7 Supplemental Appendix to "Credible Deviations from Signaling Equilibria" by Eső and Schummer

This "working paper appendix" is **not for journal-publication**. It substantiates a few claims that are made without proof in the paper, and contains everything "available upon request."

Intuitive Criterion vs Credibility

Proposition 1 *When there are only two types, if an equilibrium is immune to Credible Deviations, then it satisfies the Intuitive Criterion.*

Proof: If an equilibrium violates the Intuitive Criterion, then there exists a type in $\{\theta_1, \theta_2\}$, say θ_1, and out-of-equilibrium message m' such that

$$u_S^*(\theta_1) < \min_{a \in BR(\Theta \setminus J, m')} u_S(\theta_1, m', a)$$

where

$$J = \{\theta \in \{\theta_1, \theta_2\} : u_S^*(\theta) > \max_{a \in BR(\Theta, m')} u_S(\theta, m', a)\}.$$

We show $J = \{\theta_2\}$. The first inequality implies $\theta_1 \notin J$. Incentive compatibility of the equilibrium requires $u_S^*(\theta_1) \geq u_S(\theta_1, m', a')$ where a' is the Receiver's response to m' under equilibrium strategies. Since $a' \in BR(\Theta, m')$ by definition of equilibrium, this implies $J \neq \emptyset$ (otherwise the first inequality would be contradicted). Hence $J = \{\theta_2\}$, so

$$u_S^*(\theta_2) > \max_{a \in BR(\Theta, m')} u_S(\theta_2, m', a).$$

This means that θ_2 cannot belong to any "club" C that satisfies eqn. (1). Therefore, to show that $C = \{\theta_1\}$ is the unique C to satisfy eqn. (1) requires only that

$$u_S^*(\theta_1) < \min_{a \in BR(\{\theta_1\}, m')} u_S(\theta_1, m', a)$$

which is true under the first inequality since $J = \{\theta_2\}$. □

Proposition 2 *Consider a game where (i) there are only two types, (ii) after each message the Receiver has only two actions, and (iii) any two* *message-action pairs yield strictly different payoffs to any given type. (The last assumption is generic.) Then a pure-strategy equilibrium is immune to Credible Deviations if and only if it satisfies the Intuitive Criterion.*

Proof: Due to Prop. 1, we only have to show one direction: if an equilibrium is Vulnerable to Credible Deviations, then it violates the Intuitive Criterion. Given an equilibrium, suppose there is a unique set of types C that satisfies (1) for some out of equilibrium m. Clearly $C \neq \{\theta_1, \theta_2\}$, otherwise the types would break the equilibrium by playing m. Without loss, suppose $C = \{\theta_1\}$.

Denote the Receiver's possible responses to m by $\{a_1, a_2\}$, and suppose a_1 is the equilibrium response. Generically, both types *strictly* lose deviating to m if the Receiver would respond with a_1. Since $C = \{\theta_1\}$, the Receiver's best response believing only θ_1 is deviating is to play a_2. Furthermore, θ_2 must be (generically strictly) worse off in that case, too, otherwise he would want to "join the club."

But this means that θ_2 is strictly worse off no matter how the Receiver responds to m. Hence the Intuitive Criterion forces the Receiver to believe that m was sent by θ_1, breaking the equilibrium. $\square$

With three or more types, there can exist an equilibrium which is immune to Credible Deviations but does *not* satisfy the Intuitive Criterion. Consider the following game (priors are not relevant).

	m_1	m_2: a_1	m_2: a_2	m_2: a_3
H	1,1	0,0	2,1	2,2
M	1,1	0,0	0,2	2,1
L	1,1	0,2	0,0	0,0

Figure 5: Intuitive Criterion stronger than Credibility.

In one equilibrium, all Sender types send m_1 and the Receiver, believing that only (or very likely) type L would send m_2, replies to m_2 with a_1. This equilibrium fails the Intuitive Criterion. To see this, note that type L cannot possibly gain from deviating to m_2, therefore this type must be ruled out as a deviator according to the Intuitive Criterion. Restricting the Receiver's beliefs to $\{H, M\}$, the best response is either a_2 or a_3 (or a mixture of the two) since a_1 is dominated. But type H benefits from deviation if such replies are anticipated.

On the other hand, this equilibrium is not Vulnerable to Credible Deviations; i.e. no Deviators' Club C satisfies (1) with respect to m_2. To see this, note that since L would strictly lose by deviating, $L \notin C$. This implies, as above, that any best response to some C would be some mixture of a_2 and a_3. This would imply $H \in C$.

If $C = \{H\}$ then the Receiver's best response would be a_3, in which case type M would want to join C, a contradiction. However, $C = \{H, M\}$ would violate (1) because type M is worse off than in equilibrium when the Receiver chooses the best response of a_2 (when believing only M deviated). Therefore there is no Deviators' Club.

Pooling on "no MBA" satisfies Stablility in Fig. 1

An equilibrium component is called stable à la Kohlberg-Mertens (see p. 443 of Fudenberg-Tirole (1991)), if for all $\delta > 0$ there exists $\varepsilon > 0$ such that for all i (indexing agents of players), s_i (denoting strategies), and for all $\varepsilon(s_i) \in (0, \varepsilon)$, if each agent i is constrained to play each strategy s_i with a probability of at least $\varepsilon(s_i)$, then the modified game has an equilibrium that is δ-close (in the space of strategies) to some equilibrium in the component.

Consider our example and its "pooling on m_1" equilibrium component. Recall that in this equilibrium component both types of the Sender (Q and P) play m_1, and the Receiver responds to m_2 by playing a mixture $(\alpha, 1-\alpha)$ of the strategies $a_1 = HR$ and $a_3 = Asst$ such that $\alpha \geq 1/2$. Denote the prior probability that the Sender's type is Q by π.

Denote ε_Q and ε_P respectively the minimal probability that types Q and P have to put on playing m_2. Let ε_k denote the probability that the Receiver has to put on action a_k, $k = 1, 2, 3$. All these probabilities are bounded by ε from above. Now we will find nearby equilibria for ε sufficiently low.

Case 1: $\varepsilon_Q\pi/(\varepsilon_Q\pi + \varepsilon_P(1-\pi)) \leq 2/5$. This is the case when $\varepsilon_P \gg \varepsilon_Q$.

We claim that a nearby equilibrium is where both types still "try" to play

m_1 (so type Q plays m_2 with probability ε_Q and P plays it with probability ε_P), and the Receiver tries to reply with a_1 (but also plays a_2 and a_3 with probabilities ε_2 and ε_3, respectively because he is constrained). Given the hypothesized play of the Sender, the Receiver's posterior is that the deviator is of type Q with probability $\varepsilon_Q\pi/(\varepsilon_Q\pi + \varepsilon_P(1-\pi)) \leq 2/5$, therefore his reply is rationalized. Given that the Receiver mostly plays a_1, for sufficiently low ε, the Sender has an incentive to play m_1 no matter what his type is.

Case 2. $\varepsilon_Q\pi/(\varepsilon_Q\pi + \varepsilon_P(1-\pi)) > 2/5$. This is the case when $\varepsilon_Q \gg \varepsilon_P$.

The equilibrium of Case 1 does not work because if both types try to play m_1 then the Receiver believes it is type Q that "most likely" deviates to m_2, in which case a_1 is not his best response. However, there is another "nearby" equilibrium, in which the Receiver plays a_1 and a_3 with probability $(1-\varepsilon_2)/2$ each. For ε sufficiently low, type Q strictly prefers m_1 to m_2, while P is indifferent between m_1 and m_2. Therefore Q plays m_2 with probability ε_Q (i.e., he "tries" to play m_1), while type P may randomize. Let type P play m_2 with probability $\rho = (5/2)\varepsilon_Q\pi/(1-\pi)$. Then the Receiver's posterior on type Q given m_2 is $\varepsilon_Q\pi/(\varepsilon_Q\pi + \rho(1-\pi)) = 2/5$, which rationalizes mixing a_1 and a_3 and sustains the equilibrium. This equilibrium can be made arbitrary close to the one where both types of the Sender play m_1 by setting ρ and ε sufficiently low.

Non-existence

The following cheap-talk example is based on one given by Farrell (1993). Let there be an arbitrary number of messages (at least two), m_1, m_2,... Cheap-talk means the message does not affect payoffs; hence only one matrix is needed to explain the game.

	a_1	a_2	a_3	π
θ_1	2, 10	−1, 0	0, 6	1/2
θ_2	1, 0	0, 10	2, 6	1/2

m_i

Figure 6: No (pure-strategy) equilibrium is immune to Credible Deviations.

There is a pooling equilibrium outcome where both Sender types send the same message m_i, and the Receiver responds to *any* message by playing a_3. This can be supported if the Receiver's posterior following any message is his prior. The pooling equilibrium is Vulnerable to Credible Deviations, however, since $C = \{\theta_1\}$ would be the unique deviators' club following any out of equilibrium message. There is no separating equilibrium outcome because θ_2 would rather be perceived as θ_1 than as himself, and messages are costless.

As happens in all cheap talk games, this equilibrium outcome can be supported by mixed strategies—babbling—where both types randomizing equally over *all* messages. In this case there are no out-of-equilibrium messages, and there can be no Credible Deviation. Thus, no *pure* equilibrium is immune to Credible Deviations, but a babbling equilibrium is.

This leads us to the game in Fig. 7. Attributing it to Farrell's 1984 paper, van Damme (1991) (Fig. 10.5.5) uses it to demonstrate the non-existence of Perfect Sequential Equilibrium.

	a_1	a_2	a_3	a_1	a_2	a_3	π
θ_1	$2+\epsilon$, 3	$1+\epsilon$, 0	$0+\epsilon$, 2	2, 3	1, 0	0, 2	2/3
θ_2	1, 0	0, 3	2, 2	$1+\epsilon$, 0	$0+\epsilon$, 3	$2+\epsilon$, 2	1/3
		m_1			m_2		

FIGURE 7: No equilibrium is immune to Credible Deviations.

The unique equilibrium outcome is to pool on m_1, but this is Vulnerable to a Credible Deviation.

PSE vs. Credibility in the MBA example

This example is mentioned in footnote 5 of the paper.

Suppose that the Sender's choice is not simply whether or not to earn an MBA, but also *how good a degree to get*. There are three different schools indexed by $e \in \{1, 2, 3\}$; a higher index represents a "better" MBA degree (meaning a higher cost for the Sender and more benefits for the Receiver). We denote the action "no MBA" by $e = 0$; this yields a default payoff of $(4, 4)$ for all types of the Sender (the Receiver has no choice to make).

The payoffs from choosing one of the three positive MBA levels is the following (Q stands for the "Quant" type and P for the "non-Quant" type):

	HR	CFO	A
Q	0, 0	$22-4e$, $4+2e$	$15-3e$, $3+e$
P	2, $4+2e$	$15-5e$, 0	$15-3e$, $3+e$

$e \in \{1, 2, 3\}$

In order to make the calculations easier we provide the entire payoff matrix of the game:

	A
Q	4, 4
P	4, 4

$e = 0$

	HR	CFO	A
Q	0, 0	18, 6	12, 4
P	2, 6	10, 0	12, 4

$e = 1$

	HR	CFO	A
Q	0, 0	14, 8	9, 5
P	2, 8	5, 0	9, 5

$e = 2$

	HR	CFO	A
Q	0, 0	10, 10	6, 6
P	2, 10	0, 0	6, 6

$e = 3$

Assume the prior is 50-50%. (The example in the paper restricts to $e \in \{0, 3\}$, doubling the payoffs.) The Receiver's best response to $e \in \{1, 2, 3\}$ is HR if the Sender is believed to be of type P, CFO if he is believed to be of type Q, and A if he is believed to be either type with 50-50% chance.

There are five pure-strategy sequential equilibria.

There exists a single separating equilibrium: P chooses $e = 0$, Q picks $e = 3$. Similar separating equilibria where Q picks $e = 2$ (or $e = 1$) do not exist because P would imitate Q ending up at CFO with a higher payoff. Clearly, in any separating equilibrium P plays $e = 0$, therefore there are no other separating equilibria.

There exist four pooling equilibria, one for each $e \in \{0, 1, 2, 3\}$. A pooling equilibrium can be supported by the Receiver believing that a deviator is of type P and replying HR.

All pooling equilibria are Vulnerable to Credible Deviations; the separating equilibrium is not. Pooling on $e = 0$ (yielding 4 for both Sender types) is ruled out by a deviation to $e = 3$ and $C = \{Q\}$ (type Q gets 10 while P gets 0 if the speech "I am Q" is believed). Pooling on $e = 1$ (yielding 12 to both types) is ruled out by a deviation to $e = 2$ and $C = \{Q\}$ (type Q gets

14 while P gets 5 if "I am Q" is believed). Pooling on $e = 2$ (yielding 9 to both types) is ruled out by a deviation to $e = 3$ and $C = \{Q\}$ (Q gets 12 while P gets 0). Finally, pooling on $e = 3$ (yielding 6 to both types) is ruled out by a deviation to $e = 2$ and $C = \{Q\}$.

In contrast, **PSE eliminates all equilibria**. The pooling equilibria can be ruled out the same way as in our refinement, by Q's deviation to a different level of e. But, PSE also rules out the separating equilibrium where P plays $e = 0$ and Q plays $e = 3$. Consider a deviation to $e = 1$: If both types are equally likely to have deviated then the Receiver's best response is A, which indeed makes both types better off and self-confirms the Receiver's 50-50% belief.

The point we make in the paper regarding the difference between our concept and PSE is that we do not find the latter reasoning convincing. Starting from the separating equilibrium, if a deviation to $e = 1$ occurs, we believe it is too restrictive to assume that the Receiver has (near) 50-50% beliefs. If she happens to believe that type P is much more likely to have deviated than type Q (e.g., because type P gains more from $e = 1$ followed by A) then she should respond with HR, which in turn scares off both types. If the Sender cannot be sure what (reasonable) beliefs the Receiver may form upon a deviation, and he is sufficiently ambiguity-averse, then this equilibrium should not be considered Vulnerable. $\square$