Appendix A. Refinement issues

A.1. **Intuitive Criterion.** Since in our game, the role of the receiver has been reduced to that of rewarding the sender her expected quality, we simplify the definition of the Intuitive Criterion. Our notation is from Fudenberg and Tirole (1991).

Definition 1. Let $u^*(q)$ be a vector of Perfect Bayesian Equilibrium payoffs for the sender. For any $s \in S$, define

$$J(s) = \{q \mid u^*(q) > H - c(s,q)\}. \tag{A.1}$$

A pure strategy Perfect Bayesian Equilibrium satisfies the *Intuitive Criterion* if and only if there does not exist $s \in S$ and $q \in Q$ such that

$$u^*(q) < \min\{Q \setminus J(s)\} - c(s,q) \tag{A.2}$$

for $\{Q \setminus J(s)\} \neq \emptyset$.

As defined, $J(s)$ is the set of types that could never do better by deviating from the equilibrium and sending signal s. According to the Intuitive Criterion, out of equilibrium beliefs must therefore put zero probability on the event that some $q \in J(s)$ sends signal s. As a result, a type $q \in Q$ agent who sends signal s can never expect to get less than $\min\{Q \setminus J(s)\} - c(s,q)$. If, for some agent, this is greater than her expected equilibrium payoff then the equilibrium fails to survive the Intuitive Criterion.

As is well known, when there are more than two types the Intuitive Criterion does not imply a unique signaling equilibrium. As we will now see, it cannot always eliminate countersignaling equilibria. In the following proposition, we characterize the set of countersignaling equilibria. It is then easy to show that under tighter conditions, though qualitatively similar to those set out in Proposition 2, this set is non-empty.

In order to characterize the set of countersignaling equilibria that satisfy the Intuitive Criterion, we first define the following additional notation. For some countersignaling equilibrium, (s^*, s_M^*, s^*), let $\dot{s}_L^*(s^*)$, $\dot{s}_M^*(s_M^*)$ and $\dot{s}_H^*(s^*)$ solve $\bar{q}_{\{L,H\}}(L) - c(s^*, L) = H - c(\dot{s}_L^*, L)$, $M - c(s_M^*, M) = H - c(\dot{s}_M^*, M)$ and $\bar{q}_{\{L,H\}}(H) - c(s^*, H) = H - c(\dot{s}_H^*, H)$. These critical values represent the highest signal agents of any type would ever be willing to send (*i.e.*, if they were believed to be of type H with probability 1).

Proposition 5. *A countersignaling equilibrium, (s^*, s_M^*, s^*), survives the Intuitive Criterion if and only if $\bar{q}_{\{L,H\}}(H) - c(s^*, H) \geq H - c(\dot{s}_M^*(s_M^*), H)$.*

Proof. ($\Rightarrow$) Given some countersignaling equilibrium (s^*, s_M^*, s^*), $J(s)$ is the set of types that satisfy $\bar{q}_{\{L,H\}}(q) - c(s^*, q) > H - c(s, q)$ for $q = L, H$ and $M - c(s_M^*, M) > H - c(s, M)$. To consider this, note that by the single-crossing property, $\dot{s}_L^*(s^*) < \dot{s}_M^*(s_M^*)$. Since for sufficiently large $\bar{q}_{\{L,H\}}(H)$, $\dot{s}_H^*(s^*) \to 0$, there are three relevant, mutually exclusive cases: a) $\dot{s}_L^*(s^*) < \dot{s}_M^*(s_M^*) < \dot{s}_H^*(s^*)$, b) $\dot{s}_L^*(s^*) \leq \dot{s}_H^*(s^*) \leq \dot{s}_M^*(s_M^*)$ and c) $\dot{s}_H^*(s^*) < \dot{s}_L^*(s^*) < \dot{s}_M^*(s_M^*)$.

Under case a),

$$J(s) = \begin{cases} \emptyset & \text{if } s \leq \dot{s}_L^*(s^*) \\ \{L\} & \text{if } \dot{s}_L^*(s^*) < s \leq \dot{s}_M^*(s_M^*) \\ \{L, M\} & \text{if } \dot{s}_M^*(s_M^*) < s \leq \dot{s}_H^*(s^*) \\ \{L, M, H\} & \text{if } s > \dot{s}_H^*(s^*) \end{cases}$$

Only the second and third cases impose any restrictions on equilibrium beliefs.

For signals $s \in (\dot{s}_M^*(s_M^*), \dot{s}_H^*(s^*))$, the right hand side of (A.2) is $H - c(s, q)$. By definition, Highs are indifferent between sending signal $\dot{s}_H^*(s^*)$ for a payoff of $H - c(\dot{s}_H^*(s^*), H)$ and sending signal s^*. But $s < \dot{s}_H^*(s^*)$ and beliefs which satisfy the Intuitive Criterion put probability 1 on the event $q = H$. Thus, since s is less expensive than $\dot{s}_H^*(s^*)$, Highs must strictly prefer s to s^* (*i.e.*, $H - c(s, H) > H - c(\dot{s}_H^*(s^*), H) = \bar{q}_{\{L,H\}}(H) - c(s^*, H)$). Therefore *any* countersignaling equilibrium where $\dot{s}_H^*(s^*) > \dot{s}_M^*(s_M^*)$ fails the Intuitive Criterion and case a) cannot hold in any countersignaling equilibrium that survives the Intuitive Criterion.

Under case b),

$$J(s) = \begin{cases} \emptyset & \text{if } s < \dot{s}_L^*(s^*) \\ \{L\} & \text{if } \dot{s}_L^*(s^*) \leq s < \dot{s}_H^*(s^*) \\ \{L, H\} & \text{if } \dot{s}_H^*(s^*) \leq s < \dot{s}_M^*(s_M) \\ \{L, M, H\} & \text{if } s \geq \dot{s}_M^*(s_M^*) \end{cases}$$

Again, only the second and third cases impose restrictions on out of equilibrium beliefs.

In both cases (*i.e.*, either $s \in [\dot{s}_L^*(s^*), \dot{s}_H^*(s^*))$ or $s \in [\dot{s}_H^*(s^*), \dot{s}_M^*(s_M)))$, the right hand side of (A.2) becomes $M - c(s, q)$. Since $s_M^* < \dot{s}_L^*(s^*) < \dot{s}_H^*(s^*)$—the first inequality follows from the single-crossing property and the second by assumption—no type will ever deviate to s. That is, $u^*(q) \geq M - c(s_M^*, q) > M - c(s, q)$ for $q \in Q$ (the first inequality follows by the fact that $\mathbf{s}^*$ is an equilibrium and the second by the fact that $s \geq \dot{s}_L^*(s^*) > s_M^*$).

Under case c),

$$J(s) = \begin{cases} \emptyset & \text{if } s < \dot{s}_H^*(s^*) \\ \{H\} & \text{if } \dot{s}_H^*(s^*) < s \leq \dot{s}_L^*(s^*) \\ \{L, H\} & \text{if } \dot{s}_L^*(s^*) < s \leq \dot{s}_M^*(s_M^*) \\ \{L, M, H\} & \text{if } s > \dot{s}_M^*(s_M^*) \end{cases}$$

Again, only the second and third cases affect beliefs. Since for the second case, the right hand side of (A.2) becomes $L - c(s, q)$, no type will deviate to such an s. For the third case, the right hand side of (A.2) becomes $M - c(s, q)$. Again, since $s > \dot{s}_L^*(s^*) > s_M^*$, no type will ever deviate to such a signal.

In sum, a countersignaling survives the Intuitive Criterion if and only if $\dot{s}^*_M(s^*_M) \geq \dot{s}^*_H(s^*)$. Since $H - c(\dot{s}^*_H(s^*), H) = \bar{q}_{\{L,H\}}(H) - c(s^*, H)$, this is true if and only if $\bar{q}_{\{L,H\}}(H) - c(s^*, H) \geq H - c(\dot{s}^*_M(s^*_M), H)$.

($\Leftarrow$) Follows by reversing previous argument. ■

Thus the existence of countersignaling equilibria under the Intuitive Criterion requires that the exogenous information should further separate the Highs from the Lows.[1]

A.2. ***Divinity*-like refinements.** Consider the Cho and Kreps (1987) equilibrium refinements, D1 and D2. Clearly since the exogenous information is irrelevant to the signaling equilibria, they will never be eliminated by D1 or D2. On the other hand, we would like to determine whether

countersignaling equilibria can also survive the more stringent equilibrium concepts D1 and D2. The definition of D1 requires defining the sets of rationalizable gross payoffs that would give the sender a greater net payoff than candidate equilibrium payoff for sending signal s.

$$D(q, s) = \{\pi \mid L \leq \pi \leq H \text{ and } u^*(q) < \pi - c(s, q)\} \tag{A.3}$$

$$D^0(q, s) = \{\pi \mid L \leq \pi \leq H \text{ and } u^*(q) = \pi - c(s, q)\} \tag{A.4}$$

where $u^*(q)$ is the equilibrium payoff of a type q sender.

Definition 2. The criterion D1 puts probability zero on any type q sending signal s if there is some other type q' such that

$$\{D(q, s) \cup D^0(q, s)\} \subset D(q', s). \tag{A.5}$$

The argument being that if there are a wider range of receiver payments rationalizable by some beliefs, then type q' should be infinitely more willing to send s than type q.[2]

Let $\mathbf{s}^* = (s^*, s^*_M, s^*)$ be some countersignaling equilibrium. Define $\underline{\pi}(q, s)$ to solve:

$$\begin{aligned} \underline{\pi}(L, s) - c(s, L) &= u^*(L) \\ \underline{\pi}(M, s) - c(s, M) &= u^*(M) \\ \underline{\pi}(H, s) - c(s, H) &= u^*(H) \end{aligned} \tag{A.6}$$

The $\underline{\pi}(q, s)$ should be interpreted as the minimum out of equilibrium payoff that would be necessary to induce an agent of type q to send signal s and

[1] Compare $\bar{q}_{\{L,H\}}(H) \geq H - c(\dot{s}^*_M(\tilde{s}^*_M(0)), H)$ to $\bar{q}_{\{L,H\}}(H) \geq M - c(\tilde{s}^*_M(0), H)$ from the proof of Proposition 2. Since $\dot{s}^*_M(\tilde{s}^*_M(0))$ is further up the Medium indifference curve, it must be the case that $H - c(\dot{s}^*_M(\tilde{s}^*_M(0)), H) > M - c(\tilde{s}^*_M(0), H)$.

[2] For our model, D1 and D2 are identical because criterion D2 puts probability zero on any type q sending signal s if $\{D(q, s) \cup D^0(q, s)\} \subset \cup_{q' \neq q} D(q', s)$. D1 and D2 are equivalent here because for any q', $D(q', s)$ is an interval of the form $(\pi, H]$ and any finite union of such intervals is equal to one of its members.

are simply given by agents' equilibrium indifference curves. Given these definitions, it is easy to see that:

$$D(q,s) = \{\pi \mid H \geq \pi > \underline{\pi}(q,s)\}$$
$$D(q,s) \cup D^0(q,s) = \{\pi \mid H \geq \pi \geq \underline{\pi}(q,s)\} \tag{A.7}$$

To analyze the stategy/type combinations which can be eliminated under D1, define $\ddot{s}_{qq'}$, for $q < q'$, to solve $\underline{\pi}(q,s) = \underline{\pi}(q',s)$. Given the single-crossing property, each of these values is unique and exists. Since $\mathbf{s}^*$ is a countersignaling equilibrium, the single-crossing property further implies that $\ddot{s}_{LM} \leq \ddot{s}_{LH} \leq \ddot{s}_{MH}$.

PROPOSITION 6. *Countersignaling equilibrium* (s^*, s^*_M, s^*) *survives criterion D1 (and hence D2) if and only if* $s^*_M = \tilde{s}_M(s^*)$ *and* $\bar{q}_{\{L,H\}}(H) - c(s^*, H) \geq H - c(\dot{s}^*_M(s^*_M), H)$.

Proof. ($\Rightarrow$) Let $s \neq s^*, s^*_M$. If $s \in [0, \ddot{s}_{LM})$ then $\underline{\pi}(L,s) < \underline{\pi}(M,s) < \underline{\pi}(H,s)$ so that $\{D(H,s) \cup D^0(H,s)\} \subset \{D(M,s) \cup D^0(M,s)\} \subset D(L,s)$. Thus, out of equilibrium beliefs must assign zero probability to the event that a Medium or a High sends such a signal so that $\mu(L|s,x) = 1$. Using similar arguments it is easy to show that if $s = \ddot{s}_{LM}$ then $\mu(H|s,x) = 0$, if $s \in (\ddot{s}_{LM}, \ddot{s}_{MH})$ then $\mu(M|s,x) = 1$, if $s = \ddot{s}_{MH}$ then $\mu(L|s,x) = 0$ and if $s > \ddot{s}_{MH}$ then $\mu(H,s,x) = 1$.

Since $\mathbf{s}^*$ is a countersignaling equilibrium, the only deviations that need to be considered are non-equilibrium deviations (*i.e.*, $s \neq s^*, s^*_M$). It is clear that no type will ever deviate to $s \in [0, \ddot{s}_{LM}]$ since the worst possible belief in this case is $\mu(L|s,x) = 1$ and the associated deviation payoff would be $L - c(s,q) \leq u^*(q)$ for every q.

Consider $s \in (\ddot{s}_{LM}, \ddot{s}_{MH}]$. The worst (from the sender's viewpoint)D1-consistent receiver beliefs for deterring deviations in are $\mu(M|s,x) = 1$. Mediums have an incentive to deviate to s if and only if $s < s^*_M$. Thus if $\mathbf{s}^*$ survives criterion D1, it must be that $\ddot{s}_{LM} \geq s^*_M$. But it can be seen that in any countersignaling equilibrium, $\ddot{s}_{LM} \leq s^*_M$, with equality only if $s^*_M = \tilde{s}^*_M(s^*)$ (*i.e.*, equilibria that survive D1 must have Mediums sending the lowest possible signal, given s^*). Henceforth, assume that $s^*_M = \tilde{s}^*_M(s^*)$. Since $s^*_M = \tilde{s}^*_M(s^*)$, by definition, neither Lows nor Highs have an incentive to deviate to s.

Now consider $s > \ddot{s}_{MH}$. In this case, $\mu(H|s,x) = 1$. Thus type q senders can profitably deviate to some signal s if $\ddot{s}_{MH} < \dot{s}^*_q(s^*_q)$. Since $\dot{s}^*_L(s^*) < \dot{s}^*_M(s^*_M)$, countersignaling equilibria that survive criterion D1 must have $\ddot{s}_{MH} \geq \max\{\dot{s}^*_M(s^*_M), \dot{s}^*_H(s^*)\}$. We now need to show that this condition is equivalent to $\bar{q}_{\{L,H\}}(H) - c(s^*, H) \geq H - c(\dot{s}^*_M(s^*_M), H)$. Consider if this latter condition holds with equality. By definition of $\dot{s}^*_H(s^*)$, $\dot{s}^*_H(s^*) = \dot{s}^*_M(s^*_M)$ so the former condition holds with equality. Now consider if the latter condition is a strict inequality. Since it is readily shown that $d\dot{s}^*_H(s^*)/d\bar{q}_{\{L,H\}}(H) < 0$, $d\dot{s}^*_M(s^*_M)/d\bar{q}_{\{L,H\}}(H) = 0$ and $d\ddot{s}_{MH}/d\bar{q}_{\{L,H\}}(H) > 0$, the former condition still holds.

($\Leftarrow$) Follows similarly. ■

Although D1 further restricts the set of countersignaling equilibria (there are fewer equilibrium signals available to Mediums), the condition for their existence is identical to that under the Intuitive Criterion. As a result, we can conclude that the Pareto dominant countersignaling equilibrium under D1 and D2 is identical to that under the Intuitive Criterion and therefore strongly Pareto dominates all signaling equilibria.

Furthermore, unlike the standard signaling model, not only does D1 not select a unique equilibrium, there are many countersignaling equilibria which survive D1. At first glance, this is somewhat puzzling because sender preferences are completely standard and clearly satisfy conditions (i), (ii) *and* (iii) of the Cho-Sobel Theorem (as defined in Fudenberg and Tirole (1991, p. 458)). Cho-Sobel fails because our assumption about the receiver's priors are different. In Cho-Sobel, D1-consistent beliefs are identical across all senders prior to observing the signal s, so within any group of pooling sender types the highest type would prefer to signal her type at a marginally higher cost. In our model after observing x the receiver has different beliefs for each sender which are "on average" correlated with the sender's actual type. In any equilibrium in which two or more types pool, the highest type need not prefer to break away since the exogenous information ensures that the highest type already expects to receive a higher type estimate than the lower types.

As a final note, the out of equilibrium beliefs required to support these equilibria suggest a weakness in some of the Divinity-like refinement concepts. D1 and stronger beliefs (stronger refinements include Universal Divinity but exclude Divinity) involve a discontinuity which does not seem entirely plausible. In particular, upon observing (s^*, x), the receiver believes that the sender is of type L with probability $g(x|L)f(L)/(g(x|L)f(L)+g(x|H)f(H))$ and of type H with probability $g(x|H)f(L)/(g(x|L)f(L) + g(x|H)f(H))$. However for signals, $s \in N(s^*, \epsilon) \setminus s^*$, these D1 beliefs jump discontinuously to probability 1 that the sender is of type L.

References

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.

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