Meaning and Credibility in Experimental Cheap-Talk Games*

Ernest K. Lai
Department of Economics
Lehigh University
kwl409@lehigh.edu

Wooyoung Lim†
Department of Economics
The Hong Kong University of Science and Technology
wooyoung@ust.hk

February 5, 2015

Abstract

We design four simple cheap-talk games to experimentally investigate the refinement concept of neologism-proofness. All four games admit fully revealing equilibrium, but whether the equilibrium is neologism-proof varies across the games. We find that neologisms played an evident role in how subjects played the games. Overall, fully revealing equilibria that are robust in the sense of being neologism-proof were played more often. Senders and receivers were, however, affected differently by neologisms. The mere existence of meaningful neologisms, even though non-credible, attracted deviating behavior on senders’ part. Receivers’ behavior, on the other hand, was affected by whether the neologisms were credible or not, with credible neologisms attracting more deviating behavior from separating strategies.

Keywords: Neologism-Proofness; Cheap Talk; Equilibrium Selection; Laboratory Experiments

JEL classification: C72; C92; D82; D83

*We are grateful to Andreas Blume, Syng-Joo Choi, Navin Kartik, Barry Sopher and Joel Sobel for their valuable comments and suggestions. We also thank conference and seminar participants at HKUST, Rutgers University, the 2nd Haverford Meeting on Behavioral and Experimental Economics, the 16th KAEA-KEA International Conference and the 5th Annual Xiamen University International Workshop on Experimental Economics for valuable discussions. This study is supported by a grant from the Research Grants Council of Hong Kong (Grant No. ECS-699613).

†Corresponding author. Address: HKUST Department of Economics, LSK Business Building, Clearwater Bay, Kowloon, Hong Kong. Phone: (852) 2358 7628.
1 Introduction

Cheap-talk games are a type of signaling games in which messages are costless. Such a message property has profound implications for the treatment of equilibrium. With messages being costless, meaning can only be established by use (in equilibrium) and not by introspection. And for every equilibrium with unsent messages, there exists another equilibrium whose outcome is equivalent to the original one, in which all messages are used and thus an off-equilibrium path does not exist. Accordingly, standard refinement arguments powerful for costly signaling games, which are based on how messages should be interpreted off-the-equilibrium path, do not restrict the set of equilibria for cheap-talk games.

Several *ad hoc* equilibrium refinements to cheap talk games have been proposed. Farrell’s (1993) notion of *neologism-proofness* is the first such refinement, providing an important foundation for later work. Matthews, Okuno-Fujiwara, and Postlewaite (1991) develop the concept of announcement-proofness, which generalizes neologism-proofness. Chen, Kartik, and Sobel (2008) relate their criterion, NITS (No Incentive To Separate), to neologism-proofness by showing that an equilibrium survives NITS if it is neologism-proof. More recently, De Groot Ruiz, Offerman, and Onderstal (2012) propose a behavioral refinement, ACDC (Average Credible Deviation Criterion), that further extends on neologism-proofness and announcement-proofness.\(^1\)

Unlike the refinement concepts for costly signaling games, which have been experimentally investigated (Brandts and Holt, 1992; Banks, Camerer and Porter, 1994), neologism-proofness has not been sufficiently scrutinized and its empirical validity unambiguously proven. The objective of this study is to investigate how well this refinement concept, which serves as the headwater of the refinement literature in cheap talk, predicts play in a lab setting.

To our knowledge, the first experimental study that touches on neologism-proofness is Blume, DeJong, Kim and Sprinkle (2001). With the primary objective of testing the validity of the Partial Common Interest (PCI) criterion proposed by Blume, Kim and Sobel (1993), Blume et al. (2001) compare PCI with other selection criteria including neologism-proofness. They show that sometimes these other criteria give better predictions than does PCI, although their empirical performance is not always reliable. They consider four games, and neologism-proofness shares the same predictions as Pareto efficiency in three of them and rejects all equilibria in one of them. Their findings therefore do not distinctively reveal how useful neologism-proofness, which is not their primary focus to begin with, is in predicting play. Two other experimental studies related to cheap-talk refinements are Kawagoe and Takiawa (2008) and De Groot Ruiz, Offerman, and Onderstal (2012) propose a behavioral refinement, ACDC (Average Credible Deviation Criterion), that further extends on neologism-proofness and announcement-proofness.\(^1\)

\(^1\)Other non-belief based refinement concepts include Partial Common Interest by Blume, Kim, and Sobel (1993) and recurrent mop by Rabin and Sobel (1996). See also Eső and Schummer’s (2009) criterion of immunity to credible deviations which has selection power for Crawford and Sobel’s (1982) cheap-talk model where the cheap talk is replaced by costly messages.
Offerman, and Onderstal (2012). Kawagoe and Takizawa (2008) compare equilibrium refinements and level-$k$ analysis for their experimental sender-receiver games. They show that level-$k$ analysis could better organize their data. Their design, however, does not provide a rich environment to address neologism-proofness, because neologisms are absent in their design. De Groot Ruiz, Offerman, and Onderstal (2012) experimentally investigate their newly developed behavioral criterion, ACDC. But they also do not directly explore neologism-proofness.

A central element behind neologism-proofness is the notion of credible neologisms. Farrell (1993) assumes that messages have literal meaning and that out-of-equilibrium messages always exist. Loosely speaking, when the set of sender types who want to convey the literal meaning using out-of-equilibrium messages (the neologism) exactly coincides with the set of types expressed by the literal meaning, the neologism is credible or self-signaling. A neologism-proof equilibrium is one in which a credible neologism does not exist. We design four cheap-talk games that admit fully revealing equilibrium. We induce neologisms by controlling the message space, such that whether the fully revealing equilibrium is neologism-proof varies across the games. We observe the frequencies with which fully revealing equilibrium is played under the four games (treatments).

Our experimental findings show that neologisms played an evident role in how subjects played the cheap-talk games. Overall, fully revealing equilibria that are robust in the sense of being neologism-proof were played more often. Senders and receivers were, however, affected differently by neologisms. The mere existence of meaningful neologisms, even though non-credible, attracted deviating behavior on senders’ part. Receivers’ behavior, on the other hand, was affected by whether the neologisms were credible or not, with credible neologisms attracting more deviating behavior from separating strategies.

Other than the papers discussed above, our paper is also 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; Hurkens and Kartik, 2009; Wang, Spezio, and Camerer, 2010).

---

2A feature of neologism-proofness that distinguishes it from other standard signaling refinements is its natural language requirement. Perfect sequential equilibrium developed by Grossman and Perry (1986) for costly signaling games does not have the same natural language requirement but is otherwise closely related to the concept of neologism-proofness. Sharing the natural language requirement, Rabin’s (1990) nonequilibrium notion of “credible message rationalizability” builds on neologism-proofness by combining it with rationalizability. Sopher and Zapater (2000) find experimental support for credible message rationalizability.

3Blume et al. (2008) document that restricting the message space expedites convergence in sender-receiver games with a priori meaningless messages. Lai, Lim and Wang (2014) highlight the role of the message space in facilitating information transmission in multidimensional settings. Serra Garcia et al. (2013) study the effect of limiting the message space in public goods games with cheap-talk communication about private information. Our contribution is to provide a novel channel (relative to these studies)—neologism-proofness—through which the message space plays a role in the information transmission.

4See also Crawford (1998) for a survey of earlier studies and a discussion on the connection between cheap-talk
One common finding of this literature is the observation of “over-communication” or “lying aversion,” where subjects communicated more than what the equilibria predicted. Our paper differs from these papers in that we are interested not only in whether subjects play according to equilibria but also in whether they play according to robust equilibria.

The other literature to which our paper is related is the experimental investigation of equilibrium refinements for costly signaling games. Brandts and Holt (1992) test the intuitive criterion by Cho and Kreps (1987). They find that both message-level and action-level data supported the intuitive equilibrium, although this equilibrium was vulnerable to self-enforcing assignment (suggesting another equilibrium play) by an outside authority. Banks, Camerer and Porter (1994) design games to separate various refinements including the Nash equilibrium, sequential equilibrium, intuitive criterion, divine, universal divine, and never-weak-best-reply. They show that subjects’ behaviors converged to the more refined equilibrium up to the intuitive criterion.

Before we delve into the details of our experimental games, it may be constructive to discuss the limitations of neologism-proofness as a refinement concept and how those relate to our study. First, neologism-proofness lacks a general existence property, and therefore we do not know what it predicts when existence fails. Second, it lacks a complete formalization regarding the presence of unsent messages with natural meaning. Third, it falls short of fully addressing our concern over the usefulness of natural language, as neologisms arise off the equilibrium path but languages on the path are still arbitrary. The lack of theoretical predictions when existence fails may provide room for experimental studies to inform theory. Our study helps evaluate whether the lack of a general existence property should be the reason to discard the refinement concept even for the games in which it has the power to predict play. Our study does not currently address the second limitation, although it provides a framework that can potentially be augmented to do so. As for the third limitation, our study allows us to see empirically whether subjects use messages in a way that their natural meaning matches with the strategic meaning on the equilibrium path.

The rest of the paper proceeds as follows. Section 2 lays out our experimental cheap-talk games and analyzes their equilibria. Section 3 discusses our experimental hypotheses and procedures. Section 4 reports our findings. Section 5 concludes.
2 The Games and the Equilibrium Analysis

We design four cheap-talk games. In each game, there are two players, a sender (he) and a receiver (she). The sender is privately informed about his type \( \theta \in \{s, t\} \), where the common prior is that the two types are equally likely. After observing his type, the sender sends a message, \( m \in M \), to the receiver, where \( |M| = 2 \) or 3. Upon receiving the message, the receiver takes an action \( a \in \{L, C, R\} \).

Table 1 presents the payoffs of the games. In each cell, the first number is the sender’s payoff and the second number the receiver’s when an action is taken in a state.

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>C</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>30, 20</td>
<td>20, 30</td>
<td>0, 8</td>
</tr>
<tr>
<td>t</td>
<td>30, 20</td>
<td>8, 0</td>
<td>20, 30</td>
</tr>
</tbody>
</table>

(a) Game 1

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>C</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>50, 20</td>
<td>20, 30</td>
<td>0, 8</td>
</tr>
<tr>
<td>t</td>
<td>10, 20</td>
<td>8, 0</td>
<td>20, 30</td>
</tr>
</tbody>
</table>

(b) Game 2

Table 1: Payoffs of the Games

We design the games for their simplicity and out of three other considerations. First, Game 1 and Game 2 both have multiple perfect Bayesian equilibrium (thereafter “equilibrium”) outcomes with no Pareto ranking; they therefore provide room for selection yet rule out Pareto dominance as a potential confounding selection criterion. As long as \(|M| \geq 2\), there are two equilibrium outcomes, babbling and fully revealing. In the babbling equilibrium outcome, the receiver ignores the sender’s message, taking the \( ex-ante \) ideal action \( L \) under the prior. In the fully revealing equilibrium outcome, the receiver takes distinct actions after receiving the sender’s distinct messages sent for distinct types. There is no Pareto dominance, because the sender strictly prefers the babbling outcome to the fully revealing outcome, whereas the opposite is true for the receiver. Second, in an effort to ensure any difference between the two games in subjects’ choices of which equilibrium to play is not influenced by the sender’s payoff, the sender’s expected payoffs from the babbling equilibrium outcome (30) and the fully revealing equilibrium outcome (20) are controlled to be the same across the games. Finally, the sender’s expected payoff for one equilibrium outcome is the receiver’s expected payoff for another and vice versa so that any consideration of other-regarding preferences such as fairness does not play a role in equilibrium selection.

---

7 Game 1 shares the same qualitative structure as the “I Won’t Tell You” game in Farrell (1993), Game \( \Gamma_2 \) 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 the payoff structure of the game.

8 Another consideration we have in our design is to create an environment in which the neologism-proofness prediction can be separated from that of the level-\( k \) model of bounded rationality, which is commonly applied to explain findings from communication games as in Kawagoe and Takizawa (2008). Note that the only difference between Game 1 and Game 2 is the type-\( t \) sender’s ideal actions, whereas the expected payoff for each outcome
2.1 Neologism-Proofness Applied

A neologism-proof equilibrium is an equilibrium in which a credible neologism does not exist. According to Farrell (1993), a credible neologism requires that a) a neologism relative to the putative equilibrium exists for every non-empty subset $K$ of types, where a neologism is an unsent message in the putative equilibrium that literally says “my type is in $K$,” and b) the sender’s types in $K$ strictly prefer the outcome achieved when this neologism is believed over the putative equilibrium outcome, whereas the types not in $K$ would weakly prefer to stay in the putative equilibrium (self-signaling).

In designing our experimental environment, we look for games that, while capturing the essence of neologism-proofness, allow us to apply the concept in the simplest possible setting. We achieve this by minimizing the number of out-of-equilibrium messages. We introduce what we call limited neologisms, in which, departing from Farrell (1993), for a fixed equilibrium, a neologism exists for some but not all subsets of types.

We take Game 1 and Game 2 and pair each of them with a message space containing three elements: $M = \{ \text{“my type is } s\text{”, “my type is } t\text{”, “I won’t tell you my type”} \}$. The resulting games are called Game $1M3$ and Game $2M3$. We characterize the neologism-proof equilibria of the two games as follows:

**Proposition 1.** The fully revealing equilibrium outcome in Game $1M3$ cannot be supported by a neologism-proof equilibrium whereas that in Game $2M3$ can. The babbling equilibrium outcome in Game $1M3$ can be supported by a neologism-proof equilibrium whereas that in Game $2M3$ cannot.

Note that in the presence of our limited neologisms, the approach of showing that an equilibrium outcome can be supported as neologism-proof is different from that in Farrell (1993). In Farrell (1993), since for any equilibrium a neologism exists for every subset of types, when a neologism-proof supporting equilibrium exists, it automatically covers all neologisms, i.e., under no neologism does self-signaling occur to compromise the equilibrium (outcome). When there are limited neologisms, however, different supporting equilibria may be associated with different (sets of) neologisms. Even if self-signaling does not occur under one neologism, it may occur under another for another supporting equilibrium. Accordingly, under limited neologisms we need to show that all supporting equilibria with different configurations of neologism(s) are neologism-proof in order to cover all neologisms. In the following, we argue that all babbling equilibria in Game $1M3$ and all fully revealing equilibria in Game $2M3$ are neologism-proof, while providing for each player and the receiver’s ideal actions are made the same across the two games. Accordingly, for a given message space size, very limited configurations of the level-\(k\) model would generate different predictions for the two games. For more discussion on level-\(k\) analysis and its prediction, see Appendix A.
examples of non-neologism-proof babbling equilibria for Game 2M3 and non-neologism-proof fully revealing equilibria for Game 1M3.

To illustrate that the fully revealing equilibrium outcome in Game 1M3 cannot be supported as a neologism-proof equilibrium, consider the truth-telling equilibrium in which s and t send, respectively, “my type is s” and “my type is t,” each with probability one. Here, “I won’t tell you my type,” which literally means my type is in \{s, t\}, is the only neologism. If the receiver believes the literal meaning of the neologism, she will take the ex-ante optimal action L. Given that 30 > 20, both s and t—the sender’s types corresponding to the literal meaning of the neologism—prefer the outcome achieved when the neologism is believed over the putative equilibrium outcome; the neologism is thus self-signaling.

For the truth-telling equilibrium in Game 2M3, when the receiver believes the literal meaning of the message “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 thus not self-signaling. For the remaining cases in which either “my type is s” or “my type is t” is the neologism in the non-truth-telling fully revealing equilibria, the corresponding sender type would not strictly prefer the outcome achieved when the neologism is believed, because the on- and off-equilibrium path payoffs are the same. The neologisms are thus also not self-signaling. Absent a credible neologism, the fully revealing equilibrium outcome in Game 2M3 can thus be supported as neologism-proof.

Since both sender’s types in Game 1M3 receive their maximum payoffs in the event of babbling equilibrium outcome, it is straightforward to show that the outcome can be supported as neologism-proof. To illustrate that the babbling equilibrium outcome in Game 2M3 cannot be supported as neologism-proof, consider the babbling equilibrium in which “my type is t” is one of the unsent messages. Given that 10 < 20, type t strictly prefers the outcome that could be achieved when the neologism is believed (in which the receiver takes R) over the putative equilibrium outcome, whereas type s prefers to stay in the equilibrium; the neologism is thus self-signaling.

To generate richer comparative statics for our experimental hypotheses, we introduce another two games, Game 1M2 and Game 2M2, which are Game 1 and Game 2 with a binary message space \( M' = \{\text{“my type is s”}, \text{“my type is t”}\} \). The neologism-proof equilibria of the games are characterized as follows:

**Proposition 2.** Both the fully revealing and babbling equilibrium outcomes in Game 1M2 can be supported by neologism-proof equilibria. Only the fully revealing equilibrium outcome in Game 2M2 can be supported by a neologism-proof equilibrium.

\(^9\)Note that the truth-telling equilibrium is a fully revealing equilibrium in which literal meanings are used on the equilibrium path.
A fully revealing equilibrium outcome can be supported by a truth-telling equilibrium or by another equilibrium in which the use of the literal meaning of messages is reversed. The neologism-proofness of the fully revealing equilibrium outcomes is a simple consequence of the binary message space so that in either equilibrium there is no neologism; we have thus taken the limited neologisms to the extreme by eliminating them.\textsuperscript{10} The babbling equilibrium outcomes in both games are characterized in the same way as those in Games \textsuperscript{1}M\textsubscript{3} and \textsuperscript{2}M\textsubscript{3}.

A couple of remarks on our design of message spaces in relation to Farrell’s (1993) notion of “natural language” are in order. Farrell’s notion of natural language has two main ingredients: 1) \textit{common language}, i.e., each message has a literal meaning that is associated with a type in the type space, and 2) \textit{rich language}, i.e., the message space is large enough (relative to the type space) so that for any subset \(K\) of the sender’s types, 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 that we induce limited neologisms, given the binary types of our games, the message space \(M\) is rich enough to describe all subsets of types, even though some of these messages are used in equilibrium.

An additional assumption in Farrell’s (1993) development of the refinement concept is the existence of unsent messages. Farrell argues that a mixed-strategy equilibrium in which the sender randomizes over messages with the same equilibrium meaning (so that there is no unsent message) is implausible. This type of equilibria will be implausible when the sender has a preference for short, simple and straightforward messages.\textsuperscript{11} Our design with two different sizes of message spaces provides us with an opportunity to evaluate whether the assumed existence of unsent messages is an empirically plausible assumption. If randomization is a natural empirical tendency, having an additional message in \(M\) relative to \(M'\) should not affect the communication outcomes.

\textsuperscript{10}Game \textsuperscript{1}M\textsubscript{2} is in the same class of games as Game 2 in Kawagoe and Takiawa (2008). Our conclusion on the neologism-proofness of the fully revealing equilibrium outcome is, however, different from theirs. With binary types and binary messages, which is also the case in Kawagoe and Takiawa (2008), there will be no unsent messages in a fully revealing equilibrium. Accordingly, a neologism does not exist, and the equilibrium outcome trivially survives neologism-proofness. This difference in characterization will have implications for the interpretation of experimental findings. As will be shown below, the majority of plays in Game \textsuperscript{1}M\textsubscript{2} conformed to the fully revealing equilibrium, which is, accordingly to our view, consistent with the prediction of neologism-proofness. Considering that only the babbling equilibrium survives the refinement, Kawagoe and Takiawa (2008) attribute their similar findings to over-communication.

\textsuperscript{11}On this point, Farrell (1993) states that: “For example, if type \(t\) wants (and is expected) to reveal himself, and if both the English sentences, ‘I am \(t\),’ and ‘I am either \(u\) or \(v\),’ are interpreted in equilibrium as meaning ‘I am \(t\),’ then \(S\) [the sender] will prefer the former” (pp. 518, Section 4. Are There Unused Messages?).
3 Experimental Hypotheses and Procedures

3.1 Hypotheses

The variations in the payoff structures and the message spaces give us the $2 \times 2$ design summarized in Table 2.

Table 2: Experimental Treatments

<table>
<thead>
<tr>
<th></th>
<th>Game 1</th>
<th>Game 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>M</td>
<td>= 3$</td>
</tr>
<tr>
<td>$</td>
<td>M'</td>
<td>= 2$</td>
</tr>
</tbody>
</table>

We formulate our experimental hypotheses in terms of the frequencies with which fully revealing equilibrium is played. The hypotheses are based on the premise that a robust equilibrium in the sense of being neologism-proof is more likely to be played. Propositions 1 and 2, which are also summarized in Table 2, serve to inform our hypotheses.

Starting from Game 2M2 and moving to Game 2M3, we create a neologism that is not credible. Our first hypothesis is about the effect of the existence of non-credible neologisms:

**Hypothesis 1. Effect of the Existence of (Non-Credible) Neologisms:** The frequency of a fully revealing equilibrium in Game 2M2 is the same as that in Game 2M3.

Starting from Game 2M3 and moving to Game 1M3, we make the neologism credible. Our second hypothesis is about the effect of the credibility of neologisms:

**Hypothesis 2. Effect of the Credibility of Neologisms:** The frequency of a fully revealing equilibrium is higher in Game 2M3 than in Game 1M3.

Starting from Game 1M3 with a credible neologism and moving to Game 1M2, we get rid of any neologism. Our third hypothesis is about the joint effect of the existence and credibility of neologisms:

---

12There are two fundamental differences between Game 2M3 and 1M3, the sender’s payoffs under $L$ and the size of the message space. But, given that the sender’s expected payoffs under $L$ are made the same, in an ex-ante sense the only difference between the two games is the size of the message space. In fact, since the credibility of a neologism is related to whether a profitable deviation exists when the neologism is believed by the receiver, which is payoff related, one can view the two fundamental differences as having merged into one in the credibility of the neologism.
Hypothesis 3. **Effect of the Existence and Credibility of Neologisms:** The frequency of a fully revealing equilibrium is lower in Game 1M3 than in Game 1M2.

### 3.2 Procedures

Our experiment was conducted in English using z-Tree (Fischbacher, 2007) at the Hong Kong University of Science and Technology. A total of 160 subjects having no prior experience with our experiment were recruited from the undergraduate population of the university. Upon arrival at the laboratory, subjects were instructed to sit at separate computer terminals. Each received a copy of the experimental instructions. Appendix B contains the instructions for Game 1M3; the instructions for the other games are similar. The instructions were read aloud using slide illustrations as an aid, and a comprehension quiz was held immediately after.

Two sessions were conducted for each game. Random matching was used. Two matching groups participated in each of these sessions. A matching group consisted of 10 subjects, five as senders (Member A) and five as receivers (Member B). Viewing each matching group as an independent observation, we thus have four observations per game.

In all sessions, subjects participated in 20 rounds of decision making under a single treatment condition (between-subject design). At the beginning of a session, half of the subjects were randomly assigned the role of Member A and the other half the role of Member B. The role assignment remained fixed throughout the session. We employed the strategy method in eliciting subjects’ decisions. Subjects were told that there was a random variable, X, whose integer value ranged equally likely from 1 to 100. At the beginning of each round, Member A was asked what message he/she would send to the paired Member B if X > 50 and if X ≤ 50. The available messages were “X is bigger than 50,” “X is smaller than or equal to 50,” and “I won’t tell you.” Member B was asked what action he/she would take, namely L, C, or R, upon receiving each of the three messages from the paired Member A. Once the subjects had stated their strategies, the value of X was realized. Subjects’ strategies were then implemented based on the realized X, and the reward for that round was determined. Member A was also asked to predict Member B’s action upon seeing the implemented message. A correct prediction was rewarded with 2 payoff points. Throughout the session, the corresponding payoff in Table 1 was shown on the subjects’ screens. Feedback on choices and rewards but not strategies was provided at the end of each round.

---

13For the sake of simplicity in the experimental instructions and design, we did not allow players to randomize explicitly over different messages/actions for a given contingency. Note that a strict and unique best response action exists for a babbling message and for any truth-telling message so that allowing explicit randomization does not create any gain in generality.
We randomly selected two rounds out of a total of 20 rounds that were played and calculated the payment for each subject. The sum of the payoffs a subject earned in the two randomly selected rounds was 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 was also paid. Subjects on average earned HK$76.8 (≈ US$9.85).\textsuperscript{14} A session on average lasted about an hour.

4 Results

In Section 4.1, we first report our findings and evaluate the hypotheses using on-path aggregate behavior, i.e., we take the value of $X$ in each instance of play and apply subjects' strategies to it to establish the aggregate frequencies of interests. We will examine individuals’ strategies and offer further, supporting evidence in Section 4.2.

4.1 On-Path Aggregate Behavior

Figure 1 presents the round-by-round frequencies of fully revealing equilibrium outcomes for the four games, in which we measure how often receivers’ ideal actions were taken.

\textbf{Finding 1. Effect of the Existence of (Non-Credible) Neologisms: Consistent with Hypothesis 1, there was no significant difference in the frequency of fully revealing equilibrium outcomes between Game 2M2 and Game 2M3.}

Using data from the last 10 rounds for each matching group as independent observations, we failed to reject the null hypothesis of there being no difference in the frequencies of fully

\textsuperscript{14}Under Hong Kong’s currency board system, the HK dollar is pegged to the US dollar at the rate of HK $7.8 to US$1.
revealing equilibrium outcomes between Game 2M2 and Game 2M3, which were 57% and 53% respectively (two-sided \( p = 0.5614 \), Mann-Whitney test). While the existence of a non-credible neologism had no impact on the overall frequencies of fully revealing equilibrium outcomes, it did affect senders’ and receivers’ behavior. Figure 2 presents the frequencies of messages conditioned on types and the frequencies of actions conditioned on messages.

![Figure 2: Senders’ and Receivers’ Behavior](image)

Compared with Game 2M2, the introduction of the non-credible neologism (relative to the truth-telling equilibrium) “I won’t tell you” in Game 2M3 attracted deviating behavior on senders’ part. In Game 2M2, senders exhibited truth-telling behavior, where types \( s \) sent “my type is \( s \)” 93% of the time and types \( t \) sent “my type is \( t \)” 83% of the time. In Game 2M3, the former frequency dropped significantly to 39% (\( p = 0.0143 \), Mann-Whitney test) and the latter not significantly to 61% (\( p = 0.1 \), Mann-Whitney test). Furthermore, in Game 2M3 the \( s \) and \( t \) types used the neologism “I won’t tell you” 59% and 36% of the time respectively. In response to “I won’t tell you,” receivers in Game 2M3 took \( L \) 50% of the time. Even though in theory “I won’t tell you” was not a self-signaling neologism, types \( s \) sent it to induce—to a certain extent successfully—receivers to take the pooling action for 50 payoff points.

Unlike in the case of senders, the presence of a non-credible neologism did not lead to more deviating behavior from receivers. In fact, receivers’ responses to “my type is \( s \)” and “my type is \( t \)” in Game 2M3 were more in line with the truth-telling equilibrium behavior than with their responses in Game 2M2. The frequencies of action \( C \) conditioned on “my type is \( s \)” increased significantly from 64% in Game 2M2 to 82% in Game 2M3 (\( p = 0.0571 \), Mann-Whitney test); the frequencies of action \( R \) conditioned on “my type is \( t \)” increased from 65% in Game 2M2 to 80% in Game 2M3, though not significantly (\( p = 0.1714 \), Mann-Whitney test). In terms of generating deviating behavior, the non-credible neologism affected senders but not receivers. And

---

\(^{15}\)Not much convergence was observed in subjects’ behavior. We nevertheless used data from the last 10 rounds for our statistical tests and other frequency figures to account for any convergence of behavior. Unless otherwise indicated, the \( p \) values are from one-sided tests.
the senders’ differing behavior in the two games roughly offset the receivers’ differing behavior to generate the same overall frequencies of fully revealing equilibrium outcomes.

Finding 2. **Effect of the Credibility of Neologisms:** Consistent with Hypothesis 2, the frequency of fully revealing equilibrium outcomes was significantly higher in Game 2M3 than in Game 1M3.

The frequency of fully revealing equilibrium outcomes in Game 1M3 was 32%, significantly lower than the 53% in Game 2M3 ($p = 0.0571$, Mann-Whitney test). Overall, the result was consistent with our premise that robust equilibria in the sense of being neologism-proof are more likely to be played. However, the credibility of the neologism had varying and insignificant impacts on senders’ uses of it. For types $t$, the frequencies of “I won’t tell you” increased insignificantly from 36% in Game 2M3 to 49% in Game 1M3 ($p = 0.1$, Mann-Whitney test). For types $s$, the frequencies of “I won’t tell you” decreased insignificantly from 59% to 46% ($p = 0.2429$, Mann-Whitney test). For receivers, on the other hand, the frequencies of truth-telling equilibrium behavior decreased significantly when the neologism was credible. The frequency of $C$ contingent on “my type is $s$” decreased from 82% in Game 2M3 to 51% in Game 1M3 ($p = 0.0147$, Mann-Whitney test). The frequency of $R$ contingent on “my type is $t$” decreased from 80% in Game 2M3 to 52% in Game 1M3 ($p = 0.0286$, Mann-Whitney test). Unlike in the case of mere existence, in terms of generating deviating behavior the credibility of a neologism affected receivers but not senders. And the lower adherence to fully revealing equilibrium outcomes came from receivers.

Finding 3. **Effect of the Existence and Credibility of Neologisms:** Consistent with Hypothesis 3, the frequency of fully revealing equilibrium outcomes was significantly lower in Game 1M3 than in Game 1M2.

The frequency of fully revealing equilibrium outcomes in Game 1M2 was 66%, significantly higher than the 32% in Game 1M3 ($p = 0.0143$, Mann-Whitney test). Getting rid of any neologism increased the frequency of truth-telling equilibrium behavior for both senders and receivers. For senders, the frequency with which types $s$ sent “my type is $s$” increased from 47% in Game 1M3 to 82% in Game 1M2 ($p = 0.0143$, Mann-Whitney test); the frequency with which types $t$ sent “my type is $t$” increased from 40% in Game 1M3 to 85% in Game 1M2 ($p = 0.0143$, Mann-Whitney test). For receivers, the frequency of $C$ contingent on “my type is $s$” increased from 51% in Game 1M3 to 75% in Game 1M2 ($p = 0.0551$, Mann-Whitney test); the frequency of $R$ contingent on “my type is $t$” increased from 52% in Game 1M3 to 80% in Game 1M2 ($p = 0.0571$, Mann-Whitney test).
4.2 Individuals’ Strategies and Senders’ Beliefs

Using our data obtained via the strategy method, in this subsection we examine the strategy adopted by each subject in each round and provide the resulting aggregate frequencies. Our objective is to provide strategy-level data that further confirm our findings in Section 4.1. Senders’ beliefs regarding receivers’ responses will also be examined.

For senders, we record each instance of strategy decision that fits into the following four categories, some of which are not mutually exclusive: literal babbling (both types s and t sending “I won’t tell you”), non-revealing (both types s and t sending the same message), truth telling (type s sending “My type is s” and type t sending “My type is t”), and fully revealing (types s and t sending different messages). For receivers, we record each instance of strategy decision that fits into the following two categories: pooling (taking L upon receiving all available messages), and separating (taking C upon receiving “my type is s,” taking R upon receiving “my type is t,” and, for three-message games, taking any action upon receiving “I won’t tell you”). Figure 3 presents the frequencies of the strategy categories.

For senders in Game 2M2, almost all fully revealing strategies were truth telling (84% vs. 82%). This is to be contrasted with senders in Game 2M3, in which the frequencies of fully revealing strategies and truth-telling strategies were 64% and 33% respectively. It suggests that in Game 2M3, some senders used “I won’t tell you” and another message (“My type is t” according to the analysis in Section 4.1), which effectively revealed their types. The fact that the non-credible neologism attracted deviating behavior on senders’ part was also apparent from the significant drop in the frequency of truth-telling strategies from Game 2M2 to Game 2M3 ($p = 0.0143$, Mann-Whitney test).

Consistent with the fact that the fully revealing equilibrium outcome in Game 1M3 cannot be supported as neologism-proof, there were more instances of literal babbling in Game 1M3 than
in Game 2M3 (39% vs. 31%). The difference was, however, not significant \((p = 0.4429, \text{Mann-Whitney test})\), echoing the finding that the effect of the credibility of a neologism on senders was present but overall not significant. Without any neologism, the frequencies of fully revealing and truth-telling strategies in Game 1M2 were as high as those in Game 2M2 (82% and 75%), and significantly higher than those in Game 1M3 (54% and 34%; \(p \leq 0.021, \text{Mann-Whitney tests}\)), again echoing the findings in Section 4.1.

Turning our attention to receivers’ behavior, we note that the existence of a non-credible neologism led to greater adoption of separating strategies from 57% in Game 2M2 to 75% in Game 2M3. The difference was, however, not very significant \((p = 0.0732, \text{Mann-Whitney test})\). This echoes the finding that the existence of a non-credible neologism has a less significant effect on receivers than on senders. Consistent with the finding that the credibility of the neologism led to deviating behavior from receivers, in Game 1M3 we observed a lower frequency of separating strategies at 39% \((p = 0.0295, \text{Mann-Whitney test})\) than in Game 2M3. The frequencies of pooling strategies accordingly increased significantly from 12% in Game 2M3 to 34% in Game 1M3 \((p = 0.0407, \text{Mann-Whitney test})\). Finally, the frequency with which separating strategies were adopted in Game 1M2 was 73%, significantly higher than the 39% in Game 1M3 \((p = 0.0295, \text{Mann-Whitney test})\). Overall, the analysis suggests that the on-path behavior presented in Section 4.1 was supported by individuals playing corresponding strategies.

Figure 4: Senders’ Beliefs about Receivers’ Responses

Figure 4 presents senders’ beliefs about receivers’ responses conditioned on senders’ strategy categories. To varying degrees, senders’ strategies were consistent with their reported beliefs. Except for the truth-telling and fully revealing strategies in Game 1M3, more than 50% of the reported beliefs were best responses to the senders’ strategies, ranging from 54% to 93%. This finding suggests that our data are largely explained by the fact that some subjects went for the truth-telling/fully revealing equilibrium strategies whereas others chose the babbling equilibrium strategies.\(^{16}\)

\(^{16}\)Refer to Appendix A for a level-\(k\) analysis in which we explore, as an alternative, a potential rationalization
From Game 1M3, we obtain the unique finding that senders anticipated more pooling than separating responses when adopting the truth-telling and fully revealing strategies. In other words, senders anticipated that receivers would ignore their messages. Recall that Game 1M3 is the only game in which the fully revealing equilibrium is not neologism-proof. Our finding thus suggests that senders anticipated receivers’ understanding that senders have no incentive to communicate in the presence of credible neologisms. Yet senders told the truth which is also optimal given their reported beliefs. This points to the possibility that they were lying averse as documented in other experimental studies discussed in the introduction.

5 Concluding Remarks

In this paper, we experimentally explore a cheap-talk refinement concept known as neologism-proofness. In the experimental literature of communication games, behavioral phenomena such as overcommunication are prevalent, and they have been quite successfully explained by behavioral concepts or models such as lying aversion and the level-\(k\) analysis. The same behavioral factors may have been driving some of our findings. However, by focusing on the differences among our four treatments, we have largely controlled for these factors, which allows us to focus on the predictive power of neologism-proofness.

We find that whether or not a fully revealing equilibrium is neologism-proof affects how often it is played, lending support to the predictive power of the refinement concept. We also obtain a finding that is not covered by the theory of neologism-proofness. In particular, we find that even non-credible neologisms can cause senders to deviate. Credible neologisms, on the other hand, will lead to a lower frequency of fully revealing equilibrium outcomes through receivers’ behavior.

Farrell (1993) discusses the possibility of there being no pre-existing common language that is rich enough to communicate neologisms and therefore the meaning of a neologism must evolve. Our experimental games provide an apt environment to investigate this issue, in which we could endow subjects with a message space containing messages with no a priori meaning and see how meaning evolves (see also Blume et al., 1998, 2001). We leave this to future research.
Appendix A – Level-$k$ Analysis

Motivated by the over-communication phenomenon well-documented in the literature of experimental cheap-talk games and following the convention in the literature (Crawford, 2003; Cai and Wang, 2006), we specify the following level-$k$ model. The model starts with a credulous $L_0$ receiver and a truthful $L_0$ sender. The sender of the lowest level of sophistication, $L_0$, simply sends the truthful message all the time. In response to the $L_0$ sender, the $L_0$ receiver trusts the sender and always chooses the ideal action given the belief that is consistent with the literal meaning of a message. The model further assumes that $L_{k\geq 1}$ senders best-respond to $L_{k-1}$ receivers and $L_{k\geq 1}$ receivers best-respond to $L_k$ senders. Tables A1 and A2 report the level-$k$ predictions for Game 1M3 and Game 2M3, respectively. Tables A3 reports the level-$k$ predictions for Game 1M2 and Game 2M2 where the predictions coincide.

Table A1: Level-$k$ Predictions for Game 1M3

<table>
<thead>
<tr>
<th>Sender's Strategy</th>
<th>Receiver's Strategy</th>
<th>&quot;My type is s.&quot;</th>
<th>&quot;My type is t.&quot;</th>
<th>&quot;I won’t tell you.&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_0$</td>
<td>&quot;My type is s.&quot;</td>
<td>$C$</td>
<td>$R$</td>
<td>$L$</td>
</tr>
<tr>
<td>$L_1$</td>
<td>&quot;I won’t tell you.&quot;</td>
<td>$C$</td>
<td>$R$</td>
<td>$L$</td>
</tr>
<tr>
<td>$L_{k\geq 2}$</td>
<td>&quot;I won’t tell you.&quot;</td>
<td>$C$</td>
<td>$R$</td>
<td>$L$</td>
</tr>
</tbody>
</table>

Table A2: Level-$k$ Predictions for Game 2M3

<table>
<thead>
<tr>
<th>Sender's Strategy</th>
<th>Receiver's Strategy</th>
<th>&quot;My type is s.&quot;</th>
<th>&quot;My type is t.&quot;</th>
<th>&quot;I won’t tell you.&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_0$</td>
<td>&quot;My type is s.&quot;</td>
<td>$C$</td>
<td>$R$</td>
<td>$L$</td>
</tr>
<tr>
<td>$L_1$</td>
<td>&quot;I won’t tell you.&quot;</td>
<td>$C$</td>
<td>$R$</td>
<td>$C$</td>
</tr>
<tr>
<td>$L_{k\geq 2}$</td>
<td>&quot;My type is s.&quot; or &quot;I won’t tell you.&quot;</td>
<td>$C$</td>
<td>$R$</td>
<td>$C$</td>
</tr>
</tbody>
</table>

Table A3: Level-$k$ Predictions for Game 1M2 and Game 2M2

<table>
<thead>
<tr>
<th>Sender's Strategy</th>
<th>Receiver's Strategy</th>
<th>&quot;My type is s.&quot;</th>
<th>&quot;My type is t.&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_0$</td>
<td>&quot;My type is s.&quot;</td>
<td>$C$</td>
<td>$R$</td>
</tr>
<tr>
<td>$L_1$</td>
<td>&quot;My type is s.&quot;</td>
<td>$C$</td>
<td>$R$</td>
</tr>
<tr>
<td>$L_{k\geq 2}$</td>
<td>&quot;My type is s.&quot;</td>
<td>$C$</td>
<td>$R$</td>
</tr>
</tbody>
</table>

First of all, like neologism-proofness, the model predicts a babbling outcome for Game 1M3. Second, like neologism-proofness, the model essentially predicts a fully revealing outcome for Game 2M3. Similarly, the model predicts a fully revealing outcome for Game 1M2 and Game 2M2. However, unlike neologism-proofness, the model fails to predict the systematic difference between receivers’ strategies in Game 1M3 and Game 2M3. Recall that neologism-proofness
predicts that when the neologism becomes credible in Game 1M3, the receiver would use the pooling strategy in which action $L$ is taken regardless of the message received. Our data indicate that the frequency of the pooling strategy being used by receivers is indeed significantly higher in Game 1M3 than in Game 2M3. On the other hand, the specified level-$k$ model organizes the data from Game 2M3 quite well. In particular, the model predicts that the message “I won’t tell you” was paired with another message “My type is $t$” to reveal senders’ types effectively, which is indeed observed in our strategy-level data.

A few remarks regarding alternative specifications of $L_0$ players are in order. One plausible alternative is to specify that the $L_0$ receiver uniformly randomizes over the three actions regardless of the message. But this alternative specification is somewhat unnatural given our adopted message space in which each message has a clear literal meaning, and secondly the specification essentially generates the same fully revealing outcomes from Game 1M3 and Game 2M3, which do not match our data.

We believe that the level-$k$ model and the equilibrium notion of neologism proofness should be complements rather than substitutes in explaining the experimental data. Our result indicates that neologism-proofness is crucial to explain the observed treatment effects or the difference among the four treatments. However, the finding does not imply that the level-$k$ model has no power to explain the data. Our simple analysis in this section demonstrates that the level-$k$ model can explain some feature of our data reasonably well.
Appendix B – Experimental Instructions for Game 1M3

Welcome to the experiment. This experiment studies decision making between two individuals. In the following hour or so, you will participate in 20 rounds of decision making. Please read the instructions below carefully; the cash payment you will receive at the end of the experiment depends on how you make your decisions according to these instructions.

Your Role and Decision Group

There are 20 participants in today’s session. Prior to the first round, one half of the participants will be randomly assigned the role of Member A and the other half the role of Member B. Your role will remain fixed throughout the experiment. In each round, one Member A and one Member B will be randomly and anonymously paired to form a group, with a total of 10 groups.

Regarding how players are matched, the 10 groups are equally divided into 2 classes so that there are 5 groups in each class with 10 participants, 5 Members A and 5 Members B; in each and every round, you will be randomly matched with a participant in the other role in your class. Thus, in a round you will have an equal, 1 in 5 chance of being paired with a participant in the other role in your class. You will not be told the identity of the participant you are matched with, nor will that participant be told your identity—even after the end of the experiment.

Your Decision in Each Round

Member A’s Decision

In each round and for each group, the computer randomly selects, with equal chance, an integer X from 1 to 100. (Each number therefore has probability \( \frac{1}{100} \) to be selected.)

Figure 5 shows Member A’s decision screen. You as Member A have two kinds of decisions to make:

1. what to tell Member B if it turns out that \( X > 50 \); and
2. what to tell Member B if it turns out that \( X \leq 50 \).

Note that you are making a decision plan (what to do if this happens and what to do if that happens), where you will make these decisions without knowing the actually selected X. For each decision, you can choose from the following three options: 1) “X is bigger
than 50”; 2) “X is smaller than or equal to 50”; and 3) “I won’t tell you”. Your choices for different decisions could be different or the same.

Your decision for the round is completed after the two choices, which will then enter into the determination of rewards at a later stage of the round.

**Member B’s Decision**
Figure 6 shows Member B’s decision screen. You as Member B have three kinds of decisions to make:

1. what action to take if Member A says “X is bigger than 50”;  
2. what action to take if Member A says “X is smaller than or equal to 50”; and  
3. what action to take if Member A says “I won’t tell you”.

Note that you are making a decision plan (what to do if this happens and what to do if that happens and etc.), where you will make these decisions without knowing what Member A actually says. For each decision, you can choose from the following three options: 1) action L; 2) action C; and 3) action R. Your choices for different decisions could be different or the same.

Your decision for the round is completed after the three choices, which will then enter into the determination of rewards at a later stage of the round.

**Your Reward in Each Round**

After Member A’s and Member B’s decisions, the computer will proceed to draw the integer X randomly from 1 to 100. Then the realization of X will be revealed to Member A and Member B. Member A’s and Member B’s rewards will be determined according to the reward table in Figures 5 and 6 and what choices they have made. In each cell of the reward table, the first number represents the reward in HKD to Member A and the second number the reward in HKD to Member B. (The relevant numbers for each role are highlighted in Blue.)

The reward procedure, in which the computer draw determines the relevant row of the reward table and Member B’s action the relevant column, may best be illustrated with an example.

Consider the following choices made by the two members. Member A chooses to say

1. “X is bigger than 50” if X > 50;  
2. “I won’t tell you” if X ≤ 50.

Member B chooses to take

1. action L if Member A says “X is bigger than 50”;  
2. action C if Member A says “X is smaller than or equal to 50”;
3. action R if Member A says “I won’t tell you”.

Suppose the computer randomly draws $X = 27$. According to the choices, Member A says “I won’t tell you” and Member B takes action R. Given that $X \leq 50$, Member A will receive 20 HKD and Member B will receive 30 HKD. On the other hand, if the computer draws $X = 79$, according to the choices Member A says “$X$ is bigger than 50” and Member B takes action L. Given that $X > 50$, Member A will receive 30 HKD and Member B will receive 20 HKD.

**Prediction Reward by Member A.** If you are Member A, you have an opportunity to earn an extra reward. After the computer draws $X$, you will be asked to predict what Member B will take after what you choose to say to him/her for the relevant range of $X$. If your prediction is correct, you will receive an extra 2 HKD.

**Information Feedback**

During the course of a round, you will be informed about the drawn $X$, what Member A chooses to say for the relevant range of $X$, and what Member B chooses to take for what Member A chooses to say. (The information provided does not include the whole decision plan of the members.)

**Your Cash Payment**

The experimenter randomly selects 2 rounds out of 20 to calculate your cash payment. (So it is in your best interest to take each round seriously.) Your total cash payment at the end of the experiment will be the sum of HKD you earned in the 2 selected rounds plus a 30 HKD show-up fee.

**Quiz and Practice**

To ensure your comprehension of the instructions, we will provide you with a quiz and a practice round. We will go through the quiz after you answer it on your own. You will then participate in 1 practice round. At the beginning of the practice round, you will be randomly assigned the role of either Member A or Member B. Your role in the official rounds is the same as that in the practice round.

Once the practice round is over, the computer will tell you “The official rounds begin now!”

**Adminstration**
Your decisions as well as your monetary payment will be kept confidential. Remember that you have to make your decisions entirely on your own; please do not discuss your decisions with any other participants.

Upon finishing the experiment, you will receive your cash payment. You will be asked to sign your name to acknowledge your receipt of the payment (which will not be used for tax purposes). You are then free to leave.

If you have any question, please raise your hand now. We will answer your question individually. If there is no question, we will proceed to the quiz.

**Quiz**

1. True or False: I will remain as a Member A or Member B in all 20 rounds of decision-making. Circle one: True / False

2. True or False: I will be matched with the same player in the other role in all 20 rounds. Circle one: True / False

3. True or False: My decisions involve making plans for different scenarios but not specific choice for single scenario. Circle one: True / False

4. True or False: At the end of the experiment, I will be paid my earnings in HKD from two randomly chosen rounds in addition to 30 HKD show-up fee. Circle one: True / False.
References


