited range. Then, the set $\mathcal{P}(p, \mathcal{S})$ corresponds to a connected portion of the U-shaped curve in panel (d). Hence the high type optimal outcome lies at either extreme point of the curve depending on the range of precisions. This means that a small technical change leading to the availability of slightly more precise tests, or a strengthening of regulation imposing a higher minimal precision, could lead to a dramatic change of outcome from the least informative test to the most informative test.

Appendix

A Three Refinements and a Proof

In this section I define the refinements for general sender-receiver games. The type set is denoted by T, $p(\cdot)$ denotes the prior belief, and for any set $S \subseteq T$, $p(\cdot|S)$ denotes the restriction of the prior to S. Then I prove Proposition 3.

$\mathcal{R}_1$. The idea of this family of refinements is that an out-of-equilibrium action should be interpreted as an attempt by a player to signal that she would prefer to coordinate on another equilibrium in which this action is played. If the beliefs of the players in the original equilibrium do not anticipate this, the original equilibrium should be discarded. This idea was developed independently in Mailath et al. (1993) and Umbhauer (1994) with small differences. Formally, in a sender-receiver game with general action set for the sender, if T is the type set of the sender and e is the original equilibrium, consider an equilibrium e' and an action of the sender a' which is on the equilibrium path of e' but off the equilibrium path of e'. Then let $T^+ \subseteq T$ be the set of types that strictly prefer e' to e and use action a' in e', and $T^0 \subseteq T$ be the set of types who are indifferent between e and e' and use a' in e'. Then the belief of the receiver that follows the use of a' in the initial equilibrium e should be in the convex hull of $p(.|T^+)$ and $p(.|T^+ \cup T^0)$. This means that the receiver should believe that all types in T^+ send a' while types in T^0 may send a' with positive probability. If this does not hold, then e is defeated by e'. The refinement retains only undefeated equilibria. Unlike Mailath et al. (1993) or Umbhauer (1994), I do not require all types who use a' in e' to prefer e' to e, or even some best response of the receiver to her belief in following a' in e'.

This refinement raises an issue absent under the original definitions. In the model of this paper, a deviation to an information system used in a pooling equilibrium that only the high type prefers is attributed to the high type. The problem is that such an attribution would make the low type want to use this deviation as well. It does not seem to be particularly problematic in this game because, if the receiver were to attribute this deviation to both types equally, then

the high type (and only the high type) would still find the deviation profitable. But the logic of this attribution may seem unsatisfying (note however that this problems is absent in any equilibrium satisfying the refinement). The next two refinements tackle this issue.

$\mathcal{R}_2$. This refinement is inspired from Myerson (1983). The difference is that the sender announces an information structure instead of a mechanism. When the sender announces a mechanism, his announcement includes the suggestion of a course of action for the receiver, and it is natural to restrict potentially destabilizing mechanisms to be incentive compatible given the beliefs they may generate. When the sender merely announces an information structure, the receiver should best respond given her beliefs, but that does not entail any natural restriction on the announcements of the sender. In order to define core outcomes, I consider a sender-receiver game, with general finite type set T for the sender, and where the sender announces an information system π (defined for the general type set T). Then an information system π is a core information system if it is an equilibrium and there is no other information system $\pi' \neq \pi$ and set $S \subseteq T$, such that for every belief $p(\cdot|S')$ of the receiver, where $S \subseteq S' \subseteq T$, any type $t \in S$ strictly prefers the outcome obtained when the receiver best responds to π' to the initial equilibrium outcome. The motivation is as follows. Suppose that π is not a core information system. Then there exists a subset of types S that would benefit from any beliefs that restricts the prior to any superset of S. Then any type in S could credibly announce π', and tell the receiver "my type is in S." The receiver does not have to believe that the sender is indeed in S, but she should account for the fact that all types in S are strictly better off as long as she believes that they make this statement.

This refinement tackles the logical difficulty with $\mathcal{R}_1$ since a deviation must be profitable to those who initiate it if it is correctly attributed to them, but also if it is attributed to any larger set of types, thus anticipating the fact that some types may try to pool on the deviation.

$\mathcal{R}_3$. As in the former paragraph. I describe the refinement for a sender-receiver game with general type set T, and where the sender announces an information system π. As in Farrell (1993), I assume that statements of the kind "my type is in S" are available for every $S \subseteq T$.

Consider an equilibrium e and an information system π' which is never played in e. When deviating to π', the sender can also announce that her type belongs to some set $S_0 \subseteq T$. Then let S_1 be the set of types that strictly benefit from the best response of the receiver to π' under the belief $p(\cdot|S_0)$ relative to the initial equilibrium, and so on, so that S_{k+1} is the set of types that strictly benefit from the best response of the receiver to π' under the belief $p(\cdot|S_k)$ relative to the initial equilibrium. The sequence stops if the empty set is ever reached. The types in $\bigcup_k S_k$ are those who could be tempted to use the deviation π' together with the announcement "my type is in S_0." Therefore the initial equilibrium is deemed unreasonable if it can only be supported by a belief $q \in \Delta(T)$ that does not lie in the convex hull of the set $\{p(\cdot|S_1), p(\cdot|S_2), p(\cdot|S_3), \cdots\}$. If the sequence is empty ($S_1 = \emptyset$), then all beliefs are allowed. Note the difference with Farrell (1993), which would require the existence of a set S_0 such that $S_1 = S_0$ (in Farrell (1993), the deviation is the announcement itself, whereas here it consists in a choice of a different information system accompanied with the announcement).

This refinement tackles the logical difficulty with $\mathcal{R}_1$ since a deviation must be profitable to those who initiate it if it is correctly attributed to them, but also if it is attributed to a larger set of types that may pool on the deviation. The difference with $\mathcal{R}_2$ is that $\mathcal{R}_3$ is more selective about assessing the types that would pool on the deviation.

Proof of Proposition 3. First consider $\mathcal{R}_1$. Let (π, Σ) be an information system such that the associated equilibrium outcome in $\mathcal{P}(p, \mathcal{S})$ is not in $\mathcal{H}(p, \mathcal{S})$ (so full revelation must not be available).Then consider any information system $\pi' \neq \pi$ such that the associated equilibrium outcome is in $\mathcal{H}(p, \mathcal{S})$. Then the high type must prefer the new outcome to the original. Suppose first that the low type prefers the original equilibrium outcome. Then after observing the deviation π', the receiver who, according to $\mathcal{R}_1$, is assumed to interpret it as an attempt to coordinate on the new equilibrium must believe that this message comes from the high type. If she did, however, the original equilibrium would not be an equilibrium as both types would benefit by deviating to π'. Suppose now that the low type is weakly prefers the new equilibrium. Then after observing the deviation π', the receiver must believe that she faces the high type

with probability $p' \geq p$. However the original equilibrium cannot be supported by such a belief, since by deviating to π' the high type would get both a more favorable belief p' and a more favorable information system. This shows that all selected equilibrium outcomes lie in $\mathcal{H}(p, \mathcal{S})$. To show that the two sets are in fact equal, consider an information system π that leads to an equilibrium outcome in $\mathcal{H}(p, \mathcal{S})$. Since the high type cannot improve her situation, the refinement does not prevent from believing that any deviation is originated by the low type, and such beliefs clearly support the equilibrium.

Now consider $\mathcal{R}_2$. Let π be an information system such that the associated equilibrium outcome in $\mathcal{P}(p, \mathcal{S})$ is not in $\mathcal{H}(p, \mathcal{S})$. Consider another information system π' with an associated equilibrium outcome in $\mathcal{H}(p, \mathcal{S})$. Then let $S = \{H\}$. Clearly the high type prefers the outcome associated to the information system π' and the belief $p(\cdot|S)$ since the latter must put probability one on the high type. Now consider $S' = \{H, L\}$. Then $p(\cdot|S')$ is simply the prior, and since π' is in $\mathcal{H}(p, \mathcal{S})$, the best response to π' when the belief is the prior leads to a better outcome than the equilibrium associated with π. Therefore the initial equilibrium is not a core equilibrium. This proves that all core equilibrium outcomes lie in $\mathcal{H}(p, \mathcal{S})$. Now consider an information system π such that the associated equilibrium outcome lies in $\mathcal{H}(p, \mathcal{S})$. It can be supported by the belief that any other choice is due to the low type. Suppose that this equilibrium is not a core equilibrium. Then the high type would have to strictly prefer the outcome associated with a different information system π' under the belief $p(.|T)$, which is simply the prior. But then that would contradict the fact that π together with the prior leads to a high type optimal outcome.

Now consider $\mathcal{R}_3$. Let π be an information system such that the associated equilibrium outcome in $\mathcal{P}(p, \mathcal{S})$ is not in $\mathcal{H}(p, \mathcal{S})$. Consider another information system π' with an associated equilibrium outcome in $\mathcal{H}(p, \mathcal{S})$. Suppose that the receiver deviates from the original equilibrium by choosing π' and at the same time suggests to the receiver that she is the high type, so $S_0 = \{H\}$. The receiver must realize that both types would benefit if she were to believe the suggestion, so $S_1 = \{H, L\}$, and the corresponding belief is exactly the prior. If that is indeed the receiver's belief, she will reproduce the outcome associated with π'. This

outcome makes the high type strictly better of. There are two cases. First, if it is does not make the low type strictly better of, then $S_2 = \{H\}$, and the sequence generated is therefore $S_k = \{H\}$ for every even k, and $S_k = \{L, H\}$ for every odd k. Second, if both types are better of under the outcome obtained when the receiver best responds to π' with a belief equal to the prior. Then $S_2 = S_1 = \{L, H\}$, and that pins down the sequence $S_k = \{L, H\}$ for every $k \geq 1$. In both cases, the possible beliefs that support the initial equilibrium following a deviation to π' must lie in $[p, 1]$, but that clearly makes this deviation profitable for the high type, so the initial equilibrium cannot satisfy the refinement. Note that we could have used the suggestion $S_0 = \{L, H\}$ to get the same result. This proves that all equilibrium outcomes that satisfy $\mathcal{R}_3$ lie in $\mathcal{H}(p, \mathcal{S})$. Now consider an information system π such that the associated equilibrium outcome lies in $\mathcal{H}(p, \mathcal{S})$. Consider any deviation π'. If the suggestion of the receiver is $\{L\}$, then $S_1 = \emptyset$ and the belief that puts all the weight on the low type is allowed following this deviation and it supports the original equilibrium. If the suggestion is $\{L, H\}$, then the associated belief is the prior, but since the original equilibrium is high type optimal, the set S_1 is either $\{L\}$ or the empty set. If $S_1 = L$, then $S_2 = \emptyset$. In both cases, the belief that puts all the weight on the low type is allowed and supports the original equilibrium. Finally, suppose the suggestion of the receiver is $S_0 = \{H\}$. Then $S_1 = \{L, H\}$, and S_2 is either $\{L\}$ or the empty set. So the prior is a possible belief in both cases, and it supports the original equilibrium. □

B Remaining Proofs

Proof of Proposition 1. Suppose that there exists a fully separating equilibrium in which the low type plays π and the high type plays $\pi' \neq \pi$. Then the high type is validated with probability 1 and the low type with probability 0. If the low type deviates to π', she is validated with probability 1 unless π' is fully revealing. So for this to be a separating equilibrium, π' must be fully revealing. But then the same outcome is obtained in a pooling equilibrium in which both types choose π', which can be supported by believing that any deviation can only be initiated by the low type. □

Proof of Proposition 2. If perfect revelation is available, the high type can ensure validation with probability 1 by deviating to full revelation, hence any equilibrium must satisfy $\nu = 1$. Any outcome in $\mathcal{P}(p, \mathcal{S})$ such that $\nu = 1$ can clearly be supported as an equilibrium which concludes the proof of 1.

Let $(1, \hat{\nu})$ be as in the proposition. If $\hat{\nu} = 1$, full revelation must be available and we are back to 1, so suppose $\hat{\nu} < 1$. I look for information systems that can generate the outcome $(1, \hat{\nu})$. The only way for the low type to be rejected with probability 1 while the high type is validated with positive probability is if the information system partially separates the two types: there must exist some signal realizations that only the high type can send, and following which the receiver validates, and some signal realizations that can be sent by both types or only the low type and following which the receiver rejects with probability 1. Furthermore, the probability that a signal that can be generated only by the high type occurs must be exactly $\hat{\nu}$. Hence there must exist an information system that proves the high type with probability $\hat{\nu}$. But then the high type can always ensure a validation probability of $\hat{\nu}$ by deviating to this information system, so any perfect Bayesian equilibrium must give her a validation probability of at least $\hat{\nu}$. Now it must also be the case that no deviation can give the high type a validation probability $\nu > \hat{\nu}$ for that would mean that $(1, \nu) \in \mathcal{P}(p, \mathcal{S})$, a contradiction. Clearly, then, every outcome in $(\rho), \nu) \in \mathcal{P}(p, \mathcal{S})$ can be supported as a pooling equilibrium.

To prove the last point, note by 1. and 2. that there cannot exist any information system such that the high type can prove her type with positive probability. Therefore any outcome in $\mathcal{P}(p, \mathcal{S})$ can be supported as an equilibrium if the receiver believes that any deviation comes from the low type exclusively. □

C D1: An Example

In this example the sender only has two information systems π_1 and π_2 available, and the sets of possible outcomes generated by these information systems under all beliefs, $\cup_{\beta \in [0,1]} \mathcal{P}(\beta, \pi_1)$ and $\cup_{\beta \in [0,1]} \mathcal{P}(\beta, \pi_2)$, are represented respectively by the lower and the upper curve on Figure 2.

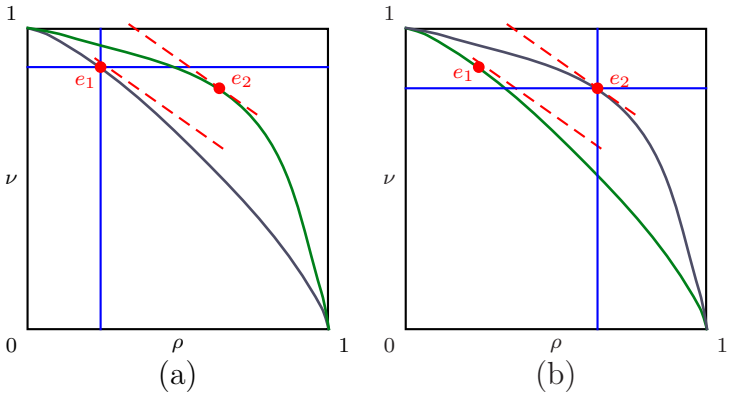

Figure 2: Example – The high type optimal equilibrium is e_1, but the only equilibrium that satisfies D1 is e_2. (a) shows why e_1 cannot satisfy D1, while (b) shows why e_2 satisfies D1.

The receiver is pro validation, and her indifference sets over outcomes are represented by the dashed red lines. So the two possible equilibria are e_1 and e_2. Consider e_1 and a deviation in which the sender announces π_2 instead of π_1. The best responses of the receiver that are preferred by the high type given this deviation correspond to the outcomes that lie on the portion of the higher curve that is above the horizontal blue line that goes through e_1 in panel (a), while those that are preferred by the low type are the ones that lie on the portion of the higher curve that is to the left of the vertical blue line. Clearly, according to D1, the receiver should attribute the deviation to the high type, but e_1 cannot be supported if that is the case. Now consider e_2 and a deviation in which the sender announces π_1. The best responses of the receiver that are preferred by the high type given this deviation are the ones that lie on the portion of the lower curve that is above the horizontal blue line in panel (b), while those that are preferred by the low type are the ones that lie on the portion of the lower curve that is to the left of the vertical blue line. According to D1, the receiver should attribute the deviation to the low type, and since this belief is compatible e_2, this equilibrium passed the test. However it is easy to see that e_1 is the unique high type optimal equilibrium.

References

Banks, J. S. and J. Sobel (1987): "Equilibrium Selection in Signaling Games," *Econometrica*, 55, 647–661.

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

Cho, I.-K. 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.

Gill, D. and D. Sgroi (2008): "Sequential Decisions with Tests," *Games and Economic Behavior*, 63, 663–678.

——— (2012): "The Optimal Choice of Pre-Launch Reviewer," *Journal of Economic Theory*, 147, 1247–1260.

Grossman, S. J. (1981): "The Informational Role of Warranties and Private Disclosure about Product Quality," *Journal of Law and Economics*, 24, 461–483.

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

Kamenica, E. and M. Gentzkow (2011): "Bayesian Persuasion," *American Economic Review*, 101, 2590–2615.

Kohlberg, E. and J.-F. Mertens (1986): "On the Strategic Stability of Equilibria," *Econometrica*, 54, 1003–1037.

Mailath, G. J., M. Okuno-Fujiwara, and A. Postlewaite (1993): "Belief-Based Refinements in Signaling Games," *Journal of Economic Theory*, 60, 241–276.

Maskin, E. and J. Tirole (1990): "The Principal-Agent Relationship with an Informed Principal: The Case of Private Values," *Econometrica*, 58, 379–409.

——— (1992): "The Principal-Agent Relationship with an Informed Principal, II: Common Values," *Econometrica*, 60, 1–42.

Milgrom, P. (1981): "Good News and Bad News: Representation Theorems and Applications," *Bell Journal of Economics*, 12, 380–391.

Miyamoto, S. (2013): "Signaling by Blurring," Working Paper.

Myerson, R. B. (1983): "Mechanism Design by an Informed Principal," *Econometrica*, 51, 1767–1797.

Mylovanov, T. and T. Troeger (2012): "Informed Principal Problems in Generalized Private Value Environments," *Theoretical Economics*, 7, 465–488.

Perez-Richet, E. and D. Prady (2012): "Complicating to Persuade?" Working Paper.

Rayo, L. and I. Segal (2010): "Optimal Information Disclosure," *Journal of Political Economy*, 118, 949–987.

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

Umbhauer, G. (1994): "Forward Induction, Consistency, Preplay Communication and Epsilon Perturbations," Mimeo Beta, Univesité Louis Pasteur, Strasbourg, France.