Quantitative Economics 9 (2018), 1453–1487 1759-7331/20181453

Meaning and credibility in experimental cheap-talk games

Ernest K. Lai
Department of Economics, Lehigh University

Wooyoung Lim
Department of Economics, The Hong Kong University of Science and Technology

We design experimental games to evaluate the predictive power of the first cheap-talk refinement, *neologism-proofness*. In our first set of treatments designed to evaluate the refinement with its usual emphasis on literal meanings, we find that a fully revealing equilibrium that is neologism-proof is played more often; senders deviate from an equilibrium in a way that can be predicted by the credibility of the neologism; and receivers' behavior indicates that they understand senders' deviating incentives. Our second set of treatments evaluates neologism-proofness from an evolutionary perspective in the absence of a common language. We find that the proportion of observations in which the meaning of a neologism evolves to disrupt a prevailing fully revealing equilibrium is lower when the equilibrium is neologism-proof. Our findings shed light on the capabilities and limitations of the refinement concept in predicting laboratory behavior under different language environments.

Keywords. Neologism-proofness, cheap talk, equilibrium refinement, evolution of meanings, laboratory experiment.

JEL classification. C72, C92, D82, D83.

1. Introduction

A cheap-talk game is a signaling game in which messages have no direct payoff consequences. This costless nature of messages has profound implications for the treatment of equilibria. In using cheap talk to communicate, the meaning of a message can only be established by use (in equilibrium) and cannot be learned by introspection. Furthermore, for every equilibrium with unused messages, one can construct another outcome-

Ernest K. Lai: kwl409@lehigh.edu
Wooyoung Lim: wooyoung@ust.hk

We are grateful to Andreas Blume, Syng-Joo Choi, Navin Kartik, Barry Sopher, Joel Sobel, the coeditor, and the anonymous referees for their valuable comments and suggestions. We also thank conference and seminar participants at HKUST, Rutgers University, Sogang University, the 2nd Haverford Meeting on Behavioral and Experimental Economics, the 16th KAEA-KEA International Conference, the 5th Annual Xiamen University International Workshop on Experimental Economics, the 90th Western Economic Association International Annual Conference, the 26th International Conference on Game Theory, the 2016 and 2017 Asian Meetings of the Econometric Society for valuable discussions. This study is supported by a grant from the Research Grants Council of Hong Kong (Grant No. ECS-699613). Lai gratefully acknowledges the financial support from the Office of the Vice President and Associate Provost for Research and Graduate Studies at Lehigh University. We thank Oliver Lai for useful little research assistance.

 https://doi.org/10.3982/QE683

equivalent equilibrium in which all messages are used.[1] Consequently, the standard refinement concepts for costly signaling games, which select equilibria by restricting the interpretations of unused, out-of-equilibrium messages, cannot rule out any cheap-talk equilibrium outcome. The otherwise powerful intuitive criterion (Cho and Kreps (1987)), for example, has no selection power for cheap-talk games.

Farrell's (1993) notion of *neologism-proofness*, a belief-based refinement that appeals to natural language for conferring meanings to out-of-equilibrium messages, pioneers the selection of equilibria for cheap-talk games. A number of refinement concepts inspired by or related to neologism-proofness have since been developed. Rabin's (1990) nonequilibrium notion of credible message rationalizability (CMR) combines a feature of neologism-proofness with rationalizability; Matthews, Okuno-Fujiwara, and Postlewaite (1991) generalize neologism-proofness with their concept of announcement-proofness; Chen, Kartik, and Sobel (2008) relate their selection criterion, no incentive to separate (NITS), to neologism-proofness by showing that an equilibrium survives NITS if it is neologism-proof; de Groot Ruiz, Offerman, and Onderstal (2015) propose a behavioral refinement, average credible deviation criterion (ACDC), which further extends on neologism-proofness and announcement-proofness.[2]

Despite the fact that neologism-proofness serves as the headwater of the literature and has been developed for more than 20 years, no experimental effort has been devoted exclusively to evaluating its predictive power. This is unlike the case of costly signaling games, where the major refinement concepts have been subject to thorough experimental investigations at an early stage of the literature (e.g., Brandts and Holt (1992), Banks, Camerer, and Porter (1994)). In fact, only a handful of earlier experimental studies have studied cheap-talk refinements (e.g., Sopher and Zapater (2000), Blume, Dejong, Kim, and Sprinkle (2001)), and the concepts covered have been limited.[3] This study contributes to filling this void of the literature by experimentally evaluating the very first concept in the history of cheap-talk refinements.

[1]This can be done, for example, by having the sender randomize over a support comprised of an equilibrium message and all unused messages. The receiver's posterior belief upon receiving any of the originally unused messages will now coincide with the belief after receiving the equilibrium message. With messages being costless, the sender only cares about the receiver's beliefs and the resulting actions; it is thus a best response for the sender to randomize as prescribed, leading to a new equilibrium that shares the same outcome (mapping from states to actions) as the original equilibrium. For a more formal discussion, see Farrell (1993, p. 517).

[2]Cheap-talk refinements can be broadly divided into two classes: belief-based and nonbelief-based. Other than CMR and NITS, nonbelief-based refinements include evolutionary stability (Blume, Kim, and Sobel (1993), Rabin and Sobel's (1996) recurrent mop, Blume and Sobel's (1995) communication-proof equilibrium, and Blume's (1993) perturbed message persistence and effective language equilibrium. Belief-based refinements, which include neologism-proofness and its two derivatives, announcement-proofness and ACDC, are closer to the tradition of costly-signaling refinements. A case in point is the perfect sequential equilibrium developed by Grossman and Perry (1986). The equilibrium concept does not have the same natural language element but is otherwise closely related to neologism-proofness. On the link between costly-signaling refinements and cheap-talk games, it is also worth mentioning the immunity to credible deviations developed by Eső and Schummer (2009). The criterion has selection power for a Crawford and Sobel's (1982) type of model where the original cheap-talk messages are replaced by costly messages.

[3]There are some recent attempts (e.g., Kawagoe and Takizawa (2008), de Groot Ruiz, Offerman, and Onderstal (2014)), which will be discussed in the literature review.

Meaning and credibility are two building blocks of neologism-proofness. In the context of a communication game, a neologism, from the Greek for "new word," refers to an out-of-equilibrium message. Despite the fact that a neologism—an unused message—has no meaning in light of established use, Farrell (1993) argues that when players share a preexisting common language (e.g., English) a neologism should still be understood, and it is by its literal meaning. Farrell (1993) goes on to tackle credibility, introducing the concept of *credible neologisms*. Suppose that there are certain sender's types (and no other else) who want to deviate from a putative equilibrium by sending a particular neologism. If the literal meaning of that neologism is precisely that "I am of one of these certain types who want to send this neologism," it is credible or self-signaling. A neologism-proof equilibrium is one in which no credible neologisms exist.

The above definition of neologism-proofness, with its emphasis on literal meanings, informs the design of our first set of treatments. We design two cheap-talk games with binary type and literally meaningful messages. We vary the cardinality of the message spaces from two to three, where the additional message represents the neologism that we induce with respect to a putative equilibrium. This manipulation of the message spaces results in a total of four treatments. All four games each admit a fully revealing equilibrium. But the games' payoff structures and/or message spaces vary so that only the fully revealing equilibrium in some of the games is neologism-proof, and this serves as our treatment variation.

Findings from our first set of treatments indicate that neologisms and their credibility play an evident role in how subjects communicate. Overall, fully revealing equilibria that are neologism-proof are played more often than those that are not. The qualitative patterns of senders' deviating behavior are predicted by the self-signaling properties of the neologisms. Receivers' behavior also suggests that they understand the different deviating incentives of the senders under credible and noncredible neologisms.

While appealing to a preexisting language to generate meanings for unused messages, Farrell (1993) also provides an evolutionary interpretation of neologisms, invoking the fact that an equilibrium can be viewed as an evolutionarily stable outcome. An environment with a priori meaningless messages brings to the fore this evolutionary interpretation. In such an environment, the meaning of a neologism, which contributes to determine its self-signaling property, must evolve endogenously. An experimental study with subjects participating in repeated interactions provides a valuable opportunity to study neologism-proof equilibria from this evolutionary perspective. It allows us to investigate empirically how the evolved meaning of a neologism may affect the communication outcomes.

Our second set of treatments takes the two basic games in the first set of treatments and endow them with meaningless, symbolic messages, initially with only two messages. Unlike the first set of treatments where the presence of a neologism is a between-subject treatment variable, in these treatments the introduction of a neologism, that is, a third meaningless message, is within subject. The message is introduced in the middle of an experimental session, after endogenous meanings have supposedly been developed for the two initial messages. We investigate whether introducing a neologism in

this manner may disrupt any (fully revealing) equilibrium play and whether it is as predicted by the neologism-proof property of the equilibrium.[4]

Findings from this set of treatments reveal that, in the absence of a preexisting common language, there is—unsurprisingly—a greater variation in individual group behavior. Focusing then on individual groups rather than aggregate behavior, we find that separations occur and distinct meanings evolve for the two initial messages in the majority of individual groups.[5] Among these selected groups, we find that the percentage of groups in which the meaning of the third message evolves to disrupt the separations is lower for the game in which the fully revealing equilibrium is neologism-proof.

Before we proceed to review the literature, we discuss three limitations of neologism-proofness and argue how our study may provide responses to two of the limitations. The discussion serves to further underscore the contribution of our experiment in light of the theory. Neologism-proofness lacks a general existence property and, therefore, we do not know what it predicts when existence fails. It also does not provide a complete formalization for the presence of unsent messages with natural meanings. Finally, it falls short of fully addressing the concern over the usefulness of natural language, as neologisms arise off the equilibrium path but languages on the path are still arbitrary.[6]

Regarding the first limitation, our experiment helps assess from an empirical vantage point whether the lack of a general existence property should be a reason to discard the refinement concept even for the games for which it has predictive power.[7] While our study does not address the second limitation, it provides an experimental response to the third limitation by exploring whether there is an empirical tendency for literal meanings to be followed on the equilibrium path. Furthermore, our treatments with meaningless messages offer a contrasting point to demonstrate the importance of natural language in a laboratory setting.

Related literature

To our knowledge, the first experimental study that involves neologism-proofness is Blume et al. (2001). They evaluate the validity of the partial common interest criterion (Blume, Kim, and Sobel (1993)), comparing it with other selection criteria including

[4]We also include a control treatment to address the potential experimenter demand effect associated with the introduction of a third message during the experiment.

[5]By individual group, we mean a matching group in which a few sender-subjects randomly match with a few receiver-subjects to form groups of two. Statistically, a matching group constitutes an independent observation.

[6]We thank Joel Sobel for sharing these points with us.

[7]Regarding the lack of existence, for instance, no equilibrium outcome in the leading example of Crawford and Sobel (1982)—the uniform quadratic formulation—is neologism-proof. Yet there are other instances where neologism-proofness is relied upon to obtain sharper predictions (e.g., Gertner, Gibbons, and Scharfstein (1988), Farrell and Gibbons (1988, 1989), Austen-Smith (1990), Lim (2014), Kim and Kircher (2015)); see also the discussion in Farrell and Rabin (1996). A major question we pose in this paper is therefore: for the games in which neologism-proofness has some bite, how well the refinement predicts play in the lab.

neologism-proofness. In their experimental games, neologism-proofness either shares the same prediction as Pareto efficiency or rejects all equilibria. Their findings therefore do not distinctively reveal how useful neologism-proofness is in predicting play, which is not their primary objective to begin with. As will be discussed below, avoiding the potential confound of Pareto efficiency as a selection criterion is a major consideration behind the design of our games.

Two other experimental studies related to cheap-talk refinements are Sopher and Zapater (2000) and Kawagoe and Takizawa (2008). Sopher and Zapater (2000) find experimental support for the CMR proposed by Rabin (1990). Kawagoe and Takizawa (2008) compare the effectiveness of equilibrium refinements and level-k models in explaining their data, concluding that the level-k approach outperforms. Their design, however, does not provide a rich enough environment to address neologism-proofness as neologisms are absent in their games.

More recently, de Groot Ruiz, Offerman, and Onderstal (2015) propose a new cheap-talk refinement, ACDC, in which the stability of an equilibrium is measured by the frequency and size of credible deviations. They show that ACDC successfully organizes the data from several prior experiments. Most related to our study is their another paper (de Groot Ruiz, Offerman, and Onderstal (2014)), which further evaluates ACDC with their own experiment. They find that neologism-proofness performs well when its prediction is unique, in which case ACDC predicts the same as neologism-proofness. The ACDC equilibrium is found to perform the best when neologism-proofness has no selection power. Our paper differs from theirs on focus, design, and the issue addressed. While they use the inability of neologism-proofness to predict as a contrasting point to demonstrate the power of ACDC, our study evaluates the usefulness of neologism-proofness by focusing exclusively on the cases where the concept selects an equilibrium. The use of a priori meaningless messages to investigate the evolution of meaning of a neologism is also unique to our study.

Our paper is related to two other strands of literature. The first is the literature on experimental communication games (e.g., Dickhaut, McCabe, and Mukherji (1995), Blume et al. (1998, 2001), Gneezy (2005), Cai and Wang (2006), Sánchez-Pagés and Vorsatz (2007, 2009), Hurkens and Kartik (2009), Wang et al. (2010)).[8] A robust finding of this literature is the observation of "over-communication" or "lying aversion," where subjects communicate more than is predicted by equilibria. Our paper differs from these papers in that we are interested not only in whether subjects play according to an equilibrium but also in whether they play according to a refined equilibrium.

The investigation of the evolution of meanings using a priori meaningless messages is a subject matter of Blume et al. (1998, 2001) and Blume, Dejong, and Sprinkle (2008), who address the question using common-interest or partially-common-interest games. Our study of meanings in reference to neologism-proof equilibria represents a new inquiry. The introduction of an additional message in the middle of an experimental ses-

[8]See Blume, Lai, and Lim (2017) for a survey of this literature. See also Crawford (1998) for an earlier survey and a discussion on the connection between cheap-talk and costly-signaling refinements.

sion is also, to our knowledge, not explored before.[9] Furthermore, we focus exclusively on games in which the sender and the receiver have conflicting preferences over different equilibria.[10]

Another literature to which our paper is related is the experimental investigation of equilibrium refinements for costly signaling games. Brandts and Holt (1992) find support for the intuitive criterion by Cho and Kreps (1987). Banks, Camerer, and Porter (1994) design games to separate various refinements, which include the Nash equilibrium, sequential equilibrium, intuitive criterion, divine, universal divine, and never-weak-best-reply. They find that subjects' behavior converges to the more refined equilibrium up to the intuitive criterion.

The rest of the paper proceeds as follows. Section 2 lays out our experimental games and performs the equilibrium analysis. Section 3 discusses our experimental hypotheses and procedures. Section 4 reports our findings. Section 5 concludes.

2. The experimental cheap-talk games

2.1 *Four games with literal messages*

Our first set of treatments consists of four games endowed with literally meaningful messages. For each treatment, subjects participate in 20 rounds of the game. Putting aside the variations in message spaces, the four games reduce to two games with different payoff structures, which we call Game 1 and Game 2. We first describe the basic features of Games 1 and 2 and will provide details of the message spaces when we discuss neologisms below.

There are two players, a sender (he) and a receiver (she). The sender is privately informed about his type $\theta \in \{s, t\}$. The common prior is that the two types are equally likely. After observing the realized θ, the sender sends a cheap-talk message, $m \in M$, to the receiver. The receiver then takes an action $A \in \{L, C, R\}$.

Table 1 presents the payoffs. For each pair of numbers, the first entry is the sender's payoff and the second entry the receiver's payoff for the corresponding state-action combination.

We design these games out of simplicity and three other considerations. First, as long as $|M| \geq 2$, Games 1 and 2 both have multiple perfect Bayesian equilibrium outcomes not Pareto ranked. This creates rooms for equilibrium selection, and yet Pareto

[9]Other studies that adopt between-subject manipulations of message spaces include Blume et al. (1998), who document that restricting the message space expedites convergence in the presence of meaningless messages; Lai, Lim, and Wang (2015), who highlight the role of message spaces in facilitating information transmission in a multidimensional setting; and Serra-Garcia, van Damme, and Potters (2013), who study the effect of limiting the message space in public good games with communication of private information. Our novel contribution relative to these studies lies in providing a new channel—a refinement of equilibria—through which message spaces may play a role in information transmission.

[10]Duffy, Lai, and Lim (2017) study how meanings develop in the Battle of the Sexes game, also a game with conflicting preferences over equilibria. Their study, however, is not in a sender-receiver setting and is about how subjects learn to play a correlated equilibrium aided by an external correlation device sending meaningless messages.

TABLE 1. Payoffs.

	(a) Game 1			(b) Game 2		
	L	*C*	*R*	*L*	*C*	*R*
s	30, 20	20, 30	0, 8	50, 20	20, 30	0, 8
t	30, 20	8, 0	20, 30	10, 20	8, 0	20, 30

dominance is ruled out as a confounding selection criterion.[11] There are two salient equilibrium outcomes, babbling and fully revealing. In the former, the receiver ignores the sender's message, taking the ex ante ideal action L; in the latter, the receiver takes different actions, C or R, after receiving different messages. In both games, the sender strictly prefers the babbling outcome to the fully revealing outcome, while the opposite is true for the receiver.

Our second consideration is to minimize the possibility that subjects play different equilibria in different games entirely out of payoff considerations. We control the sender's expected payoffs from the two equilibrium outcomes so that in both games the payoffs are 30 for the babbling and 20 for the fully revealing outcomes.

Our last consideration is to ensure that other-regarding preferences, if any, do not play a role in the equilibrium selection. We make the equilibrium payoff profiles in our games similar to those found in the Battle of the Sexes game, where a player's expected payoff from one equilibrium outcome is the other player's expected payoff from the other outcome.[12]

Message spaces and neologism-proofness A neologism-proof equilibrium is one in which a credible neologism does not exist. To set the ground for defining credible neologisms, Farrell (1993) first assumes that for every relevant equilibrium, a neologism, being an unsent message with literal meaning "my type is in K," exists for every nonempty subset K of the type space. Relative to a putative equilibrium, a neologism with meaning "my type is in K" is then credible or self-signaling if (a) the sender's types in K strictly prefer the outcome achieved by having the neologism believed over the putative equilibrium outcome, and (b) the types not in K weakly prefer to stay in the putative equilibrium.

In designing a laboratory environment that is easy for subjects to comprehend, we look for games that allow us to apply neologism-proofness in the simplest possible set-

[11]Game 1 shares the same qualitative structure as the games in several existing papers: the "I Won't Tell You" game in Farrell (1993), Example 3 in Rabin (1990), Game Γ_2 in Matthews, Okuno-Fujiwara, and Postlewaite (1991), Game 2 in Kawagoe and Takizawa (2008), and Game 1 in Sobel (2013). Refer to Sobel (2013) for a justification for this kind of payoff structures. Game 2 is qualitatively the same as Example 2 in Rabin (1990).

[12]We also strive to separate the prediction of neologism-proofness as much as possible from that of level-k analyses, which are commonly used to rationalize findings from communication games. Note that the only difference between Games 1 and 2 is the ideal actions of type-t sender. The expected payoff from each equilibrium outcome for each player, as well as the receiver's ideal actions, is the same across the two games. Accordingly, for a given size of the message space, very limited configurations of level-k models can generate different predictions for the two games. Refer to Appendix B in the Supplementary Material (Lai and Lim (2018)) for an auxiliary level-k analysis.

ting. Note that Farrell's (1993) requirement that a neologism exists for every nonempty subset of the type space is not part of the definition of credible neologisms. Accordingly, neologism-proofness can still be applied when, for a given equilibrium, a neologism exists for some but not all collections of types. The balance between a simple design and the applicability of the concept guides the design of our message spaces, which feature what we call *limited neologisms*.

We pair Games 1 and 2 each with a message space that contains three literal messages: $M = \{$"my type is s," "my type is t," "I won't tell you my type"$\}$. The resulting games are called Game $1M3$ and Game $2M3$. Applying neologism-proofness to the babbling and fully revealing outcomes of the games, we obtain the following characterization.

PROPOSITION 1. *The fully revealing outcome in Game* $1M3$ *cannot be supported as neologism-proof, whereas that in Game* $2M3$ *can be so supported. The babbling outcome in Game* $1M3$ *can be supported as neologism-proof, whereas that in Game* $2M3$ *cannot be so supported.*

To show that the fully revealing outcome in Game $1M3$ cannot be supported as neologism-proof, it suffices to show that one of the fully revealing equilibria is not neologism-proof. Consider the truth-telling equilibrium in which s and t send, respectively, "my type is s" and "my type is t."[13] The payoffs for both types in this equilibrium are 20. "I won't tell you my type," which literally means that the sender's type is in $\{s, t\}$, is the neologism. If the receiver believes the literal meaning of this neologism, she will take the ex ante ideal action, L, resulting in a payoff of 30 for the sender regardless of his type. Given that $30 > 20$, both s and t—precisely the types that correspond to the literal meaning of the neologism—prefer the outcome achieved by sending the neologism over the putative equilibrium outcome: the neologism is credible.

For the truth-telling equilibrium in Game $2M3$, note that when the receiver believes the literal meaning of "I won't tell you my type" and takes action L, only s but not t strictly prefers the neologism to be believed: the neologism is not credible. Consider further the remaining two fully revealing equilibria that are not truth-telling, in which either "my type is s" or "my type is t" is the neologism. In either case, the type expressed by the literal meaning of the neologism would not strictly prefer the outcome achieved by the neologism, because the on- and off-equilibrium-path payoffs are the same for the type. The neologisms in these nontruth-telling fully revealing equilibria are therefore also not credible. Consequently, the fully revealing outcome in Game $2M3$ can be supported as neologism-proof.[14]

[13]The truth-telling equilibrium is a fully revealing equilibrium in which literal meanings are used on the equilibrium path.

[14]While neologism-proofness could be applied under our limited neologisms, a different approach in verifying a neologism-proof outcome is used. In Farrell (1993), since for any equilibrium a neologism exists for every subset of the type space, when a neologism-proof equilibrium exists to support an outcome, it automatically covers all neologisms. With limited neologisms, however, different equilibria that support an outcome may be associated with different (sets of) neologisms. Even if self-signaling does not occur under the neologism(s) available in one supporting equilibrium, it may occur under another neologism in another supporting equilibrium. To show that a given outcome is neologism-proof, one therefore needs

In Game $1M3$, since both sender's types receive their maximum payoffs in the babbling outcome, it is straightforward that no credible neologisms can exist. Finally, to show that the babbling outcome in Game $2M3$, which gives s a payoff of 50 and t a payoff of 10, cannot be supported as neologism-proof, it suffices to consider a babbling equilibrium in which "my type is t" is a neologism (there may be another neologism in the equilibrium). In this case, t strictly prefers the outcome achieved by the neologism over the putative equilibrium outcome; s, on the other hand, prefers to stay in the equilibrium. The different moves preferred by the two types render the neologism credible.[15]

To generate richer treatment variations so as to identify the effects of neologisms, we introduce two more games, Game $1M2$ and Game $2M2$, which are Games 1 and 2 endowed with a binary message space $M' = \{$"my type is s," "my type is t"$\}$. The neologism-proofness of the babbling and the fully revealing outcomes of these games is characterized as follows.

PROPOSITION 2. *Both the fully revealing and babbling outcomes in Game* $1M2$ *can be supported as neologism-proof. Only the fully revealing outcome in Game* $2M2$ *can be supported as neologism-proof.*

For these games with binary messages, the fully revealing outcomes are either supported by a truth-telling equilibrium or by another equilibrium in which the literal meanings of the messages are not followed on the equilibrium path. Their neologism-proofnesses are a trivial consequence of the fact that in any such supporting equilibrium all messages are used. We have thus taken the limited neologisms to the extreme by eliminating neologisms altogether.[16] The neologism-proofness of the babbling outcomes in Games $1M2$ and $2M2$ is characterized in the same way as that in Games $1M3$ and $2M3$.[17]

to show that all its supporting equilibria with different neologisms are neologism-proof so as to cover all possible neologisms.

[15]Following the convention of the literature, our characterizations focus on the babbling and the fully revealing equilibria, which can be obtained with sender's pure strategies and receiver's unique pure-strategy best responses. Each of our games admits multiple behavior-strategy equilibria. For example, the following constitutes an equilibrium in Game $1M3$: s sends message m with probability $\frac{2}{19}$ and $m' \neq m$ with probability $\frac{17}{19}$, while t sends m' with probability one; the receiver takes action C after receiving m and randomizes between L and R with probabilities $\frac{2}{3}$ and $\frac{1}{3}$ after receiving m'. One can show that, for any such partially revealing equilibrium in any of our game, the equilibrium outcome is neologism-proof if and only if the babbling outcome is neologism-proof. With the asymmetric moves by different types and the precise randomization required, these partially revealing equilibria would be less salient in the lab. In fact, as will be discussed in Section 3.2, we use the strategy method for our games with literal messages; for the sake of simple experimental instructions, we do not allow subjects to randomize explicitly for a given contingency.

[16]Our Game $1M2$ is of the same class as Game 2 in Kawagoe and Takizawa (2008). Our characterization is, however, different from theirs. Given that there are no unused messages in any fully revealing equilibrium in our Game $1M2$ or their Game 2, we consider that the fully revealing outcomes trivially survive neologism-proofness, while they consider that only the babbling outcome survives. This difference in our characterizations would have implications for the interpretation of experimental findings. As will become apparent below, subjects' behavior in our Game $1M2$ conforms to the fully revealing equilibrium, which, accordingly to our view, is consistent with the prediction of neologism-proofness. Under the conviction that only the babbling outcome survives neologism-proofness, Kawagoe and Takizawa (2008) attribute their similar findings to over-communication.

[17]Other cheap-talk refinements provide varying selections for Game 1. The two other belief-based concepts, announcement-proofness and ACDC, as well as the nonbelief-based CMR, share the same predic-

A couple of remarks about how the design of our message spaces relates to Farrell's (1993) notion of natural language and postulated existence of neologisms is in order. There are two main ingredients in Farrell's (1993) natural language: (a) *common language*, that is, each message has a literal meaning that is associated with a type in the type space, and (b) *rich language*, that is, the message space is large enough so that for any subset K of the type space, a message with the literal meaning "my type is in K" exists. Our message spaces fulfill the first requirement: being framed in English, the messages have clear and commonly understood literal meanings. While by most standards our message spaces cannot be considered as large and despite the limited neologisms that we induce, the message space M is rich enough to describe all subsets of the binary type space, even though some of the messages are used in equilibrium.

In assuming the existence of neologisms, Farrell (1993) argues that a behavior-strategy equilibrium in which the sender randomizes over messages with the same equilibrium meaning is implausible. This will be so when, for example, the sender has a preference for short and simple messages. Our design with two different sizes of message spaces allows us to evaluate whether the assumed existence of unused messages is an empirically plausible assumption; if it is a natural tendency to use all available messages, with certain distinct messages conveying the same meaning, having an additional message in M relative to M' should not affect observed behavior.[18]

2.2 *Three games with a priori meaningless messages*

Notwithstanding the important role of common language in neologism-proofness, Farrell (1993) also provides an evolutionary interpretation of neologisms in an environment without preexisting language, in which the meaning of a neologism must evolve. To investigate the predictive power of neologism-proofness from this perspective, we design a second set of treatments with a priori meaningless messages. It consists of three additional games, Game $1E$, Game $2E$, and Game $1E^d$, where the label "E" refers to "*e*volution." For each of these treatments, subjects participate in 40 rounds of the game.

Games $1E$ and $2E$ are, respectively, Games 1 and 2 endowed with an *initial* binary message space $M_e = \{\$, \%\}$. In each of these games, a third meaningless message, "&," is—without using deception—unexpectedly introduced and made available to

tions with neologism-proofness, selecting differently for the two incarnations of Game 1. This contrasts with a few nonbelief-based concepts, which do not offer varying predictions across the two treatments. For example, evolutionary stability rules out the babbling and the fully revealing outcomes for both Games $1M2$ and $1M3$, while communication-proofness and perturbed message persistence select both outcomes. Including the CMR, which, despite being nonbelief-based, is natural-language reliant, the refinements that select differently for Games $1M2$ and $1M3$ are based on a variant of or related to credible neologism. For Games $2M2$ and $2M3$, these refinements also select the same as neologism-proofness. One can therefore argue that our experiment is not about evaluating the predictive power of neologism-proofness per se but more broadly the notion of credible neologism.

[18]Despite our exclusion of *explicit* randomization in the experimental design (see Section 3.2), subjects may still randomize *internally*, in which behavior observed over repetitions of the games would allow us to assess whether different messages are used to convey the same meaning.

the sender-subjects after the 20th round. These treatments therefore involve a within-subject variation of the message spaces, where there are two messages in the first 20 rounds and three messages in the last 20 rounds.[19]

Without proving additional propositions, we discuss what behavior may emerge in Games 1E and 2E, leveraging the characterization in Proposition 1 and the anticipation that meanings will emerge in one way or the other.[20] We start with two plausible conjectures about how play may evolve before and *immediately* after the new message is introduced. First, given the payoff structures of the games, it is plausible that some separation may occur under the binary message spaces, which results in distinct meanings being emerged for the two initial messages, "$" and "%."[21] When the third message, "&," is introduced and sent, it is considered as a neologism. In the absence of any literal meaning or precedential use of the message, a natural response of the receiver is to ignore it; our second conjecture is that the receiver considers that the neologism, immediately after it is introduced, does not provide any information, and her initial response is to take the ex ante ideal action L.

The above conjectures predict different evolutions of play in Games 1E and 2E. Recall that in Game 1E, both s and t obtain 20 in a fully revealing equilibrium and 30 from the receiver's taking L. Accordingly, it pays for both types to send the neologism when the receiver's initial response is L. As they do, a meaning that does not exist in the prevailing fully revealing equilibrium emerges for "&," which is exactly that the sender can be of either type. This in turn reinforces the receiver to choose L. The neologism is thus credible with respect to its evolved meaning, and we expect to see less separation after it is introduced, echoing the fact that the fully revealing outcome in the corresponding Game 1M3 cannot be supported as neologism-proof.

In Game 2E, the receiver's initial response of L to "&" will only attract s to send the neologism. As "&" becomes a choice of message of s but not t, its meaning evolves to be that the sender's type is s, to which the receiver's best response is C. The incentive for s to send the neologism is then weakened, because the type receives the same treatment in the prevailing fully revealing equilibrium. The neologism also does not convey any new meaning. Note further that regardless of whether the receiver's response to "&" is L (initially) or C (eventually), it does not pay for t to send the neologism. Given that no type strictly prefers to send "&," the neologism is not credible; separation remains even with the third message, echoing the fact that the fully revealing outcome in the corresponding Game 2M3 can be supported as neologism-proof.

[19]Accordingly, the games should be more properly understood as an experimental design rather than one-shot games in the formal sense; the variation of the message space in a particular "game" (treatment) involves, strictly speaking, two different games.

[20]Farrell's (1993) evolutionary interpretation of neologism-proofness (pp. 526–527) is based on an example and is rather informal. Our design is not to empirically evaluate the exact argument in his example, partly because the example is based on a game with a different equilibrium property. Rather, our objective is to investigate, from the evolutionary perspective suggested by Farrell (1993), how well neologism-proofness might predict in the absence of a preexisting language.

[21]As an empirical precedence supporting this conjecture, Blume et al. (1998) find that meaningful communication with separation endogenously emerges in a common-interest communication game with meaningless messages, where, as in our case, the type and the message spaces are both binary.

Our introduction of a new message in the midst of the experiment may create an experimenter demand effect: subjects may feel compelled to use the third message independent of the incentives induced. While this may present a confound, note that any such effect would have existed in both Games 1E and 2E and thus should not intervene with the interpretation of any difference in the findings from the two games. Nevertheless, it would be useful to gauge the extent to which behavior is affected by the *mere* introduction of a new message, which we pursue with our last game, Game 1E^d. This control game designed to address the potential *d*emand effect is the same as Game 1E, except that its third message is reverted back to literal: in the last 20 rounds, its initial message space, $M'_e = \{\$, \%\}$, is augmented to include "my type is s."

The initial prediction for Game 1E^d is the same as that for Game 1E. The prediction with respect to the introduction of the third message is, however, different. Given the premise that a neologism is to be understood by its literal meaning (when one is available), the introduction of "my type is s" in Game 1E^d would have no impact on the prevailing equilibrium outcome. The literal meaning of "my type is s" would coincide with the equilibrium meaning of one of the existing symbols. This different predicted play, if observed, would gain us confidence that it is the meaning of the neologism, either literally given or evolved, not its mere introduction via the experimenter demand effect, that is driving the observed behavior.[22]

3. Experimental hypotheses and procedures

3.1 *Hypotheses*

The four configurations of message spaces and the two payoff structures, with one of the message configurations used only under one payoff structure, give rise to a $3 \times 2 + 1$ treatment design. Table 2 summarizes the properties of these seven treatments.

In formulating experimental hypotheses, we perform comparisons *within* each set of treatments. For the treatments with literal messages, we compare the frequencies of fully revealing outcomes across games. Informed by Propositions 1 and 2, we evaluate how the introduction of neologisms or change in their self-signaling properties affects the attainments of fully revealing outcomes, based on the premise that a fully revealing equilibrium surviving neologism-proofness is more likely to be played. We compare Game 2M2 with Game 2M3, Game 2M3 with Game 1M3, and Game 1M3 with Game 1M2.[23]

[22]We thank an anonymous referee for pointing out the potential experimenter demand effect, which prompted us to design Game 1E^d as a control game. Our choice of "my type is s" as the literal third message is largely a random decision; using "my type is t" (but not "I won't tell you my type") would serve the same purpose. Our objective in designing a control game with but one of meaningless messages is to obtain a different predicted play from its all-meaningless-message counterpart. As discussed in footnote 14, there may exist in our setting a neologism-proof equilibrium as a supporting equilibrium for a nonneologism-proof outcome, while all the supporting equilibria of a neologism-proof outcome must be neologism-proof. Accordingly, Game 1E is the choice of game for which we should provide a control: given that the fully revealing outcome in Game 2E is neologism-proof, we would not be able to find a control for Game 2E that has a different predicted play.

[23]Note that for Games 1M2 and 2M3, two of the games that we do not compare, two different neologism-proof equilibria exist in the former and only one exists in the latter. The coordination problem on which

TABLE 2. Experimental Treatments.

	Game 1	Game 2
Literal M $\lvert M\rvert = 3$	Game 1$M3$ neologism-proof: babbling	Game 2$M3$ neologism-proof: fully revealing
Literal M' $\lvert M'\rvert = 2$	Game 1$M2$ neologism-proof: both	Game 2$M2$ neologism-proof: fully revealing
Meaningless M_e $\lvert M_e\rvert = 2 \to 3$	Game 1E neologism-proof: both $\to$ babbling	Game 2E neologism-proof: fully revealing $\to$ fully revealing
Mixed M'_e $\lvert M'_e\rvert = 2 \to 3$	Game 1E^d neologism-proof: both $\to$ fully revealing	

We use Game 2$M2$ as the starting point of our comparative statics. From Game 2$M2$ to Game 2$M3$, we create a neologism that is not credible; our first hypothesis concerns the effect of the *existence* of noncredible neologisms:

HYPOTHESIS 1 (Effect of the Existence of Noncredible Neologisms). *The frequency of fully revealing outcome in Game* 2$M2$ *is the same as that in Game* 2$M3$.

From Game 2$M3$ to Game 1$M3$, we make the neologism credible; our second hypothesis concerns the effect of the *credibility* of neologisms:

HYPOTHESIS 2 (Effect of the Credibility of Neologisms). *The frequency of fully revealing outcome is higher in Game* 2$M3$ *than in Game* 1$M3$.

From Game 1$M3$ with a credible neologism to Game 1$M2$ with only two messages, we get rid of any neologism; our third hypothesis concerns the *joint effect* of the existence and credibility of neologisms:

HYPOTHESIS 3 (Effect of the Existence and Credibility of Neologisms). *The frequency of fully revealing outcome is lower in Game* 1$M3$ *than in Game* 1$M2$.[24]

We turn to the treatments with meaningless messages for our last hypothesis, which addresses the evolved meanings, uses, and effects of the third messages as implied by the

neologism-proof equilibrium is to be played, which is harder in the two-message game, provides a criterion for comparing Games 1$M2$ and 2$M3$ (a similar argument can be made for comparing Games 1$M2$ and 2$M2$). However, since there are two variations in the game environments and to avoid distraction from our focus on the predictive power of neologism-proofness, we do not compare along the diagonals in the upper part of Table 2. Given that there is no change in the message spaces or the property of the neologisms, we also do not compare Games 1$M2$ and 2$M2$.

[24]Note that the same type of coordination issue discussed in footnote 23 exists in the comparison between Games 1$M2$ and 1$M3$. It does not, however, result in a detrimental confound. The harder coordination problem in Game 1$M2$ provides an unfavorable ground for its fully revealing equilibrium to be played; if the hypothesized treatment effect is observed, which, as will be reported below, is indeed the case, the more frequent fully revealing outcome in Game 1$M2$ will be net of the impact of the differential coordination difficulties.

analysis in Section 2.2. We refer "the sender is equally likely to be of either type" as the "pooling meaning" and will later on use the empirical frequencies of types conditional on messages to establish the endogenous meanings of messages.

HYPOTHESIS 4 (Effects of the Evolution of Meanings of Credible and Noncredible Neologisms).

(a) *Evolved Meanings of the Third Messages*: *In Game* $1E$*, the endogenous meaning of "&" evolves to be pooling, which is a new meaning not present before the message is introduced*; *in Game* $2E$*, the endogenous meaning of "&" does not evolve to be pooling*; *in Game* $1E^d$*, the endogenous meaning of "my type is* s*" coincides with its literal meaning and is therefore never pooling.*

(b) *Uses of the Third Messages*: *The frequency of the third message being sent is higher in Game* $1E$ *than in Games* $2E$ *and* $1E^d$.

(c) *Effects of the Introduction of the Third Messages*: *In Game* $1E$*, the frequency of fully revealing outcome after the introduction of "&" is lower than that before it is introduced*; *in Games* $2E$ *and* $1E^d$*, there are no differences in such frequencies before and after the respective third messages, "&" and "my type is* s*," are introduced.*

3.2 *Procedures*

Our experiment is conducted in English using z-Tree (Fischbacher (2007)) at the Hong Kong University of Science and Technology. A total of 292 subjects participate in seven treatments. The subjects, all without prior experience with our experiment, are recruited from the undergraduate population of the university. Upon arrival at the laboratory, subjects are instructed to sit at separate computer terminals. Each receives a copy of the experimental instructions. The instructions are read aloud using slide illustrations as an aid. A comprehension quiz and a practice round follow.

The two sets of treatments follow different procedures and have different instructions. In the following, we describe the essences of the instructions.[25]

Treatments with literal messages Two sessions are conducted for each game with literal messages. Two matching groups participate in a session. A matching group consists of 10 subjects, five as senders (Member As) and five as receivers (Member Bs). *Random matching* allowing repeating partners is used in each matching group, where senders and receivers are matched within groups. Viewing each matching group as an independent observation, we thus have four observations per game, providing sufficient statistical power for nonparametric tests.

In each session, subjects participate in 20 rounds of decision making under a single treatment condition, that is, *between-subject* design is used for these treatments. At the beginning of a session, half of the subjects are randomly assigned the role of Member A

[25]Appendix C in the Supplementary Material (Lai and Lim (2018)) contains the instructions for Games $1M3$ (with literal messages) and $1E$ (with meaningless messages). The instructions for the other games are similar.

and the other half the role of Member B. The role assignment remains fixed throughout the session.

In order to elicit a rich set of information regarding how subjects make decisions, we employ the *strategy method* and elicit beliefs. Subjects are told that there is a random variable, X, whose integer value range from 1 to 100 with equal probabilities. At the beginning of each round, the Member A in each group is asked what message he/she would send to the paired Member B if $X > 50$ (corresponding to $\theta = s$) and if $X \leq 50$ (corresponding to $\theta = t$). For the games with three messages, the available messages are "X is bigger than 50," "X is smaller than or equal to 50," and "I won't tell you" (the last message is omitted in the games with two messages). The Member B in each group is asked what action he/she would take, namely, L, C, or R, upon receiving each of the two or three messages from the paired Member A.

After the members furnish their strategies, the value of X will be realized and revealed to them. Their strategies are then implemented based on the realized X, and the rewards for the round will be determined. Having known the implemented message, Member A in each group is further asked to predict the paired Member B's action choice. A correct prediction is rewarded with two payoff points. Feedback on choices and rewards, but not on strategies, is provided at the end of each round.

For these treatments with 20 rounds, we randomly select two rounds for payments. The sum of the payoffs a subject earns in the two selected rounds is converted into Hong Kong dollars at a fixed and known exchange rate of HK$1 per payoff point. A show-up fee of HK$30 is also provided. Subjects on average earn HK$76.8 (≈US$9.85) by participating in a session that lasts less than an hour.[26]

Treatments with a priori meaningless messages Two sessions are conducted for each game with meaningless messages. Four to six matching groups participate in a session.[27] In order to expedite the emergence of meanings while maintaining the use of random matching, we reduce the size of a matching group to four subjects, two as Member As and two as Member Bs. The smaller matching groups also allow us to have more observations without increasing the number of sessions per game. Viewing each matching group as an independent observation, we have 8–12 observations per game in these treatments.

In each session, subjects participate in 40 rounds of decision making under a single treatment condition with a *within-subject variation of the message spaces* after 20 rounds. The role-assigning procedure is the same as that for the treatments with literal messages.

Given the focus on the emergence of meanings, we adopt the *choice method* for these treatments, which allows us to collect sufficient information for the purpose. We also do not elicit beliefs.[28] Subjects are told that there is a random variable, X, whose inte-

[26]Under the Hong Kong currency board system, the HK dollar is pegged to the US dollar at the rate of HK$7.8 to US$1.

[27]All sessions of Games $1E$ and $2E$ are participated by six matching groups, while one session of Game $1E^d$ is participated by five matching groups and the other by four matching groups.

[28]While we draw analogy between the two sets of treatments, we do not perform direct comparative statics between the games with and without literal messages. The different use of decision-elicitation method

ger value will be drawn from 1 to 100 with equal probabilities. At the beginning of each round, the Member A in each group is privately informed about whether the randomly drawn X is larger than 50 (corresponding to $\theta = s$) or less than or equal to 50 (corresponding to $\theta = t$). In each of the 1st to the 20th round, the Member A is asked to send one of the two messages, "$" or "%," to the paired Member B after seeing the realized range of X.[29]

In the paper instructions, subjects are told that further instructions will be given for the last 20 rounds. Subjects are thus aware of the different instructions for the later rounds, but their decisions in the first 20 rounds will not be influenced by the details of the forthcoming instructions. After the 20th round, further instructions are given on subjects' screens. Subjects are told that an additional message, "&" ("my type is s" for Game $1E^d$), becomes available for Member As to send.[30] Everything else remains the same as in the first 20 rounds.

In each round, after receiving a message from the paired Member A, the Member B in each group chooses one of the three actions: L, C, or R. The rewards for the round are then determined based on the realized X and the Member B's action choice. Feedback on choices and rewards is provided at the end of each round.

Since subjects participate in 40 rounds of decision making in these treatments, we commensurately increase their average payments to reflect the extra time they spend. We randomly select three rounds, and the sum of the payoffs a subject earns in these three rounds is converted into Hong Kong Dollars at a fixed and known exchange rate of HK$1 per payoff point. A higher show-up fee of HK$40 is also provided. Subjects on average earn HK$105.88 (≈US$13.57) by participating in a session that lasts less than two hours.[31]

---

and the omission of belief elicitations therefore do not jeopardize the integrity of our design. Note that the different procedure used in these treatments with meaningless messages is also driven by a practical consideration. Since subjects are making 20 more rounds of decision, using the strategy method and eliciting beliefs will extend the time allocated to a session beyond the limit stipulated by our lab policy.

[29]In order to avoid subjects using the positions of the symbols "$" and "%" on their screens and in the instructions as focal points for their choices of messages, we randomize the relative positions of the two symbols, and this is made known to the subjects. On a Member A's decision screen, the positions of the buttons "$" and "%" are randomized in each round. We also prepare two sets of instructions, in which the two symbols are stated in different orders. We randomly distribute the instructions so that half of the subjects receives one set and the other half receives the other set.

[30]Figure C.5 in Appendix C in the Supplementary Material (Lai and Lim (2018)) presents the on-screen instruction.

[31]While our payment may appear to be on the low side, for these treatments subjects are participating in sessions that last less than two hours. Our average payment of US$13.57 is therefore close to a US benchmark: the current federal minimum wage at US$7.25 per hour. For the treatments with literal messages, in which a session lasts less than an hour, the average payment of US$9.85 is even higher. The relatively lower payments for the treatments with meaningless messages reflect that the efficient fully revealing outcomes are harder to obtain when meanings have to emerge endogenously, not necessarily that, ex ante, monetary incentives are not sufficiently provided. For a local reference, a regular full meal at HKUST costs about US$4; we believe that two full meals per an hour time should provide reasonable incentives to motivate our student-subjects.

4. Experimental findings

Section 4.1 reports findings from our first set of treatments with literal messages. Section 4.2 is devoted to the treatments with a priori meaningless messages. We present our main findings by stating them in summary forms, which are each followed by supporting evidence.

4.1 *Games with literal messages*

For the games with literal messages for which the strategy method is used, we report two sets of findings: first examining the *on-path plays* and then the *elicited strategies and beliefs*.

On-path plays To extract on-path plays from our data, we apply the elicited strategies to the realized X to determine the realized path of play in each instance of the decision making. We then use these realized plays to establish the frequencies of fully revealing outcomes, which will be used to evaluate the experimental hypotheses. In addition, senders' and receivers' behavior on the realized paths will also be separately analyzed.

Figure 1 presents the round-by-round frequencies of fully revealing outcomes. The frequencies are obtained by measuring how often the receivers best respond to senders' *truthful* messages on the realized paths, taking their ex post ideal actions.[32] We start by

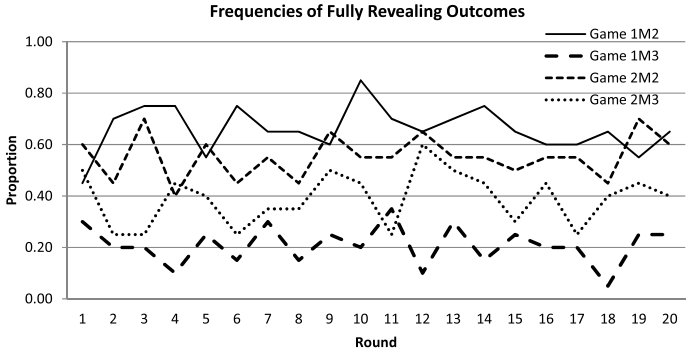


Figure 1. Frequencies of fully revealing outcomes, Games $1M2$, $1M3$, $2M2$, and $2M3$.

[32]For equilibria in cheap-talk games, a mapping from types to actions can typically be supported by different equilibria with different uses of messages. This inessential multiplicity has led practitioners to define equilibrium outcomes in cheap-talk games as mappings from types to actions, omitting messages. A parallel approach to measure information transmission in the lab would entail counting just how often receivers take their ex post ideal actions. With our use of literal messages, we nevertheless count the more restrictive scenarios that take messages into account. The fully revealing outcomes so defined, which should be more accurately called the *truth-telling* outcomes, reflect our expectation that subjects utilize the literal meanings. The more restrictive measure also allows us to gauge *how* information transmission is achieved. Refer to Figure A.1 in Appendix A.1 in the Supplementary Material (Lai and Lim (2018)) for the frequencies of fully revealing outcomes that exclude messages. An explicit distinction between truth-telling and fully revealing will be made in our analysis of elicited strategies below.

comparing Games $2M2$ and $2M3$. Supporting Hypothesis 1, our first finding suggests that the introduction of a noncredible neologism brings no statistically significant effect on information transmission outcomes:

FINDING 1 (Effect of the Existence of Noncredible Neologisms). *Consistent with Hypothesis* 1, *there is no significant difference in the frequencies of fully revealing outcome between Games* $2M2$ *and* $2M3$.

Using the independent group-level-all-round data, we fail to reject the null hypothesis of there being no difference in the frequencies of fully revealing outcome between Games $2M2$ and $2M3$, which are, respectively, 55% and 39% (two-sided $p = 0.2$, Mann–Whitney test).[33,34]

We next compare Game $2M3$ and $1M3$. Consistent with our premise that equilibria surviving neologism-proofness are more likely to be played in the laboratory, our second finding supports Hypothesis 2:

FINDING 2 (Effect of the Credibility of Neologisms). *Consistent with Hypothesis* 2, *the frequency of fully revealing outcome is significantly higher in Game* $2M3$ *than in Game* $1M3$.

The frequency of fully revealing outcome is 21% in Game $1M3$, lower than the 39% in Game $2M3$ with statistical significance ($p = 0.041$, Mann–Whitney test).

Our last finding on realized information transmission outcomes, which compares Games $1M3$ and $1M2$, supports Hypothesis 3:

FINDING 3 (Effect of the Existence and Credibility of Neologisms). *Consistent with Hypothesis* 3, *the frequency of fully revealing outcome is significantly lower in Game* $1M3$ *than in Game* $1M2$.

The frequency of fully revealing outcome in Game $1M2$ is 66%, significantly higher than the 21% in Game $1M3$ ($p = 0.015$, Mann–Whitney test).

Despite supporting the hypotheses, these frequencies of fully revealing outcome are admittedly on the low side. This may call into question the predictive power of equilibrium and its refinement, the very subject matter of our inquiry. Our defenses are two. First, we note that even with reasonably high instances of equilibrium behavior by the senders and the receivers, the resulting frequencies of fully revealing outcomes will be substantially lower.[35] Second, we emphasize that our focus is on comparative statics,

[33]As Figure 1 reveals, there is no notable learning and convergence in subjects' behavior. We therefore use all-round data for our statistical tests. We consider a comparison with $p \leq 0.05$ as being statistically significant. Unless otherwise indicated, the reported p values are from one-sided tests.

[34]The effect of the introduction of a noncredible neologism is even more insignificant when messages are excluded from the calculation of the frequencies of fully revealing outcome: the frequency rises to 51% in Game $2M3$, compared to the same 55% in Game $2M2$ (two-sided $p = 0.486$, Mann–Whitney test).

[35]For example, suppose that the senders truthfully reveal 70% of the time and the receivers optimally respond 70% of the time. The resulting frequency of fully revealing outcome will be 49%.

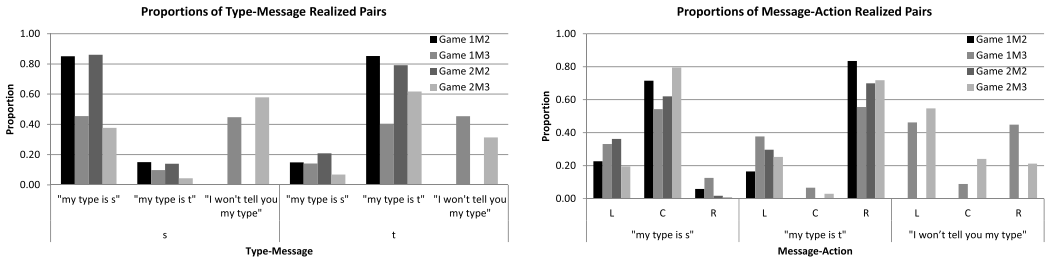
not the absolute levels of how often fully revealing equilibria are played; while some subjects might be driven by factors other than the equilibrium incentives, the observed variations across treatments are presumably obtained by holding any random or otherwise driven behavior constant. In fact, as the ensuing analysis of subjects' behavior shows, the observed information transmission outcomes are overall a result of purposeful behavior.

We proceed to analyze realized behavior, starting with the senders. We obtain the following finding:

FINDING 4 (Senders' Realized Behavior). *The introduction of a neologism, credible or not, attracts senders' deviations from truth-telling behavior. The profiles of the deviations are, however, consistent with the self-signaling properties of the neologisms: a noncredible "I won't tell you my type" attracts uneven deviations from the two types, while the two types deviate with the same frequencies under a credible "I won't tell you my type."*

Figure 2(a) presents the frequencies of messages conditional on types.[36] Using the observation from Game 1$M2$ as the point of reference, the credible neologism introduced in Game 1$M3$ attracts senders' deviations. The senders in Game 1$M2$ engage in truth-telling, where s-types send "my type is s" and t-types send "my type is t," both at a frequency of 85%. In Game 1$M3$, the frequency of s-types sending "my type is s" drops to 45%, and the frequency of t-types sending "my type is t" drops to 40%. Furthermore, the frequencies of the two types' sending the neologism "I won't tell you my type" are the same at 45%.

A similar comparison between Games 2$M2$ and 2$M3$ indicates that the neologism introduced in Game 2$M3$, though not credible, also attract senders' deviations. The frequencies of truthful messages in Game 2$M2$ are 86% for s-types and 79% for t-types. In Game 2$M3$, the corresponding frequencies drop to 38% and 61%; at the same time, the



(a) Type-Message (b) Message-Action

FIGURE 2. Senders' and receivers' realized behavior, Games 1$M2$, 1$M3$, 2$M2$, and 2$M3$.

[36]To illustrate how these frequencies are calculated based on on-path plays, suppose that in a round a Member A (sender) furnishes the following strategy: send "X is bigger than 50" if X $>$ 50 and send "X is smaller than or equal to 50" if X $\leq$ 50. If it turns out that in this round the realized X $>$ 50, then one data point of "my type is s" being sent by s will be recorded.

frequencies of s-types and t-types sending "I won't tell you my type" are, respectively, 58% and 31%.

The different effects of credible and noncredible neologisms on senders' behavior can be demonstrated by directly comparing Games $2M3$ and $1M3$. Although the deviations occurred in Game $2M3$ are not predicted by neologism-proofness, the different deviation profiles observed in the two games are consistent with the self-signaling properties of their respective neologisms. Theoretically, we obtain the following characterization: in Game $2M3$ only s has an incentive to send the noncredible "I won't tell you my type," while both types in Game $1M3$ have an incentive to send the credible version of the neologism. Empirically, we obtain the following observation: in Game $2M3$ s-types send "I won't tell you my type" more often than t-types, while in Game $1M3$ both types send the neologism with the same frequencies.

Turning to receivers' realized behavior, we obtain the following finding:

FINDING 5 (Receivers' Realized Behavior). *The introduction of a credible neologism attracts receivers' deviations from truth-telling responses, while the introduction of a noncredible neologism results in more truth-telling responses. Furthermore, receivers' behavior is consistent with senders' different deviating incentives under credible and noncredible neologisms.*

Figure 2(b) presents the frequencies of actions conditional on messages.[37] Using the observation from Game $1M2$ as the point of reference, the credible neologism introduced in Game $1M3$ attracts receivers' deviations. High frequencies of truth-telling responses, in which the receivers best respond to the literal meanings of "my type is s" and "my type is t," are observed in Game $1M2$: the frequencies of action C conditional on "my type is s" and of R conditional on "my type is t" are, respectively, 71% and 84%. In Game $1M3$, the corresponding frequencies drop to 54% and 56%. The higher frequency of the receivers' truth-telling responses in Game $1M2$, coupled with the senders' more frequent truth-telling behavior, accounts for the higher frequency of fully revealing outcome in Game $1M2$ than in Game $1M3$ (Finding 3).

A similar comparison between Games $2M2$ and $2M3$ indicates that the noncredible neologism introduced in Game $2M3$ does not attract deviating behavior by the receivers like it does for the senders. In fact, the introduction of "I won't tell you my type" contributes to more truth-telling responses: the frequencies of C conditional on "my type is s" and of R conditional on "my type is t" are, respectively, 62% and 70% in Game $2M2$, while in Game $2M3$ the corresponding frequencies increase to 80% and 72%.

While it may appear puzzling that the extra choice of messages in Game $2M3$ leads to more equilibrium-consistent behavior, note that s-types' deviating incentive to send

[37] To illustrate how these frequencies are calculated based on on-path plays, suppose that in a round a Member B (receiver) furnishes the following strategy: take action C after receiving message "X is bigger than 50," take R after receiving "X is smaller or equal to 50," and take L after receiving "I won't tell you." If it turns out that in this round the realized $X > 50$ and the Member A in the group chooses to send "X is bigger than 50" for this contingency, then one data point of action C being taken after "my type is s" is received will be recorded.

"I won't tell you my type" makes "my type is s" more trustworthy.[38] The apparently puzzling finding in fact reveals that the receivers understand this incentive. To the extent that actions are a proxy of beliefs, the higher frequencies of C in Game $2M3$ than in Game $2M2$ suggest that the receivers in Game $2M3$ are more likely to believe that an implemented "my type is s" is sent by an s-type. No comparable difference is observed in the case of "my type is t," consistent with the fact that t-types have no similar deviating incentive to send the noncredible neologism.

To reconcile subjects' behavior in Games $2M2$ and $2M3$ with the observed information transmission outcomes, we note that while the receivers in Game $2M3$ best respond to the literal meanings more often than the receivers in Game $2M2$, the senders in Game $2M3$ less frequently encode the messages using the literal meanings. Consequently, the receivers in Game $2M3$ identify senders' types via truth-telling messages less often than the receivers in Game $2M2$, although the difference is not statistically significant (Finding 1).

The "difference-in-difference" in the findings between Games $2M2$–$2M3$ and Games $1M2$–$1M3$ points to the contrasting effects of credible and noncredible neologisms on receivers, which can be further highlighted by directly comparing Games $2M3$ and $1M3$. When the neologism is not credible, the frequencies of the receivers' best responding to the literal meanings are 80% for "my type is s" and 72% for "my type is t"; when the neologism is credible, the corresponding frequencies drop to 54% and 56%. Despite the different profiles of senders' deviations (Finding 4), the total frequencies of senders' truth-telling behavior are close in the two games (43% in Game $1M3$ and 50% in Game $2M3$). This suggests that the less frequent fully revealing outcome observed in Game $1M3$ than in Game $2M3$ (Finding 2) is attributed to the more frequent receivers' deviations in Game $1M3$.

We conclude our on-path-play analysis by examining subjects' payoffs, which provide an additional gauge of realized plays to evaluate our hypotheses. Hypothesis 1 suggests that both senders and receivers should receive similar payoffs across Games $2M2$ and $2M3$, which is indeed observed: senders' average payoffs are 21.92 in Game $2M2$ and 22.54 in Game $2M3$ (two-sided $p = 0.886$, Mann–Whitney test); receivers' average payoffs are 23.66 in Game $2M2$ and 22.94 in Game $2M3$ (two-sided $p = 0.686$, Mann–Whitney test).

Suppose that whenever not playing a nonneologism-proof fully revealing equilibrium, subjects play a babbling equilibrium. Hypotheses 2 and 3 then suggest that the senders in Game $1M3$ should receive higher payoffs than those in Games $2M3$ and $1M2$, while the opposite should be true for the receivers. The lower receivers' payoffs in Game $1M3$ are observed: receivers' average payoff is 19.29 in Game $1M3$, significantly lower than the 22.94 in Game $2M3$ and the 24.96 in Game $1M2$ ($p \leq 0.029$, Mann–Whitney tests). The higher senders' payoffs in Game $1M3$ are, however, not observed: senders' average payoff is 19.48 in Game $1M3$, in fact lower than the 22.54 in Game $2M3$ and the

[38] In Game $2M3$, the frequency of L conditional on "I won't tell you my type" is 55%. This best response to the literal meaning of the message gives s-types a payoff of 50, the maximum they can get in the game; even though in theory "I won't tell you my type" is not a credible neologism, s-types' observed deviations are incentivized by a "success rate" of more than 50%.

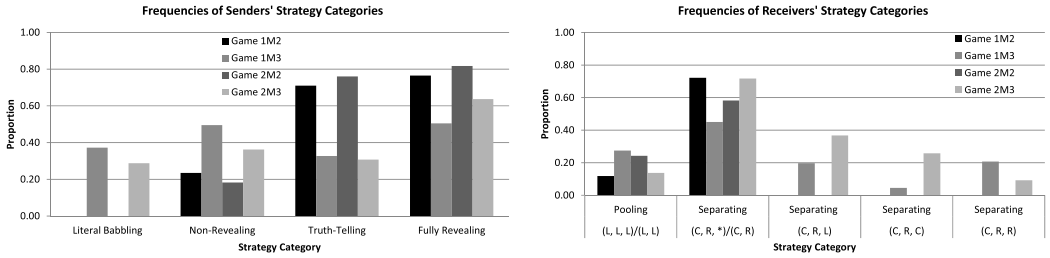
19.75 in Game 1M2. When some subjects in Game 1M3 fail to play a fully revealing equilibrium, they miscommunicate instead of playing a babbling equilibrium. The frequencies of nonideal actions, R conditional on s and C conditional on t, are, respectively, 34% and 18% in Game 1M3, compared to 15% and 12% in Game 2M3 and 16% and 10% in Game 1M2. This miscommunication, in which s-types receive 0 and t-types receive 8, accounts for the lower payoffs of the senders in Game 1M3.[39]

Elicited strategies and beliefs An analysis of elicited strategies provide further support for the on-path-play findings by including data on unrealized, hypothetical choices. Examining the frequencies of different kinds of strategies adopted by subjects, we obtain the following finding:

FINDING 6 (Elicited Strategies). *An analysis of elicited strategies indicates that the realized on-path plays are results of subjects playing the corresponding strategies.*

We classify senders' strategies into four categories: *literal babbling* ("I won't tell you my type" chosen for both s and t), *nonrevealing* (the same message chosen for both s and t), *truth-telling* ("My type is s" chosen for s and "My type is t" chosen for t), and *fully revealing* (different messages chosen for s and t).[40] For receivers' strategies, we classify them into two categories: *pooling* (action L chosen for all available messages) and *separating* (C chosen for "my type is s," R chosen for "my type is t," and, for three-message games, any one of the actions chosen for "I won't tell you my type").

Figure 3(a) presents the frequencies of senders' strategy categories. The on-path-play finding that both credible and noncredible neologisms attract senders' deviations is reflected in the lower frequencies of truth-telling strategies in the three-message games. The frequencies are 33% in Game 1M3 and 31% in Game 2M3, compared to, respectively, 71% in Game 1M2 and 76% in Game 2M2. The frequencies of nonrevealing strategies are commensurately higher in Games 1M3 and 2M3 (50% and 36%) than in Games 1M2 and 2M2 (24% and 18%).



(a) Senders' Strategy Categories (b) Receivers' Strategy Categories

FIGURE 3. Senders' and receivers' strategies, Games 1M2, 1M3, 2M2, and 2M3.

[39]Although the lower average payoffs of the receivers in Game 1M3 compared to those in Games 2M3 and 1M2 are consistent with the payoff implications of Hypotheses 2 and 3, the miscommunication, in which the receivers similarly receive either 0 or 8, also accounts for part of the lower receivers' payoffs.

[40]Note that some of these categories are not mutually exclusive: a literal babbling strategy is also nonrevealing, while a truth-telling strategy is also fully revealing.

Two aspects of senders' strategies echo the different profiles of deviations observed under credible and noncredible neologisms. First, in Game 2M3, there is a relatively large gap between the frequencies of truth-telling and fully revealing strategies (31% vs. 64%). The difference suggests that some senders choose to send "I won't tell you my type" for one type and another message for the other type, which is consistent with the on-path-play finding that in the majority of realized plays s-types send the neologism and t-types send "My type is t." Second, there is more frequent uses of literal babbling strategy in Game 1M3 than in Game 2M3 (37% vs. 29%). This is consistent with the on-path-play finding that the neologism is more evenly sent by the two types in Game 1M3 than in Game 2M3.

Figure 3(b) presents the frequencies of receivers' strategy categories.[41] The on-path-play finding that the introduction of a credible neologism attracts receivers' deviations is reflected in the lower frequency of separating strategies in Game 1M3 than in Game 1M2 (45% vs. 72%). The contrasting finding that a noncredible neologism results in more truth-telling responses is also echoed in the higher frequency of separating strategies in Game 2M3 than in Game 2M2 (72% vs. 58%). The additional contingency for "I won't tell you my type" in Game 2M3 facilitates the receivers' adoptions of separating strategies.

Note that the different on-path plays as well as strategies observed under two alternative sizes of message spaces undermine the idea that randomization is a natural empirical tendency. This lends force to the empirical plausibility of Farrell's (1993) assumed existence of unsent messages, which, as discussed in Section 2.1, is an essential element of the refinement concept.[42]

We proceed to analyze senders' elicited beliefs. We obtain the following finding:

FINDING 7 (Elicited Beliefs). *An analysis of senders' elicited beliefs indicates that observed behavior is overall a manifestation of purposeful decisions*: *senders' strategies are, to varying degrees, best responses to their beliefs, and their beliefs are in turn qualitatively consistent with receivers' actual responses in the cases of information revelation*.

Figure 4 presents senders' predictions of receivers' responses. To illustrate how accurate the predicted responses are, the realized receivers' responses are also reported side by side. The responses are classified into either separating or pooling, and their frequencies are calculated conditional on the senders' strategy categories.[43]

[41]Other than the two broad categories, the figure further breaks down the separating strategies into three cases, depending on which action is chosen for "I won't tell you my type."

[42]An analysis at the individual level further reveals that the majority of subjects quickly converges to a particular strategy (or class of strategies) without engaging in systematic randomizations over rounds. For senders, 90% of the subjects in Games 1M2 and 2M2 and 65% of the subjects in Games 1M3 and 2M3 stay with the same contingency plans in at least 8 rounds out of the last 10 rounds. For receivers, 90% of the subjects in Game 1M2, 75% of the subjects in Game 2M2, 70% of the subjects in Game 1M3, and 80% of the subjects in Game 2M3 stay with the same classes of contingency plans (with possible variations with respect to the actions chosen for "I won't tell you my type") in at least 8 rounds out of the last 10 rounds.

[43]Senders' beliefs are elicited after they are informed about their types, where each sender is asked to predict the paired receiver's response to the implemented message. A predicted response of either C or R is classified as separating, and a predicted L is classified as pooling. To maintain consistency in the comparison, the realized separating and pooling responses are defined similarly using receivers' on-path responses.

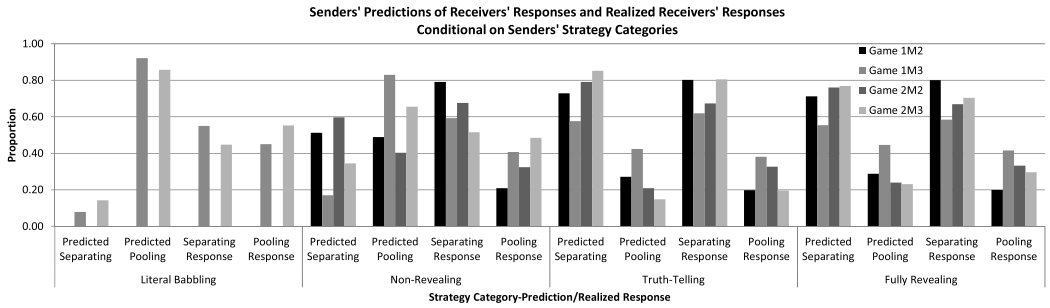


FIGURE 4. Senders' predictions of receivers' responses and realized receivers' responses, Games $1M2$, $1M3$, $2M2$, and $2M3$.

Senders' adopted strategies are, to varying degrees, consistent with their elicited beliefs. In all four games, more than 50% of predicted receivers' responses, ranging from 58% to 92%, coincide with the best responses to the senders' adopted literal babbling or truth-telling strategies. In the cases where the senders adopt truth-telling and fully revealing strategies, the profiles of predicted responses also follow similar patterns as those of realized receivers' responses.

We note a unique finding from Game $1M3$, which shows how best responding combines with a lexicographic preference for truth-telling to shape certain senders' behavior. Figure 4 reveals that substantial proportions up to 45% of the predicted responses in Game $1M3$ are pooling when the senders adopt truth-telling or fully revealing strategies. Further analysis shows that when the implemented messages are "my type is s" and "my type is t," the frequencies of predicted L are, respectively, 54% and 41%. These pooling predictions indicate that some senders in Game $1M3$ anticipate that the receivers will ignore their messages under the credible neologism. Even though they adopt revealing strategies, these senders are still best responding, because sending any message is a best response when it is expected to be ignored.[44] The senders' truthful behavior in these cases suggests that a lexicographic preference for truth-telling operates to break the indifference when different messages are expected to induce the same action.[45]

The above finding suggests that a secondary preference for truth-telling may complement the equilibrium incentives of neologisms in driving subjects' behavior. In light of prior experimental evidence, we conclude this section by further discussing whether a related behavioral concept—lying aversion—may substitute for neologism-proofness in explaining our data. Sánchez-Pagés and Vorsatz (2009) find that in a sender-receiver game with two types and two messages, senders engage in truth-telling; when the message space is augmented with a third message presented as "no message," parallel to our "I won't tell you my type," senders deviate from truthful behavior. Such a change in

[44]The logic can be best exemplified by a babbling equilibrium in which the sender randomizes uniformly over all available messages. In this equilibrium, the receiver ignores the message, and the sender—being then indifferent over all available messages—randomizes over them.

[45]For a few recent theoretical treatments of lexicographic preferences for truth-telling, see, for example, Demichelis and Weibull (2008), Ellingsen and Östling (2010), and Kartik, Tercieux, and Holden (2014).

senders' behavior in the presence of a new message choice, the use of which does not constitute outright lying, provides evidence for lying aversion at work.

While lying aversion provides a behavioral motive for the sender, neologism-proofness is a more comprehensive, equilibrium concept that covers the sender's and the receiver's behavior. Although lying aversion predicts the same as neologism-proofness for the more frequent truth-telling behavior in our two-message games, the behavioral concept is silent on how the receivers' would respond. More importantly, even if we restrict attention to just the senders, lying aversion applied to our findings is limited to accounting for the less truthful behavior in the presence of three messages; neologism-proofness, on the other hand, provides a finer prediction accounting for the different qualitative profiles of deviations by the two sender's types observed across our games. We thus view lying aversion as a complementary but not a substituting concept for interpreting our findings.[46]

4.2 *Games with a priori meaningless messages*

Our report of the findings from the second set of treatments revolves around the three predicted items in Hypothesis 4. It should not be surprising that in the absence of a pre-existing common language, there is a greater variety in the uses and interpretations of messages across the different observations (i.e., matching groups) of a treatment. Amid this heterogeneity, rather than seeking a general regularity in aggregate behavior, our main focus is on individual group behavior, where we look for the *existence* of observations that are consistent with the predictions.

We begin by addressing Hypothesis 4(a)–(b), evaluating the meanings and uses of the third messages. In order to investigate whether a new meaning emerges for a message introduced in the midst of the experiment, it is imperative to first ascertain the endogenous meanings of the initial messages. For each game, we first identify the matching groups in which separation occurs and distinct meanings emerge for "\$" and "%" in the first 20 rounds. Evaluating whether the theory works when it should, we then restrict attention to these singled-out groups and examine whether subjects' behavior from the last 20 rounds are consistent with the predictions.[47]

For the initial meanings, we obtain the following finding:

FINDING 8 (Meanings in the First 20 Rounds). *In the first* 20 *rounds of the games with a priori meaningless messages, separation occurs and distinct meanings can be established for messages "*\$*" and "%" in* 50–78% *of the observations. Furthermore, alternative meanings for a given message are observed in different matching groups.*

[46]Yet another behaviorial model is the level-k reasoning. Appendix B in the Supplementary Material (Lai and Lim (2018)) provides a level-k analysis in which we explore an alternative rationalization of the findings from our first set of treatments. While again complementarity exists between neologism-proofness and the level-k approach, a specification of level-k model commonly found in the communication-game literature falls short of explaining some of our findings from Games 1M3 and 2M2.

[47]We thank an anonymous referee for suggesting that we focus on the groups with meanings emerged in the initial rounds as the "denominator" for evaluating our prediction with respect to the introduction of the third message.

TABLE 3. Frequencies of Messages Conditional on Types and of Types Conditional on Messages, Games $1E$, $2E$, and $1E^d$, First 20 Rounds.

Game/Observation	Frequency(Message|Type)				Frequency(Type|Message)					
(Session-Group)	s		t		"\$"			"%"		
Game 1*E*	"\$"	"%"	"\$"	"%"	Sent	s	t	Sent	s	t
1-1	0.83	0.17	0.50	0.50	0.65	0.58	0.42	0.35	0.21	0.79
1-2	0.62	0.38	0.63	0.37	0.62	0.52	0.48	0.38	0.53	0.47
1-3	0.55	0.45	0.44	0.56	0.50	0.60	0.40	0.50	0.50	0.50
1-4	0.11	0.89	0.77	0.23	0.48	0.11	0.89	0.52	0.76	0.24
1-5	0.24	0.76	0.52	0.48	0.40	0.25	0.75	0.60	0.54	0.46
1-6	0.30	0.70	0.71	0.29	0.48	0.37	0.63	0.52	0.76	0.24
2-1	0.33	0.67	0.27	0.73	0.30	0.50	0.50	0.70	0.43	0.57
2-2	0.00	1.00	0.85	0.15	0.43	0.00	1.00	0.57	0.87	0.13
2-3	0.56	0.44	0.79	0.21	0.70	0.32	0.68	0.30	0.58	0.42
2-4	0.22	0.78	0.71	0.29	0.43	0.29	0.71	0.57	0.78	0.22
2-5	0.67	0.33	0.32	0.68	0.50	0.70	0.30	0.50	0.35	0.65
2-6	0.10	0.90	0.42	0.58	0.25	0.20	0.80	0.75	0.63	0.37
Mean	–	–	–	–	0.48	–	–	0.52	–	–

Game/Observation	Frequency(Message|Type)				Frequency(Type|Message)					
(Session-Group)	s		t		"\$"			"%"		
Game 2*E*	"\$"	"%"	"\$"	"%"	Sent	s	t	Sent	s	t
1-1	0.24	0.76	0.78	0.22	0.55	0.18	0.82	0.45	0.72	0.28
1-2	0.14	0.86	0.67	0.33	0.38	0.20	0.80	0.62	0.76	0.24
1-3	0.05	0.95	0.68	0.32	0.35	0.07	0.93	0.65	0.77	0.23
1-4	0.05	0.95	0.89	0.11	0.45	0.06	0.94	0.55	0.91	0.09
1-5	0.10	0.90	0.95	0.05	0.50	0.10	0.90	0.50	0.95	0.05
1-6	0.00	1.00	0.95	0.05	0.45	0.00	1.00	0.55	0.95	0.05
2-1	0.26	0.74	0.76	0.24	0.48	0.32	0.68	0.52	0.81	0.19
2-2	0.40	0.60	0.56	0.44	0.50	0.30	0.70	0.50	0.45	0.55
2-3	0.25	0.75	0.62	0.38	0.48	0.21	0.79	0.52	0.57	0.43
2-4	0.52	0.48	0.41	0.59	0.48	0.63	0.37	0.52	0.52	0.48
2-5	0.00	1.00	1.00	0.00	0.50	0.00	1.00	0.50	1.00	0.00
2-6	0.00	1.00	1.00	0.00	0.62	0.00	1.00	0.38	1.00	0.00
Mean	–	–	–	–	0.48	–	–	0.52	–	–

Game/Observation	Frequency(Message|Type)				Frequency(Type|Message)					
(Session-Group)	s		t		"\$"			"%"		
Game 1*E*^{d}	"\$"	"%"	"\$"	"%"	Sent	s	t	Sent	s	t
1-1	0.89	0.11	0.29	0.71	0.57	0.74	0.26	0.43	0.12	0.88
1-2	0.29	0.71	0.89	0.11	0.57	0.26	0.74	0.43	0.88	0.12
1-3	0.58	0.42	0.29	0.71	0.43	0.65	0.35	0.57	0.35	0.65
1-4	1.00	0.00	0.05	0.95	0.52	0.95	0.05	0.48	0.00	1.00
2-1	0.05	0.95	0.55	0.45	0.30	0.08	0.92	0.70	0.68	0.32
2-2	0.11	0.89	0.86	0.14	0.52	0.10	0.90	0.48	0.84	0.16
2-3	0.47	0.53	1.00	0.00	0.75	0.30	0.70	0.25	1.00	0.00
2-4	0.00	1.00	1.00	0.00	0.52	0.00	1.00	0.48	1.00	0.00
2-5	0.40	0.60	0.60	0.40	0.50	0.40	0.60	0.50	0.60	0.40
Mean	–	–	–	–	0.52	–	–	0.48	–	–

Note: "Frequency(Message|Type)" measures how senders use messages. "Frequency(Type|Message)" is used to determine the endogenous meanings of messages implied by senders' uses of messages. The additional columns "Sent" provide the total frequency at which the particular message is sent by the senders (received by the receivers). Means are reported only for these "Sent" frequencies, because the alternative uses of messages in different matching groups render the means of other frequencies not informative about average behavior. The shaded observations represent those in which distinct meanings are established.

For each matching group in Games $1E$, $2E$, and $1E^d$, Table 3 reports the first-20-round frequencies of messages conditional on types and of types conditional on messages. The former frequency captures how senders use messages, while the latter helps determine the endogenous, empirical meanings. We apply a simple criterion to the latter frequency to determine whether distinct meanings emerge for the two initial messages.[48]

In Game $1E$, distinct meanings can be established in 50% of the observations (6 out of 12 matching groups). The singled-out observations are highlighted in the top panel of Table 3. Consider the strongest case, Observation 2-2 (Session-Group). Message "\$" is sent exclusively by t-types—the frequency of t conditional on "\$" is 100%—and 87% of the time message "%" is sent by s-types, which strongly establish the empirical meaning of "\$" to be t and that of "%" to be s.[49] In Game $2E$, distinct meanings can be established in 75% of the observations (9 out of 12 groups), which are highlighted in the middle panel of Table 3. Consider one of the strongest cases, Observation 2-5. Message "\$" is sent exclusively by t-types and "%" exclusively by s-types, which establish perfect, distinct meanings for the two messages. Finally, in Game $1E^d$, distinct meanings can be established in about 78% of the observations (seven out of nine groups), which are highlighted in the bottom panel of Table 3. The strongest case, Observation 2-4, also has distinct meanings perfectly established.

While the meaning mapping ["%" $\mapsto s$ and "\$" $\mapsto t$] is the dominant finding, the alternative mapping ["\$" $\mapsto s$ and "%" $\mapsto t$] is also observed. The dominant mapping accounts for 86% of all observations in which meanings can be established. The alternative mapping is established in Observation 2-5 from Game $1E$ and Observations 1-1 and 1-4 from Game $1E^d$. The existence of alternative meanings highlights the endogenous nature of meanings, especially when there is no preexisting common language.[50]

[48]Our criterion, though rather arbitrary, requires reasonably clear differentiation between the uses of the two messages while at the same time providing leeway for heterogeneous behavior. For each message, the type with the higher conditional frequency is considered the candidate for the meaning of the message. The *necessary condition* for distinct meanings is that the two candidates, one for each message, cannot be the same. (We use the empirical frequencies of s and t to calculate the conditional frequencies, which makes it possible that the higher conditional frequencies for both messages rest on the same type). In concluding whether the distinct candidates are indeed deemed to be the meanings, we further require the *sufficient condition* that for the two conditional frequencies, one for each candidate, the lower one is at least 60% and the higher at least 70%.

[49]Note how this can be reconciled with the frequencies of messages conditional on types via Bayes' rule. Table 3 reveals that 100% of the messages sent by s-types is "%" and 85% of the messages sent by t-types is "\$." Since s-types never send "\$" while t-types do, the frequency of t conditional on "\$" must be 100%. On the other hand, since "%" is sent by both types, but the frequency by s is higher, the frequency of s conditional on "%" should also be higher than that of t conditional on "%."

[50]The mapping ["%" $\mapsto s$ and "\$" $\mapsto t$] becomes the dominant meanings despite our effort to randomize the distributions of alternative instructions and the decision buttons on the screen. We speculate that the relative positions of "\$" and "%" on the keyboards, though not involved in the inputs of decisions, may provide a focal point for subjects to map "\$" to $X \leq 50$ ($\theta = t$) and "%" to $X > 50$ ($\theta = s$). This potential focal point would have been hard to avoid even if we had used other symbols for the messages. It does not, however, undermine the purpose of our design, which is to avoid focal points to be developed directly from the literal meanings of messages.

With these findings on initial meanings, we proceed to examine the meanings and uses of the third messages. Restricting our attention to the singled-out groups, we evaluate how often among these groups the pooling meaning emerges for the third message under a frequent use. We obtain the following finding, which is in line with Hypothesis 4(a)–(b):

FINDING 9 (Meanings and Uses of Messages in the Last 20 Rounds).

(a) *Evolved Meanings of the Third Messages*: *In Game* $1E$, *the pooling meaning is established for the third message "&" in* 50% *of the observations in which distinct meanings are established for the two initial messages* (*singled-out observations*); *in Game* $2E$, *the pooling meaning is established for "&" in only* 11% *of the singled-out observations*; *in Game* $1E^d$, *no pooling meaning is ever established for "my type is s" in the singled-out observations*.

(b) *Uses of the Third Messages*: *The average frequency of the third message being sent is higher in Game* $1E$ *than in Games* $2E$ *and* $1E^d$.

Table 4 reports the last-20-round frequencies of messages conditional on types and of types conditional on messages. We apply a simple criterion to the latter frequency to determine whether the endogenous meanings of the third messages are pooling.[51]

Recall that in Game $1E$ "&" is predicted to be credible with respect to its evolved meaning. Given that separation has occurred with the two initial messages, it pays for both s and t to send "&" when it becomes available, and this gives rise to a new endogenous, pooling meaning for "&". In the data, the predicted pooling meaning is established for "&" in 3 out of the 6 singled-out observations, which are highlighted with a darker shade in the top panel of Table 4. Consider the strongest case, Observation 2-2. The frequencies of s and t conditional on "&" are within 5% of the uniform distribution (resp., 55% and 45%), and the message is sent 49% of the time. These two frequencies are also clearly different from their counterparts conditional on, respectively, "%" (87%) and "$" (100%) in the first 20 rounds, indicating that the pooling meaning is not present before "&" is introduced.

Recall the different prediction for Game $2E$. Given that separation has occurred with the two initial messages, "&" is not preferred by t and is only weakly preferred by s, depriving the neologism of the pooling meaning. In the data, the pooling meaning is established for "&" in only one out of the nine singled-out observations (Observation 1-3, highlighted with a darker shade in the middle panel of Table 4). In the remaining eight observations, either the frequency of the neologism is no greater than 30% (as low as 3% in Observations 1-2 and 2-5), or its established meaning coincides with the meaning

[51]To ensure that the classification is obtained with enough observations, our criterion for the pooling meaning requires not only reasonably close conditional frequencies of the two types but also a sufficiently frequent use of the third message, where we allow some limited tradeoff between these two subcriteria. We consider the endogenous meaning of a third message to be pooling if the frequencies of s and t conditional on the message are within $50\% \pm 10\%$ and the message is used (by both types) at least as often as it would be under complete random behavior, that is, $\frac{1}{3}$ of the time; if the conditional frequencies of s and t are beyond $50\% \pm 10\%$ but within $50\% \pm 20\%$, we require that the message be used at least $\frac{2}{3}$ of the time.

TABLE 4. Frequencies of Messages Conditional on Types and of Types Conditional on Messages, Games $1E$, $2E$, and $1E^d$, Last 20 Rounds.

Game/Observation (Session-Group)	Frequency(Message|Type)						Frequency(Type|Message)								
	s			t			"$"			"%"			"&"		
Game 1*E*	"$"	"%"	"&"	"$"	"%"	"&"	Sent	s	t	Sent	s	t	Sent	s	t
1-1	0.32	0.00	0.68	0.62	0.00	0.38	0.48	0.32	0.68	0.00	N/A	N/A	0.53	0.62	0.38
1-2	0.59	0.18	0.23	0.33	0.33	0.33	0.47	0.68	0.32	0.25	0.40	0.60	0.28	0.45	0.55
1-3	0.85	0.10	0.05	0.05	0.55	0.40	0.44	0.94	0.06	0.33	0.15	0.85	0.23	0.11	0.89
1-4	0.05	0.14	0.82	0.50	0.00	0.50	0.25	0.10	0.90	0.08	1.00	0.00	0.67	0.67	0.33
1-5	0.05	0.38	0.57	0.32	0.42	0.26	0.18	0.14	0.86	0.40	0.50	0.50	0.42	0.71	0.29
1-6	0.00	0.45	0.55	0.67	0.06	0.28	0.30	0.00	1.00	0.28	0.91	0.09	0.42	0.71	0.29
2-1	0.15	0.23	0.62	0.21	0.29	0.50	0.18	0.57	0.43	0.25	0.60	0.40	0.57	0.70	0.30
2-2	0.00	0.39	0.61	0.59	0.00	0.41	0.33	0.00	1.00	0.18	1.00	0.00	0.49	0.55	0.45
2-3	0.06	0.35	0.59	0.22	0.09	0.69	0.15	0.17	0.83	0.20	0.75	0.25	0.65	0.38	0.62
2-4	0.06	0.67	0.28	0.41	0.00	0.59	0.25	0.10	0.90	0.30	1.00	0.00	0.45	0.28	0.72
2-5	0.32	0.09	0.59	0.33	0.17	0.50	0.33	0.54	0.46	0.13	0.40	0.60	0.54	0.59	0.41
2-6	0.04	0.42	0.54	0.44	0.31	0.25	0.20	0.13	0.87	0.38	0.67	0.33	0.42	0.76	0.24
Mean	–	–	–	–	–	–	0.30	–	–	0.23	–	–	0.47	–	–

Game/Observation (Session-Group)	Frequency(Message|Type)						Frequency(Type|Message)								
	s			t			"$"			"%"			"&"		
Game 2*E*	"$"	"%"	"&"	"$"	"%"	"&"	Sent	s	t	Sent	s	t	Sent	s	t
1-1	0.05	0.62	0.33	0.63	0.11	0.26	0.33	0.08	0.92	0.37	0.87	0.13	0.30	0.58	0.42
1-2	0.00	0.95	0.05	0.75	0.25	0.00	0.38	0.00	1.00	0.59	0.79	0.21	0.03	1.00	0.00
1-3	0.00	0.36	0.64	0.62	0.00	0.38	0.40	0.00	1.00	0.13	1.00	0.00	0.47	0.47	0.53
1-4	0.00	0.76	0.24	0.83	0.00	0.17	0.47	0.00	1.00	0.33	1.00	0.00	0.20	0.50	0.50
1-5	0.00	0.57	0.43	1.00	0.00	0.00	0.65	0.00	1.00	0.20	1.00	0.00	0.15	1.00	0.00
1-6	0.04	0.70	0.26	0.76	0.18	0.06	0.35	0.07	0.93	0.47	0.84	0.16	0.18	0.86	0.14
2-1	0.12	0.59	0.29	0.69	0.09	0.22	0.45	0.11	0.89	0.30	0.83	0.17	0.25	0.50	0.50
2-2	0.16	0.42	0.42	0.38	0.13	0.49	0.25	0.40	0.60	0.30	0.83	0.17	0.45	0.56	0.44
2-3	0.16	0.68	0.16	0.66	0.05	0.29	0.43	0.18	0.82	0.35	0.93	0.07	0.23	0.33	0.67
2-4	0.65	0.30	0.05	0.25	0.45	0.30	0.44	0.72	0.28	0.38	0.40	0.60	0.18	0.14	0.86
2-5	0.00	1.00	0.00	0.96	0.00	0.04	0.59	0.00	1.00	0.38	1.00	0.00	0.03	0.00	1.00
2-6	0.00	0.24	0.76	0.68	0.00	0.32	0.33	0.00	1.00	0.13	1.00	0.00	0.54	0.73	0.27
Mean	–	–	–	–	–	–	0.42	–	–	0.33	–	–	0.25	–	–

Game/Observation (Session-Group)	Frequency(Message|Type)						Frequency(Type|Message)								
	s			t			"$"			"%"			"s"		
Game 1*E*^{d}	"$"	"%"	"s"	"$"	"%"	"s"	Sent	s	t	Sent	s	t	Sent	s	t
1-1	0.41	0.11	0.48	0.54	0.46	0.00	0.44	0.61	0.39	0.23	0.33	0.67	0.33	1.00	0.00
1-2	0.05	0.00	0.95	0.42	0.16	0.42	0.23	0.11	0.89	0.08	0.00	1.00	0.69	0.71	0.29
1-3	0.29	0.13	0.58	0.50	0.31	0.19	0.38	0.47	0.53	0.20	0.38	0.62	0.43	0.82	0.18
1-4	0.26	0.26	0.48	0.12	0.88	0.00	0.20	0.75	0.25	0.52	0.29	0.71	0.28	1.00	0.00
2-1	0.13	0.62	0.25	0.63	0.29	0.08	0.43	0.12	0.88	0.43	0.59	0.41	0.14	0.67	0.33
2-2	0.00	0.00	1.00	1.00	0.00	0.00	0.45	0.00	1.00	0.00	N/A	N/A	0.55	1.00	0.00
2-3	0.55	0.20	0.25	0.85	0.05	0.10	0.69	0.39	0.61	0.13	0.80	0.20	0.18	0.71	0.29
2-4	0.27	0.27	0.46	0.86	0.07	0.07	0.47	0.37	0.63	0.20	0.87	0.13	0.33	0.92	0.08
2-5	0.18	0.12	0.70	0.78	0.22	0.00	0.52	0.14	0.86	0.18	0.29	0.71	0.30	1.00	0.00
Mean	–	–	–	–	–	–	0.42	–	–	0.22	–	–	0.36	–	–

Note: "Frequency(Message|Type)" measures how senders use messages. "Frequency(Type|Message)" is used to determine the endogenous meanings of messages implied by senders' uses of messages; "N/A" indicates that this conditional frequency cannot be calculated because the message in question is not used at all. The additional columns "Sent" provide the total frequency at which the particular message is sent by the senders (received by the receivers). Means are reported only for these "Sent" frequencies, because the alternative uses of messages in different matching groups render the means of other frequencies not informative about average behavior. For Game $1E^d$, "s" refers to the message "my type is s." The lightly shaded observations represent those in which distinct meanings are established for the two messages in the first 20 rounds; the darkly shaded observations are the nested observations in which the pooling meaning is established for the third message in the last 20 rounds.

of one of the initial messages (perfect cases in Observations 1-2, 1-5, and 2-5), or both. These observed uses of "&" are consistent with the noncredible nature of the neologism, in which there is no opportunity for profitable deviations.

The finding that the pooling meaning is established for the third message in 50% of the singled-out observations in Game $1E$ but only in 11% of those in Game $2E$ reflects the different self-signaling properties of the neologisms. Consider further the control Game $1E^d$. In all its seven singled-out observations, the endogenous meaning of "my type is s" coincides with its literal meaning (perfect cases in Observations 1-1, 1-4, and 2-2). This very contrast with Game $1E$ provides evidence that our finding above is driven

by the meaning of the neologism, either literally given or evolved, not its mere introduction via the experimenter demand effect.

The varying self-signaling properties of the third messages are also reflected in how often they are sent. The average frequency of the credible "&" being sent is 47% in Game $1E$, significantly higher than the 25% of the noncredible "&" in Game $2E$ ($p < 0.01$, Mann–Whitney test). With marginal statistical significance, it is also higher than the 36% of the literal "my type is s" in Game $1E^d$ ($p = 0.058$, Mann–Whitney test).

We proceed to evaluate the information transmission outcomes.[52] We obtain the following finding:

FINDING 10 (Effects of the Introduction of the Third Messages). *In Game* $1E$, *the average frequency of fully revealing outcome in the first* 20 *rounds is higher than that in the last* 20 *rounds*; *in Games* $2E$ *and* $1E^d$, *the average frequency of fully revealing outcome in the first* 20 *rounds is lower than that in the last* 20 *rounds.*

Table 5 reports the frequencies of fully revealing outcomes. The frequencies, which are aggregated across the first 20 rounds and the last 20 rounds for each matching group, are obtained by measuring how often the receivers took their ex post ideal actions.[53]

We first illustrate with individual matching groups. Consider Observation 2-2 in Game $1E$, in which the frequency of fully revealing outcome is 70% in the first 20 rounds, the highest among all observations in the game. In the last 20 rounds after the third message "&" is introduced, the frequency drops to 55%. Contrast this with Observation 1-4 in Game $2E$. The frequency of fully revealing outcome in the first 20 rounds is 85%. In the last 20 rounds, the frequency increases to 95%. The neologism in Game $1E$, being credible with respect to its evolved meaning, upset the equilibrium play observed in the initial rounds in Observation 2-2. By contrast, the noncredible neologism in Game $2E$ does not result in less frequent play of fully revealing equilibrium after it is introduced in Observation 1-4.[54]

The qualitative differences observed in the matching groups highlighted above are also observed in the aggregate data, although the differences are not statistically significant to support Hypothesis 4(c). In Game $1E$, the average frequency of fully revealing outcome in the first 20 rounds is 49%, higher than the 45% in the last 20 rounds ($p = 0.194$, Wilcoxon signed-rank test). In Game $2E$, the average frequency of fully revealing outcome in the first 20 rounds is 53%, lower than the 56% in the last 20 rounds

[52]Given the senders' uses of messages, the attainment of an information transmission outcome depends on how the receivers respond to the implied meanings. See Appendix A.2 in the Supplementary Material (Lai and Lim (2018)) for an analysis of receivers' behavior in the games with meaningless messages. As in the case of senders, receivers' behavior reflects the different self-signaling properties of the third messages.

[53]Unlike the case of the games with literal messages, here we omit messages in measuring the fully revealing outcomes. With the meaningless messages, truthful messages or truth-telling equilibria are not defined. Refer to footnote 32 for the relevant discussion.

[54]The frequency of fully revealing outcome is in fact higher in the last 20 rounds, which points to a learning effect. If there is also some learning in Game $1E$ toward the fully revealing equilibrium, the true disrupting effect of the credible neologism will be larger than the observed disrupting effect.

TABLE 5. Frequencies of Fully Revealing Outcomes, Games $1E$, $2E$, and $1E^d$.

Game 1*E*	First 20 Rounds	Last 20 Rounds
1-1	0.33	0.23
1-2	0.50	0.28
1-3	0.55	0.65
1-4	0.53	0.45
1-5	0.48	0.43
1-6	0.53	0.68
2-1	0.48	0.30
2-2	0.70	0.55
2-3	0.35	0.40
2-4	0.53	0.68
2-5	0.30	0.20
2-6	0.58	0.55
Mean	0.49	0.45
Game 2*E*	First 20 Rounds	Last 20 Rounds
1-1	0.45	0.63
1-2	0.28	0.23
1-3	0.60	0.53
1-4	0.85	0.95
1-5	0.85	0.90
1-6	0.58	0.63
2-1	0.43	0.53
2-2	0.40	0.38
2-3	0.38	0.63
2-4	0.20	0.38
2-5	0.75	0.73
2-6	0.58	0.28
Mean	0.53	0.56
Game 1*E*^{d}	First 20 Rounds	Last 20 Rounds
1-1	0.40	0.40
1-2	0.65	0.48
1-3	0.28	0.50
1-4	0.68	0.48
2-1	0.05	0.33
2-2	0.80	1.00
2-3	0.55	0.45
2-4	0.98	0.80
2-5	0.25	0.60
Mean	0.51	0.56

Note: The frequency of fully revealing outcome is measured by the frequency at which the receivers took the ex post ideal actions, C for type s and R for type t. The shaded observations represent those in which distinct meanings are established for the two initial messages in the first 20 rounds.

($p = 0.127$, Wilcoxon signed-rank test). Note also that in the control Game $1E^d$, the average frequency of fully revealing outcome in the first 20 rounds is 51%, lower than the 56% in the last 20 rounds ($p = 0.181$, Wilcoxon signed-rank test).

With the wide variety of individual group behavior in the absence of a preexisting common language, we are unable to obtain aggregate behavior that supports the hypotheses with statistical significance. The fully revealing equilibrium are nevertheless able to predict *some* observed behavior when meanings have to be evolved. Furthermore, we find that the theory works reasonably well when it should: among the cases

where the fully revealing equilibrium predicts senders' uses of messages in the initial rounds, the self-signaling properties of the introduced neologisms predict senders' uses of messages in the later rounds.

5. Concluding remarks

We experimentally evaluate the first cheap-talk refinement, neologism-proofness. In our first set of treatments featuring literally meaningful messages, we find that whether a fully revealing equilibrium is neologism-proof affects how often it is played, lending support to the predictive power of the refinement concept. The self-signaling property of a neologism also predicts the qualitative profiles of senders' and receivers' behavior. In our second set of treatments featuring a priori meaningless messages, unsurprisingly we find that subjects have a harder time playing a fully revealing equilibrium and responding to the incentives of the neologisms. We nonetheless obtain observations that are consistent with the theoretical predictions, in particular the predicted evolution of meanings of credible and noncredible neologisms. These conforming observations demonstrate that sophisticated coordinated plays can be achieved by human subjects, in this case the evolution of equilibrium plays and the disruption to the equilibrium as predicted by a refinement concept, all in an environment without the anchor of a common language.

Our findings shed light on the capabilities and limitations of neologism-proofness as a predictive theory. While we view it as a remarkable finding that the refinement is able to predict behavior in the absence of a preexisting language, it happens only in selected matching groups. By contrast, individual group behavior in the presence of literal messages conforms fairly consistently to the major predictions of the refinement. This contrast validates from an empirical vantage point the importance of literal meanings in the applications of neologism-proofness.

A weakness of neologism-proofness is its silence about how messages will be used in equilibrium. One of our findings is that lexicographic preferences for truth-telling play a role in how subjects use messages. The empirical regularity suggests that a secondary preference defined with respect to literal meanings may warrant a place in the predictive theory of cheap-talk refinements. It may also provide a direction to address the shortcomings of neologism-proofness. We hope that our study can help inform future research in this regard.

References

Austen-Smith, D. (1990), "Information transmission in debate." *American Journal of Political Science*, 34, 124–152. [1456]

Banks, J. S., C. F. Camerer, and D. Porter (1994), "Experimental tests of Nash refinements in signaling games." *Games and Economic Behavior*, 6, 1–31. [1454, 1458]

Blume, A. (1993), "Neighborhood stability in sender-receiver games." *Games and Economic Behavior*, 13, 2–25. [1454]

Blume, A., D. V. Dejong, Y.-G. Kim, and G. B. Sprinkle (1998), "Experimental evidence on the evolution of the meaning of messages in sender-recevier games." *American Economic Review*, 88, 1323–1340. [1457, 1458, 1463]

Blume, A., D. V. Dejong, Y.-G. Kim, and G. B. Sprinkle (2001), "Evolution of communication with partial common interest." *Games and Economic Behavior*, 37, 79–120. [1454, 1456, 1457]

Blume, A., D. V. Dejong, and G. B. Sprinkle (2008), "The Effect of Message Space Size on Learning and Outcomes in Sender-Receiver Games." In *Handbook of Experimental Economics Results*, Elsevier.[1457]

Blume, A., Y.-G. Kim, and J. Sobel (1993), "Evolutionary stability in games of communication." *Games and Economic Behavior*, 5, 547–575. [1454, 1456]

Blume, A., E. K. Lai, and W. Lim (2017), "Strategic information transmission: A survey of experiments and theoretical foundations." Report. [1457]

Blume, A. and J. Sobel (1995), "Communication-proof equilibria in cheap-talk games." *Journal of Economic Theory*, 65, 359–382. [1454]

Brandts, J. and C. A. Holt (1992), "An experimental test of equilibrium dominance in signaling games." *American Economic Review*, 82, 1350–1365. [1454, 1458]

Cai, H. and J. T.-Y. Wang (2006), "Overcommunication in strategic information transmission games." *Games and Economic Behavior*, 56, 7–36. [1457]

Chen, Y., N. Kartik, and J. Sobel (2008), "Selecting cheap-talk equilibria." *Econometrica*, 76, 117–136. [1454]

Cho, I.-K. and D. M. Kreps (1987), "Signaling games and stable equilibria." *Quarterly Journal of Economics*, 102, 179–221. [1454, 1458]

Crawford, V. (1998), "A survey of experiments on communication via cheap talk." *Journal of Economic Theory*, 78, 286–298. [1457]

Crawford, V. P. and J. Sobel (1982), "Strategic information transmission." *Econometrica*, 50, 1431–1451. [1454, 1456]

de Groot Ruiz, A., T. Offerman, and S. Onderstal (2014), "For those about to talk we salute you: An experimental study of credible deviations and ACDC." *Experimental Economics*, 17, 173–199. [1454, 1457]

de Groot Ruiz, A., T. Offerman, and S. Onderstal (2015), "Equilibrium selection in experimental cheap talk games." *Games and Economic Behavior*, 91, 14–25. [1454, 1457]

Demichelis, S. and J. Weibull (2008), "Language, meaning, and games: A model of communication, coordination, and evolution." *American Economic Review*, 98, 1292–1311. [1476]

Dickhaut, J. W., K. A. McCabe, and A. Mukherji (1995), "An experimental study of strategic information transmission." *Economic Theory*, 6, 389–403. [1457]

Duffy, J., E. K. Lai, and W. Lim (2017), "Coordination via correlation: An experimental study." *Economic Theory*, 64, 265–304. [1458]

Ellingsen, T. and R. Östling (2010), "When does communication improve coordination?" *American Economic Review*, 100, 1695–1724. [1476]

Eső, P. and J. Schummer (2009), "Credible deivations from signaling equilibria." *International Journal of Game Theory*, 38, 411–430. [1454]

Farrell, J. (1993), "Meaning and credibility in cheap-talk games." *Games and Economics Behavior*, 5, 514–531. [1454, 1455, 1459, 1460, 1462, 1463, 1475]

Farrell, J. and R. Gibbons (1988), "Cheap talk, neologisms, and bargaining." Report. [1456]

Farrell, J. and R. Gibbons (1989), "Cheap talk can matter in bargaining." *Journal of Economic Theory*, 48, 221–237. [1456]

Farrell, J. and M. Rabin (1996), "Cheap talk." *Journal of Economic Perspectives*, 10, 103–118. [1456]

Fischbacher, U. (2007), "z-tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics*, 10, 171–178. [1466]

Gertner, R., R. Gibbons, and D. Scharfstein (1988), "Simultaneous signalling to the capital and product markets." *The RAND Journal of Economics*, 19, 173–190. [1456]

Gneezy, U. (2005), "Deception: The role of consequences." *American Economic Review*, 95, 384–394. [1457]

Grossman, S. J. and M. Perry (1986), "Perfect sequential equilibrium." *Journal of Economic Theory*, 39, 97–119. [1454]

Hurkens, S. and N. Kartik (2009), "Would I lie to you? On social preferences and lying aversion." *Experimental Economics*, 12, 180–192. [1457]

Kartik, N., O. Tercieux, and R. Holden (2014), "Simple mechanisms and preferences for honesty." *Games and Economic Behavior*, 83, 284–290. [1476]

Kawagoe, T. and H. Takizawa (2008), "Equilibrium refinement vs. level-k analysis: An experimental study of cheap-talk games with private information." *Games and Economics Behavior*, 66, 238–255. [1454, 1457, 1459, 1461]

Kim, K. and P. Kircher (2015), "Efficient competition through cheap talk: The case of competing auctions." *Econometrica*, 83, 1849–1875. [1456]

Lai, E. K. and W. Lim (2018), "Supplement to 'Meaning and credibility in experimental cheap-talk games'." *Quantitative Economics Supplemental Material*, 86, https://doi.org/10.3982/QE683. [1459, 1466, 1468, 1469, 1477, 1482]

Lai, E. K., W. Lim, and J. T.-Y. Wang (2015), "An experimental analysis of multidimensional cheap talk." *Games and Economic Behavior*, 91, 114–144. [1458]

Lim, W. (2014), "Communication in bargaining over decision rights." *Games and Economic Behavior*, 85, 159–179. [1456]

Matthews, S., M. Okuno-Fujiwara, and A. Postlewaite (1991), "Refining cheap-talk equilibria." *Journal of Economic Thoery*, 55, 247–273. [1454, 1459]

Rabin, M. (1990), "Communication between rational agent." *Journal of Economic Theory*, 51, 144–170. [1454, 1457, 1459]

Rabin, M. and J. Sobel (1996), "Deviations, dynamics and equilibrium refinements." *Journal of Economic Theory*, 68, 1–25. [1454]

Sánchez-Pagés, S. and M. Vorsatz (2007), "An experimental study of truth-telling in sender-receiver game." *Games and Economic Behavior*, 61, 86–112. [1457]

Sánchez-Pagés, S. and M. Vorsatz (2009), "Enjoy the silence: An experiment on truth-telling." *Experimental Economics*, 12, 220–241. [1457, 1476]

Serra-Garcia, M., E. van Damme, and J. Potters (2013), "Lying about what you know or about what you do?" *Journal of the European Economic Association*, 11 (5), 1204–1229. [1458]

Sobel, J. (2013), "Ten possible experiments on communication and deception." *Journal of Economic Behavior and Organization*, 93, 408–413. [1459]

Sopher, B. and I. Zapater (2000), "Communication and coordination in signalling games: An experimental study." Report. [1454, 1457]

Wang, J. T.-Y., M. Spezio, and C. F. Camerer (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. [1457]

Co-editor Karl Schmedders handled this manuscript.

Manuscript received 25 February, 2016; final version accepted 26 February, 2018; available online 12 March, 2018.