RIETI Discussion Paper Series 05-E-018

Why Lying Pays:
Truth Bias in the Communication with Conflicting Interests

KAWAGOE Toshiji
RIETI

TAKIZAWA Hirokazu
RIETI

The Research Institute of Economy, Trade and Industry
http://www.rieti.go.jp/en/

RIETI Discussion Paper Series 05-E-018

Why Lying Pays: Truth Bias in the Communication with Conflicting Interests*

Toshiji Kawagoe
Department of Complex Systems, Future University - Hakodate
116-2 Kameda Nakano cho, Hakodate, Hokkaido, 041-1112, Japan.
Phone: +81-138-34-6424, Fax: +81-138-34-6301
E-mail: kawagoe@fun.ac.jp

Hirokazu Takizawa
Research Institute of Economy, Trade and Industry
1-3-1 Kasumigaseki, Chiyoda-ku, Tokyo, 100-8901, Japan.
Phone: +81-3-3501-8275, Fax: +81-3-3501-8416
E-mail: takizawa-hirokazu@rieti.go.jp

15 January 2005

Abstract

We conduct experiments of a cheap-talk game with incomplete information in which one sender type has an incentive to misrepresent her type. Although that Sender type mostly lies in the experiments, the Receiver tends to believe the Sender's messages. This confirms "truth bias" reported in communication theory in a one-shot, anonymous environment without nonverbal cues. These results cannot be explained by existing refinement theories, while a bounded rationality model explains them under certain conditions. We claim that the theory for the evolution of language should address why truthful communication survives in the environment in which lying succeeds.

Keywords: Cheap talk, Communication, Private information, Experiment, Equilibrium refinement, Bounded rationality, Truth bias

JEL Classification: C72, C92, D82

*We would like to thank Masahiko Aoki, Colin Camerer, Akihiko Matsui, Richard McKelvey, and Charles Plott for their helpful comments and encouragement. We are also grateful to Tatsuyoshi Saijo, and Takehiko Yamato for their comments at the 2nd Experimental Economics Conference. The comment given by Vincent Crawford at the 2003 ESA meeting at Pittsburgh is especially acknowledged. A part of this research was supported by the Grant-in-aid of Japan Society for the Promotion of Science project since 1998 in collaboration with Yuji Aruka and Sobei Oda.

1 Introduction

Verbal communication can occur even between senders and receivers with conflicting interests, and is often accompanied by lying and suspicion. Some communication-theoretic literature reports that, even in such situations, although senders usually lie, most receivers believe senders' messages; this is called "truth bias," the receiver's intrinsic presumption that the senders are telling the truth (McCornack and Parks, 1986).

The purpose of this paper is to present experimental results of a cheap-talk game with incomplete information in which one sender type has an incentive to misrepresent her type, confirm the existence of truth bias, and find out a theoretical framework explaining this behavior. Kawagoe and Takizawa (1999) report the experimental results of cheap-talk games with different payoff characteristics. They examine how communication between the Sender and the Receiver is affected when the degree of preference alignment between them is changed, whereas the current paper focuses on theoretical explanation for the experimental result in Game 3 experimented there [1].

In a cheap-talk game with incomplete information, the Sender first announces a message based on her private information about her own type, and then the Receiver takes an action contingent on the Sender's announcement. The payoffs to both players are determined by the combination of the Sender's type and the Receiver's action, and do not depend on the Sender's message (thus costless communication). While the Receiver tries to guess the Sender's type via her message to choose the right action, the Sender wants to influence the Receiver's choice of action by her message. Thus this class of games can be regarded as the simplest possible representation of the strategic interpersonal communication that may involve persuasion, lying, deception, believing, and suspicion.

To conduct experiments on this class of games, we make the games as simple as possible. Specifically, we let the type space be $T = \{A, B\}$ with the prior distribution being equiprobable for each type. The action space for the Receiver is $C = \{X, Y, Z\}$. To consider the situation with common language, we let the message space be $M = \{$"I am type A", "I am type B"$\}$ (we hereafter denote these messages by a and b respectively, as long as no confusion may arise). Note that this message space creates the situation where each message corresponds to truth-telling or lying.

In order to focus on cases of interest, we assume that X and Y are the best action for the Receiver when Sender types are A and B respectively. Z is introduced to identify the case where the Receiver's belief over Sender types is nearly equiprobable; it is the best action for the Receiver when the belief is near 1/2 for both types [2]. We sometimes denote by (m_1, m_2) the Sender's pure strategy to send message m_1 in case of type A and m_2 in case of type B, where $m_1, m_2 \in M$, and denote by (c_1, c_2) the Receiver's pure strategy to play c_1 receiving message a and c_2 receiving message b.

Even these simplest possible settings encompass diverse incentive situations between the Sender and the Receiver. To specify the payoffs of games used in the experiments, Kawagoe and Takizawa (1999) adopt three general incentive situations as follows:[3]

Case 1 Both Sender type A and B want to be correctly identified, inducing the Receiver to choose action X and Y respectively;

Case 2 Both Sender types want the Receiver to play Z, that is, they want to confuse the Receiver;

Case 3 Type A Sender wants to be correctly identified, while type B Sender wants to misrepresent herself as type A.

Table 1 shows the payoffs of the games we actually used in our experiment. Rows indicate Sender types; columns actions of the Receiver. The left number in each cell indicates the Sender's payoff, while the right number the Receiver's.

As is well known, every cheap-talk game has an uninformative equilibrium in which Sender's messages convey no information about her true type, the so-called "babbling equilibrium." Babbling equilibrium arises because all Sender types sending the same message does not allow the Receiver to update his prior belief and the Receiver play the best response to this distribution. Throughout the paper we will call an equilibrium play in which the Receiver plays Z a babbling equilibrium, since, in our games, Z is introduced as the Receiver's best response to the prior distribution

[1] Their experimental results including Game 1 and 2 are summarized in the Appendix B. They are also cited in Camerer (2003, Ch.7)

[2] As will be explained in the Appendix B, the labels for the Receiver's action we used in the experiments were A, B, and C for X, Y, and Z respectively in Session 1, and they were permuted from Session 2 on. However, we will use X, Y, Z as indicated in the text throughout the paper, because we need to classify the Receiver's play according to his belief.

[3] See the Appendix A for more details.

Game 1

		Action		
		X	Y	Z
type	A	4, 4	1, 1	3, 3
	B	1, 1	4, 4	3, 3

Game 2

		Action		
		X	Y	Z
type	A	3, 4	2, 1	4, 3
	B	2, 1	3, 4	4, 3

Game 3

		Action		
		X	Y	Z
type	A	4, 4	1, 1	2, 3
	B	3, 1	2, 4	4, 3

Table 1: Cheap-talk Games Experimented in Kawagoe and Takizawa (1999)

Both Game 1 and Game 2 have separating equilibria in which each Sender type sends a distinct message and the Receiver plays differently in best response to each message. Almost all the refinement concepts mentioned in Section 3 agree to predicting separating equilibrium plays for Game 1, and babbling equilibrium plays for Game 2 and 3. Kawagoe and Takizawa (1999) reports the following experimental results: [4]:

1. Quick convergence to a separating equilibrium play was observed in Game 1 ;
2. Game 2 also showed convergence to a separating equilibrium play;
3. In Game 3, Sender subjects tended to play (a, a), while Receiver subjects tended to play X or Z in response to message a and Y in response to b.

The current paper focuses on the last result above. The reason is fourfold. First, the experimental results of their Game 3 can be regarded as an anomaly that cannot be explained by the standard equilibrium concepts, such as sequential equilibrium, and their refinements. Therefore, it is meaningful to find out a theoretical framework that can explain the data.

Second, the payoff structure of Game 3 is such that there is a conflict of interests between the Sender and the Receiver as well as between Sender types. This can be regarded as an abstract situation that the study of lying in communication theory has long focused on. Communication theory has so far focused its main attention on the communication in richer environments with voice, facial expression and so on such as in face-to-face conversation. However, the role of nonverbal cues in spotting a lie has now proven to be limited (Vrij, 2000), and the focus of analysis has shifted to the controlled exchange of message *per se* rather than nonverbal cues. Kawagoe and Takizawa (1999)'s experiment is also unique in the context of communication-theoretic literature because it confirms "truth bias," a well-known phenomenon in communication-theoretic experiments, in a one-shot anonymous environment with no nonverbal cues and simplest possible messages where the conflict of interests is common knowledge between the Sender and the Receiver.

Third, Game 3 is different from the cheap-talk games that have been closely examined in experimental studies on cheap-talk games. For example, Dickhaut, McCabe, and Mukherji (1995) and Game 1 and 2 in Kawagoe and Takizawa (1999) show that informative communication arises when the Sender and the Receiver have common language and their interest is sufficiently aligned, as is predicted by various refinement theories. Blume, DeJong, Kim, and Sprinkle (1998a) also conduct experiment on various cheap-talk games under common language to show that "partial common interests"(Rabin and Sobel, 1996) gives a good prediction of actual plays in those games. However, our Game 3 does not have partial common interests, and thus this concept does not have any force in predicting behaviors in it. Put differently, Game 3 is a good material for observing actual plays without partial common interests[5].

Fourth, there seems to be a renewed interest in the working of communication among economists (Glazer and Rubinstein, 2004; Crawford, 2003). However, to the best of our knowledge, there are few experimental results that

[4]For the experimental results in Game 1 and Game 2, see Tables 9 and 10 in the Appendix B. For Game 3, see Table 3 in Section 2 and Table 11 in the Appendix B.

[5]Blume, DeJong, Kim, and Sprinkle (1998b) conducts experiments on cheap-talk games *without* common language to show that informative communication arises when Senders' and Receivers' interests are sufficiently aligned, *i.e.*, the meaning of signs evolves endogenously. See Crawford (1998) and (Camerer, 2003, Ch.7) for a survey of cheap-talk game experiments.

focus on the communication in conflicting situations. Yamamori, Kato, Kawagoe, and Matsui (2004) study how the cheap-talk preplay communication affects the actual play in a dictator game, while the current paper focuses on communication in a game with incomplete information.

The organization of the paper is as follows. Section 2 explains the experimental procedures and presents the experimental results. It is shown that Sender subjects tend to play (a,a) and Receiver subjects tend to play X or Z in response to message a and Y in response to message b. Section 3 examines the predictions of various refinement concepts developed for cheap-talk games. A new theoretical framework given by Stahl and Wilson (1995) and Crawford (2003) is also considered and applied. It is shown that a specific application of Crawford (2003) to our model with incomplete information can explain the experimental results. Section 5 summarizes the results, locate the results in the communication-theoretic literature, and suggests future directions of research.

2 Experiments

As part of a series of cheap-talk game experiments, the third session at Kyoto Sangyo University on 14 July 1999 and the fourth session at Toyo University on 21 December 1999 were focused on Game 1 and 3. In these sessions, half of the subjects who were assigned Game 3 repeatedly played that game only all over the rounds. The player's roles in the game, i.e., the Sender or the Receiver, and the opponent for each subject were not informed in advance but were assigned, according to the schedule designed by the experimenters, in order to eliminate any repeated game or reputation effect. The schedule was designed to guarantee that each subject played with different subject in each round and experienced each player's role as equally often possible. Subjects were also instructed that the role and the opponent were randomly assigned in each round. Average reward was about three thousands yen in the third session. In the fourth session, since the participation fee and the multiplier used to calculate a monetary reward from the number of payoff was halved, the average reward also halved. Instructions and practice time took about an hour and session time was about two hours in each session[6].

Subjects were told that the experiment proceeded according to the steps described below.

1. In each round, subjects were shown payoff table of the game they face in the current round, and were told whether they were the Sender or the Receiver. They could not know with whom they are matched throughout the session.

2. Assignment of games, roles in the games to each subject, and who matched with whom were randomly determined.

3. In each room, twelve out of thirteen subjects actually participated (i.e. made decisions) in the experiment and one subject waited until the next round[7].

4. The Sender was assigned one of two types, "A" or "B," randomly with probability $1/2$. The Sender type was only shown to the Sender and the Receiver could not know the Sender's type before the payoffs for both subjects were determined.

5. The Sender was told to choose between two messages, "I am type A" or "I am type B."

6. The Receiver was shown the Sender's message and was told to choose one of three actions, *A*, *B*, or *C*.

7. Payoffs for both players were determined by the Sender's true type and the Receiver's action according to the payoff tables. After all subjects had made decision, the Sender's true type, the sent message, and the action taken by the Receiver, payoffs for both were revealed separately on the blackboard.

8. A session consisted of thirteen rounds.

9. In the direct reward condition, reward was calculated as fifty times the sum of payoffs earned by each subject throughout the session and paid to her/him in cash. Participation fee was also given to each subject.

10. Prior to the actual experiment, three rounds of practice experiment were conducted, where equilibria and payoffs of the games were different from those used in the actual experiment. Payoffs earned in these practice rounds did not count for final reward calculation.

[6] The detailed descriptions of experimental procedures, including those for Game 1 and 2, are shown in Appendix B, C, and D.

[7] This is because of the nature of matching procedure we adopted. We devised random matching so that each subject plays both player roles and both Sender types as equally often as possible, matched with different subject at each round.

The above procedure was also explained in the written instructions [8].

Next we show our experimental results with respect to outcomes predicted by sequential equilibria. As we noted, there exists no sequential equilibrium other than babbling equilibria (all the sequential equilibria are shown in Table 5).

Table 2 shows the frequency of pure strategy babbling equilibrium plays and the other out of equilibrium outcomes. Recall that we regard the outcome in which the Receiver played Z as babbling equilibrium.

	Session 3	Session 4	Total
Babbling equilibria	43	35	78
out of equilibrium plays	35	43	78
Total	78	78	156

Table 2: The Frequency of Equilibrium Play in Game 3

One can easily observe from Table 2 that pure strategy babbling equilibria were played about half the time (43 (35) out of 78 times in Session 3 (4)). Figure 1 and Figure 2 also show time series data of babbling equilibria and out of equilibrium plays. These figures clearly shows no tendency of convergence to the pure strategy babbling equilibria.

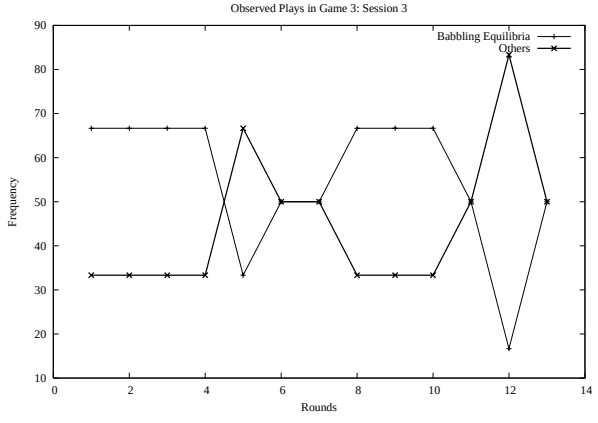

Figure 1: Time series data for Game 3 (Session 3)

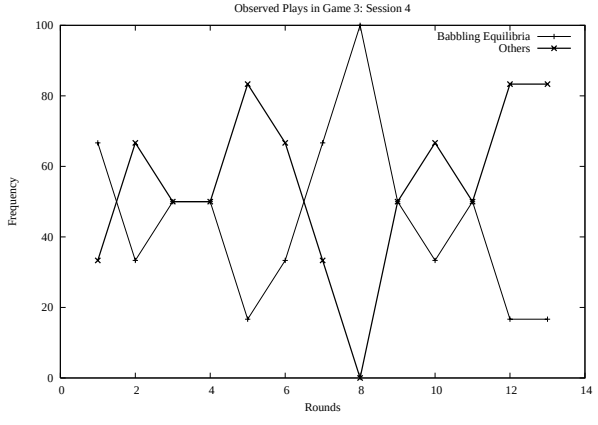

Figure 2: Time series data for Game 3 (Session 4)

Thus, equilibrium plays were not observed as frequently as expected. Next, we would like to show the data

[8] The instructions used for direct reward condition are shown in Appendix C.

arranged by each information set in order to consider which player, the Sender or the Receiver, deviated from equilibrium plays. See Table 3.

	Sender				Receiver					
	$t = A$		$t = B$		$m = a$			$m = b$		
	a	b	a	b	X	Y	Z	X	Y	Z
Session 3	31	8	27	12	19	4	35	3	9	8
Session 4	33	6	27	12	21	9	31	2	11	5
Total	64	14	54	24	40	13	66	5	20	13

Table 3: Data Arranged by Information Set in Each Session

The Sender and the Receiver's choice frequencies in each session shown in Table 3 are also depicted in Figures 3 and 4. Apparently from these figures, most Senders of both types tended to send message a in both sessions. That is, their plays were almost consistent with a pure strategy babbling equilibrium. On the other hand, the Receivers receiving message a tended to choose action Z most frequently and X with the second frequency, and the Receivers receiving message b action Y with the most and Z with the second frequency. One can see from these results that the Receivers who received message b tended to regard the message truthful to choose action Y as a best response. The Receivers who received message a also tended to regard the message truthful in a certain proportion. Hence it is clear that the Receiver's behavior was the main reason for the deviation from pure strategy babbling equilibria in our experiments.

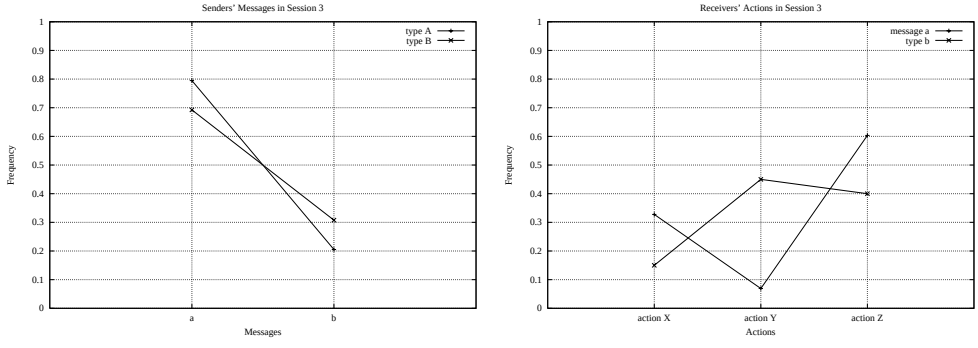

Figure 3: Plays in Session 3

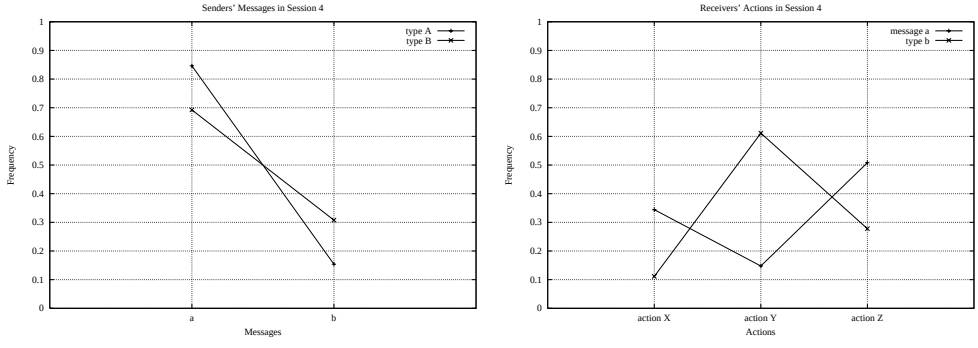

Figure 4: Plays in Session 4

In these experiments, all the subjects played both players' roles: the Sender and the Receiver. So it is worthwhile to see the relation between plays as the Sender and plays as the Receiver in each individual's behavior. We now turn to each individual's data in Session 4[9]. Table 4 gives choice frequency of thirteen subjects, arranged by the cases that they were the Sender or the Receiver.

[9]Unfortunately, the individual data in Session 3 were lost.

The data allow us to identify several prominent patterns in the subjects' behaviors. As for the the Sender, two types can be identified: *aa* type who sent message *a* regardless of the types, and *ab* type who sent message *a* in the case of type *A* and *b* in the case of type *B*. As for the Receiver, there were three types: *ZZ* type who chose action *Z* regardless of messages, *XY* type who chose action *X* upon receipt of message *a* and chose action *Y* with message *b*, and XY' type who had a mixed characteristics of both *ZZ* and *XY* types. Other subjects' behaviors are difficult to characterize.

Subject No.	As sender				As receiver					
	type A		type B		message a			message b		
	a	b	a	b	X	Y	Z	X	Y	Z
1	4	0	2	0	1	1	3	0	1	0
2	4	0	0	2	5	0	0	0	1	0
3	2	0	2	2	2	0	4	0	0	0
4	1	1	2	2	0	2	1	1	2	0
5	4	0	2	0	2	0	3	0	1	0
6	4	0	1	1	3	0	2	0	1	0
7	3	0	3	0	4	0	1	0	1	0
8	3	0	3	0	1	0	4	0	1	0
9	3	0	3	0	1	0	1	0	2	2
10	0	3	2	1	0	2	4	0	0	0
11	0	0	2	4	2	0	2	1	1	1
12	3	2	1	0	0	4	1	0	0	1
13	2	0	4	0	0	0	5	0	0	1
Total	33	6	27	12	21	9	31	2	11	5

Table 4: Individual Behaviors in Session 4

Among the thirteen subjects, Subject No.13 was the only subject who was *aa* type and *ZZ* type when he was sender and receiver respectively, namely, who played according to a pure strategy babbling equilibrium. Subject No.2 was also the only subject who was *ab* type and *XY* type when he was sender and receiver respectively, namely, who played as a truth-teller/believer. The most interesting were the subjects who were the *aa*-type Sender and XY'-type Receiver. That is, these players had quite similar characteristics as shown in the aggregated data. Four players, subjects No.5, 7, 8 and 9, belonged to this class. They are subjects who tended to believe the senders' message to be truthful while they chose messages to hide their types when they were senders. Anyway, except for subjects who were not easily classified, about half of the subjects tended to believe senders' message.

3 Theoretical Predictions

This section reviews various equilibrium refinement concepts developed for cheap-talk games with incomplete information, and examines the prediction of those concepts for the play in Game 3.

3.1 Sequential Equilibrium and its Refinements

Recall that the type space $T = \{A, B\}$, the message space $M = \{a, b\}$, the action space $C = \{X, Y, Z\}$, and the prior distribution is $\pi(A) = \pi(B) = 1/2$ in our game. Figure 5 depicts the game tree of Game 3. Let p (q) be the probability for the type A (B) Sender to send message a. Also let r_1, r_2, r_3 (s_1, s_2, s_3) denote the probability for the Receiver receiving message a (b) to play X, Y and Z respectively. $\beta(A|m)$ denotes the Receiver's belief that the Sender is type A after receiving message m. Table 5 shows all the sequential equilibria in Game 3[10]. In all the sequential equilibrium in the table, the outcome function $o : T \to C$ is constant valued with $o(t) = Z$ for all $t \in T$, that is, babbling equilibria. In this game, if a strategy profile is an equilibrium, then another strategy profile in which the role of messages is interchanged for both Sender and Receiver is also an equilibrium. Equilibria 1 and $1'$, 2 and $2'$ as well as 3 and $3'$ in Table 5 indicate such pairs.

First of all, let us compare the experimental data in the previous section with the predictions given by sequential equilibria. Since most Senders play (a, a) in the experiment, we may focus attention to equilibria 1, 2 and $3'$ where

[10] In equilibrium 3 and $3'$, r_2, r_3, s_2 and s_3 are all positive. In equilibrium 4, p and q satisfy $\frac{1}{3} \leq \frac{p}{p+q}, \frac{1-p}{2-p-q} \leq \frac{2}{3}, p+q \neq 0, p+q \neq 2$.

the Sender plays (a,a). It may appear that 1 or $3'$ is close to the experimental data because the aggregated data of the Receiver's action appear to show that the Receivers tend to play (Z,Y). However, closely inspecting table 4, there is no subject who played (Z,Y) and the Receivers usually play the mixture of X and Z receiving message a and Y receiving message b. The remarkable point here is that the Receiver plays X in response to message a in an unignorable proportion, which never constitutes a sequential equilibrium in this game.

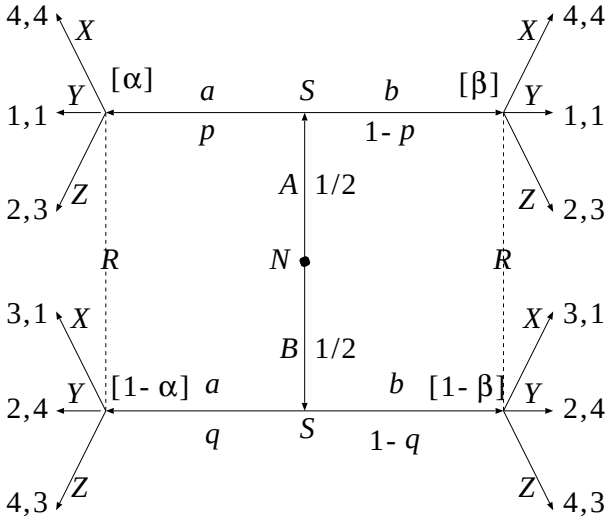

Figure 5: Extensive form of Game 3

	p	q	$\beta(A\|a)$	(r_1,r_2,r_3)	$\beta(A\|b)$	(s_1,s_2,s_3)	outcome
(1)	1	1	$\frac{1}{2}$	$(0,0,1)$	$[0,\frac{1}{3}]$	$(0,1,0)$	(Z,Z)
$(1)'$	0	0	$[0,\frac{1}{3}]$	$(0,1,0)$	$\frac{1}{2}$	$(0,0,1)$	(Z,Z)
(2)	1	1	$\frac{1}{2}$	$(0,0,1)$	$[\frac{1}{3},\frac{2}{3}]$	$(0,0,1)$	(Z,Z)
$(2)'$	0	0	$[\frac{1}{3},\frac{2}{3}]$	$(0,0,1)$	$\frac{1}{2}$	$(0,0,1)$	(Z,Z)
(3)	0	0	$\frac{1}{3}$	$(0,r_2,r_3)$	$\frac{1}{2}$	$(0,0,1)$	(Z,Z)
$(3)'$	1	1	$\frac{1}{2}$	$(0,0,1)$	$\frac{1}{3}$	$(0,s_2,s_3)$	(Z,Z)
(4)	p	q	$[1/3,2/3]$	$(0,0,1)$	$[1/3,2/3]$	$(0,0,1)$	(Z,Z)

Table 5: Sequential Equilibria in Game 3

That there is no sequential equilibria explaining experimental data means that any equilibrium refinement concept does not provide any plausible explanation. See the Appendix E for the various equilibrium concepts we examined. However, among others, Rabin and Sobel (1996)'s argument of deviation dynamics seems to be relevant to the explanation for our experimental data. To appreciate this, it is worthwhile to review the basic idea of equilibrium refinement for cheap-talk games with incomplete information.

Roughly put, the argument of refinement theories for cheap-talk games runs as follows:

1. Pick an equilibrium;
2. In this equilibrium, a subset K of Sender types sends a message that is not used in the equilibrium;
3. Tentatively suppose that the Receiver believes this message, updates his belief accordingly, and acts optimally in response to the new message. Let K' denote the set of types that can earn a higher payoff with the induced action of the Receiver than in the original equilibrium payoff. If $K = K'$, the new message is regarded as credible;
4. If there is no credible messages for deviation, the original equilibrium can be said to be robust.

This argument usually assumes that the message space is sufficiently large that a new message is always available for potential deviant types. Also note that this test for robustness returns the same result for two different equilibria

with the same outcome function, because it only checks if it is possible for any type to have higher payoff than the original payoff.

Now let us see how this test works for our Game 3. First pick equilibrium $((a,a),(Z,Z))$ for instance. In this equilibrium, type A Sender earns 2, type B gets 4, and the Receiver obtains 3 regardless of the type he matches. Suppose that type A Sender sends a new message other than a with the meaning "I am type A" and that the Receiver believes this, and play X in response to this message. Type A Sender will earn 4, which is higher than the original payoff, while type B Sender, if she sends the same message as type A, obtains 3, which is lower than the original payoff. Therefore the set of types who can benefit from this deviation attempt is $\{A\}$ and the new message for deviation is credible. Thus the original equilibrium, therefore all the sequential equilibria in table 5, is not robust.

This way of checking robustness is always plagued by the following criticism that is called Stiglitz critique (Cho and Kreps, 1987; Mathews, Okuno-Fujiwara, and Postlewaite, 1991; Rabin and Sobel, 1996). Suppose that, as the result of deviation as above, type A Sender sends a new message and the Receiver believes it. It is too early to stop the argument there. The Receiver receiving the original message will necessarily deduce that the message is from type B Sender. As a result, he will play Y in response to message a, bringing type B Sender 2. Then, however, type B Sender will find it more profitable to use the same new message used by type A and the Receiver will now optimally play Z in response to the new message as in the original equilibrium.

Mathews, Okuno-Fujiwara, and Postlewaite (1991) respond to this criticism by making the requirements for a valid deviation announcement more stringent. If there is an equilibrium in which type A and type B use different messages, the deviation announcement by type A can be regarded as announcement of her intention to play a new equilibrium. In this equilibrium, it is not optimal for type B to use the same message as used by type A, and the deviation announcement by type A can be justified. In other words, they required the deviation announcement by type A constitute a part of another equilibrium. Rabin and Sobel (1996) pushes ahead with this idea and proposes to consider the dynamics triggered by deviation announcement more explicitly.

Consider a game $\langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ and let $\sigma^* = (\sigma_1^*, \sigma_2^*)$ be an equilibrium of this game. Pick an equilibrium refinement theory and let the set of possible deviations from this equilibrium according to the theory be denoted by $Q(\sigma^*) = (Q_1, Q_2)$. Next construct a deviation correspondence by adding best responses to the deviation as follows. First let $Z_j \subseteq \Delta(S_j)$ denote a set of mixed strategies for player j and let $SBR_i(Z_j)$ be the set of pure strategies in S_i that can be a best response to some distribution over Z_j that assigns positive probability to each element in Z_j. Define the first step of deviation dynamics as follows:

$$\Sigma_i(0) = \begin{cases} \sigma_i^* & \text{if} \quad Q_i = \emptyset \\ Q_i & \text{if} \quad Q_i \neq \emptyset. \end{cases}$$

Using this, further define

$$\Sigma_i(n) = \begin{cases} \sigma_i^* & \text{if} \quad \Sigma_{-i}(n-1) = \sigma_{-i}^* \\ \Sigma_i(n-1) \cup SBR_i(\Delta(\Sigma_{-i}(n-1)) & \text{if} \quad \Sigma_{-i}(n-1) \neq \sigma_{-i}^*. \end{cases}$$

If S_i is finite, there exists some n^* such that for every i and every k, $\Sigma_i(n^*) = \Sigma_i(n^*+k)$, which we denote as Σ_i^*. A deviation correspondence is (Σ_S^*, Σ_R^*) thus constructed, and is denoted by $D(\sigma^*)$. The set of equilibria in $D(\sigma^*)$ be denoted by $ED(\sigma^*)$. A stable equilibrium σ^* with respect to the refinement theory sastifies $D(\sigma^*) = \{\sigma^*\}$ by definition. They weaken this condition to define a quasi-stable equilibrium by $ED(\sigma^*) = \{\sigma^*\}$. They further consider the dynamics among equilibria triggered by a deviation. A set of equilibria $\mathbf{E}$ is said to be a recurrent set if $ED(\sigma) \subseteq \mathbf{E}$ for every $\sigma \in \mathbf{E}$ and it is the minimal set among the set satisfying this condition. Each element in a recurrent set is said to be a recurrent equilibrium.

Let us now return to our game and consider how a deviation process evolves [11]. Suppose first that equilibrium $((b,b),(\cdot,Z))$ is played. Since message a is off the equilibrium path and the Receiver's best response depend on his belief upon receiving a, we do not specify his play here. Suppose next that type A Sender succeed in making the Receiver believe it is type A by sending message a. We thus have the first step in the deviation dynamics, $\Sigma_S(0) = (a,b), \Sigma_R(0) = (X,Z)$. This further evolves as $\Sigma_S(1) = \{(a,b)\}$, $\Sigma_R(1) = \{(X,Z),(X,Y)\}$, $\Sigma_S(2) = \{(a,b),(a,a)\}$, $\Sigma_R(2) = \{(X,Z),(X,Y)\}$, $\Sigma_S(3) = \{(a,b),(a,a)\}$, $\Sigma_R(3) = \{(X,Z),(X,Y),(Z,\cdot)\}$. See also Figure 6. This means that all the pure strategy sequential equilibria in Table 5 forms a recurrent set and all the pure strategy equilibria are recurrent equilibria. However, it is worth noting that $((a,b),(X,Y))$ and $((a,a),(X,Y))$ appear in this course of deviation process.

[11] We apply Farrell (1993)'s neologism-proofness concept here. However, the following argument does not change even if we consider Mathews, Okuno-Fujiwara, and Postlewaite (1991)'s weakly coredible announcement or credible announcement. See the Appendix E for the details of these concepts

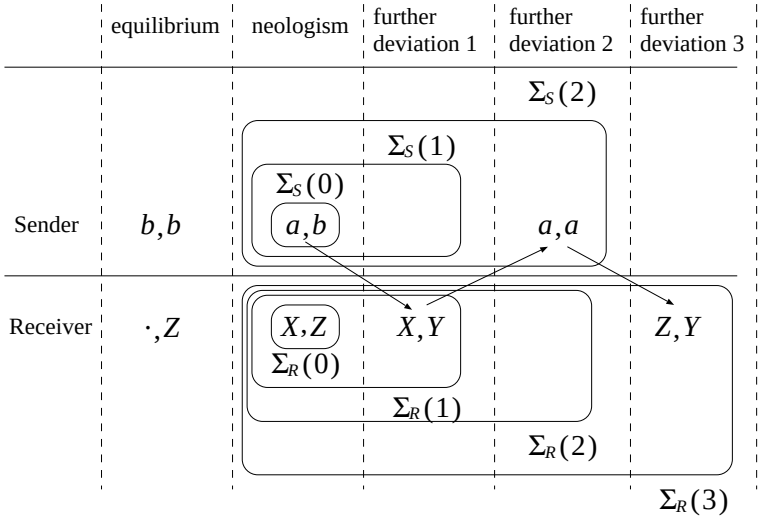

Figure 6: A Deviation Dynamics in Game 3

Restricting focus on cheap-talk games, Rabin and Sobel (1996) defines a game with "partial common interests" to characterize recurrent equilibria by the payoff structure. A cheap-talk game is said to have partial common interests if there exists a partition $J_1, \cdots, J_j$ of the type space T such that the following three conditions hold.

1. Suppose the Receiver Bayes-updates the prior distribution, knowing the Sender's type is in J_i, and choose a best response. Then every type $t_i \in J_i$ earns strictly higher payoff than in the babbling equilibrium.

2. For every type in J_i, the minimum payoff she obtains when the Receiver chooses a best action to some belief concentrated on J_i is strictly greater than the maximum payoff she obtains when the Receiver chooses a best action to some belief concentrated on $J_k (k \neq i)$.

3. Suppose L satisfies $L \cap T_k$ for at least two k's. Then there exists some type $t_i \in L \cap J_k$ whose minimum payoff she obtains when the Receiver with belief concentrated on J_k choose a best response is strictly greater than the maximum payoff she obtains when the Receiver with belief concentrated on L chooses a best response.

Rabin and Sobel (1996) shows that, in a game with partial common interests, a babbling equilibrium is not recurrent with respect to the concept of weakly credible announcement defined by Mathews, Okuno-Fujiwara, and Postlewaite (1991). So players will not play a babbling equilibrium forever. Blume, DeJong, Kim, and Sprinkle (1998a)'s experimental study confirms the validity of the concept of partial common interests. It is easy to see that our Game 3 violates the condition 2 above. Thus, our Game 3 does not have partial common interests and every pure strategy sequential equilibrium is recurrent.

To summarize, the existing refinement theories for cheap-talk game cannot explain our experimental results [12]. The most intriguing part of those discussions is that the strategy profile $((a,a),(X,Y))$ and $((a,b),(X,Y))$ appears in a deviation dynamics proposed by Rabin and Sobel (1996). The next subsection considers a theoretical framework that take into consideration such off the equilibrium plays.

3.2 A Model with Boundedly Rational Types

Crawford (2003) analyses a game that rational players play believing that their opponent may be boundedly rational types with some probability, to explain how the Allies succeeded in thwarting German army's expectation in the Operation Fortitude (D-day). His model is basically the same as the well studied models with some behavioral types as in Kreps and Wilson (1982) and Milgrom and Roberts (1982). However, while these preceding papers are concerned with the set of equilibria as the fraction of boundedly rational behavioral types approaches zero, Crawford (2003) focuses on equilibria when there is substantial probability that the opponent is of boundedly

[12] McKelvey and Palfrey (1995, 1998) propose an equilibrium concept with noisy best response, called Quantal Response Equilibrium (QRE), assuming that players play an action with higher payoff with higher probability, but the play is always accompanied by noise. We examined how this concept fares to explain our data. However, it turned out that this concept is plagued by multiplicity when applied to cheap-talk games. We are preparing another paper for the detailed analysis of QRE in our games.

rational types. Since the game he analyzes has the structure that players try to outguess the opponent's play and every strategy is a best response to some strategy of the opponent, it is also an application of Stahl and Wilson (1995)'s model that presumes the existence of players with various level of bounded rationality.

We first have to decide on the boundedly rational types to be considered in our game. Recall now that the message space in our game is intentionally made "naïve" to create the situation where each message corresponds to truth-telling or lying. On the Sender's side, we focus attention on the following two strategies among $(a,a),(a,b),(b,a),(b,b)$ of Sender's pure strategies.

(a,a) pooling type: tells the truth when she is type A and pretends to be type A when type B.

(a,b) separating type: always tells the truth.

Note that our messages have literal meaning in themselves and are easily identified as truth-telling or lying. Note also, while type B has an incentive to pretend to be type A, type A has no such an incentive. These considerations led us to exclude bb and ba. In fact, there is no subject who sent b when type A and sent a when type B. Let s_p and s_t denote the probabilities for the Sender to be pooling and separating types respectively. Let the probability for the Sender to be rational (sophisticated) be denoted by s_s. We assume that $s_p+s_t+s_s=1$ and $s_p,s_t,s_s>0$.

The Receiver's pure strategies are $(X,X),(X,Y),(X,Z),(Y,X),(Y,Y),(Y,Z),(Z,X),(Z,Y)$ and (Z,Z). Among these, (X,X) and (Y,Y) are never best responses. In choosing plausible boundedly rational strategies, we focus on whether the Receiver believes the Sender's message to be truthful or not. The strategy to believe the Sender's message to be truthful is (X,Y). To question the credibility of message a is also a plausible strategy, since only type B has an incentive to tell a lie making message a incredible. Among the candidate strategies with a form $(Z,\cdot)$, we choose to pick (Z,Y). Thus boundedly rational types we decided to consider are the following two types.

(X,Y) naïve type: believes both messages to be truthful.

(Z,Y) suspicious type: disbelieves only message a.

Let r_b and r_i denote the probability for the Receiver to be naïve type and suspicious type respectively. Also let r_s be the probability for the Receiver to be sophisticated. We assume $r_b+r_i+r_s=1$ and $r_b,r_i,r_s>0$.

Suppose now that a sophisticated Sender and a sophisticated Receiver play the game, believing that the opponent is boundedly rational types with some probability. As an example, suppose that the sophisticated Sender plays aa, while the sophisticated Receiver plays XY. On the Sender's side, the pooling type and the separating type of type A, and the sophiticated Sender of both types A and B are sending message a. Therefore the probability for the sophisticated Receiver to receive message a is $\frac{1}{2}s_t+s_p+s_s=1-\frac{1}{2}s_t$, and the conditional probability that the Sender's true type is A and B upon receiving a is $\frac{1}{2-s_t}$ and $\frac{1-s_t}{2-s_t}$ respectively. Thus the Receiver's expected payoff when he plays X upon receiving a is

$$4\cdot\frac{1}{2-s_t}+1\cdot\frac{1-s_t}{2-s_t}=\frac{5-s_t}{2-s_t}.$$

Similarly, the probability of receiving message b is calculated as $\frac{1}{2}s_t$ and the conditional probability that the Sender's true type is A and B upon receiving b is 0 and 1 respectively. Therefore the Receiver's expected payoff when he plays Y upon receiving b is 4. Taken together, the expected payoff to the sophisticated Receiver when he plays XY is

$$(1-\frac{1}{2}s_t)\frac{5-s_t}{2-s_t}+\frac{1}{2}s_t\cdot 4=2.5+1.5s_t.$$

On the other hand, the sophisticated Sender who plays aa is faced with X with probability r_b+r_s, with Z with probability r_i. Thus when she is type A, she obtains 4 with probability r_b+r_s and 2 with probability r_i, while she earns 3 and 4 with the same probabilities when she is type B. Her expected payoff is

$$\frac{1}{2}(4(r_b+r_s)+2r_i)+\frac{1}{2}(3(r_b+r_s)+4r_i)=3.5(r_b+r_s)+3r_i=3+0.5(r_b+r_s).$$

Thus calculated expected payoffs to the sophisticated Sender and the sophisticated Receiver for all the possible pure strategy profiles are shown in Table 6. The left number refers to the payoff to the Receiver, the right number the Sender. Note that both messages a and b are necessarily used with positive probability, because of the presence of boundedly rational types.

Table 7 summarizes all the pure strategy sequential equilibria of the game between the sophisticated Sender and Receiver. Note equilibria $((a,a),(Z,Y))$ and $((a,a),(X,Y))$ exist with no restrictions on r_b or r_s; $((a,a),(Z,Y))$

	aa		ab		ba		bb	
XX	2.5	$3+0.5(r_b+r_s)$	2.5	$2+r_b+1.5r_s$	2.5	$2+1.5r_s+0.5r_i$	2.5	$1+2r_s$
XY	$2.5+1.5s_t$	$3+0.5(r_b+r_s)$	$2.5+1.5(s_t+s_s)$	$2+r_b+r_s$	$1+1.5s_p+3s_t$	$2+0.5r_i$	$2.5+1.5s_t$	1.5
XZ	$2.5+s_t$	$3+0.5(r_b+r_s)$	$2.5+(s_t+s_s)$	$2+r_b+2r_s$	$2+0.5s_p+1.5s_t$	$2+0.5r_s+0.5r_i$	$2.5+s_t+0.5s_s$	$1.5+1.5r_s$
YX	$1+1.5(s_p+s_s)$	$1.5+2r_b+1.5r_i$	$1+1.5s_p$	$2+r_b$	$1+1.5s_p+3s_s$	$2+0.5r_i+r_s$	$1+1.5(s_p+s_s)$	$1.5+2r_s$
YY	2.5	$1.5+2r_b+1.5r_i$	2.5	$1.5+1.5r_b+0.5r_i$	2.5	$1.5+0.5r_b+r_i$	2.5	1.5
YZ	$2+0.5(s_p+s_s)$	$1.5+2r_b+1.5r_i$	$2+0.5s_p$	$2+r_b+0.5r_s$	$2+0.5s_p+1.5s_s$	$2+0.5r_i$	$2+0.5s_p+s_s$	$1.5+1.5r_s$
ZX	$2+(s_p+s_s)$	$3+0.5r_b$	$2+s_p$	$2+r_b+0.5r_s$	$2+s_p+1.5s_s$	$2+0.5r_i+2r_s$	$2+s_p+0.5s_s$	$1.5+2r_s$
ZY	$3+0.5s_t$	$3+0.5r_b$	$3+0.5(s_t+s_s)$	$2+r_b$	$2+s_p+1.5s_t$	$2+0.5(r_i+r_s)$	$2.5+s_t+0.5s_p$	1.5
ZZ	3	$3+0.5r_b$	3	$2+r_b+r_s$	3	$2+0.5r_i+r_s$	3	$1.5+1.5r_s$

Table 6: Payoff Table of the Reduced Form Games Played between Sophisticated Players

arises when $s_t < 1/2$ and $((a,a),(X,Y))$ when $s_t > 1/2$. The other two equilibria arise with strong restrictions on r_b and r_s, and only if $s_t < 1/3$. Equilibria $((a,a),(Z,Y))$ and $((a,a),(X,Y))$ are also relevant to our experimental results. In these equilibria, the type *A* Sender tells the truth, while the type *B* Sender tells a lie, which coincides with the Senders' behaviors in our experimental data. On the Receiver's side, the Receiver believes or disbelieves message *a*, while he believes message *b*, which also coincides with the Receivers' behaviors in our experimental results[13]. These equilibria not only show that the sophisticated Receiver can play X in the face of message *a*, but also that that he plays Y for sure upon receiving message *b*.

Equilibria	Conditions for Parameters
(aa,XY)	$s_t > 1/2$
(aa,ZY)	$s_t < 1/2$
(ba,ZX)	$2s_s - 1 < s_t < 0.5s_s, r_b < 1.5r_s - 0.5$
(ba,YX)	$s_t < 2s_s - 1, r_b < 2 - 3r_s, r_b < 2r_s - 0.5$

Table 7: Pure Strategy Sequential Equilibria

Equilibrium $((a,a),(X,Y))$ arises when the sophisticated Receiver believes that the Sender is a separating type with substantial probability, even if the sophisticated Sender plays (a,a). Thus what is important is the Receiver's belief that the Sender is a boundedly rational type with some probability. This assumption seems to be plausible when we consider a one-shot game. However, it is not clear why the sophisticated players continue to believe that boundedly rational types are present in the opponent players in the long run supporting the above equilibria. This question seems to await further analysis with the evolutionary formulation. Potentially important for such analyses are the payoffs that those boundedly rational types earn in the equilibria. In both $((a,a),(X,Y))$ and $((a,a),(Z,Y))$, the pooling type (a,a) earns the same payoff as the sophisticated Sender, while the payoff to the separating type *ab* is lower than that to the sophisticated Sender. In equilibrium $((a,a),(X,Y))$, naïve type (X,Y) earns as mush as the sophisticated Receiver and in equilibrium $((a,a),(Z,Y))$, suspicious type obtains as much as the sophisticated Receiver.

4 Concluding Remarks

4.1 Summary of the Experimental Results and Analysis

We conducted, and presented the result of, experiments of a cheap-talk game with incomplete information in which one sender type has an incentive to misrepresent her type. The game can be regarded as representing a simplest possible situation of communication with conflicting interests: it has two possible types and two possible messages for the Sender, and three actions for the Receiver; a common language is shared between the Sender and the Receiver; and the message space is intentionally made "naïve" to create the situation where each message corresponds to truth-telling or lying.

The experimental results reported and analyzed in this paper are different from the existing literature. According to the previously reported experimental results of cheap-talk games under a common language environment, the concept of "partial common interests" has proven to give a good prediction of actual plays in those games. However, the concept of partial common interests does not give a sharp prediction for our Game 3 because there is no partition of the type space in which all the Sender types in each partition set feels comfortable belonging to that set. Our game is a good material for exploring what behaviors the Sender and the Receiver show in a game without partial common interests.

It was observed that the Sender type with an incentive to misrepresent her type mostly lied, leading to the loss of credibility of the Sender's messages. This is not particularly surprising since it is consistent with the Sender's strategy in a sequential equilibrium of the game. The most surprising part of our experimental results is that although the credibility of messages was lost, the Receivers mostly believed the Senders' messages to be truthful. As we examined in Section 3, these results cannot be explained by the existing refinement concepts such as neologism-proofness, while the model incorporating some fraction of boundedly rational types, recently proposed by Crawford (2003) and Stahl and Wilson (1995), explains these results under certain conditions.

[13] As is stated in the previous section, there is no subject who played *ZY* in the experimental data. They usually played the mixture of *X* and *Z* upon receiving *a*. However, it is conceivable that the subject adopted *XY* and *ZY* as their belief fluctuates.

4.2 Relation to the Communication-Theoretic Literature

Our experimental results also seem to be relevant to a field of communication theory that has studied lying/deception. As stated above, one sender type have an incentive to misrepresent her type in the game we study. It is then of interest to us whether the Sender actually lies and/or whether the Receiver succeeds in spotting lies. In this sense, we share some interest with communication researchers, although our experimental environment is unique in the sense to be stated more precisely below.

In order to locate our experiments in the context of communication theory, it would be worthwhile to briefly review the communication-theoretic literature on deception[14]. Previous research in communication theory that has focused on deception has centered on nonverbal behaviors associated with uncontrollable psychological processes (Vrij, 2000). However, these studies show that various nonverbal cues, such as the pitch of a voice and eye movement, are not necessarily reliable signs for detecting deception, and even well-trained specialists cannot distinguish between truth-telling and lies with more than 60% accuracy rate (Burgoon, Buller, Guerrero, Afifi, and Feldman, 1996). In an alert and suspicious environment, a truth teller's adaptation to a false accusation strikes the respondent as devious, which is called "Othello error"(Ekman, 1985).

To advance deception studies a step further, certain communication theorists have turned their focus on verbal behaviors or controlled message activity in the laboratory and classified several types of deceptive messages. There seem to be two approaches in this strand of study: thinking of deceptive messages as distinct strategies, or thinking of deceptive messages as message forms resulting from the manipulation of information in different way.

Among the researchers who have taken the latter approach, McCornack (1992), based on Grice (1989)'s "cooperative principle," defines deception as a violation of one or more maxims of the cooperative principle, and proposes Information Manipulating Theory (IMT). Buller and Burgoon (1996) independently proposes a similar theory, Interpersonal Deception Theory (IDT), based on two-way communication model as a criticism against previous communication research that has considered one-way communication. Three major categories of deceptive message, falsification, concealment and equivocation, are identified by these theories according to the amount of information contained in the message.

On the other hand, McCornack and Parks (1986) coined the word "truth bias," the presistent presumption that the partners are telling the truth. As part of a hypothesis that the relational development leads to decreases in the accuracy of deception detection, McCornack and Parks (1986) propose the hypothesis that increases in the confidence in truth/lie judgement lead to increases in the presumption of honesty, truth bias. The subjects of their experiment were premarital romantic couples with varying degree of relational development. One partner was asked to tell the truth or lie about several questions with some explanation, and this was recorded in a videotape. The other partner was shown this vedeotape containing a series of truthful and deceptive statements, and was told to give truth/lie judgement. They confirmed the existence of truth bias in this environment.

Our experimental results also confirm the existence of the "truth bias," but in quite a distinct environment from theirs. First, our experiment was conducted in a one-shot anonymous environment with no room for relational development. Second, there was no nonverbal cues available to the Receiver, and the Sender was restricted to use the simplest possible messages. Third, the situations (the game structure and payoffs) were made common knowledge between the Sender and the Receiver in our experiment. Matching was designed so that all subjects experience both player's roles and both Sender types as equally often as possible. So, they should be able to understand the strategic situation they face both as the Sender and the Receiver. Fourth, in our experiment, the Sender strategically chose to tell the truth or lie.

Note especially the first and the third point above. That truth bias was observed even in this enviromnent means that truth bias was confirmed in a very strong sense. Without any relational development and with the conflicting situation being common knowledge, truth bias exists as long as common language are shared between the Sender and the Receiver, which may be called "fundamental truth bias." Although the Receiver can infer from the situation that the Sender may lie, he may be deceived by her message.

Our experimental results are also related to "truth detection bias." Burgoon, Buller, Guerrero, Afifi, and Feldman (1996)'s experimental results show that falsification is most difficult to detect, while equivocation is easiest to detect. Burgoon, Buller, Ebesu, and Rockwell (1994) show that for novices as well as experts, such as military intelligence instructors, accuracy rate of detection was much higher on truthful messages than on deceptive ones. This is called "truth detecting bias," a well-known but still disputable phenomenon (Vrij, 2000; Holm, 2004). We can also measure the rate of truth and lie detection in our experimental data, but no conclusive conclusion can be drawn[15].

[14] See Griffin (2003, Ch.7) for a survey of the interpersonal deception theory and its variants.

[15] We say that truth-telling is detected if the type A Sender sends message a and the Receiver responds with action X or if the type B Sender

4.3 Future Direction of Research

The model proposed by Crawford (2003) explains our experimental results rather well. However, there remains the question why boundedly rational types in the model could persist in the long run. To answer this question, we need to develop another evolutionary model that involves conflicting interests between the Sender and the Receiver.

Although, in the economic and game-theoretic literature, Matui (1991), Wärneryd (1993), Blume, DeJong, Kim, and Sprinkle (1998b), Rubinstein (2000, Ch.2) and others have made attempts to model the evolution of meaning of a language, they only considered the situation with common interests between the Senders and the Receivers. The experimental results and analysis presented in the paper suggest that theoretical explanation for the evolution of language should take into account the situation with conflicting interests and should address why truthful communication survives in the environment in which lying is successful. We might be able to have a different picture of the evolution of language if we do so. Our language may be necessarily vague as a result of the equilibrium of this process (Lipman, 2001).

A key could be the basic idea in Grice (1989)'s cooperative principle that communication is an attempt to determine truth value through the statements made in conversation and the conversation is intrinsically a cooperative task. Another research will be needed to explore this point, however.

sends message *b* and the Receiver responds with action *Y*. Similarly, we say that a lie is detected if the type A Sender sends message *b* and the Receiver responds with action *X* and if the type B Sender sends message *a* and the Receiver responds with action *Y*. Then, for type A Senders, truth were detected in 12 out of 39 cases in Session 3 and 16 out of 39 cases in Session 4, and lies were detected 0 out 39 cases in Session 3 and 2 out of 39 cases in Session 4. For type B Senders, truth were detected in 7 out of 39 cases in Session 3 and 1 out of 39 cases in Session 4, and lies were detected 11 out of 39 cases in session 3 and 11 out of 39 cases in Session 4. Thus, as the frequency of truth and lies detection are asymmetric for type A and B, we cannot say that neither detection bias was observed clearly.

References

BERG, J., L. DALEY, J. DICKHAUT, AND J. O'BRIEN (1986): "Controlling Preferences for Lotteries on Units of Experimental Exchange," *Quarterly Journal of Economics*, 101, 281–306.

BLUME, A., D. DEJONG, Y. KIM, AND G. SPRINKLE (1998a): "Evolution of Communication with Partial Common Interest," *Games and Economic Behavior*, 37, 79–120.

——— (1998b): "Experimental Evidence on the Evolution of Meaning of Messages in Sender-Receiver Games," *American Economic Review*, 88, 1323–1340.

BULLER, D. B., AND J. K. BURGOON (1996): "Interpersonal Deception Theory," *Communication Theory*, 6, 203–242.

BURGOON, J. K., D. B. BULLER, A. S. EBESU, AND P. ROCKWELL (1994): "Inter Personal Deception: V. Accuracy in Deception Detection," *Communication Monographs*, 61, 303–325.

BURGOON, J. K., D. B. BULLER, L. K. GUERRERO, W. A. AFIFI, AND C. M. FELDMAN (1996): "Interpersonal Deception: XII. Information Management Dimensions underlying Deceptive and Truthful Messages," *Communication Monographs*, 63, 50–69.

CAMERER, C. (2003): *Behavioral Game Theory*. Princeton University Press, Princeton, NJ.

CHO, I.-K., AND D. KREPS (1987): "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics*, 102, 179–221.

COOPER, R., D. DEJONG, R. FORSYTHE, AND T. ROSS (1992a): "Communication in Coordination Games," *Quarterly Journal of Economics*, 107, 739–71.

——— (1992b): "Communication in the Battle of Sexes Game," *RAND Journal of Economics*, 20, 568–87.

CRAWFORD, V. (1998): "A Survey of Experiments on Communication via Cheap Talk," *Journal of Economic Theory*, 78, 286–98.

CRAWFORD, V. (2003): "Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions," *American Economic Review*, 93(1), 133–149.

CRAWFORD, V., AND J. SOBEL (1982): "Strategic Information Transmission," *Econometrica*, 50, 1431–51.

DICKHAUT, J., K. MCCABE, AND A. MUKHERJI (1995): "An Experimental Study of Strategic Information Transmission," *Economic Theory*, 6, 389–403.

EKMAN, P. (1985): *Telling Lies*. W. W. Norton & Company, New York.

FARRELL, J. (1993): "Meaning and Credibility in Cheap Talk Games," *Games and Economic Behavior*, 5, 514–531.

GLAZER, J., AND A. RUBINSTEIN (2004): "On Optimal Rules of Persuation," *Econometrica*, 72(6), 1715–1736.

GRICE, P. (1989): *Studies in the Way of Words*. Harvard University Press, Cambridge, MA.

GRIFFIN, E. (2003): *A First Look at Communication Theory*. McGraw-Hill, New York.

HOLM, H. (2004): "Biases in Bluffing — Theory and Experiments," Lund University.

KAWAGOE, T., AND H. TAKIZAWA (1999): "Instability of Babbling Equilibrium in Cheap-Talk Games," mimeo.

KREPS, D., AND R. WILSON (1982): "Reputation and Imperfect Information," *Journal of Economic Theory*, 27(2), 253–79.

LIPMAN, B. L. (2001): "Why is Language Vague?," mimeo.

MATHEWS, S., M. OKUNO-FUJIWARA, AND A. POSTLEWAITE (1991): "Refining Cheap-talk Equilibria," *Journal of Economic Theory*, 55, 247–273.

MATUI, A. (1991): "Cheap-Talk and Cooperation in a Society," *Journal of Economic Theory*, 54, 245–258.

MCCORNACK, S. (1992): "Information Manipulation Theory," *Communication Monographs*, 59, 1–16.

MCCORNACK, S., AND M. PARKS (1986): "Deception Detection and Relationship Development: The Other Side of Trust," in *Communication Yearbook 9*, ed. by McLaughlin. Sage Publications, Beverly Hills, CA.

MCKELVEY, R., AND T. PALFREY (1995): "Quantal Response Equilibria for Normal Form Games," *Games and Economic Behavior*, 10, 6–38.

——— (1998): "Quantal Response Equilibria for Extensive Form Games," *Experimental Economics*, 1, 9–41.

MILGROM, P., AND J. ROBERTS (1982): "Predation, Reputation, and Entry Deterrence," *Journal of Economic Theory*, 27(2), 280–312.

RABIN, M., AND J. SOBEL (1996): "Deviations, Dynamics and Equilibrium Refinement," *Journal of Economic Theory*, 68, 1–25.

ROTH, A., AND W. MALOUF (1979): "Game-Theoretic Models and the Role of Bargaining," *Psychological Review*, 86, 574–94.

RUBINSTEIN, A. (2000): *Economics and Language*. Cambridge University Press, Cambridge, UK.

STAHL, D., AND P. WILSON (1995): "On Players' Model of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, 10, 218–254.

VRIJ, A. (2000): *Detecting Lies and Deceit*. John Wiley & Sons, West Sussex, UK.

WÄRNERYD, K. (1993): "Cheap Talk, Coordination, and Evolutionary Stability," *Games and Economic Behavior*, 5, 532–546.

YAMAMORI, T., K. KATO, T. KAWAGOE, AND A. MATSUI (2004): "Voice Matters in a Dictator Game," COE Discussion Paper COE-F-44 (September 2004), University of Tokyo.

Appendices

The purpose of these appendices is to cover the points that we could not touch on in detail in the text. Section A shows the characterization of the three games used in our experiment within a general framework of cheap-talk games with two types and three actions. Section B explains more details of the experiments we conducted, including the experimental results for Game 1 and 2 as well as the experimental data for Game 3 in Sessions 1-2. The instructions and the recording sheet for direct reward condition we used in the experiments are found in Sections C and D. Finally Section E explains refinement theories for cheap-talk games in more depth than in the text.

A Characterization of the Three Games Used in the Experiments

Even the simplest possible cheap-talk game with two types and three actions can encompasses diverse incentive situations between the Sender and the Receiver. Two distinct dimensions of incentives seems to be identified in this setting, although they are closely related to each other. On the one hand, each Sender type may or may not have aligned preference over actions with the Receiver. On the other hand, each Sender type may or may not prefer to disguise herself as a different type.

To see this point more clearly, it is instructive to look at the payoff functions adopted by Crawford and Sobel (1982). In their model, the payoff to the Sender of type t from the Receiver's action y is $-(y-(t+d))^2$ and the Receiver's payoff is $-(y-t)^2$, where t is drawn from the unit interval. This is a situation where the Receiver wants to choose action that is equal to the Sender's type, while the Sender wants the Receiver to take an action that equals to the sum of his type and d. Thus d is a single parameter that expresses the degree of preference alignment between the Sender and the Receiver. Note that d is a constant, which means all the Sender types have the same incentives to be regarded as a type that is larger than the true type by d. Their analysis concentrated on how the alignment of preferences influences equilibrium behavior in cheap-talk games, abstracting away the interaction between different Sender types.

We wanted to study situations with a more complicated incentive interaction between different Sender types. Suppose, in a generic cheap-talk game with two types and three actions, each type has strict (ordinal) preference relation over the set of Receiver's actions. Since each Sender type has six possibilities of such preference relation, there are thirty-six possible combinations of both types' preference relations. Focusing on the possibly different incentive of each Sender type to represent herself, we decided to consider three basic cases as follows:

Case 1 Both type A and B want to be correctly identified, inducing the Receiver to choose X and Y respectively;

Case 2 Both types want the Receiver to play Z, that is, they want to confuse the Receiver;

Case 3 Type A wants to be correctly identified, while type B wants to misrepresent herself as type A.

Thus, based on Crawford and Sobel's payoff functions, we created three cheap-talk games that correspond to the above cases by making d type-dependent as in Blume, DeJong, Kim, and Sprinkle (1998a). We discretized the type space of Crawford and Sobel such that $t = 1/4$ for type A and $t = 3/4$ for type B. Then, Case 1 can be characterized by setting $d = 0$ for both types. While Case 2 can be created by setting $d = 1/4$ for type A and $d = -1/4$ fot type B, in Case 3 $d = 0$ for type A and $d = -1/3$ for type B. See Figure 7. Thus obtained payoffs were converted by affine transformation and some adjustment was made to have round numbers.

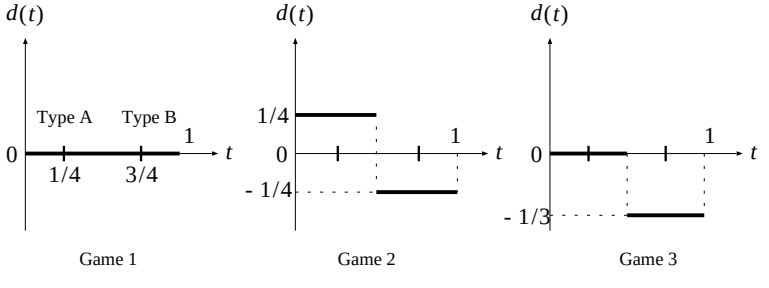

Figure 7: Incentives of different types

Table 8: Differences of Experimental Designs in Sessions 1-4

Session	1	2	3	4
Location	Chuo Univ.	Saitama Univ.	Kyoto Sangyo Univ.	Toyo Univ.
Date	Feb. 2, 1999	May 18, 1999	July 14, 1999	Dec. 21, 1999
Sample	26 undergrads	13 undergrads	26 undergrads	26 undergrads
Rewarding Method	direct/lottery	direct	direct	direct
Labeling	as in Table 1	premuted	permuted	permuted
Experimented Game	1, 2, 3	1, 2, 3	1, 3	1, 3

B Details of Experiments

Our experiment consists of four sessions, which were conducted at different time and location and with different subjects. In the first session, we carefully followed the procedure adopted by Cooper, DeJong, Forsythe, and Ross (1992b) and Cooper, DeJong, Forsythe, and Ross (1992a). Although they only deal with cheap-talk games with complete information, the feature of their experimental design seems to have become a "standard" in conducting a cheap-talk game experiment in general in the following senses. First, they apply lottery reward procedure that was first developed by Roth and Malouf (1979) and further extended by Berg, Daley, Dickhaut, and O'Brien (1986) to induce risk-neutral utility function from subjects. We randomly assigned subjects into two groups of an equal size: direct reward condition and lottery reward condition. In the lottery reward condition, subjects were faced with a two-prize lottery where the payoff they earn was proportionally reflected in the probability of winning the higher prize. Secondly, Cooper, DeJong, Forsythe, and Ross (1992b) and Cooper, DeJong, Forsythe, and Ross (1992a) carefully constructed a procedure for matching subjects.

The experimental result of the first session led us to change the experimental procedure in three directions:

1. Since the experimental results in direct and lottery reward conditions did not differ significantly in the first session, we decided to adopt direct reward method from the second session on;

2. In the payoff tables used in the first session, Receiver's action A (B) was the best response when the Sender's true type is A (B) respectively. To prevent the labels for the Receiver's action from working as a coordination device, we permuted the labels in sessions 2-4. For example, the action which was the best response for the Receiver to type A (B) Sender was relabeled as action B (C) respectively and the best response to the prior distribution was relabeled as A;

3. In the first and second sessions, each subject was randomly assigned Games 1, 2, and 3. From the third session on, we redesigned experiment so that each subject in a group faced the same game (Game 1 or 3) throughout the session, although we tried to give each subject as equal opportunities as possible to be the Sender or the Receiver, and to be type A or type B.

Table 8 summarizes those differences of experimental design by sessions. Instructors other than the authors of the paper read aloud the instructions and conducted experiments manually. The instructors knew nothing about the equilibria of the games.

Table 9 summarizes the experimental data for Game 1 through Sessions 1-4. As noted in the Introduction in the text, obvious tendency for a separating equilibrium play is observed. Table 10 summarizes the experimental data for Game 2 through Sessions 1-2. Obviously, there is a tendency for a separating equilibrium play in this game too. In the text, we used experimental data in Sessions 3 and 4 for the analysis of Game 3, because those sessions are focused on Game 3. Table 11 summarizes the experimental data for Game 3 in Sessions 1-2. It is easy to see that the same tendency prevails in Sessions 1 and 2 as in Sessions 3 and 4.

C Instructions

This is an experiment on economic decision making. You can earn some amount of money in cash in this experiment, if you make appropriate choices according to what is explained below.

In this experiment, each group consists of two persons, one of whom we call "S-player" and the other "R-player." Scores for both players are determined by choices of both players. We will not inform you who are "S-players (R-players)" or who are matched with whom at each round. Matching is determined at random at each round. In each round, one of you has to "wait" and do nothing until the next round.

	Sender				Receiver					
	$t=A$		$t=B$		$m=a$			$m=b$		
	a	b	a	b	X	Y	Z	X	Y	Z
Session 1, direct	12	1	0	13	11	0	1	0	13	1
Session 1, lottery	13	0	0	13	13	0	0	0	12	1
Session 2	12	1	0	13	10	0	2	1	1	12
Session 3	37	2	3	36	32	0	8	0	33	5
Session 4	38	1	1	38	36	0	3	0	37	2
Total	112	5	4	113	102	0	14	1	96	21

Table 9: Experimental Results for Game 1

	Sender				Receiver					
	$t=A$		$t=B$		$m=a$			$m=b$		
	a	b	a	b	X	Y	Z	X	Y	Z
Session 1, direct	10	3	0	13	5	1	4	2	11	3
Session 1, lottery	10	3	2	11	12	0	0	0	13	1
Session 2	11	2	2	11	7	0	6	4	0	9
Total	31	8	4	35	24	1	10	6	24	13

Table 10: Experimental Results for Game 2

	Sender				Receiver					
	$t=A$		$t=B$		$m=a$			$m=b$		
	a	b	a	b	X	Y	Z	X	Y	Z
Session 1, direct	12	1	7	6	14	0	5	0	5	2
Session 1, lottery	12	1	9	4	11	2	8	0	3	2
Session 2	10	3	7	6	8	5	4	2	1	0
Total	34	5	23	16	33	7	17	2	9	4

Table 11: Experimental Results for Game 3 in Sessions 1-2

We will repeat such an experimental round several times. When all the rounds finish, the instructors will tell you the end of experiment. Your reward is finally determined based on the score you earned all over the rounds. More detailed experimental procedure follows.

C.1 Experimental Procedure

In this experiment, each round proceeds as follows:

1. Each of you are told whether you are an "S-player" or an "R-player" at this round.
2. If you are an "S-player," you are also told whether you are type A or type B at this round.
3. "S-player" chooses between two alternatives "I am a type A" or "I am a type B."
4. "R-player," informed of the choice of "S-player" who is your matched opponent, chooses from among three alternatives "A," "B," and "C."
5. The score is determined according to the type of "S-player," which is assigned at the beginning of this round, and the choice by "R-player."
6. The final reward is determined based on the score you earned all over the rounds, and then paid in cash.

Let us see the details of each stage more closely.

Stage 1.

Each pair of subjects participate in each decision making, so there are 6 pairs and 1 person has to wait. One subject of a pair is called "S-player," while the other subject "R-player." Throughout the experiment, you are never told who and who match to form a pair. All that you are told is the number assigned to the pair to which you belong and whether you are an "S-player" or "R-player." All of these are predetermined according to some random matching rule by the experimenters.

More specifically, at each round a "Payoff table" is distributed to each of those who participate in the experiment. On the table, you will find a payoff table and the number assigned to the pair to which you are belonging at this round. We will later explain how to read the payoff table in more detail. If you are an "S-player," "Answer sheet" will also be distributed.

Fill in the blank of your "Recording sheet" with the number of your pair that you have found on the "Answer sheet." Circle the letter "S" in the Player field of your "Recording sheet" if you are an "S-player," "R" if "R-player."

If you are told to wait at this round, write "wait" in the Pair field of your "Recording sheet," and wait silently until the next round.

Stage 2

Look at the upper half of your "Answer sheet." If you are told to be an "S-player" in Step 1, you are also told whether you are type A or type B. Throughout the experiment, the probabilities of being type A and type B are equal. No one except you knows whether you are type A or type B.

If you are an "S-player" and your type is A, circle the letter "A" in the Type field of your "Recording sheet," likewise for the case that your type is B.

Stage 3

Those who are told to be an "S-player" in the Step 1 choose between "Alternative A" or "Alternative B."

Alternative A: "I am a type A."
Alternative B: "I am a type B."

The choice is completely up to you. While the type of which you are informed in Step 2 will not be known to the matched "R-player," the choice you made in Step 2 will be known to the opponent.

If you choose "Alternative A," circle the letter A in the Alternative field on your "Recording sheet," likewise for the case that you choose "Alternative B." Also do the same for the "Choice of S-player" field in the lower half of your "Answer sheet" and hand it to the instructors.

Stage 4

"R-player" chooses among "Alternative A," "Alternative B," and "Alternative C" knowing the choice made by "S-player" in Step 3. You can find the choice of the matched "S-player" on the "Answer sheet."

If you choose "Alternative A," circle the letter A in the Alternative field on your "Recording sheet," likewise for the case that you choose "Alternative B" or "Alternative C." Also do the same for the "Choice of R-player" field on the "Answer sheet" handed to you.

Stage 5

Both players' scores are determined according to the choice made by "R-player" in Step 4 and *the type revealed to "S-player" in Step 2.* Note that *the choice by "S-player" in Step3 does not affect scores*.

The score table shows you how both players' scores are determined. The scores that both players get will be shown separately on the blackboard, so ensure your score at each round. After ensuring your score, write it in the Score field on your "Recording sheet."

Example.

Suppose you are distributed a payoff table as follows:

	type A		type B	
Alternative A	S	R	S	R
	90	20	60	30
Alternative B	S	R	S	R
	50	10	10	90
Alternative C	S	R	S	R
	80	70	30	50

If "S-player" is assigned type A in Step 2, look down under the column "type A" on this table. If "S-player" is assigned type B, then look down under the column "type B." The left digit in each cell indicates S-player's score and the right R-player's.

For example, suppose "S-player" is told that his type is type A and "R-player's" choice is "Alternative A," then "S-player" gets 90 and "R-player" gets 20 according to this payoff table. If "S-player" is told that his type is type B and "R-player's" choice is "Alternative B," then "S-player" gets 10 and "R-player" gets 90.

Also suppose that "S-player" is told that his type is A and "S-player" chooses "Alternative B." In this case, if "R-player" chooses "Alternative A," then "S-player" gets 90 and "R-player" gets 20. Next suppose that "S-player" is told that his type is B and "S-player" chooses "Alternative B." In this case, if "R-player" chooses "Alternative A," then "S-player" gets 60 and "R-player" gets 30.

Stage 6

Steps 1-5 complete a round of the experiment. Your reward in cash in this round is fifty Yen times the score you get in this session. Fill in the Reward field on your "Recording sheet" with the number that is 50 times as large as the score in this round. The total reward in the experiment is the sum of each round's reward plus participation fee, a thousand Yen.

C.2 Notices

- Please be quiet throughout the experiment. You might be expelled if the instructor thinks it necessary. In that case, you might not be rewarded.
- You cannot leave the room throughout the experiment in principle.
- Please turn off your pocket bell or cellular phone.
- Do not take anything used in the experiment with you.

C.3 Questions

If you have any question concerning the procedure of experiment, raise your hand quietly. An instructor will answer your question in person. In some cases, the content of your question might disallow the instructor to answer it, however.

C.4 Practice

Before conducting the experiment, we have three sessions for practice. These are purely for practice and the results therein will not be counted in your reward. You can always refer to these instructions throughout the experiment.

Please take out "Recording sheet (Practice)" from your envelope and fill in your name and student ID.

We will distribute "Answer sheets (Practice)" and "Score table (Practice)" to those who are to be "S-players" in this session. To those who are to be "R-players" in this session, only "Score table (Practice)" will be distributed.

"S-players" should now circle the letter S in the Player field of the "Recording sheet (Practice)" and "R-players" the letter R.

"S-players" now make their choice looking at your own type on the "Answer sheet (Practice)" and the "Payoff table (Practice)." Mark your own type in the Type field of your "Recording sheet (Practice)" and also mark your choice in the Choice field of the "Recording sheet (Practice)." Next mark your choice on the "Answer sheet (Practice)" too. "Answer sheet (Practice)" will be collected later.

Then the lower half of the "Answer sheet (Practice)," on which "S-players" have already marked their choices, will be distributed to the matched "R-players." "R-players" can thus see the choice of "S-players," but not their true types. "R-players" should now make choice by examining the score table and mark your choice in the Choice field of your "Recording sheet (Practice)." Also mark your choice on the "Answer sheet (Practice)."

Let us now turn to actual experiment. Please fill in your name and student ID on your "Recording sheet."

D Recording Sheet

Recording Sheet

Name( ) Student ID( )

Round	Pair No.	Player	Type	Alternative	Payoff	Reward
1		S / R	A / B	A / B / C		
2		S / R	A / B	A / B / C		
3		S / R	A / B	A / B / C		
4		S / R	A / B	A / B / C		
5		S / R	A / B	A / B / C		
6		S / R	A / B	A / B / C		
7		S / R	A / B	A / B / C		
8		S / R	A / B	A / B / C		
9		S / R	A / B	A / B / C		
10		S / R	A / B	A / B / C		
11		S / R	A / B	A / B / C		
12		S / R	A / B	A / B / C		
13		S / R	A / B	A / B / C		

* Multiply the number in the Payoff field by 50 and put the resultant number into the Reward field.

E Various Refinement Concepts for Cheap-Talk Games

First we set notations for describing a general cheap-talk game. A generic cheap-talk game can be represented by a sextuple $\langle T, \pi, M, C, u_S, u_R \rangle$, where T is the finite set of the Sender's types, π is a prior distribution over T, M is the set of messages, C is the set of actions for the Receiver, and $u_S : C \times T \to \mathbb{R}$ and $u_R : C \times T \to \mathbb{R}$ are payoff functions for the Sender and the Receiver respectively. Let ΔM and ΔC denote probability distributions over the set of messages and the set of actions respectively. A strategy of the Sender is a function $\mu : T \to \Delta M$. Let Σ_S denote the set of Sender's strategies. A strategy of the Receiver is a function $\rho : M \to \Delta C$. Let Σ_R denote the set of the Receiver's strategies. The Sender of type t's expected payoff in strategy profile (μ, ρ) is

$$u_S(\mu, \rho | t) = \sum_{m \in M} \sum_{c \in C} \mu(m|t) \rho(c|m) u_S(c, t),$$

while the expected payoff to the Receiver is

$$u_R(\mu, \rho) = \sum_{t \in T} \sum_{m \in M} \sum_{c \in C} \pi(t) \mu(m|t) \rho(c|m) u_R(c, t).$$

Let $\beta(m)$ denote the Receiver's belief upon receiving message m. We also use $\beta(t|m)$ to denote the probability that $\beta(m)$ assigns to type t. Let the set of best responses for the Receiver with this belief be denoted by $BR_R(\beta(m))$.

Definition E.1 (Sequential Equilibrium) *(μ, ρ, β) is a sequential equilibrium if for every $m \in \mu(T)$, β satisfies*

$$\beta(t|m) = \frac{\pi(t)\mu(m|t)}{\sum_{s \in T} \pi(s)\mu(m|s)},$$

for every $m \in M$, $\rho(m) \in BR_R(\beta(m))$, and for every t and for every $\hat{\mu} \in \Sigma_S$, $u_S(\mu, \rho|t) \geq u_S(\hat{\mu}, \rho|t)$.

Assuming that there is a common language between the Sender and the Receiver, for every empty subset K of T, there exists a message 'K' with literal meaning that "I belong to K," Farrell (1993) proposes the following refinement theory. Consider an equilibrium and call a message that is not used in equilibrium a "neologism." Let the original sequential equilibrium be (μ, ρ, β). When the set of types $K \subset T$ send a neologism 'K' to the Receiver, and the Receiver believes this message, it will induce the following belief β('K').

$$\beta(t|\text{'K'}) = \begin{cases} \pi(t) / \sum_{s \in K} \pi(s) & \text{if } t \in K \\ 0 & \text{if } t \notin K \end{cases}$$

A neologism 'K' is said to be credible relative to equilibrium (μ, ρ, β) if the following conditions hold.

$$\begin{aligned} \text{C1':} \quad & u_S(\rho|t) > u_S(\mu, \rho|t) \quad \text{for all} \quad t \in K \text{ and } \rho \in BR_R(\beta(\text{'K'})), \\ \text{C2':} \quad & u_S(\rho|t) \leq u_S(\mu, \rho|t) \quad \text{for all} \quad t \notin K \text{ and } \rho \in BR_R(\beta(\text{'K'})). \end{aligned}$$

Definition E.2 (neologism-proofness) *A neologism-proof equilibrium is an equilibrium relative to which there exists no credible neologism.*

By definition, if an equilibrium is neologism-proof, then all the equilibria with the same outcome function are also neologism-proof. In all the sequential equilibria in Game 3, type A Sender can send a credible neologism with the meaning that "I'm type A." Thus they are not neologism-proof. Therefore Game 3 has no neologism-proof equilibria.

In neologism-proofness, it is assumed that all the types that attempt to deviate, called the set of deviant types, send the same single message. Mathews, Okuno-Fujiwara, and Postlewaite (1991) improve upon neologism-proof by considering deviation as a map from the set of deviant types to the set of deviation messages and imposing consistency conditions on this mapping.

An announcement strategy is a pair $d = \langle \delta, D \rangle$, where D is a nonempty set of deviant types and $d : D \to \Delta M$ is a talking strategy. Let the set of messages that are sent with positive probability by δ be denoted as $\delta(D)$. An anouncement is a pair $\langle m, d \rangle$, where $m \in \delta(D)$. Denote the Receiver's belief when he believes an announcement $\langle m, d \rangle$ be denoted as $\beta(m, d)$. For each $t \in T$ and each $m \in \delta(D)$, this is defined as

$$\beta(t|m, d) = \begin{cases} \pi(t)\delta(m|t) / \sum_{s \in D} \pi(s)\delta(m|s) & \text{if } t \in D \\ 0 & \text{if } t \notin D \end{cases}$$

Suppose the Receiver with the above belief plays a best response and let the minimum and maximum payoff to the type t Sender be denoted by

$$\underline{u}_S(m,d|t) = \min\{u_S(\rho|t) : \rho \in BR_R(\beta(m,d))\}, \text{and}$$
$$\bar{u}_S(m,d|t) = \max\{u_S(\rho|t) : \rho \in BR_R(\beta(m,d))\}$$

respectively. Consider an equilibrium (μ,ρ,β) and an announcement strategy $d = \langle \delta, D\rangle$, and suppose they together satisfy the following conditions.

C1 For each $t \in D$ and for each $m \in \delta(\{t\})$, $\underline{u}_S(m,d|t) \geq u_S(\mu,\rho|t)$ and strict inequality holds for some $t \in D$ and $m \in \delta(\{t\})$.

C2 For each $t \in T \setminus D$ and for each $m \in \delta(D)$, $\bar{u}_S(m,d|t) \leq u_S(\mu,\rho|t)$.

C3 For each $t \in D$, for each $m \in \delta(\{t\})$, and for each $\hat{m} \in \delta(D) \setminus \{m\}$, $\underline{u}_S(m,d|t) \geq u_S(\hat{m},d|t)$.

A announcement strategy $d = \langle \delta, D\rangle$ and a corresponding announcement $\langle m,d\rangle$ are said to be weakly credible relative to equilibrium (μ,ρ,β) if they satisfy C1-C3.

Definition E.3 (strong announcement-proofness) *A strongly announcement-proof equilibrium is an equilibrium relative to which there exists no weakly credible announcement.*

When $\delta : D \to M$ is constant valued, the above definition coincides with that of a credible neologism. Therefore, a credible neologism is a weakly credible announcemnet. This means a strongly announcement-proof equilibrium is neologism-proof. Therefore our Game 3 does not have a strongly announcement-proof equilibrium.

A stronger notion of credibility can be defined by further requiring C4 below. An announcement strategy $d = \langle \delta, D\rangle$ and a corresponding announcement $\langle m,d\rangle$ is said to be credible relative to (μ,ρ,β) if they together satisfy C1-C4.

C4 If $d' = \langle \delta', D'\rangle$ also satisfies C1-C3 relative to (μ,ρ,β), then for each $t \in D \cap D'$, for each $m \in \delta(\{t\})$, and for each $m' \in \delta'(\{t\})$, $\underline{u}_S(m,d|t) \geq \bar{u}_S(m',d'|t)$.

Definition E.4 (announcement-proofness) *An announcement-proof equilibrium is an equilibrium relative to which there exists no credible announcement.*

Obviously, a strongly announcement-proof equilibrium is announcement-proof. In Game 3, there is just one announcement strategy that satisfies C1-C3, and thus C4 is satisfied trivially. Therefore all the sequential equilibria in Game 3 are not announcement-proof.

As is stated in the text, Mathews, Okuno-Fujiwara, and Postlewaite (1991) proposes another concept in response to Stiglitz critique. This requires a credible message to satisfy further the following condition.

C3A There exist a strategy $\hat{\delta} : T \setminus D \to \Delta M$ with $\delta(D) \cap \hat{\delta}(T \setminus D) = \emptyset$, the Receiver's strategy $\hat{\rho}$ and belief $\hat{\beta}$ such that $(\hat{\mu},\hat{\rho},\hat{\beta})$ is an equilibrium, where $\hat{\mu} = (\delta, \hat{\delta})$.

C4' If $d' = \langle \delta', D'\rangle$ also satisfies C1-C3 and C3A relative to (μ,ρ,β), for every $t \in D \cap D'$, for every $m \in \delta(\{t\})$, and for every $m' \in \delta'(\{t\})$, $\underline{u}_S(m,d|t) \geq \bar{u}_S(m',d'|t)$.

An announcement strategy $d = \langle \delta, D\rangle$ and its corresponding announcement $\langle m,d\rangle$ is said to be strongly credible relative to equilibrium (μ,ρ,β) if they satisfy C1-C3, C3A and C4'.

Definition E.5 (weak announcement-proofness) *A weakly announcement-proof equilibrium is an equilibrium relative to which there exists no strongly credible announcement.*

By definition, an announcement-proof equilibrium is also weakly announcement-proof. Obviously, the deviation announcement by type A in our Game 3 does not satisfy C3A. Thus babbling equilibria in Game 3 are all weakly announcement-proof.

The concepts of recurrent set and recurrent equilibrium proposed by Rabin and Sobel (1996) are explained in some detail in the text. Here we give a rigorous definition of partial common interests that we omitted in the text.

Let $BR_R(K)$ denote the set of Receiver's actions that can be a best response to some probability distribution over a nonempty subset $K \subseteq T$. Also let $\underline{u}_S(t,L)$ denote the minimum expected payoff that the type t Sender earns when the Receiver choose among $BR_R(K)$. Let $BR_R(K,\pi)$ denote the set of Receiver's best actions when he Bayes-update the prior belief concentrating on K. Denote by $u_S^P(t)$ the type t Sender's expected payoff in a babbling equilibrium.

Definition E.6 (partial common interests) *A cheap-talk game has partial common interests if there exists a partition $J_1, \cdots, J_j$ of T such that:*

1. *for all $t_i \in J_i$, and for all $J_k \neq J_i$, $\underline{u}_S(t_i, J_i) > \max\{u_s(t_i, a_k) : a_k \in BR_R(J_k)\}$;*
2. *for each i, there exists $a_i \in BR_R(J_i, \pi)$ such that $u_S(t_i, a_i) > u_s^P(t)$ for all $t_i \in J_i$; and*
3. *If $L \cap J_k \neq \emptyset$ for at least two k, then for each $a \in BR_R(L)$ there exists an i and $t_i \in L \cap J_i$ such that $u_S(t_i, J_i) > u_S(t_i, a)$.*

It is obvious that Game 3 does not have partial common interests. Note that the condition 2 above does not allow the trivial partition $\{T\}$ to be a qualified partition.