TI 2015-012/VII
Tinbergen Institute Discussion Paper

Equilibrium Selection in Experimental Cheap Talk Games

Adrian de Groot Ruiz[1]
Theo Offerman[2]
Sander Onderstal[2]

[1] *Radboud University Nijmegen, the Netherlands;*
[2] *Amsterdam School of Economics, University of Amsterdam, and Tinbergen Institute, the Netherlands.*

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Equilibrium Selection in Experimental Cheap Talk Games*

Adrian de Groot Ruiz[a], Theo Offerman[b] and Sander Onderstal[b]

January 14, 2015

In the past, many refinements have been proposed to select equilibria in cheap talk games. Usually, these refinements were motivated by a discussion of how rational agents would reason in some particular cheap talk games. In this paper, we propose a new refinement and stability measure that is intended to predict actual behavior in a wide range of cheap talk games. According to our Average Credible Deviation Criterion (ACDC), the stability of an equilibrium is determined by the frequency and size of credible deviations. ACDC organizes the results from several cheap talk experiments in which behavior converged to equilibrium, even in cases where other criteria do not make a prediction.

KEYWORDS: cheap talk, neologism proofness, credible deviation, refinement, ACDC, experiment

[*] We are grateful to Navin Kartik, Stefano Dellavigna and Charlie Holt for insightful advise and Joel Sobel and two anonymous referees for extensive comments. In addition, we would like to thank seminar participants at the Universities of Aachen, Amsterdam, Bonn, Cologne, East Anglia, and at LMU Munich and New York University for helpful comments. Financial support from the Dutch National Science Foundation (NWO-VICI 453-03-606) and from the University of Amsterdam Research Priority Area in Behavioral Economics is gratefully acknowledged.

[a] Department of Economics, Institute for Management Research, Radboud University Nijmegen, P.O, Box 9108, 6500 HK Nijmegen, the Netherlands; a.grootruiz@fm.ru.nl.

[b] Faculty of Economics and Business, University of Amsterdam, and Tinbergen Institute, Roetersstraat 11, 1018 WB, the Netherlands; T.J.S.Offerman@uva.nl, onderstal@uva.nl.

1 Introduction

Crawford & Sobel (1982) showed how meaningful costless communication between an informed Sender and an uninformed Receiver can be supported in equilibrium when Sender and Receiver's preferences are not perfectly aligned.[1] Their seminal paper inspired many applications ranging from the presidential veto (Matthews, 1989), legislative committees (Gilligan & Krehbiel, 1990) and political correctness (Morris, 2001) to double auctions (Matthews & Postlewaite (1989); Farrell & Gibbons (1989)), stock recommendations (Morgan & Stocken, 2003) and matching markets (Coles, Kushnir & Niederle, 2013). These cheap talk games are characterized by multiple equilibria which differ crucially in their prediction about how much information will be transmitted.

Several refinements have been proposed to select an equilibrium in cheap talk games. Often, such refinements were based on an intuitive notion of how rational players would reason in the context of a particular set of cheap talk games.[2] For instance, Farrell (1993) and Matthews, Okuno-Fujiwara & Postlewaite (1991) formulated refinements in which equilibria are discarded if they allow senders to submit credible deviating messages.[3] Unfortunately, both Farrell's neologism proofness and Matthews et al.'s (strong) announcement proofness criteria eliminate all equilibria in many games, including the original Crawford-Sobel game.[4] Several other types of concepts have been proposed that distinguish between stable and unstable equilibria (or profiles), such as Partial Common Interest (PCI) (Blume, Kim & Sobel, 1993), the recurrent mop (Rabin & Sobel, 1996) and No Incentive To Separate (NITS) (Chen, Kartik & Sobel,

[1] Much earlier, Schelling (1960/1980) discussed how costless communication could help players to coordinate when preferences are aligned (see in particular Chapter 4 and pages 120/121). Lewis (1969) showed that meaningful costless communication emerges in a separating equilibrium of a sender-receiver game where Sender and Receiver preferences are perfectly aligned.

[2] For a comprehensive review of Sender-Receiver games, see Sobel (2013).

[3] Standard signaling refinements such as Kohlberg & Mertens' (1986) strategic stability have no bite in cheap talk games because messages are costless.

[4] Weak announcement proofness tends to eliminate too few equilibria.

2008). These criteria often select a plausible equilibrium in specific settings, but fail to discriminate successfully across a wider range of cheap talk games.

In this paper, we propose a refinement that is intended to predict actual behavior in a wide range of cheap talk applications. Our Average Credible Deviation Criterion (ACDC) is based on credible deviations but allows for a continuous instead of a binary stability concept. Its main contribution to the literature is that it makes a prediction in many games and that its predictions are validated by experimental evidence.

ACDC takes as a point of departure the theory of credible neologisms (Farrell, 1993). This theory stipulates conditions under which a message inducing a deviation from equilibrium is credible and thus upsets the equilibrium. The current approach is to assume that all equilibria that admit credible deviations are equally unstable. ACDC, however, assumes that the stability of an equilibrium is a decreasing function of its Average Credible Deviation (ACD), a measure of the frequency and intensity of credible deviations. The ACD measures the mass of types that can credibly deviate and the size of those induced deviations (as measured by the difference in Sender payoff between the equilibrium and deviating action). Comparable equilibria will perform better if they have a lower ACD on this account. In particular, we call an equilibrium that minimizes the ACD in a game an 'ACDC equilibrium.' This allows us to select equilibria, even in games where no equilibrium is completely stable.

We think ACDC provides an intuitive solution to the equilibrium selection problem for cheap talk games. Humans occasionally make errors, which implies that behavior is seldom completely in a deterministic equilibrium. In the context of a standard Crawford-Sobel game, we show that the predictions of ACDC are supported by a simple 'neologism dynamic'. The neologism dynamic is a best-response dynamic that allows Senders to send credible neologisms when an equilibrium is reached.[5]

[5] Quantal response equilibrium (McKelvey & Palfrey 1998) provides an alternative approach to account for the possibility that people sometimes err. Unlike ACDC, it is typically not

Existing criteria like neologism proofness, announcement proofness, and many of the other cheap talk refinements impose a neat binary distinction between stable and unstable equilibria. Whereas such a binary criterion is appropriate for rational agents, it may unnecessarily lose predictive power when applied to human behavior. ACDC solves this problem in two ways. It is able to select among equilibria in a wide range of games and it provides a continuous stability measure for each equilibrium.

We derive the following results. First, we show that an ACDC equilibrium exists under general conditions. In all applications we have come across, there is a unique ACDC equilibrium. Second and more importantly, the predictions of ACDC are validated by existing experiments. Wherever experimental evidence exists, the predictions of ACDC are in line with the data in two ways. First, if outcomes are consistent with some equilibrium, it is most likely to be consistent with the ACDC equilibrium. Second, comparing outcomes between games, behavior is more likely to be consistent with the ACDC equilibrium the lower its ACD. In this way, ACDC performs at least as well as other criteria, if they are predictive. In addition, it also makes predictions when other criteria are silent. We show that ACDC selects the unique maximum size equilibrium in the leading uniform-quadratic case of the Crawford-Sobel game for a large range of bias parameters. Until now, only NITS was able to select this equilibrium in the Crawford-Sobel setting (Chen, Kartik & Sobel, 2008). In addition, the maximum size equilibrium becomes more stable as the bias parameter becomes smaller according to ACDC, which is not predicted by existing criteria. Both results are supported by experimental work on (discrete) Crawford-Sobel games (Dickhaut, McCabe & Mukherji (1995), Cai & Wang (2006), Wang, Spezio & Camerer (2010)).

Furthermore, we find that ACDC organizes the main features of the experimental data of the discrete games analyzed by Blume, DeJong, Kim & Sprinkle

selective in cheap talk games because the pooling equilibrium is always a (limiting principal branch agent) quantal response equilibrium.

(2001), originally intended to test the Partial Common Interest criterion. Finally, ACDC is also successful in organizing the results of De Groot Ruiz, Offerman & Onderstal (2014). In that paper, we test the predictions of ACDC in a class of veto-threat games introduced in De Groot Ruiz, Offerman & Onderstal (2012). The data corroborate the predictions of ACDC. The ACDC equilibrium performs better in games where its ACD is smaller. In addition, in each treatment the ACDC equilibrium predicts best, also in games where all other criteria do not select a unique equilibrium.

This paper has the following structure. In section 2, we motivate, define and illustrate ACDC. In section 3, we apply ACDC to the Crawford-Sobel uniform-quadratic model and compare it to other concepts in this framework. In section 4, we review existing experimental evidence on the Crawford-Sobel game in the light of ACDC and the neologism dynamic. In section 5, we discuss other experimental evidence on ACDC. Finally, section 6 concludes.

2 Motivation, Definition, Properties, and Applications

2.1. Motivation

Applied theorists who analyze strategic information transmission face the following problem when they analyze a new cheap talk game. The model is likely to have several equilibria and so one would like a concept that selects the most plausible equilibrium, that tells one how stable that equilibrium is, and that is validated by experimental data. However, currently chances are high that one will not find such a concept for the new cheap talk game for two reasons.

First, existing selection criteria tend to select an equilibrium in specific classes of games but not in all relevant applications. Neologism proofness (Farrell, 1993) and announcement proofness (Matthews, Okuno-Fujiwara & Postlewaite, 1991) provide a strong intuition and make meaningful predictions in specific simple discrete games. However, they fail to select an equilibrium in many applied settings, such as that of Crawford and Sobel (1982). In contrast, NITS is very effective in selecting equilibria in the Crawford-Sobel model (Chen, Kartik & Sobel, 2008). The predictions by NITS are to some extent in line with experimental evidence on the Crawford-Sobel model (Dickhaut, McCabe & Mukherji (1995); Cai & Wang (2006); Wang, Spezio & Camerer (2010)). If subjects behave according to any equilibrium, it is the most informative equilibrium, as predicted by NITS. However, in other games, NITS is often not defined, as we will discuss in section 3. PCI has shown to make a prediction which is borne out by experimental data in particular discrete games (Blume, DeJong, Kim & Sprinkle, 2001), but does not select a partition in many continuous settings, such as that of Crawford-Sobel.

Second, even if current concepts select an equilibrium, they do not tell how stable it is, although experimental evidence suggest there is a considerable degree of variation in the stability of cheap talk equilibria. For instance, experiments on (discrete versions of) the Crawford-Sobel game show that the NITS equilibrium indeed performs best, but that behavior is less likely to be consistent with the equilibrium outcome when the bias parameter increases.

Game A in Table 1 illustrates these two issues of selection and stability. In Game A, the Sender sends a costless message m to the Receiver, who then takes an action from the set $\{a_1, a_2, a_3, a_4, a_5\}$.[6] The payoffs for both players depend on the Receiver's action and the Sender's type. The Sender's type is private information and is t_1 or t_2 each with probability $(1-\delta)/2$ and t_3 with probability δ.

[6] We will refer to the Sender as a 'she' and the Receiver as 'he.'

TABLE 1
GAME A

	a_1	a_2	a_3	a_4	a_5
t_1 $(\frac{1-\delta}{2})$	1, 4	0, 0	0, 0	0, 0	2, 4 – δ
t_2 $(\frac{1-\delta}{2})$	0, 0	0, 2 + δ	3, 0	4, 2	2, 1
t_3 (δ)	0, 0	0, 0	2 + ε, 3	2, 2	1, 1

Notes: The left column shows the Sender's type and between brackets the probability that it is drawn. The top row shows the Receiver's actions. The remaining cells provide the Sender's payoff in the first entry and the Receiver's payoffs in the second entry. $0 < \delta < \frac{1}{3}$ and $0 \leq \varepsilon < 1$.

Game A has two equilibrium outcomes, one resulting from pooling and the other from partial separating.[7] We say a type t induces action a if the Receiver always takes action a after any message t sends in equilibrium. In the pooling equilibrium, all Senders induce a_5. In the partially separating equilibrium, t_1 induces a_1, whereas t_2 and t_3 induce a_4.

What do credible neologisms (Farrell, 1993) do in this game? Neologisms are out-of-equilibrium messages which are assumed to have a literal meaning in a pre-existing natural language.[8] Farrell considers neologisms which literally say: "play action $\tilde{a}$, because my type is in set N." Farrell deems a neologism credible if and only if (*i*) all types t in N prefer $\tilde{a}$ to their equilibrium action $\alpha(t)$, (*ii*) all types t not in N prefer their equilibrium action $\alpha(t)$ to $\tilde{a}$ and (*iii*) the best reply of the Receiver after restricting the support of his prior to N is to play $\tilde{a}$. We will denote neologisms by $\langle \tilde{a}, N \rangle$. According to Farrell, credible deviations lead rational players to deviate from equilibrium. An equilibrium is neologism proof, and stable on this account, if and only if it does not admit any credible neologism.

[7] In the remainder of the text, we will slightly abuse terminology by referring to these equilibrium outcomes as 'the pooling equilibrium' and 'the partially separating equilibrium' respectively.

[8] See Blume, DeJong, Kim & Sprinkle (1998) and Agranov & Schotter (2012) for studies on the role of natural language in cheap talk games.

If $\varepsilon = 0$, neologism proofness provides a compelling reason why the partially separating equilibrium is more plausible than the pooling equilibrium. The pooling equilibrium admits the credible neologism $\langle a_4, \{t_2, t_3\}\rangle$. Hence, it is likely to be unstable as types t_2 and t_3 can credibly separate themselves from t_1. On the other hand, the partially separating equilibrium is neologism proof: it admits no credible neologisms and is stable. For $\varepsilon > 0$, a key limitation of neologism proofness becomes evident. In this case, the partially separating equilibrium also admits a credible neologism, to wit $\langle a_3, \{t_3\}\rangle$. This leaves us with no stable equilibrium and no prediction. For entirely rational agents, the fact that neither equilibrium is stable might be all there is to be said. When explaining or predicting human behavior, however, we feel we can go further.[9]

In game A, even though the partially separating equilibrium is not entirely stable, it seems more plausible than the pooling equilibrium if either t_3 is infrequent (δ small) or t_3 has a very small incentive to deviate (ε small). If δ is small, then the partially separating equilibrium will be upset with a small probability, whereas the pooling equilibrium will be upset almost half of the time. Similarly, if ε is small, then t_3 has a small incentive to deviate in the partially separating equilibrium and may choose to stick to it, lest she be misunderstood and get a payoff lower than she gets by sticking to equilibrium. Hence, we would expect to observe behavior close to the partially separating equilibrium more frequently than behavior close to the pooling equilibrium. This implies two things. First, it may be possible to select the most plausible equilibrium in a game, even though no equilibrium exists that is entirely stable. Second, to describe behavior in a cheap talk game one needs a continuous stability measure and not just a binary criterion.

[9] We consider equilibrium to be most meaningful in a dynamic context, where members of a group interact frequently with different other members. In this context language evolves and behavior is shaped by strategic forces in the direction of equilibrium. For a one-shot game between rational individuals without social information, an approach based on rationalizability and some focal meaning of messages, such as that in Rabin (1990), may be appropriate.

2.2. Definition and General Results

Our intuition is that, from a behavioral perspective, the stability of an equilibrium is a decreasing function of the average *intensity* of the credible deviations it admits. This depends, firstly, on the mass of types that can credibly induce a deviation and, secondly, on the intensity of the deviation, measured by the incentive the Sender has to deviate. As a consequence, if the deviating mass and the induced deviations from equilibrium are small, the equilibrium is likely to be a good predictor of behavior. We formalize this intuition in the ACDC criterion. We first provide a definition of ACDC and apply it to the Crawford-Sobel game in the following section.

We define ACDC in the following context. Nature draws the Sender type t from probability density f on T, where T is a compact metric space. The Sender then privately observes her type t and chooses a costless message $m \in M$, where M represents the set of available messages. After having observed the Sender's message, the Receiver chooses an action $a \in A$, where A is a compact metric space. [10] Let $U^R : A \times T \rightarrow \mathbb{R}$ be the utility function of the Receiver $U^S : A \times T \rightarrow \mathbb{R}$ that of the Sender. We assume both are bounded from above and below. A strategy for the Sender consists of a function $\mu : T \rightarrow M$, and a strategy of the Receiver is a function $\alpha : M \rightarrow A$. Let Σ^S be the set of Sender strategies and Σ^R the set of Receiver strategies. Let $\{\mu, \alpha\}$ be a strategy profile and Σ the set of all strategy profiles. Finally, let the Receiver have prior beliefs $\beta^0(t) = f(t)$. A pure strategy perfect Bayesian equilibrium (henceforth just 'equilibrium') $\sigma = \{\mu, \alpha, \beta\}$ is characterized by the following three conditions:

[10] This representation allows for T and A to be de facto discrete, by allowing U^S and U^R to be constant on regions of the type and outcome space.

$$\text{For each } t \in T,\ \mu(t) \in \arg\max_{m \in M} U^S(\alpha(m), t)$$

$$(1) \qquad \text{For each } m \in M, \alpha(m) \in \arg\max_{a \in A} \int_T U^R(a,t)\beta(t \mid m)dt$$

where $\beta(t \mid m)$ denotes the Receiver's posterior beliefs, which is derived from μ and β^0 using Bayes' rule wherever possible.

Let Σ^* be the set of equilibria. ACDC provides a stability measure and a selection criterion for equilibria in Σ^*. The starting point of ACDC is Farrell's (1993) theory of credible neologisms. Recall that a credible neologism, denoted by $\langle \tilde{a}, N \rangle$, satisfies the following properties: (*i*) all types *t* in *N* prefer $\tilde{a}$ to their equilibrium action $\alpha(t)$, (*ii*) all types t not in N prefer their equilibrium action $\alpha(t)$ to $\tilde{a}$ and (*iii*) $\tilde{a}$ is the Receiver's best response after restricting the type support to *N*. Following Farrell (1993), we assume that players have access to some 'natural language' that allows Senders and Receivers to agree on the meaning of out-of-equilibrium messages. An equivalent assumption is that the players have learned to coordinate on common interpretations of initially abstract messages.[11]

We associate a deviating profile $\gamma(\sigma) = \{\mu^{\gamma(\sigma)}, \alpha^{\gamma(\sigma)}\} \in \Sigma$ with each equilibrium $\sigma = \{\mu, \alpha, \beta\}$ where $\mu^{\gamma(\sigma)}(t) = \langle \tilde{a}, N \rangle$ if type t can send credible neologism $\langle \tilde{a}, N \rangle$ relative to σ,[12] and $\alpha^{\gamma(\sigma)}(t) = \tilde{a}$. If a type t is not part of a credible neologism relative to σ, then $\mu^{\gamma(\sigma)}(t) = \mu(t)$ and $\alpha^{\gamma(\sigma)}(t) = \alpha(\mu(t))$, i.e., $\gamma(\sigma)$ indicates that those Sender types stick to their equilibrium message and the Receiver to the corresponding equilibrium action. So, $\gamma(\sigma)$ specifies firstly which Sender types would deviate and in which way (by sending credible neologisms

[11] Blume (1996) shows that in cheap talk games with a finite type space (an assumption that is typically satisfied in experimental cheap talk games), a dynamic process based on perturbed games converges to effective communication under a 'rich language' condition (which may or may not be satisfied in the laboratory). Blume et al. (1998) present experimental evidence on how meaningful communication may evolve in the sense that a priori meaningless messages become informative over time.

[12] We assume that if a type can send multiple neologisms, she sends the neologism that gives her the highest payoff.

according to message strategy $\mu^{\gamma(\sigma)}$), and secondly, how the Receiver would react (by responding to credible neologisms according to $\alpha^{\gamma(\sigma)}$). Note that if an equilibrium $\sigma = \{\mu, \alpha, \beta\}$ is neologism proof, no type can send a credible neologism so that $\gamma(\sigma) = (\mu, \alpha)$.

For simplicity, we define ACDC for pure strategies. The definition of ACDC can be straightforwardly extended to also incorporate Receiver mixed strategies. Allowing for Sender mixed strategies is more problematic because by doing so any equilibrium outcome could be implemented by a strategy where the Sender uses all possible messages, leaving no room for neologisms. Farrell (1993) argues that in a setting with a natural language it is implausible for the Sender to randomize over messages. We follow Farrell (1993) by assuming that sufficiently many unused out-of-equilibrium messages exist so that any subset of types that wishes to induce the Receiver to play an out-of-equilibrium action is able to send a neologism (see also Matthews et al. (1991)). If Senders play pure strategies and if Sender types who induce the same equilibrium action use the same message, a sufficient condition for sufficiently many out-of-equilibrium message to exist is that the cardinality of M is at least as great as the cardinality of A. In that case, for any action the Receiver does not use in equilibrium, the Sender has an unused message available she could use to induce the Receiver to play that action.

Let $\Sigma^{\dagger}$ be the set of rationalizable strategy profiles. Then, $\underline{U}^S(t) \equiv \inf_{\{\alpha,\mu\}\in\Sigma^{\dagger}} U^S\big(t, \alpha\big(\mu(t)\big)\big)$ and $\bar{U}^S(t) \equiv \sup_{\{\alpha,\mu\}\in\Sigma^{\dagger}} U^S\big(t, \alpha\big(\mu(t)\big)\big)$ are the lowest and highest rationalizable payoff for Sender type t. Now, we measure the *intensity* of type t's credible deviation as

$$CD(t,\sigma) \equiv \frac{U^S\left(t, \alpha^{\gamma(\sigma)}(\mu^{\gamma(\sigma)}(t))\right) - U^S\left(t, \alpha(\mu(t))\right)}{\bar{U}^S(t) - \underline{U}^S(t)} \qquad (2)$$

if $U^S\left(t,\alpha(\mu(t))\right) > \underline{U}^S(t)$. If $U^S\left(t,\alpha(\mu(t))\right) = \underline{U}^S(t)$, then $CD(t,\sigma) \equiv 1$, as in this case the Sender has no incentive to adhere to her equilibrium strategy.

$CD(t,\sigma)$ captures the likelihood of type t deviating from her equilibrium strategy. It has the desirable properties that it is

- invariant to affine transformations of payoffs;
- increasing in the difference between the deviating and equilibrium payoff.

Additionally, it is increasing in the lowest rationalizable payoff. We believe this to be a desirable as the lowest rationalizable payoff measures how risky it is to deviate from equilibrium for the Sender. Finally, we think it is convenient to normalize it so that it is 0 if the difference between deviating and equilibrium payoff is zero and 1 if the difference between deviating and equilibrium payoff is maximal. Even though we think our specific normalization is very natural because it makes the measure invariant to affine transformations of payoffs, we are aware that there are other ways in which the incentive to send a neologism could have been normalized. We acknowledge that the specific implementation of the normalization term is important because it may affect the predictions from the theory.

We define the Average Credible Deviation (ACD) of an equilibrium σ as:

$$(3) \quad ACD(\sigma) = E_t[CD(t,\sigma)]$$

Based on the ACD, we formulate the ACD-Criterion (ACDC), which says that an equilibrium σ^* will on average predict better than equilibrium σ if $ACD(\sigma^*) < ACD(\sigma)$. In particular, based on ACDC we can formulate the following selection criterion:

Definition 1 *An equilibrium σ^* is an ACDC equilibrium if $ACD(\sigma^*) \leq ACD(\sigma)$ for all $\sigma \in \Sigma^*$.*

Note that this selection criterion selects the equilibrium that will predict best on average rather than the equilibrium that will always be played. A simple implication is that if σ is neologism proof, it is an ACDC equilibrium.

Observe that if weakly dominated Receiver actions are added to the game, $CD(t,\sigma)$ may be affected and, in turn, the ACDC predictions may change. At this moment, we cannot tell whether this is a good or bad feature of our criterion. We are not aware of existing experiment evidence that would allow us to test if behavior changes in the right direction when weakly dominated Receiver actions are added to the game.[13]

The following result is immediate.

Proposition 1 *If the number of equilibrium outcomes is finite, the cheap talk game has an ACDC equilibrium.*

Hence, existence of an ACDC equilibrium is guaranteed by a finite set of equilibrium-outcomes. This is a relevant result, as Park (1997) has shown that finite Sender-Receiver games have a finite set of equilibrium outcomes under generic conditions. Before, Crawford and Sobel (1982) showed a similar result for their setting with a continuous type-space. Even if games do not have a finite outcome set, mild conditions can be formulated in order to guarantee existence of an ACDC equilibrium:

Proposition 2 *Let s be an equilibrium outcome and $ACD(s)$ the ACD of equilibria inducing s. Suppose the equilibrium outcome set S can be represented by a finite union of compact metric spaces $S = \bigcup_{i\in N} S_i$, such that $ACD(s)$* is *continuous in s on all subsets S_i. Then, an ACDC equilibrium exists.*

[13] If future experiments indicate that adding weakly dominated Receiver actions does not affect behavior, in contrast to what ACDC would predict, the ACDC definition can be straightforwardly adapted by defining it on the basis of the game that results when all weakly dominated Receiver actions are eliminated. This would not change any of the results in this paper because the Receiver does not have weakly dominated actions in any of the games that we study.

Proof $ACD(s)$ achieves a minimum on each compact subset S_i and thus on S. Hence, $\min_{\Sigma^*} ACD(\sigma)$ is nonempty and an ACDC equilibrium exists. *Q.E.D.*

Proposition 2 is informative for continuous games for which the equilibrium outcome set is known but not finite. This proposition implies that continuous games with an equilibrium set consisting of partition equilibria that are well-behaved with respect to their ACD will have an ACDC equilibrium.[14] For instance, the class of continuous veto-threats games we introduce in De Groot Ruiz, Offerman & Onderstal (2012) has an infinite equilibrium outcome set that meets the conditions of Proposition 2.

We conclude this subsection with a motivation as to why we use Farrell's (1993) theory of credible neologisms to construct the deviation profile γ instead of (weak/strong) credible announcements as introduced by Matthews, Okuno-Fujiwara & Postlewaite (1991). We agree with the observations in Matthews, Okuno-Fujiwara & Postlewaite (1991) about the limitations of credible neologisms for rational agents. However, the motivation behind ACDC is to predict behavior and explain experimental data. Hence, our aim is somewhat different from that of most of the credible deviations literature, of which the main concern is to establish what would be credible from a rational perspective. In our view, the simplicity of credible neologisms makes it the most apt to describe the behavior of boundedly rational individuals.

2.3. Applying ACDC

We will now apply ACDC to Game A in Table 1. In a discrete game, the ACD of a pure equilibrium σ (with a pure $\gamma(\sigma)$) reduces to

[14] An equilibrium of a game with a one-dimensional type and action set is a partition equilibrium if there exists a partition $t_0 < t_1 < \cdots < t_{n-1} < t_n$ of T such that each type in $[t_{i-1}, t_i]$ induces action a_i with $a_1 < a_2 < ... < a_n$. Hence, a partition equilibrium is characterized by a vector $a = (a_1, ..., a_n)$ and a partition equilibrium outcome set can be represented by a finite union of subsets of $\mathbb{R}_1, ..., \mathbb{R}_n$.

$$ACD(\sigma)=\sum_{t\in T} f(t)\frac{U^S\left(t,\alpha^{\gamma(\sigma)}(\mu^{\gamma(\sigma)}(t))\right)-U^S\left(t,\alpha(\mu(t))\right)}{\bar{U}^S(t)-\underline{U}^S(t)},$$

where $f(t)$ is type t's prior probability. Recall that in Game A in the pooling equilibrium, all Senders induce a_5 and $\langle a_4,\{t_2,t_3\}\rangle$ is the unique credible neologism. Hence, the ACD of the pooling equilibrium is $\frac{(1-\delta)}{2}\frac{(4-2)}{4-0}+\delta\frac{2-1}{2+\varepsilon-0}=\frac{1}{4}+\delta\frac{2-\varepsilon}{8+4\varepsilon}$. In the partially separating equilibrium t_1 induces a_1, whereas t_2 and t_3 induce a_4 and $\langle a_3,\{t_3\}\rangle$ is admitted. Hence, the ACD of the partially separating equilibrium is $\delta\frac{(2+\varepsilon-2)}{2+\varepsilon-0}=\frac{\delta\varepsilon}{2+\varepsilon}$. It is readily verified that the pooling equilibrium's ACD is greater than the partially separating equilibrium's so that the latter is the ACDC equilibrium. In addition, the ACD of the partially separating equilibrium goes to zero if δ or ε go to zero. Finally, even though the partially separating equilibrium is the ACDC equilibrium, if δ or ε become large, it becomes less stable because a type can deviate frequently or deviations have a high intensity.

3 Crawford-Sobel Game

In this section, we apply ACDC to the leading uniform-quadratic case of Crawford & Sobel's (1982) cheap talk game (henceforth 'CS game'). We compare its predictions to those of existing refinements.

In the CS game, types are uniformly distributed on [0,1], the action space is [0,1], $U^R(a,t)=-(a-t)^2$ and $U^S(a,t)=-(a-(t+b))^2$, with $b>0$ capturing the Sender bias. Crawford & Sobel (1982) show that this game only has (perfect Bayesian) partition equilibria and that the maximum equilibrium size $n(b)$ is the largest integer n for which

$$(2) \qquad 2n(n-1)b < 1.$$

The game has a unique size-n equilibrium for each $n \in \{1,...,n(b)\}$. Let

$$(3) \qquad t_i^n \equiv \frac{i}{n} - 2bi(n-i).$$

for $i = 0,...,n$ and $n = 1,...,n(b)$. In the size-n equilibrium, types in $[t_{i-1}^n, t_i^n)$ send the same equilibrium message, which induces the Receiver to choose action

$$(4) \qquad a_i^n = \frac{1}{2}(t_{i-1}^n + t_i^n), \;\; i = 1,...,n.$$

We start by deriving all credible neologisms the equilibria admit. For each credible neologism $\langle \tilde{a}, N \rangle$, the set of deviating types N turns out to be an interval between some $\underline{\tau}$ and $\overline{\tau}$. Hence, we can characterize neologisms by $[\underline{\tau}, \overline{\tau}]$ alone, since the Receiver's best response is $\tilde{a} = \dfrac{\underline{\tau} + \overline{\tau}}{2}$. An equilibrium can admit three types of credible neologisms. First of all, there may be a credible neologism which includes $t = 0$. If this credible neologism exists, then it has the shape $[0, \overline{\tau}_0^n)$ where

$$\overline{\tau}_0^n = \frac{2}{3}a_1^n - \frac{4}{3}b = \frac{1}{3n} - \frac{2}{3}b\left(n+1\right).$$

Chen, Kartik & Sobel (2008) show that an equilibrium that fails NITS has a credible neologism of this kind and prove that only the size-$n(b)$ equilibrium satisfies NITS. Hence, the credible neologism $[0, \overline{\tau}_0^n)$ exists if and only if $n < n(b)$.

Second, Farrell (1993) shows that if $b < \frac{1}{2}$, the game has a credible neologism on the right-end of the type space of the form $(\underline{\tau}_n^n, 1]$ where

$$\underline{\tau}_n^n = 1 - \frac{1}{3n} - \frac{2}{3} b(n+1).$$

Finally, if $n \in \{2, ..., n(b) - 1\}$, there are $n - 1$ credible neologisms "in the middle." These take the form $(\underline{\tau}_i^n, \overline{\tau}_i^n)$ for $i = 1, ..., n-1$, where $\underline{\tau}_i^n$ $[\overline{\tau}_i^n]$ is indifferent between the equilibrium action a_i^n $[a_i^{n+1}]$ and the neologism action $\tilde{a}_i^n = (\underline{\tau}_i^n + \overline{\tau}_i^n) / 2$. We obtain for $i = 1, ..., n-1$:

$$\begin{aligned} \underline{\tau}_i^n &= \frac{3}{4} a_i^n + \frac{1}{4} a_{i+1}^n - 2b \text{ and} \\ \overline{\tau}_i^n &= \frac{1}{4} a_i^n + \frac{3}{4} a_{i+1}^n - 2b, \end{aligned} \tag{5}$$

If $n = n(b)$, the game has the same types of credible neologisms "in the middle," with the exception that the neologism $\left(\underline{\tau}_1^{n(b)}, \overline{\tau}_1^{n(b)}\right)$ need not exist.[15] Observe that $\overline{\tau}_{i-1}^n < \underline{\tau}_i^n$ for $i = 1, ..., n$, so that none of the credible neologisms overlap. Figure 1 illustrates the results for $b = \frac{1}{18}$.

It seems intuitive that the highest size equilibrium is the ACDC equilibrium, since the deviations seem to get smaller and smaller as the size increases. This indeed turns out to be the case. Although one can obtain analytical results for the ACD for specific parameter values, finding the ACDC equilibrium for

[15] If (and only if) $2bn(b)^2 \geq 1$, there is no credible neologism of the form $(\underline{\tau}_1^{n(b)}, \overline{\tau}_1^{n(b)})$ because $\underline{\tau}_1^{n(b)} = \frac{3}{4} a_1^{n(b)} + \frac{1}{4} a_2^{n(b)} - 2b = -\frac{3}{4n(b)}\left(2bn(b)^2 - 1\right) \leq 0$, which is inconsistent with all types being in the interval $[0,1]$ or the interval $\left(\underline{\tau}_1^{n(b)}, \overline{\tau}_1^{n(b)}\right) = \left(0, t_1^{n(b)}\right)$ being a neologism.

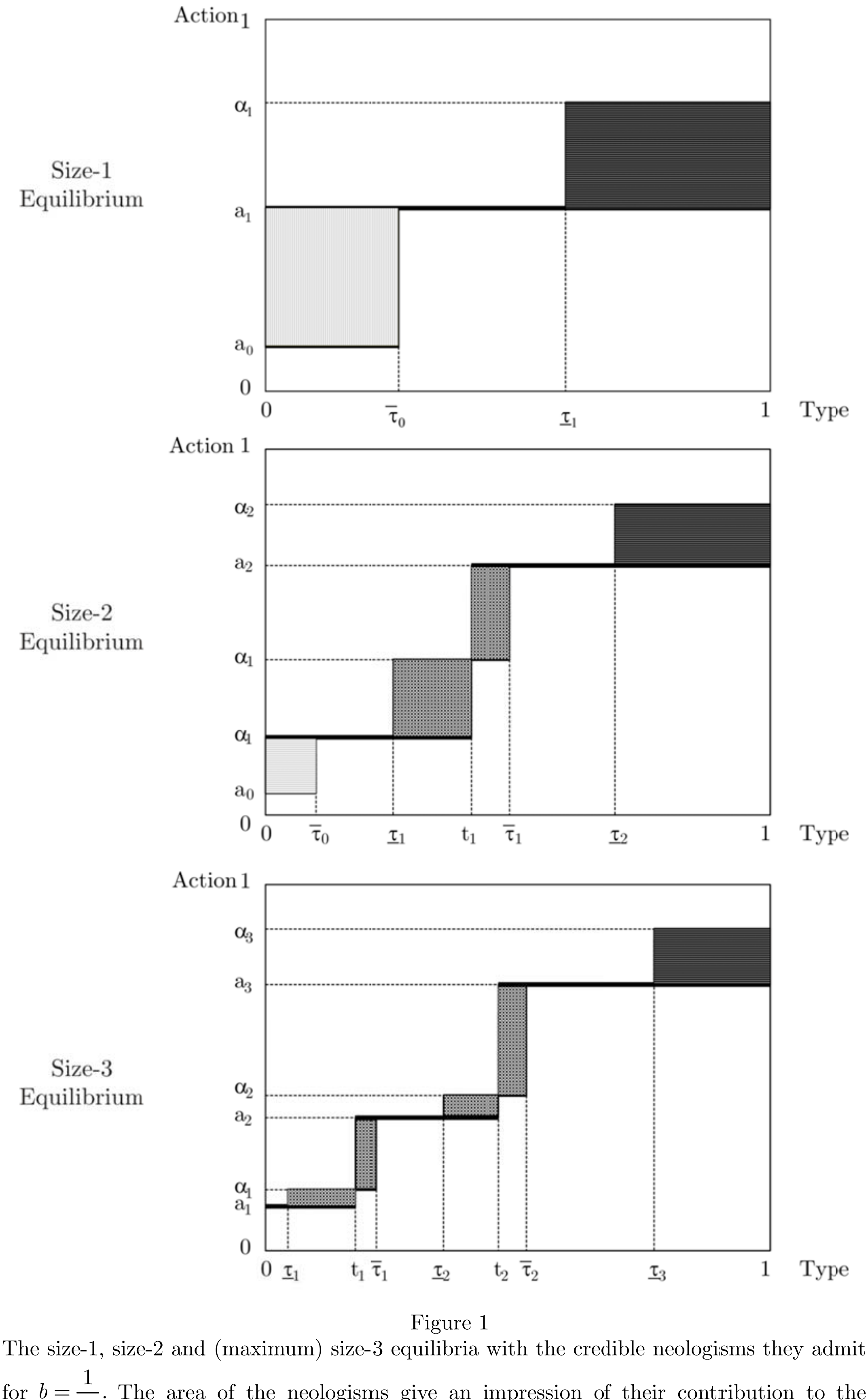

Figure 1

The size-1, size-2 and (maximum) size-3 equilibria with the credible neologisms they admit for $b = \frac{1}{18}$. The area of the neologisms give an impression of their contribution to the ACD, although their height contributes quadratically to the ACD.

general b defies an analytical approach. Hence, we calculated the ACD for a very fine grid of b and obtain the following result.

Proposition 3 *For all* $b \in \left\{\frac{1}{10000}, \frac{2}{10000}, ..., \frac{1}{4}\right\}$ *it holds that the ACD of the size-n equilibrium in the CS game is decreasing in n.*

Proof See the Appendix.

Corollary 1 *For all* $b \in \left\{\frac{1}{10000}, \frac{2}{10000}, ..., \frac{1}{4}\right\}$, *the size-n(b) equilibrium is the unique ACDC equilibrium of the CS game.*

We also derive the following property of the maximum size equilibrium (for which we do not need to calculate the ACD's for each b):

Proposition 4 *The ACD of the size-n(b) equilibrium tends to zero if b tends to zero in the CS game.*

Proof Let $\sigma(b) \equiv \sigma_b^{n(b)}$ be the maximum size equilibrium for b. Then,

$$\lim_{b \downarrow 0} ACD(\sigma(b)) \le \lim_{b \downarrow 0} E_t\left[\frac{U^S(\tilde{a}^{\sigma(b)}(t), t) - U^S(a^{\sigma(b)}(t), t)}{\min\limits_{t \in T}\{\bar{U}^S(t) - \underline{U}^S(t)\}}\right] \le \lim_{b \downarrow 0} E_t\left[\frac{0 - U^S(a^{\sigma(b)}(t), t)}{\frac{1}{4}}\right]$$

$$= -4 \cdot \lim_{b \downarrow 0} EU^S \overset{1}{=} 4 \cdot \lim_{b \downarrow 0}\left(b^2 + \frac{1}{12n(b)^2} + \frac{b^2(n(b)^2 - 1)}{3}\right)$$

$$\le 4 \cdot \lim_{b \downarrow 0}\left(b^2 + \frac{1}{n(b)^2} + b^2 n(b)^2\right) \overset{2}{\le} 4 \cdot \lim_{b \downarrow 0}\left(b^2 + \frac{4}{\left(\sqrt{2 / b + 1} - 1\right)^2} + \frac{\left(b + \sqrt{2b + b^2}\right)^2}{4}\right) = 0$$

Equality 1 follows from the specification of EU^S in Crawford & Sobel (1982). Inequality 2 follows from $n(b) = \left\lceil \frac{1}{2} + \frac{1}{2}\sqrt{2 / b + 1} \right\rceil - 1$ due to (2). The other manipulations are straightforward. *Q.E.D.*

Hence, the ACD of the maximum size equilibrium converges to zero if b approaches zero, i.e. if the interests of the players are almost perfectly aligned. This finding is intuitive because the Sender obtains almost her ideal outcome when b is close to zero, so she will not gain much in the case of deviation, and even if she deviates, the deviation will hardly change the Receiver's action.

We can now compare ACDC with other criteria. First of all, neologism proofness does not make a prediction: all equilibria admit credible neologisms and are thus unstable. Matthews, Okuno-Fujiwara & Postlewaite (1991) refine neologism proofness with three progressively stronger stability criteria: weak, ordinary and strong announcement proofness. Weak announcement proofness eliminates all equilibria for the same reasons as neologism proofness. Ordinary announcement proofness also tends to eliminate all equilibria and sometimes selects an unintuitive equilibrium. For instance, if $b \in \left(\frac{1}{24}, \frac{1}{16}\right)$, it selects the pooling equilibrium and eliminates the size-2 and size-3 equilibrium.[16] Strong announcement proofness fails to select an equilibrium as it eliminates no equilibrium.

Rabin & Sobel (1996) propose the recurrent mop criterion, which can select equilibria that, although not impervious to credible deviations, are likely to recur in the long run, because they are frequently deviated to. The authors restrict their definition of the recurrent mop to games with a finite number of actions as it may run into problems in continuous games, amongst others because the deviation correspondence may not converge in these settings.

Blume, Kim & Sobel (1993) put forward the Partial Common Interest (PCI) concept. A partition of the typeset satisfies PCI "if types in each partition

[16] For $b \in \left(\frac{1}{24}, \frac{1}{16}\right)$, the pooling equilibrium admits the weakly credible announcement composed of the neologisms at the beginning and end, characterized by the set of intervals of deviating types sending the same message $\left\{\left[0, \frac{1}{6} - \frac{4}{6}b\right], \left[\frac{5}{6} - \frac{4}{6}b, 1\right]\right\}$. In addition, however, it admits the weakly credible announcement $\left\{\left[0, \frac{1}{5} - \frac{16}{5}b\right], \left[\frac{1}{5} - \frac{16}{5}b, \frac{2}{5} - \frac{12}{5}b\right], \left[\frac{3}{5} - \frac{12}{5}b, \frac{4}{5} - \frac{16}{5}b\right], \left[\frac{4}{5} - \frac{16}{5}b, 1\right]\right\}$. Since for all weakly credible announcements deviating types exist that prefer another weakly credible announcement, none is announcement proof. The size-2 and size-3 equilibria only admit weakly credible announcements composed of the non-overlapping credible neologisms, which are thus credible. Observe that the computational demands on agents to determine whether credible announcements exist and how they look like are quite high.

element unambiguously prefer to be identified as members of that element, and there is no finer partition with that property." PCI does not make a definite prediction in the CS game, as no partition of the type space satisfies PCI (except for the type space itself). The main reason is that the highest Sender-type of a partition-element always prefers the Receiver to believe that the upper boundary is higher than the true boundary (except for type $t = 1$).

Non-equilibrium concepts exist too. Rabin (1990) introduced the concept of Credible Message Rationalizability (CMR). This non-equilibrium concept proposes conditions under which communication can be guaranteed to happen. It assumes that rational players take truth-telling as a focal point, but use the strategic incentives of the game to check whether truth-telling is rational. In the CS game, CMR is silent. CMR requires that all Sender-types who send a credible message receive an action in which they achieve their maximum payoff. This would imply that the Receiver does not best respond to credible messages, which cannot be the case under CMR.[17]

The NITS criterion (Chen, Kartik, & Sobel, 2008) is up till now the only refinement based on some notion of stability that can successfully select an equilibrium in the CS game. NITS starts by specifying a 'lowest type,' a type with the property that all other types prefer to be revealed as themselves rather than as that lowest type. An equilibrium survives NITS if the lowest type has no incentive to separate, i.e. if the lowest type prefers her equilibrium outcome to the outcome she would get if she could reveal her type. In the general CS game, only the maximum size equilibrium outcome satisfies NITS. The strength of NITS is that it can make predictions under a general monotonicity assumption, and can be justified on the basis of perturbed games with lying averse or non-strategic players.[18]

[17] Rabin also introduces an equilibrium version of CMR, Credible Message Equilibria (CME), but as a consequence of the previous analysis, neither equilibrium in the game can be a CME.

[18] A challenge for NITS is that it cannot be applied easily in cheap talk games that have no clear lowest type. For instance, in Game A, the lowest type cannot easily be defined. In section 5.2, we discuss a game that has many NITS equilibria.

Alternatively, the maximum size equilibrium in the CS game can also be selected by an approach that does not make use of stability arguments. If players coordinate on the Sender's Ex ante Preferred Equilibrium (SEPE), the equilibrium favored by the Sender when she has not yet been informed of her type, the maximum size equilibrium of the CS game remains the only equilibrium that survives (Crawford & Sobel (1982), Theorem 5).

The prediction of ACDC in Corollary 1 is thus in line with NITS and SEPE. A difference between ACDC on the one hand and NITS and SEPE on the other hand is that the predictions of the latter concepts are invariant to the alignment of the preferences b. In the CS game, SEPE's prediction of the maximum size equilibrium is simply independent of b. NITS assumes that only the lowest type can separate herself. Hence, according to NITS, the most informative equilibrium is (equally) stable regardless of b. According to ACDC, also other types can separate through credible neologisms. In fact, it predicts that the stability of the maximum size equilibrium depends on the bias parameter b. Only in the limit, if b becomes negligible, the maximum size equilibrium is expected to be stable. The experimental data discussed in the next section provides support for this prediction.

4 The Crawford-Sobel Game in the Lab

In this section, we discuss existing experimental evidence on the CS game in the light of the results obtained in the previous section. There, we concluded that both ACDC, NITS, and SEPE select the maximum size equilibrium. Experimental evidence on discrete versions of the Crawford-Sobel linear quadratic game supports this prediction (Dickhaut, McCabe & Mukherji, 1995; Cai & Wang, 2006; Wang, Spezio & Camerer, 2010).

ACDC supports another experimental finding that remains unexplained by NITS and SEPE: As the bias parameter b decreases, the maximum size equilibrium becomes more stable. Consider, for instance, the results on a discrete

Crawford-Sobel game by Cai & Wang (2006) depicted in Table 2.[19] Applying ACDC to this discrete Crawford-Sobel game is straightforward. For example, for $b = 2$, the ACD of the pooling equilibrium is

$$\frac{1}{5}\left(\frac{U^S(5,7) - U^S(5,5)}{U^S(5,7) - U^S(5,1)} + \frac{U^S(7,7) - U^S(7,5)}{U^S(7,9) - U^S(7,1)} + \frac{U^S(9,7) - U^S(9,5)}{U^S(9,9) - U^S(9,1)}\right) \approx 0.137.$$

In Table 2, we provide the ACD-measures for the equilibria of the treatments of Cai & Wang (2006). ACDC makes two predictions in line with the experimental data. First, ACDC selects the most informative equilibrium. Second, the most informative equilibrium has a lower ACD and thus becomes more stable as b decreases and the reverse holds for the pooling equilibrium. Both predictions are intuitive. The most informative equilibrium admits 'fewer' or 'smaller' credible deviations than the pooling equilibrium for all values of b in Table 2. This could provide an explanation why it predicts better.

In addition, as b increases, the most informative equilibrium admits 'more' or 'larger' credible deviations. This may explain the fact that the prediction error of the most informative equilibrium appears to become larger as b increases (and the pooling equilibrium appears to predict less bad). One particular feature of the instability of the most informative equilibrium is that unless it is perfectly separating, there appears to be overcommunication. One explanation for overcommunication could be due to lying averse Senders and/or naïve Receivers (Kartik, Ottaviani, & Squintani, 2007). Overcommunication may also emerge in non-equilibrium models of cognitive hierarchy. Indeed, Crawford, Costa-Gomes, and Iriberri (2013) argue that the overcommunication observed in the Cai & Wang (2006) experiments are well explained by a level-k model anchored on the

[19] The results of Dickhaut, McCabe & Mukherji (1995) on a Crawford-Sobel game are similar to those reported by Cai and Wang, although they do not interpret their results in terms of overcommunication. More recently, Wang, Spezio & Camerer (2010) replicate the results of Cai & Wang (2006) and find that look-up patterns of Senders (as measured by eye-tracking) reveals a significant amount of information about their type.

literal meaning of messages like the one proposed by Crawford (2003).[20] An additional explanation is that credible neologisms not only destabilize but also lead to more information transmission, as can be seen in Table 2. For instance, if $b = 4$, the unique pooling equilibrium predicts no information transmission, but credible neologisms allow types 3 to 9 to separate themselves from type 1 if equilibrium is reached.

TABLE 2
ACDC in Baseline Treatments Cai & Wang (2006)

	Pooling Equilibrium[1]			Most Informative Equilibrium			
b [2]	Credible Neologisms	Error[3]	ACD	Equilibrium[4]	Credible Neologisms	Error	ACD
0.5	$\langle 1,\{1\}\rangle,\langle 3,\{3\}\rangle,$ $\langle 7,\{7\}\rangle,\langle 9,\{9\}\rangle$	.916	.220	{1},{3},{5}, {7},{9}		−.084	0
1.2	$\langle 1,\{1\}\rangle,\langle 7,\{5,7,9\}\rangle,$ $\langle 8,\{7,9\}\rangle$	.896	.181	{1,3}, {5,7,9}	$\langle 3,\{3\}\rangle,$ $\langle 8,\{7,9\}\rangle$	.146	.074
2	$\langle 7,\{5,7,9\}\rangle$	.734	.137	{1}, {3,5,7,9}	$\langle 7,\{5,7,9\}\rangle,$ $\langle 8,\{7,9\}\rangle$	.234	.099
4	$\langle 6,\{3,5,7,9\}\rangle$	.391	.101	{1,3,5,7,9}	$\langle 6,\{3,5,7,9\}\rangle$	.391	.101

Notes: In this Sender-Receiver game payoffs are given by $U^R(a,t) = 110 - 10 \cdot |t-a|^{7/5}$ and $U^S(a,t) = 110 - 10 \cdot |t+b-a|^{7/5}$. The type set is {1,3,5,7,9}, the action set is {1,2,3,4,5,6,7,8,9}. Each type is equally likely.

[1] In the pooling equilibrium, the Receiver takes action 5 regardless of the message.

[2] The baseline treatments only differ in the size of the bias parameter b.

[3] As prediction error we take the reported difference between the actual and predicted message-type correlation. Cai and Wang also report other measures as message-action and type-action correlations, which yield a similar picture.

[4] In this column, we show the equilibrium type partition. The Receiver's action from a message coming from a partition element is the average of the types in the partition element.

In fact, a 'neologism dynamic' may explain why Cai & Wang's (2006) findings are qualitatively in line with ACDC. The neologism dynamic is a best-response dynamic with a twist. If players' strategies are not in equilibrium, the Sender simply best responds to the Receiver's strategy in the previous round. The Receiver, in turn, plays a best response to the Sender's message strategy in the

[20] Wang et al. (2010) and Kawagoe and Takizawa (2009) provide further support for level-k thinking in cheap talk games.

current round of interaction. The difference with a standard best-response dynamic is that the neologism dynamic allows Senders to send credible neologisms when the dynamic reaches an equilibrium. In Appendix C, we analyze the outcomes of the neologism dynamic for the parameters corresponding to the four treatments in Cai & Wang (2006). We do so by taking both a pooling strategy and a naïve strategy as the initial conditions. Senders who pool send out the same message regardless of their type while naïve Senders simply reveal their type.[21]

The neologism dynamic produces the following results. First, it converges to an outcome that is independent of the initial conditions. This is in contrast to a standard best-response dynamic (or a level-k analysis), where players play pooling equilibrium strategies forever if the pooling equilibrium is the initial condition while the dynamic may converge to another equilibrium if the initial condition has the Sender play a naïve strategy. Second, if $b = 0.5$, the neologism dynamic converges to the most informative equilibrium, which is neologism proof and therefore coincides with the ACDC equilibrium. Third, for higher values of b, the neologism dynamic converges to a cycle that includes the strategies of the ACDC equilibrium and none of the other equilibria. Fourth, the higher b, the more 'noisy' the cycles are. We measure noise by taking the average variance of the actions implemented for each type over the course of the cycle. Fifth, more types deviate from the equilibrium the higher b. Finally, the Sender overcommunicates in the sense that in each iteration of the cycle, senders send out at least as many messages as in the most informative equilibrium and typically more.

The outcomes of the neologism dynamic are consistent with both the ACDC predictions and the Cai & Wang (2006) results. First, the outcomes support

[21] Cai & Wang (2006) present results for a level-k analysis for $b = 4$ where they take a naïve Sender strategy as level-0. The analysis is quite similar to ours in the sense that up to level $k = 2$, the level-k prediction coincides with the k+1-th iteration in our dynamic. Cai & Wang (2006) show that the pooling equilibrium coincides with level-2 strategies for both the Sender and the Receiver. Level-k organizes their data quite well in the sense that the majority of subjects could be classified as playing level-k strategies for $k = 0,1,2$.

ACDC as a selection criterion in that it successfully predicts which equilibrium is most likely to be observed. The dynamic converges to the ACDC equilibrium if the equilibrium is neologism proof and it converges to cycles that include the ACDC equilibrium if the equilibrium is not neologism proof. In addition, the dynamic does not converge to another equilibrium. The Cai & Wang (2006) results are in line with these observations as behavior observed in their experiment is 'closest' to the ACDC equilibrium than to any other equilibrium. Second, the outcomes of the neologism dynamic are consistent with the stability of the ACDC equilibrium as measured by the equilibrium's ACD. In particular, the ACD of the ACDC equilibrium is increasing in b, which is in line with the observations that (1) more types deviate from the equilibrium if b increases, (2) there is only overcommunication for $b > 0$, and (3) the cycles become noisier for higher b. The Cai and Wang data exhibit similar patterns.

5 Other Experimental Results

The limited availability of experimental data on the CS model does not yet make it possible to judge the robustness of the predictions of ACDC. Here we discuss the experimental work on equilibrium selection in cheap talk games, in addition to that in the CS game, which we discussed in chapter 4.

5.1. Discrete games

Blume, DeJong, Kim & Sprinkle (2001) collect experimental evidence for other cheap talk games. We now turn our attention to how ACDC organizes the data in these experiments.

Blume et al. (2001) provide an experimental analysis of 4 discrete cheap talk games, in which they compare the predictive power of refinements such as neologism proofness, influentiality, and ex-ante efficiency with PCI. They find that PCI is a reliable predictor of when communication takes place and that the equilibrium refinements sometimes but not always improve on PCI. In their

Games 1 and 3, the predictions of PCI and neologism proofness (and ACDC) are very much aligned, and borne out by the data. In their Game 2 (see Table 3) neologism proofness predicts complete separation while the finest partition consistent with PCI entails partial separation. The data are in line with separation, as a clear majority of 88% of the outcomes is consistent with the separating equilibrium. One could argue that this result does not contradict PCI, because PCI allows multiple patterns including separation (see their footnote 10). As the authors note (in footnote 19), one needs to add neologism proofness to PCI to actually predict that separation happens.

TABLE 3
REPRODUCTION OF GAMES 2 AND 4 OF BLUME ET AL. (2001)

	a_1	a_2	a_3	a_4	a_5
t_1	800, 800	100, 100	0, 0	500, 500	0, 400
t_2	x, 100	y, 800	0, 0	500, 500	0, 400
t_3	0, 0	0, 0	500, 800	0, 0	0, 400

Notes: All the three types $\{t_1,t_2,t_3\}$ of the Sender are equally likely and the Receiver can implement one of the actions $\{a_1,\ldots,a_5\}$. Entry *i,j* represents $U^S(t_i,a_j),U^R(t_i,a_j)$. Games 2 and 4 are identical, except that $x = 100$, $y = 300$ in game 2, whereas $x = 300$, $y = 100$ in game 4.

In Blume et al.'s Game 4 (Table 3), no equilibrium is neologism proof while PCI selects a unique equilibrium. This game has two equilibrium outcomes. Besides the pooling equilibrium where action a_5 is induced there is a partially separating equilibrium where types t_1 and t_2 send a common message that differs from the message of t_3. Types t_1 and t_2 induce a_4 while type t_3 induces a_3. Full separation is not an equilibrium because t_2 prefers to mimic t_1. None of the equilibria satisfies neologism proofness. PCI predicts meaningful communication because the finest partition consistent with PCI is given by $\{\{t_1,t_2\},\{t_3\}\}$. The partially separating equilibrium only has a credible neologism where t_1 deviates to a_1. Thus, its ACD equals $\frac{1}{3}\frac{(800-500)}{800}=\frac{1}{8}$. The pooling equilibrium admits the neologism where t_1 and t_2 deviate to a_4 and the credible neologism where

t_3 deviates to a_3. Consequently, its ACD is $\frac{1}{3}\left(\frac{(500-0)}{800}+\frac{(500-0)}{500}+\frac{(500-0)}{500}\right)=\frac{7}{8}$. So ACDC predicts that the partially separating equilibrium will be the most observed equilibrium outcome but that it will not be completely stable.

In line with this prediction, Blume et al. find that 37% of the outcomes are consistent with the partially separating equilibrium but no outcome is consistent with the pooling equilibrium. Thus, of the two equilibria, the one with the lowest ACD performs best. Consistent with the ACD measures, much fewer outcomes are in line with the equilibrium selected by ACDC in Game 4 than in Game 2. In line with the fact that types t_1 have a credible neologism, they turn out to be the ones that are able to credibly identify themselves.[22]

Our conclusion is that our ACDC concept improves the predictions of neologism proofness and that it does at least as well as PCI in explaining the data of Blume et al. (2001). The extra mileage for ACDC with respect to PCI comes from continuous games like the CS game and the veto-threat game which we discuss in the next subsection. PCI fails to predict any communication at all in these settings, while in accordance with ACDC subjects are able to communicate meaningfully to a large extent.

5.2. ACDC in a continuous veto-threat game

In De Groot Ruiz, Offerman & Onderstal (2014), we test ACDC in fresh experiments. For this, we use games that belong to a class of veto-threat games introduced in De Groot Ruiz, Offerman & Onderstal (2012). These games are suitable to test ACDC, as they allow for a continuous manipulation of the size and frequency of credible deviations and can have a rich equilibrium set that is

[22] While across games, the ACDC equilibrium's ACD is a good predictor of the relative likelihood of outcomes being consistent with the ACDC equilibrium, it is silent about what kind of out-of-equilibrium behavior to expect. Of course, this limitation holds true for any equilibrium refinement.

difficult to refine. In Appendix B, we show that ACDC, when adapted for veto-threat games, selects a unique equilibrium in this class of games.

We briefly discuss our results for four treatments (see Table 4). Each treatment is a variation of the following game. The Sender's type is drawn from the uniform distribution over integers in the interval $[0,B]$. The Sender sends a costless message $m \in \{0,1,...,B\}$ to the Receiver, who makes a proposal $a \in \{0,1,...,B\}$. The Sender can then accept a or reject it, in which case the outcome is the disagreement point δ. Payoffs for $a \in \{0,1,...,B\}$ are $U^R(a) = 60 - \frac{2}{3}a$ and $U^S(a,t) = 60 - |t-a|$. In treatments G(120), G(130) and G(210), $U^R(\delta) = U^S(\delta,t) = 0$. These treatments only differ in the boundary parameter B.

Each of these treatments has a pooling equilibrium where the Receiver always proposes 45 and a partially separating equilibrium, where the Receiver proposes 0 or 60. The only difference is that for $B = 120$, the partially separating equilibrium is the unique neologism proof equilibrium, whereas for the other treatments neither equilibrium is neologism proof. For similar reasons as in the CS game, also neologism proofness, the recurrent mop, PCI and CMR do not select an equilibrium. The same holds true for announcement proofness because credible announcements coincide with credible neologisms in this game. Assuming the recurrent mop would converge, neither equilibrium is stable and both are recurrent.[23] CMR can only guarantee that the 0 type can send a credible message (and is silent about what other types do). The only partition that is PCI is

[23] The deviation correspondence of the pooling equilibrium (the most interesting case), for instance, contains only message strategies with three messages (say 'low', 'medium' and 'high'). In any Receiver strategy in this correspondence, the Receiver proposes 0 after 'low', 0 or 45 after 'medium' and some higher action after 'high'; furthermore, the correspondence will contain the strategy in which the Receiver proposes 45 after 'medium.' Hence, type $t = 45$ will separate and send 'medium' in any best response to a full-support strategy of the Receiver. Because the deviation correspondence only contains message strategies with three messages, it will not converge to either equilibrium. A similar reasoning holds for the separating equilibrium.

$0 = t_0 < t_1 = B$.[24] NITS selects the partially separating equilibrium for all B if one takes 0 as the lowest type.[25]

TABLE 4

FOUR TREATMENTS FROM DE GROOT RUIZ, OFFERMAN & ONDERSTAL (2013)

Treatment	$U^R(\delta)$	$U^S(\delta)$	B	Equilibrium actions[1]	ACD[2]
G(120)	0	0	120	{45}, {0, 60}**	0
G(130)	0	0	130	{45}, {0, 60}*	0.22
G(210)	0	0	210	{45}, {0, 60}*	0.50
G3Actions	0	30	120	$\{30\},\{a_1,a_1+60\}$, $\{0,a_2,a_2+60\}$** [3,4]	0

Notes: In each game, the Sender sends a message m, after which the Receiver proposes an action a. Then the Sender can accept a or reject a, in which case the outcome is the disagreement point δ. t was uniformly distributed on the integers in $[0,B]$. $U^R(x) = 60 - \frac{2}{3}x$ and $U^S(x,t) = 60 - |t-x|$. [1]An equilibrium has a * if it is ACDC and ** if it is neologism proof as well. [2]The ACD of the ACDC equilibrium. [3]$a_1 \in [0,30]$ and $a_2 \in (0,30]$. [4] Only {0,30,90} is ACDC.

ACDC selects the partially separating equilibrium in the three treatments and, in addition, predicts that the partially separating becomes less stable as B increases. In De Groot Ruiz, Offerman & Onderstal (2013), we find that the data supports the predictions of ACDC. As can be seen in Figure 2, the higher the ACD, the higher the prediction error of an equilibrium. In particular, we find that in each treatment the partially separating equilibrium performs significantly better than the pooling equilibrium. In addition, we find that for B = 130 the partially separating equilibrium performs very similar as when B = 120, supporting the notion than stability is a continuous characteristic. Finally, we find that the partially separating equilibrium performs significantly better for B = 120 or B = 130 than for B = 210. As in the CS game, SEPE selects the same

[24] The main reason is that the highest Sender-type of a partition-element always prefers the Receiver to believe that the upper boundary is higher than the true boundary (except for types $t = 0$ or $t = B$). Finally, the 'partition' 0 and $(0,B]$ is not PCI, as 0 (which is the best response if the Sender type is 0) is also a best response to some Receiver beliefs with support on the interval (0,1].

[25] All types in [0,60] are lowest types according to Chen et al.'s definition. The pooling equilibrium survives NITS relative to types in [22.5,105], whereas the separating equilibrium survives NITS relative to types in [0,30].

equilibrium as ACDC, but fails to account for the fact that the performance of the prediction depends on the ACD.

Finally, treatment G3Actions has $B = 120$, $U^R(\delta) = 0$ and $U^S(\delta, t) = 30$. The corresponding game has a continuum of size-2 and size-3 equilibria. Except for SEPE, none of the earlier refinements selects a unique equilibrium. Even influentiality (selecting the equilibrium with the maximum size) does thus not identify a unique equilibrium because there are several size-3 equilibria. Similarly, NITS is not selective as one size-2 and all size-3 equilibria survive NITS. In sum, in De Groot Ruiz, Offerman & Onderstal (2013), we find that the ACD of an equilibrium predicts how well it does in comparison to other equilibria and that the ACDC equilibrium has the lowest prediction error.[26]

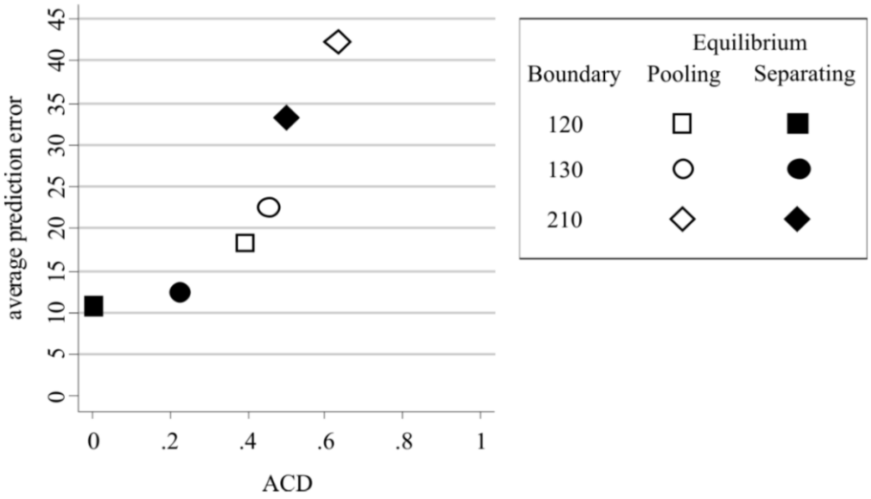

FIGURE 2

The figure plots for each equilibrium in each treatment its theoretical ACD against its empirical prediction error for the last 15 periods. Let $a^\sigma(t)$ be the equilibrium action of the Receiver given type t and $\hat{a}_i(t_i)$ the observed action for observation i. The average prediction error (for a set of n observations I) is then $\frac{1}{n}\sum_{i\in I}\left|\hat{a}_i(t_i) - a^\sigma(t_i)\right|$.

[26] In De Groot Ruiz, Offerman & Onderstal (2014), we show that a neologism dynamic explains some of the main features of the experimental data in a very similar way as in the Cai & Wang (2006) experiment.

6 Conclusion

ACDC generalizes refinements based on credible deviations, in particular neologism proofness, capturing the behaviorally relevant aspects of equilibrium stability in cheap talk games. ACDC is based on the intuition that the frequency and size of credible deviations affects equilibrium stability in a continuous rather than a binary manner. ACDC measures the (in)stability of cheap talk equilibria and determines which are most plausible. We showed that an ACDC equilibrium exists under general conditions and that it is unique in a large range of applications.

Most importantly, the predictions of ACDC organize existing experimental data well in the sense that (1) if observed behavior is in line with some equilibrium, it is mostly in line with the ACDC equilibrium, and (2) between games, behavior is more likely to be consistent with the ACDC equilibrium the lower its ACD. All in all, we believe that ACDC is an improvement of neologism proofness and that it is more successful in describing actual behavior across a large range of cheap talk games than other criteria.

A limitation that ACDC shares with other equilibrium refinements is that it does not predict how experimental subjects behave 'out of equilibrium'. In particular, it does not predict systematic overcommunication as observed in the experiments of Cai & Wang (2006) and Wang et al. (2010).

In this paper, we have defined ACDC for simple Sender Receiver cheap talk games. An interesting avenue for future research is to generalize the concept to other settings, e.g., settings with multiple Senders or multiple Receivers, or with noisy information channels.

References

Agranov, M., & Schotter, A. (2012). Ignorance is Bliss: an Experimental Study of the Use of Ambiguity and Vagueness in the Coordination Games with Asymmetric Payoffs. *AEJ: Microeconomics, 4*, 77-103.

Blume, A., DeJong, D. V., Kim, Y.-G., & Sprinkle, G. B. (1998). Experimental Evidence on the Evolution of Meaning of Messages in Sender-Receiver Games. *The American Economic Review, 88*, 1323-1340.

Blume, A., DeJong, D. V., Kim, Y.-G., & Sprinkle, G. B. (2001). Evolution of Communication with Partial Common Interest. *Games and Economic Behavior, 37*, 79-120.

Blume, A., Kim, Y.-G., & Sobel, J. (1993). Evolutionary Stability in Games of Communication. *Games and Economic Behavior, 5*, 547-575.

Cai, H., & Wang, J. T. (2006). Overcommunication in Strategic Information Transmission Games. *Games and Economic Behavior, 56*, 7-36.

Chen, Y., Kartik, N., & Sobel, J. (2008). Selecting Cheap Talk Equilibria. *Econometrica, 76*, 117-136.

Coles, P., Kushnir, A., & Niederle, M. (2013). Preference Signaling in Matching Markets. *AEJ: Microeconomics, 5*, 99-134.

Crawford, V. P. (2003). Lying for strategic advantage: Rational and boundedly rational misrepresentation of intentions. *American Economic Review*, 133-149.

Crawford, V. P., Costa-Gomes, M. A., & Iriberri, N. (2013). Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature, 51*(1), 5-62.

Crawford, V., & Sobel, J. (1982). Strategic Information Transmission. *Econometrica, 50*, 1431-1451.

De Groot Ruiz, A. W., Offerman, T., & Onderstal, S. (2012). Power and the Privilege of Clarity: An Analysis of Bargaining Power and Information Transmission. *Working paper, University of Amsterdam.*

De Groot Ruiz, A. W., Offerman, T., & Onderstal, S. (2014). For Those about to Talk We Salute You: An Experimental Study of Credible Deviations and ACDC. *Experimental Economics, 17*, 173-199.

Dickhaut, J. W., McCabe, K. A., & Mukherji, A. (1995). An Experimental Study of Strategic Information Transmission. *Economic Theory, 6*, 389-403.

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

Farrell, J., & Gibbons, R. (1989). Cheap Talk Can Matter in Bargaining. *Journal of Economic Theory, 48*, 221-237.

Gilligan, T. W., & Krehbiel, K. (1990). Organization of Informative Committees by a Rational Legislature. *American Journal of Political Science, 34*, 531-564.

Kartik, N., Ottaviani, M., & Squintani, F. (2007). Credulity, Lies and Costly Talk. *Journal of Economic Theory, 134*, 93-116.

Kawagoe, T., & Takizawa, H. (2009). Equilibrium Refinement vs. Level-k Analysis: An Experimental Study of Cheap-Talk Games with Private Information. *Games and Economic Behavior, 66*, 238-255.

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

Lewis, D. K. (1969). *Convention: A Philosophical Study.* Cambridge: Harvard University Press.

Matthews, S. A. (1989). Veto Threats: Rhetoric in a Bargaining Game. *Quarterly Journal of Economics, 104*, 347-369.

Matthews, S. A., & Postlewaite, A. (1989). Pre-Play Communication in Two-Person Sealed-Bid Double Auctions. *Journal of Economic Theory, 48*, 238-263.

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

McKelvey, R. D., & Palfrey, T. R. (1998). Quantal Response Equilibrium for Extensive Form Games. *Experimental Economics, 1*, 9-41.

Morgan, J., & Stocken, P. C. (2003). An Analysis of Stock Recommendations. *The RAND Journal of Economics, 34*, 183-203.

Morris, S. (2001). Political Correctness. *Journal of Political Economy, 109*, 231-265.

Park, I.-U. (1997). Generic Finiteness of Equilibrium Outcome Distributions for Sender-Receiver Cheap-Talk Games. *Journal of Economic Theory, 76*, 431-448.

Pearce, D. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica, 52*, 1029-1050.

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

Rabin, M., & Sobel, J. (1996). Deviations, Dynamics and Equilibrium Refinements. *Journal of Economic Theory, 68*, 1-25.

Schelling, T. C. (1960/1980). *The Strategy of Conflict.* Cambridge: Harvard University Press.

Sobel, J. (2013). Giving and Receiving Advice. In D. Acemoglu, M. Arellano, & E. Dekel, *Advances in Economics and Econometrics.* Cambridge: Cambridge University Press.

Wang, J. T., Spezio, M., & Camerer, C. F. (2010). Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games. *American Economic Review, 100*, 984-1007.

Appendix A

Proof of Proposition 3

The proof proceeds as follows. First, we obtain closed-form solutions for the ACD for all b. Second, we calculate the ACD for the specified values of b.

The ACD of equilibrium σ in the CS game is equal to

$$\begin{aligned}
ACD(\sigma) &= E_t\left[\frac{U^S(\tilde{a}^\sigma(t),t) - U^S(a^\sigma(t),t)}{\bar{U}^S(t) - \underline{U}^S(t)}\right] \\
&= \int_0^1 \frac{U^S(\tilde{a}^\sigma(t),t) - U^S(a^\sigma(t),t)}{U^S(\min\{t+b,1\},t) - \min\{U^S(0,t),U^S(1,t)\}}\,dt \\
&= \int_0^1 \frac{\left(a^\sigma - (t+b)\right)^2 - \left(\tilde{a}^\sigma - (t+b)\right)^2}{-\max\{0,t+b-1\}^2 + \max\{(t+b)^2,(t+b-1)^2\}}\,dt.
\end{aligned}$$

Note that $(t+b-1)^2 > (t+b)^2$ if and only if $t < \frac{1}{2} - b$. Suppose $a^\sigma(t)$ and $\tilde{a}^\sigma(t)$ are constant and $\bar{U}^S(t) = 0$ on the interval $[\underline{t},\bar{t}]$. Let $\hat{t} \equiv \max\{\underline{t},\min\{\bar{t},\frac{1}{2} - b\}\}$. Then, $\int_{\underline{t}}^{\bar{t}} CD(t,\sigma)dt$ is equal to

$$h(b,a^\sigma,\alpha^\sigma,\underline{t},\bar{t}) \equiv \int_{\underline{t}}^{\bar{t}} \frac{\left(a^\sigma - (t+b)\right)^2 - \left(\tilde{a}^\sigma - (t+b)\right)^2}{\max\{(t+b)^2,(t+b-1)^2\}}\,dt$$

$$= \int_{\underline{t}}^{\hat{t}} \frac{\left(a^\sigma - (t+b)\right)^2 - \left(\tilde{a}^\sigma - (t+b)\right)^2}{(t+b-1)^2}\,dt + \int_{\hat{t}}^{\bar{t}} \frac{\left(a^\sigma - (t+b)\right)^2 - \left(\tilde{a}^\sigma - (t+b)\right)^2}{(t+b)^2}\,dt$$

$$= (a^\sigma - \tilde{a}^\sigma)\left(\frac{(a^\sigma + \tilde{a}^\sigma - 2)(\hat{t} - \underline{t})}{(b-1+\hat{t})(b-1+\underline{t})} + 2\log\left[\frac{b-1+\underline{t}}{b-1+\hat{t}}\right]\right)$$

$$+(a^{\sigma}-\tilde{a}^{\sigma})\left(\frac{(a^{\sigma}+\tilde{a}^{\sigma})(\overline{t}-\hat{t})}{(b+\overline{t})(b+\hat{t})}+2\log\left[\frac{b+\hat{t}}{b+\overline{t}}\right]\right).$$

As noted in the main text, an equilibrium of size n can have a neologism in the beginning, $\tilde{a}_0^n$, a neologism at the end $\tilde{a}_n^n$ and at most $n-1$ neologisms in the middle, $\tilde{a}_i^n, i=1,...,n-1$. The size-1 equilibrium has a neologism at the beginning and at the end. The maximum size $n(b)$ equilibrium has a neologism at the end and neologisms in the middle $\tilde{a}_i^n, i=\underline{i}(b),...,n-1$, where $\underline{i}(b)=1$ if $2bn(b)^2<1$ and $\underline{i}(b)=2$ if $2bn(b)^2\geq 1$. Size-n equilibria with $1<n<n(b)$ admit all neologisms specified above. Observe that $\tau_{n-1}^n<1-b$, such that $\overline{U}^S(t)=0$ except for the highest types of the highest neologism, such that $h(b,a^{\sigma},\tilde{a}^{\sigma},\underline{t},\overline{t})$ can be used to calculate the contribution to the ACD for neologisms $\underline{i}(b)=1,...,n-1$. For the highest neologism, the contribution to the ACD is equal to

$$\begin{aligned}\overline{h}(b,n) &= h(b,a_n^n,\tilde{a}_n^n,\underline{\tau}_n^n,b-1)+\int_{b-1}^{1}\frac{\left(a_n^n-(t+b)\right)^2-\left(\tilde{a}_n^n-(t+b)\right)^2}{-(t+b-1)^2+(t+b)^2}dt\\ &= h(b,a_n^n,\tilde{a}_n^n,\underline{\tau}_n^n,b-1)+\frac{1}{2}(a_n^n-\tilde{a}_n^n)(a_n^n+\tilde{a}_n^n-1)\log[2b+1].\end{aligned}$$

Let σ_b^n be the size-n equilibrium of the game with bias parameter b. Then, the ACD of the pooling equilibrium is

$$ACD(\sigma_b^1)=h(b,a_1^1,\tilde{a}_0^1,0,\overline{\tau}_0^1)+\overline{h}(b,1).$$

The ACD of the maximum-size equilibrium is

$$ACD\left(\sigma_b^{n(b)}\right) = \sum_{i=\underline{i}(b)}^{i=n(b)-1} [h(b, a_i^{n(b)}, \tilde{a}_i^{n(b)}, \underline{\tau}_i^{n(b)}, t_i^{n(b)}) + h(b, a_{i+1}^{n(b)}, \tilde{a}_i^{n(b)}, t_i^{n(b)}, \overline{\tau}_i^{n(b)})] + \overline{h}(b, n(b)).$$

The ACD of a size-n equilibrium with $1 < n < n(b)$ is equal to

$$ACD(\sigma_b^n) = h(b, a_1^n, \tilde{a}_0^n, 0, \overline{\tau}_0^n) + \sum_{i=1}^{i=n-1} [h(b, a_i^n, \tilde{a}_i^n, \underline{\tau}_i^n, t_i^n) + h(b, a_{i+1}^n, \tilde{a}_i^n, t_i^n, \overline{\tau}_i^n)] + \overline{h}(b, n)$$

For each $b \in \left\{\frac{1}{10000}, \frac{2}{10000}, \ldots, \frac{1}{4}\right\}$, one can calculate the (closed-form) value of $ACD(\sigma_b^n)$ for all $1 \le n \le n(b)$, and verify that the ACD of the size-n equilibrium in the CS game is decreasing in n.

Appendix B: ACDC in a Veto-Threat Games

B.1. Equilibria and ACDC in veto-threat games

Consider the following game. Nature draws the Sender type t from distribution f on T, where T is a compact metric space. The Sender then privately observes her type t and chooses a message $m \in M$. After having observed the Sender's message, the Receiver chooses an action $a \in A$, where A is a compact metric space. After seeing the action, the Sender chooses between accepting ($v = 1$) or rejecting ($v = 0$) the action. If she rejects, the outcome is the disagreement point δ. If $\delta \in A$, the game has an internal veto threat and otherwise it has an external veto threat. The outcome set is $X = A \cup \{\delta\}$. Let $U^R : X \times T \to \mathbb{R}$ be the utility function of the Receiver $U^S : X \times T \to \mathbb{R}$ that of the Sender. We assume both are bounded from above and below.

A strategy for the Sender consists of a message strategy $\mu : T \to M$ and an acceptance strategy $\nu : A \times T \to \{0,1\}$.The strategy of the Receiver is an action strategy $\alpha : M \to A$. Let Σ^S be the set of Sender strategies and Σ^R the set of Receiver strategies. Let $\{\mu, \alpha, \nu\}$ be a strategy profile and Σ the set of all strategy profiles. Define $V^R(x,t;\nu) = U^R(x,t) \cdot \nu(x,t) + U^R(\delta,t) \cdot (1 - \nu(x,t))$ and $V^S(x,t;\nu) = U^S(x,t) \cdot \nu(x,t) + U^S(\delta,t) \cdot (1 - \nu(x,t))$. Finally, let the Receiver have prior beliefs $\beta^0(t) = f(t)$. A pure strategy perfect Bayesian equilibrium (equilibrium henceforth) $\sigma = \{\mu, \alpha, \beta\}$ is characterized by the following four conditions:

$$\text{For each } t \in T,\ m(t) \in \arg\max V^S(\alpha(m), t; \nu)$$

$$\text{(6)} \qquad \text{For each } m \in M, \alpha(m) \in \arg\max_{a \in A} \int_T V^R(a, t; \nu)\beta(t \mid m)dt$$

$$\nu(a,t) = 1 \text{ if } U^S(a,t) > U^S(\delta,t) \text{ and } \nu(a,t) = 0 \text{ if } U^S(a,t) < U^S(\delta,t)$$

where $\beta(t \mid m)$ denotes the Receiver's posterior beliefs, which is derived from μ and β^0 using Bayes' rule wherever possible.

Let Σ^* be the set of equilibria and $\Sigma^\dagger$ be the set of cautiously rationalizable strategy profiles. Define $\underline{V}^S(t) \equiv \inf_{\{\alpha,\mu,\nu\} \in \Sigma^\dagger} V^S\left(t, \alpha\left(\mu(t)\right); \nu\right)$ and $\bar{V}^S(t) \equiv \sup_{\{\alpha,\mu,\nu\} \in \Sigma^\dagger} V^S\left(t, \alpha\left(\mu(t)\right); \nu\right)$. Then

$$\text{(7)} \qquad CD(t,\sigma) \equiv \frac{V^S\left(t, \alpha^\gamma(\mu^\gamma(t)); \nu^\gamma\right) - V^S\left(t, \alpha(\mu(t)); \nu^\gamma\right)}{\bar{V}^S(t) - \underline{V}^S(t)}$$

if $V^S\left(t, \alpha(\mu(t)); \nu^\gamma\right) > \underline{V}^S(t)$. If $V^S\left(t, \alpha(\mu(t)); \nu^\gamma\right) = \underline{V}^S(t)$, then $CD(t,\sigma) \equiv 1$, ACDC can be now be defined analogously to the case without a veto by the Sender.

B.2. ACDC in a veto-threat game

Here we show that ACDC selects a unique equilibrium in the class of veto-threat games introduced by De Groot Ruiz, Offerman & Onderstal (2012). The games studied experimentally in De Groot Ruiz, Offerman & Onderstal (2014) belong to this class of games. We assume the Sender's type t is uniformly distributed on the interval [0,1]. We model the player's bargaining power as the payoff of the disagreement point $U^R(\delta)$ and $U^S(\delta)$, where we assume $U^S(\delta,t) = U^S(\delta)$ does not depend on t. U^R and U^S satisfy the following assumptions:

(8) U^R on $\mathbb{R}$ is twice continuously differentiable, unimodal with a peak at 0 and concave.

(9) $U^S(x,t)$ can be written as a function $f(t-x)$, for all x in $\mathbb{R}$, t in [0,1], where f is continuously differentiable, symmetric, concave, strictly increasing in $\mathbb{R}_-$ and for all $y \in \mathbb{R}$ there is a $z > 0$ such that $f(z) < y$ and $f(-z) < y$; Finally, $U^S(\delta) < f(0)$.[27]

In De Groot Ruiz, Offerman & Onderstal (2012), we show that only partition equilibria exist. Here we show that there is a unique ACDC equilibrium:

Proposition 5 *Under assumptions* (8) *and* (9), *the unique ACDC equilibrium is the maximum size equilibrium with the highest equilibrium action.*

For the proof of Proposition 5, we introduce some definitions and results from De Groot Ruiz, Offerman & Onderstal (2012) and derive two helpful lemmas.

Observe that in this game, a neologism $\langle \tilde{a}, N \rangle$ is credible relative to equilibrium σ^* if and only if

$$\tilde{a} \in \arg\max_{a\in\mathbb{R}} P\left\{U^S(a,t) \geq 0 \middle| t \in N\right\}\left(U^R(a) - U^R(\delta)\right), \text{ and}$$

$$\text{for all } k = 1,\ldots,n \text{ it holds that } t \in [t_{k-1}, t_k] \cap N \Rightarrow U^S(\tilde{a},t) > U^S(a_k,t) \text{ and}$$

$$t \in [t_{k-1}, t_k] \setminus N \Rightarrow U^S(\tilde{a},t) \leq U^S(a_k,t).$$

Lemma 1 If $\langle \tilde{a}, N \rangle$ *is a credible neologism relative to equilibrium* σ^*, *then* N *is an interval.*

[27] Observe that (9) implies assumptions (A2)-(A5) in De Groot Ruiz, Offerman & Onderstal (2012). Our assumptions here are stricter. In particular, they require a uniform type distribution and a symmetric and concave payoff function for the Sender.

Proof. The proof is by contradiction. Suppose $0 \leq t^1 < t^2 < t^3 \leq 1$, $t^1, t^3 \in N$ and $t^2 \notin N$. Suppose further that in equilibrium, type t^i obtains action a^i, $i = 1,2,3$. The fact that the a type's utility is strictly decreasing in the distance between $t - a$ implies $a^1 \leq a^2 \leq a^3$. If $\tilde{a} \leq t^2$ then it must be the case that $\tilde{a} \leq a^2$ (otherwise type t^2 would prefer $\tilde{a}$ over a^2). As a consequence, $\tilde{a} \leq t^3 \leq a^3 = a^2$ because type t^3 must prefer $\tilde{a}$ over a^3 and a^3 over a^2. A contradiction is established, because the fact that the indifference points $t - d$ and $t + d$ are strictly increasing in t implies that type t^2 strictly prefers $\tilde{a}$ over a^2. This is in conflict with the definition of a credible neologism. Analogously, $\tilde{a} > t^2$ can be ruled out, so that N is an interval. *Q.E.D.*

From (9), it follows that there is a $d > 0$ such that for all t and $a \in \mathbb{R}$, $U^S(a,t) \geq U^S(\delta)$ if and only if $a \in [t - d, t + d]$. Hence, $t - d$ and $t + d$ are the Sender's indifference points as to whether she accepts action a. From Lemmas 2 and 3 in De Groot Ruiz, Offerman & Onderstal (2012), it follows that in equilibrium

$$a_1 \geq 0, \quad t_{k-1} - d < a_k \leq t_k - d \;\; \textit{for all} \;\; k = 2,\ldots,n \;\; \textit{and} \;\; t_{k-1} + d \leq a_k \;\; \textit{for} \;\; k = 3,\ldots,n. \tag{10}$$

We can now show that under (9), it holds that

Lemma 2 *In equilibrium,* $a_k + d = t_k = a_{k+1} - d$ *for* $k = 2,\ldots,n-1$.

Proof. Due to the t being uniformly distributed and (9), the indifference points $t - d$ and $t + d$ are uniformly distributed as well. This means that if the Receiver receives a message that identifies Sender types to be in the interval $[t_k, t_{k+1}]$ $(k = 0,\ldots,n-1)$, the probability the Sender accepts an action is not higher for an action $a > t_k + d$ than for action $a' = t_k + d$, while

$U^R(a) < U^R(a')$. Hence, for the equilibrium action a_k it holds true that $a_k \leq t_{k-1} + d$ and by (10), this means $a_k = t_{k-1} + d \leq t_k - d$ for $k = 3,...,n$. Now, suppose that $t_{k-1} + d < t_k - d$ for some $k = 3,...,n$. This means that $a_k < t_k - d$ and hence $U^S(a_k, t_k) < 0$. Since $U^S(a_k, t_k) = U^S(a_{k+1}, t_k)$, this implies, however, that $a_{k+1} > t_k + d$, which for $k = 3,...,n-1$ is a contradiction with $a_k \leq t_{k-1} + d$ for $k = 3,...,n$. Hence, $a_k = t_{k-1} + d = t_k - d$ for $k = 3,...,n-1$. Consequently, $a_k + d = t_k = a_{k+1} - d$ for $k = 3,...,n-1$.

Furthermore, from the discussion above we have that $t_2 = a_3 - d$ and that $a_2 \leq t_1 + d$. In addition, from (10) it follows that $a_2 \leq t_2 - d$. Hence, a necessary condition on a_2 is that $a_2 \in \arg\max_{t_1+d \leq a \leq t_2 - d} \left(U^R(a) - U^R(\delta)\right)\left(a + d - t_1\right)$. Analogously to the discussion in the proof of Proposition 2 in De Groot Ruiz, Offerman & Onderstal (2012), one can show that this implies that a_2 must be equal to $t_2 - d$. As a result, $a_2 + d = t_2 = a_3 - d$. *Q.E.D.*

Proof of Proposition 5 Suppose that the game has more than one equilibrium outcome. If $\overline{x} \leq 2d$, then consider the equilibrium outcome σ^* with $a_1 = 0$ and a_2 such that $a_2 \in \arg\max_{a \in \mathbb{R}} U^R(a)\left(\min\{a + d, 1\} - \frac{1}{2}a_2\right)$. If $\overline{x} > 2d$, let n be the natural number for which $\overline{x} - 2dn \leq 0$ and $\overline{x} - 2d\left(n-1\right) > 0$, and consider the following $\sigma^* : a_1 = 0$; $a_1 = 0, a_k = \overline{x} - 2d(n-k-2), k = 2,...,n$. We now show that σ^* has the maximum equilibrium size and is the unique ACDC equilibrium outcome.

From Lemma 4 in De Groot Ruiz, Offerman & Onderstal (2012) and (9), it follows that there exists an $\overline{x} \in \mathbb{R}$ such that

(11) $U^R(x) - U^R(\delta) + 2dU^{R\prime}(x) \geq 0$ for all $x \in [0, \overline{x})$ and

$$U^R(x) - U^R(\delta) + 2dU^{R\prime}(x) < 0 \text{ for all } x \in (\overline{x}, 1-d].$$

where a prime ($'$) denotes a derivative with respect to x. Let $\overset{*}{a}$ denote the highest equilibrium action a_n in σ^*. Using (11), it can be verified that σ^* constitutes the highest size equilibrium, analogously to the proof of Proposition 3 in De Groot Ruiz, Offerman & Onderstal (2012). Similarly, it can be verified that the highest action a^{**} in any other equilibrium σ^{**} must be smaller than a^*:

$$a^{**} \leq a^* \leq 1-d.$$

If $a^{**} < 1-d$, σ^{**} has at least one credible neologism: Types in the interval $(\underline{\tau}^{**}, 1]$ are willing to send a credible neologism $\left\langle \tilde{a}^{**}, (\underline{\tau}^{**}, 1] \right\rangle$, where

$$\underline{\tau}^{**} = \frac{1}{2}\left(a^{**} + \tilde{a}^{**}\right), \text{ and}$$

$$\tilde{a}^{**} \in \operatorname*{arg\,max}_{a \in (a^{**}, 1-d]} \left(U^R(a) - U^R(\delta)\right) \frac{a + d - \underline{\tau}^{**}}{1 - \underline{\tau}^{**}}.$$

To prove that $\overset{*}{\sigma}$ is an ACDC equilibrium, we first show it has at most one credible neologism (claim 1) and this credible neologism, if it exists, maximizes $\underline{\tau}^{**}$ and minimizes $\tilde{a}^{**} - a^{**}$ (claim 2).

In order to prove claim 1, suppose that σ^* has another credible neologism. By Lemma 1, the set of types that send the credible neologism relative to equilibrium σ^* is an interval. We can exclude neologisms that induce the Receiver to propose $a = 0$, because $a_1 = 0$ is already an equilibrium action. Hence, the neologism $\tilde{a}$ (with supremum neologism type $\tilde{\tau}$) is in between two equilibrium actions a_{k-1} and a_k. Due to Lemma 1, $a_{k-1} < \tilde{a} < \tilde{\tau} < a_k$. This implies that $U^S(\tilde{a}, \tilde{\tau}) \leq 0$, because if $U^S(\tilde{a}, \tilde{\tau}) > 0$, action $\tilde{\tau} - d$ would be better for the

Receiver than $\tilde{a}$ after receiving the neologism. Consequently, $U^S(a_{k-1},\tilde{\tau}) < U^S(\tilde{a},\tilde{\tau}) \leq 0$ and $U^S(a_k,\tilde{\tau}) < U^S(\tilde{a},\tilde{\tau}) \leq 0$. This means that an $\varepsilon > 0$ exists such that a types in $(\tilde{\tau}-\varepsilon,\tilde{\tau}+\varepsilon)$ receive 0 payoff in equilibrium. Since this is not the case in σ^*, σ^* has no other neologisms.

The proof of claim 2 proceeds as follows. Note that $\tilde{a}^{**} = \min\{\overline{a}^{**}, 1-d\}$, where $\overline{a}^{**} = \arg\max_{a\in\mathbb{R}} \left(U^R(a) - U^R(\delta)\right)\dfrac{a+d-\underline{\tau}^{**}}{1-\underline{\tau}^{**}}$. We know $\overline{a}^{**} > a^{**}$, because the solution to $\arg\max_{a\in\mathbb{R}} \left(U^R(a) - U^R(\delta)\right)\dfrac{a+d-\underline{t}}{1-\underline{t}}$ is increasing in $\underline{t}$ and a^{**} is the solution for $\underline{t} = t_{n-1}$, and $\overline{a}^{**}$ is the solution to the problem with $\underline{t} \geq a^{**} > t_{n-1}$. Moreover,

$$U^R(\overline{a}^{**}) - U^R(\delta) + U^{R\prime}(\overline{a}^{**})\left(\overline{a}^{**} + d - \underline{\tau}^{**}\right) = U^R(\overline{a}^{**}) - U^R(\delta) + U^{R\prime}(\overline{a}^{**})\left(\frac{\overline{a}^{**} - a^{**}}{2} + d\right) = 0$$

implies that

$$\overline{a}^{**} - a^{**} = -2\frac{U^R(\overline{a}^{**})}{U^{R\prime}(\overline{a}^{**})} - 2d.$$

From the concavity of U^R it follows that $\dfrac{U^R(a)}{U^{R\prime}(a)}$ is increasing in a. Hence, $\overline{a}^{**} - a^{**}$ is decreasing in a^{**}. In particular, this implies that $\tilde{a}^{**} - a^{**}$ is decreasing in a^{**}. Moreover, $\underline{\tau}^{**}$ is increasing in a^{**}.

Finally, to show that σ^* is an ACDC equilibrium, we show that it has the lowest ACD. By Lemma 2, for equilibrium σ^{**} it must then hold that $a_1^{**} > 0$ or $a^{**} < a^{*}$. If $a_1^{**} > 0$, then a neologism $\left\langle \tilde{a}_0, [0,\overline{\tau}_0]\right\rangle$ exists with $\tilde{a}_0 < a_1^{**}$.[28] Suppose now that $a^{**} < a^{*}$. If σ^* does not admit a credible neologism, it is

[28] If $a_1^{**} \geq 2d$, $\tilde{a}_0 = d$ and $U^S(d,\tilde{\tau}_0) = U^S(a_1^{**},\tilde{\tau}_0)$. If $a_1^{**} \leq d$, $\tilde{a}_0 = 0$ and $U^S(\tilde{a}_0,\tilde{\tau}_0) = U^S(0,\tilde{\tau}_0)$. If $d < a_1^{**} < 2d$, $\tilde{a}_0 = \tilde{\tau}_0 + d$ and $U^S(\tilde{a}_0,\tilde{\tau}_0) = U^S(a_1^{**},\tilde{\tau}_0)$. This has a solution, because $U^S(\tilde{\tau}_0 - d,\tilde{\tau}_0) - U^S(a_1^{**},\tilde{\tau}_0) > 0$ for $\tilde{\tau}_0 = 0$ and $U^S(\tilde{\tau}_0 - d,\tilde{\tau}_0) - U^S(a_1^{**},\tilde{\tau}_0) < 0$ for $\tilde{\tau}_0 = a_1^{**}$.

evident that $ACD(\sigma^{*}) = 0 < ACD(\sigma^{**})$. Hence, suppose that σ^{*} admits the credible neologism $\left\langle \tilde{a}^{*}, [\underline{\tau}^{*}, 1] \right\rangle$.

We can now compare the ACD of σ^{*} and σ^{**}. First, $CD^{\sigma^{*}}(t) = 0$ for $t \in [0, \underline{\tau}^{*})$. Second, we show that $U^{S}(\tilde{a}^{**}, t) - U^{S}(a^{**}, t) > U^{S}(\tilde{a}^{*}, t) - U^{S}(a^{*}, t)$ for $t \in [\underline{\tau}^{*}, a^{**} + d)$. Due to claim 2 $\tilde{a}^{**} - a^{**} > \tilde{a}^{*} - a^{*}$ and $\underline{\tau}^{**} < \underline{\tau}^{*}$. If $t \leq \tilde{a}^{**} < \tilde{a}^{*}$, then $U^{S}(a^{**}, t) < U^{S}(a^{*}, t)$ and $U^{S}(\tilde{a}^{**}, t) > U^{S}(\tilde{a}^{*}, t)$, so that the result is immediate. Assume now that $\tilde{a}^{**} < t$. By (9), $U^{S}(a, t)$ is concave in *a,* such that for $x < y \leq t$ and $b, c > 0$ it holds that:

$$U^{S}(y, t) - U^{S}(x, t) \leq U^{S}(y - b, t) - U^{S}(x - b, t) < U^{S}(y - b, t) - U^{S}(x - b - c, t).$$

Hence, for $t \in [\underline{\tau}^{*}, \tilde{a}^{*}]$ we have that $U^{S}(\tilde{a}^{*}, t) - U^{S}(a^{*}, t) \leq U^{S}(t, t) - U^{S}(a^{*}, t) \leq U^{S}(\tilde{a}^{**}, t) - U^{S}(a^{*} - t + \tilde{a}^{**}, t) < U^{S}(\tilde{a}^{**}, t) - U^{S}(a^{**}, t)$. (Observe that $t - a^{*} < \tilde{a}^{*} - a^{*} < \tilde{a}^{**} - a^{**}$.) Similarly, for $t \in (\tilde{a}^{*}, a^{**} + d]$, $U^{S}(\tilde{a}^{*}, t) - U^{S}(a^{*}, t) \leq U^{S}(\tilde{a}^{**}, t) - U^{S}(a^{*} - \tilde{a}^{*} + \tilde{a}^{**}, t) < U^{S}(\tilde{a}^{**}, t) - U^{S}(a^{**}, t)$. As a consequence, $CD^{\sigma^{**}}(t) > CD^{\sigma^{*}}(t)$ for $t \in [\underline{\tau}^{**}, a^{**} + d)$. Finally, $CD^{\sigma^{**}}(t) = 1 \geq CD^{\sigma^{*}}(t)$ for $t \in [a^{**} + d, 1]$. Together, this implies that $ACD(\sigma^{**}) = E_{t}\left[CD^{\sigma^{**}}(t)\right] > E_{t}\left[CD^{\sigma^{*}}(t)\right] = ACD(\sigma^{*})$.

In sum, if σ^{**} is different from σ^{*}, then either $a_{0}^{**} > 0$ or $a^{**} < a^{*}$ and in both cases $ACD(\sigma^{**}) > ACD(\sigma^{*})$. Therefore, σ^{*} is the unique ACDC equilibrium. *Q.E.D.*

Appendix C: Neologism Dynamic

C.1. Description dynamic

In this section, we define the neologism dynamic. We assume that the type space $T \subset \mathbb{R}$ contains a finite number of types and that the size of the message space $M \subset \mathbb{R}$ is at least as large as the size of the type space.

We start by describing the simple best response dynamic in pure strategies (without neologisms) on which the neologism dynamic is based. In each round $r = 0,1,2,...$ of interaction, the Sender and the Receiver choose a strategy. The strategy of the Sender in round r is then given by $m_r : T \to M$ and that of the Receiver by $a_r : M \to A$, where $m_r(t)$ denotes the message Sender type t sends and $a_r(m)$ the Receiver's action after receiving message m. In each round $r = 1,2,...$, the Sender best responds to the strategy of the Receiver in the previous round and the Receiver best responds to the strategy of the Sender in the same round. For the Sender, this means that $m_r(t) \in \arg\max_{m\in M} U^S(a_{r-1}(m),t)$. If the strategies in round r constitute an equilibrium, the Sender sends the same message as in round $r-1$. We make this assumption to ensure the existence of meaningful neologisms in the neologism dynamic. If the strategies in round $r-1$ do not constitute an equilibrium and $\arg\max_{m\in M} U^S(a_{r-1}(m),t)$ contains more than one message, the Sender sends the lowest optimal message.[29] For the Receiver, best responding implies that $a_r(m) \in \arg\max_{a\in A} E_t[U^R(a,t) \mid \beta^r(m)]$, where the Receiver's belief $\beta^r(m)$ regarding the Sender's type distribution given message m is derived using Bayes' rule whenever possible. If the Sender does not use m in round r, the Receiver plays the lowest action induced by any message $m^* > m$ used in round r. If such an

[29] We use tie-breaking rules to keep the analysis simple.

action does not exist, the Receiver picks the highest action induced by any used message $m^* < m$ in round r.

The neologism dynamic differs from the best response dynamic in one crucial aspect: if the dynamic has reached an equilibrium in the previous round (i.e. if $m_{r-1}(t) \in \arg\max_{m \in M} U^S(a_{r-1}(m), t)$ for all t), then the Sender can send a credible neologism in round r. When sending a neologism, the Sender will send out of the set of messages not used in round $r-1$ that message that is closest to the action she intends to induce. If two of such messages exist, she will pick the higher of the two. If the Sender can send multiple credible neologisms, she will select a credible neologism that maximizes her utility given the Receiver's best response to it.

We restrict our analysis of the games studied in the Cai & Wang (2006) experiment to two natural initial Sender strategies: a pooling strategy (where all Senders start with the same message) and a naïve strategy (where all Senders send a message equal to their own true type). Below, we provide for the experimental parameters b=0.5, b=1.2, b=2 and b=4, (*i*) the attractors (steady states or cycles) to which the initial conditions converge, (*ii*) the first rounds of the dynamic until they converge to an attractor and (*iii*) the average variance of actions a Sender type induces in the attractor.

C.2. $b = 0.5$

As the separating equilibrium admits no credible neologisms, it is a steady state of the neologism dynamic. (The average variance of Receiver actions Sender type induces is zero in a steady state.) Note that the profile where the Sender plays a naïve strategy is also the separating equilibrium:

$b = 0.5$, naïve initial conditions

Round	**Strategy Sender**					**Strategy Receiver**					
	$m_r(1)$	$m_r(3)$	$m_r(5)$	$m_r(7)$	$m_r(9)$	$a_r(1)$	$a_r(3)$	$a_r(5)$	$a_r(7)$	$a_r(9)$	Observation
r=0	1	3	5	7	9	1	3	5	7	9	Naïve / separating equilibrium

Hence, the dynamic converges to the separating equilibrium from naïve initial conditions in a trivial way.

As can be seen below, the dynamic converges to the separating equilibrium from initial pooling conditions. In round 1, Sender types 1, 3, 7, and 9 can send a credible neologism inducing an action equal to their type.

b = 0.5, pooling initial conditions

Round	**Strategy Sender**					**Strategy Receiver**					
	$m_r(1)$	$m_r(3)$	$m_r(5)$	$m_r(7)$	$m_r(9)$	$a_r(1)$	$a_r(3)$	$a_r(5)$	$a_r(7)$	$a_r(9)$	Observation
r=0	5	5	5	5	5	5	5	5	5	5	Pooling equilibrium
r=1	1	3	5	7	9	1	3	5	7	9	Naïve / separating equilibrium

C.3. b = 1.2

For b=1.2, the dynamic has as an attractor the following 4-cycle that includes the most informative equilibrium:

b = 1.2, attractor

Round	**Strategy Sender**					**Strategy Receiver**					
	$m_r(1)$	$m_r(3)$	$m_r(5)$	$m_r(7)$	$m_r(9)$	$a_r(1)$	$a_r(3)$	$a_r(5)$	$a_r(7)$	$a_r(9)$	Observation
r	3	3	7	7	7	2	2	7	7	7	Most informative equilibrium
r+1	3	5	7	9	9	1	1	3	5	8	
r+2	5	7	7	9	9	1	1	1	4	8	
r+3	1	7	9	9	9	1	3	3	3	7	
r+4	3	3	9	9	9	2	2	7	7	7	Most informative equilibrium

To illustrate the dynamic, we will briefly explain this attractor. In round r, the Sender starts playing according to the most informative equilibrium and the Receiver responds with an equilibrium strategy. In round r+1, Sender type 1 and 5 best respond to the Receiver's strategy in r by sending m=3 and m=7 respectively. In contrast, types 3, 7, and 9 can send credible neologisms inducing a=3 for type 3 and a=8 for types 7 and 9. To send this neologism they use the highest unused message closest to the action they want to induce, which are m=5 and m=9 respectively. The Receiver knows that t=1 sends m=3, so he best responds to m=3 by playing a=1. The other actions specified follow in a similar way. (According to our specification, he plays a=1 if the unused message m=1 would be sent.)

Since the strategies played in round r+1 do not constitute an equilibrium, no Sender can send a credible neologism in r+2 and they all best respond to the Receiver's strategy in r+1. Finally, in round r+4, the dynamic is back to the most informative equilibrium (although the message strategy differs somewhat). If continued, the dynamic would effectively go on in this 4-cycle.

As the tables below show, the dynamic converges from both the pooling and the naïve initial conditions to this attractor:

b = 1.2, pooling initial conditions

	Strategy Sender					**Strategy Receiver**					
	$m_{\mathrm{T}}(1)$	$m_{\mathrm{T}}(3)$	$m_{\mathrm{T}}(5)$	$m_{\mathrm{T}}(7)$	$m_{\mathrm{T}}(9)$	$a_{\mathrm{T}}(1)$	$a_{\mathrm{T}}(3)$	$a_{\mathrm{T}}(5)$	$a_{\mathrm{T}}(7)$	$a_{\mathrm{T}}(9)$	Observation
r=0	5	5	5	5	5	5	5	5	5	5	Pooling equilibrium
r=1	1	5	7	9	9	1	3	3	5	8	
r=2	3	7	7	9	9	1	1	4	4	8	
r=3	1	5	9	9	9	1	3	3	7	7	
r=4	3	3	7	7	7	2	2	7	7	7	Most informative equilibrium

b = 1.2, naïve initial conditions

	Strategy Sender					**Strategy Receiver**					
	$m_{\mathrm{T}}(1)$	$m_{\mathrm{T}}(3)$	$m_{\mathrm{T}}(5)$	$m_{\mathrm{T}}(7)$	$m_{\mathrm{T}}(9)$	$a_{\mathrm{T}}(1)$	$a_{\mathrm{T}}(3)$	$a_{\mathrm{T}}(5)$	$a_{\mathrm{T}}(7)$	$a_{\mathrm{T}}(9)$	Observation
r=0	1	3	5	7	9	1	3	5	7	9	Naïve strategies
r=1	3	5	7	9	9	1	1	3	5	8	
r=2	5	7	7	9	9	1	1	1	4	8	
r=3	1	7	9	9	9	1	1	1	3	7	
r=4	7	7	9	9	9	2	2	2	2	7	Most informative equilibrium

Below we show the actions, the variance of the Receiver actions each Sender type induces, and the average of the variance over all Sender types in the attractor:

b = 1.2, Actions and average variance in attractor

Round	$a(m(1))$	$a(m(2))$	$a(m(3))$	$a(m(4))$	$a(m(5))$	**Average**
r	2	2	7	7	7	
r+1	1	3	5	8	8	
r+2	1	4	4	8	8	
r+3	1	3	7	7	7	
Var	3/16	1/2	27/16	1/4	1/4	**23/40**

C.3. $b = 2$

Also for b=2, the dynamic has a 4-cycle that includes the most informative equilibrium as an attractor:

$b = 2$, attractor											
	Strategy Sender					**Strategy Receiver**					
Round	$m_{\mathrm{r}}(1)$	$m_{\mathrm{r}}(3)$	$m_{\mathrm{r}}(5)$	$m_{\mathrm{r}}(7)$	$m_{\mathrm{r}}(9)$	$a_{\mathrm{r}}(1)$	$a_{\mathrm{r}}(3)$	$a_{\mathrm{r}}(5)$	$a_{\mathrm{r}}(7)$	$a_{\mathrm{r}}(9)$	Observation
r	1	5	5	5	5	1	6	6	6	6	Most informative equilibrium
r+1	1	5	7	9	9	1	3	3	5	8	
r+2	3	7	9	9	9	1	1	1	3	7	
r+3	7	7	9	9	9	2	2	2	2	7	
r+4	1	9	9	9	9	1	6	6	6	6	Most informative equilibrium

The dynamic converges from both the pooling and the naïve initial conditions to this attractor:

$b = 2$, pooling initial conditions											
	Strategy Sender					**Strategy Receiver**					
Round	$m_{\mathrm{r}}(1)$	$m_{\mathrm{r}}(3)$	$m_{\mathrm{r}}(5)$	$m_{\mathrm{r}}(7)$	$m_{\mathrm{r}}(9)$	$a_{\mathrm{r}}(1)$	$a_{\mathrm{r}}(3)$	$a_{\mathrm{r}}(5)$	$a_{\mathrm{r}}(7)$	$a_{\mathrm{r}}(9)$	Observation
r=0	5	5	5	5	5	5	5	5	5	5	Pooling equilibrium
r=1	5	5	7	7	7	2	2	2	7	7	
r=2	1	7	7	7	7	1	6	6	6	6	Most informative equilibrium

$b = 2$, naïve initial conditions											
	Strategy Sender					**Strategy Receiver**					
Round	$m_{\mathrm{r}}(1)$	$m_{\mathrm{r}}(3)$	$m_{\mathrm{r}}(5)$	$m_{\mathrm{r}}(7)$	$m_{\mathrm{r}}(9)$	$a_{\mathrm{r}}(1)$	$a_{\mathrm{r}}(3)$	$a_{\mathrm{r}}(5)$	$a_{\mathrm{r}}(7)$	$a_{\mathrm{r}}(9)$	Observation
r=0	1	3	5	7	9	1	3	5	7	9	Naïve strategies
r=1	3	5	7	9	9	1	1	3	5	8	
r=2	5	7	9	9	9	1	1	1	3	7	
r=3	7	9	9	9	9	1	1	1	1	6	Most informative equilibrium

Below we show the actions and the average variance of the actions in the attractor:

$b = 2$, Actions and average variance in attractor						
Round	$a(m(1))$	$a(m(2))$	$a(m(3))$	$a(m(4))$	$a(m(5))$	**Average**
r	1	6	6	6	6	
r+1	1	3	5	8	8	
r+2	1	3	7	7	7	
r+3	2	2	7	7	7	
Var	3/16	9/4	11/16	1/2	1/2	**33/40**

C.3. $b = 4$

For $b = 4$, the dynamic has as an attractor the following 2-cycle that includes the pooling equilibrium (which is the unique and thus also the most informative equilibrium):

$b = 4$, attractor

	Strategy Sender					Strategy Receiver					
Round	$m_{\mathrm{r}}(1)$	$m_{\mathrm{r}}(3)$	$m_{\mathrm{r}}(5)$	$m_{\mathrm{r}}(7)$	$m_{\mathrm{r}}(9)$	$a_{\mathrm{r}}(1)$	$a_{\mathrm{r}}(3)$	$a_{\mathrm{r}}(5)$	$a_{\mathrm{r}}(7)$	$a_{\mathrm{r}}(9)$	Observation
r	5	5	5	5	5	5	5	5	5	5	Most informative equilibrium
r+1	5	7	7	7	7	1	1	1	6	6	
r+2	7	7	7	7	7	5	5	5	5	5	Most informative equilibrium

The dynamic converges to the attractor from initial pooling conditions but also from initial naïve conditions:

$b = 4$, naïve initial conditions

	Strategy Sender					Strategy Receiver					
Round	$m_{\mathrm{r}}(1)$	$m_{\mathrm{r}}(3)$	$m_{\mathrm{r}}(5)$	$m_{\mathrm{r}}(7)$	$m_{\mathrm{r}}(9)$	$a_{\mathrm{r}}(1)$	$a_{\mathrm{r}}(3)$	$a_{\mathrm{r}}(5)$	$a_{\mathrm{r}}(7)$	$a_{\mathrm{r}}(9)$	Observation
r=0	1	3	5	7	9	1	3	5	7	9	Naïve strategies
r=1	5	7	9	9	9	1	1	1	3	7	
r=2	7	9	9	9	9	1	1	1	1	6	
r=3	9	9	9	9	9	5	5	5	5	5	Most informative equilibrium

Below we show the actions and the average variance of the actions in the attractor:

$b = 4$ Actions and average variance in attractor

Round	$a(m(1))$	$a(m(2))$	$a(m(3))$	$a(m(4))$	$a(m(5))$	**Average**
r	5	5	5	5	5	
r+1	1	6	6	6	6	
Var	4	1/4	1/4	1/4	1/4	**1**