Credibility in Persuasion: A Laboratory Experiment*

Yuki Shiomi[a] Nobuyuki Hanaki[b]

September 22, 2025

Abstract

We experimentally test how the interaction between the sender's commitment ability and the players' payoff structure affects information transmission within the Bayesian persuasion framework. The theory predicts that even when the sender's commitment ability is limited, meaningful information can still be transmitted (i.e., persuasion is credible) when the interests of the sender and the receiver are aligned, whereas persuasion is not credible when their interests conflict. Consistent with these predictions, our experimental results show that the presence of both a lack of commitment and conflicting interests almost completely suppresses information transmission between players. Contrary to the theoretical prediction, however, we also find that lack of commitment and conflicting interests independently exert negative effects on information transmission.

JEL Classification codes: D82, D83
Keywords: Bayesian persuasion, commitment, credibility, laboratory experiment

*The experiments reported in this paper is approved by the IRB of the Institute of Social and Economic Research, the University of Osaka (No. 20250701). We are grateful for the financial support from the grant-in-aid in Scientific Research, Japan Society for Promotion of Sciences (No. 23H00055 and 25H00388).

[a]Graduate School of Economics, University of Osaka. E-mail: u282915h@ecs.osaka-u.ac.jp

[b]Institute of Social and Economic Research, University of Osaka and University of Limassol. E-mail: nobuyuki.hanaki@iser.osaka-u.ac.jp

1 Introduction

The persuasion problem, in which a sender provides information to a decision maker (the receiver) to induce the receiver to take a desirable action, is commonly observed in many real-world environments. For example, schools disclose information about student ability to recruiting firms through grade transcripts, and used-car sellers disclose vehicle condition to buyers via condition reports.

The Bayesian persuasion framework, epitomized by Kamenica and Gentzkow (2011), has become a standard tool for analyzing such information transmission problems. A key assumption in the standard Bayesian persuasion is that the sender can fully commit ex ante to the information disclosure rule mapping states to probability distributions over messages (hereafter, the information structure). In practice, however, observability and verifiability constraints make it difficult for the sender to fully commit to the information structure. Schools typically do not announce precise grading policies—e.g., what share of high-ability students receive grade A and what share of others receive grade B or C—and even if such a policy were announced, verifying that grades are assigned accordingly is nearly impossible.

By contrast, as noted by Lin and Liu (2024), the distribution of messages is often observable in practice. For instance, some U.S. universities actually publish the distribution of students' grades. Thus, in many settings the sender's commitment power appears limited not to the information structure itself but only to the induced distribution of messages. Motivated by this observation, Lin and Liu (2024) study a model of Bayesian persuasion in which the sender can commit only to the message distribution induced by a designed information structure. They show that payoff structure between sender and receiver plays a critical role in the credibility of persuasive communication. Specifically, when the sender can commit only to the message distribution, if the sender's and receiver's marginal interests are aligned, the same equilibrium as in standard Bayesian persuasion is implementable; if they are misaligned, no informative communication occurs in equilibrium.

This paper tests the core predictions of Lin and Liu (2024) in the laboratory by exogenously manipulating the sender's commitment power and the players' payoff structure within a Bayesian persuasion setting. We implement two payoff regimes —Aligned and Misaligned— and three commitment regimes —Full Commitment, Partial Commitment, and No Commitment. Full Commitment corresponds to the standard Bayesian persuasion, in which the sender can fully commit to the information structure. Under Partial Commitment, the sender publicly announces the designed information structure but may deviate to another structure as long as the induced distribution of messages is unchanged. The receiver understands this possibility but cannot directly observe the post-deviation structure. Partial Commitment thus captures a verifiability constraint, namely that deviations leaving the message distribution unchanged are typically not verifiable to the receiver. Under No Commitment, the sender does not announce the information structure; only the induced message distribution is disclosed to the receiver. No Commitment thus captures an observability constraint in which information structures are often unobservable, whereas message distributions are often observable. In both Partial and No Commitment, the sender can commit only to the message distribution; consequently, these two regimes are theoretically equivalent in their equilibrium implications.

We cross the three commitment regimes with the two payoff regimes to obtain six experimental treatments. In the minimal environment we implement, the predictions following Lin and Liu (2024) yield sharp results: when full commitment is absent (Partial or No) and interests are misaligned, the equilibrium is the uninformative babbling equilibrium. By contrast, if the sender has full commitment even under misaligned interests, or if full commitment is absent but interests are aligned, the same equilibrium as in the standard Bayesian persuasion arises.

The experimental results show that the presence of both a lack of commitment and a conflict of interest largely eliminates informative communication, consistent with the babbling-equilibrium

Table 1: (A) Payoff structure "ALG." (B) Payoff structure "MIS."

	red	$blue$
Red	200, 150	80, 80
Blue	100, 80	80, 150

	red	$blue$
Red	100, 150	80, 80
Blue	200, 80	80, 150

Note: In each cell, the left number represents the sender's payoff and the right number represents the receiver's payoff.

prediction. However, contrary to the theoretical prediction, we also find that presence of either the lack of commitment or misaligned interests alone hinders information transmission.

2 Theoretical Framework

This section aims to present the main theoretical predictions in a concise manner, focusing only on the environment implemented in our experiment.

2.1 Setup

There are two players: a sender (S) and a receiver (R). A box contains 10 balls, 3 Red and 7 Blue. The computer randomly draws one ball from the box, and neither player can directly observe its color. The color of the ball represents the state of the world, $\theta \in \Theta = \{\text{Red, Blue}\}$, with a common prior probability $\mu_0 = \Pr(\text{Red}) = 0.3$.

The receiver chooses an action $a \in A = \{red, blue\}$ as her guess of the ball color. Each player's payoff depends on the ball color $\theta \in \Theta$ and the receiver's guess $a \in A$. Table 1 presents the payoff matrices used in our experiment. We employ two versions of the payoff structure, referred to as "Aligned (ALG)" and "Misaligned (MIS)." The only difference between them lies in the sender's payoffs when the receiver guesses red: $u_S(red, \text{Red})$ and $u_S(red, \text{Blue})$.

In both payoff structures, the receiver always has an incentive to correctly predict the ball color, while the sender has an incentive to induce the receiver to guess red. In this setting, the receiver's optimal behavior is characterized by a threshold strategy:

$$a(\mu) = \begin{cases} red & \text{if } \mu \geq \frac{1}{2}, \\ blue & \text{otherwise,} \end{cases} \tag{1}$$

where μ denotes the receiver's posterior that the state is Red. Given that $\mu_0 = 0.3 < \frac{1}{2}$, the receiver would, under the prior, always choose $a = blue$. Hence, the sender must employ some form of information transmission to persuade the receiver.

As in the standard Bayesian persuasion framework, we assume that the sender transmits information to the receiver by designing an information structure $\pi : \Theta \to \Delta M$. Here, $M = \{r, b\}$ denotes the set of messages available to the sender, where r (b) is a message that implies $\theta = \text{Red}$ (Blue).

2.2 Theoretical Predictions

Full commitment. When the sender can fully commit to the information structure she designs, the equilibrium outcome coincides with that in Kamenica and Gentzkow (2011) (hereafter, BP equilibrium). This holds under both payoff structures implemented in our experiment, "ALG" and

"MIS." In BP equilibrium, the sender's optimal information structure is

$$\pi(r|\text{Red}) = 1,\ \pi(r|\text{Blue}) = \frac{3}{7}. \tag{2}$$

The receiver's optimal strategy and belief are

$$a(m) = \begin{cases} red & \text{if } m = r, \\ blue & \text{if } m = b, \end{cases} \qquad \mu(m) = \begin{cases} \frac{1}{2} & \text{if } m = r, \\ 0 & \text{if } m = b. \end{cases} \tag{3}$$

That is, in the BP equilibrium, the sender commits to truthfully sending $m = r$ when $\theta = \text{Red}$, and to sending the false message $m = r$ with probability $3/7$ when $\theta = \text{Blue}$. In this case, even though the receiver recognizes that the sender's message may be false with positive probability, it is still optimal for her to follow the sender's message.

Without Full commitment. Following Lin and Liu (2024), we describe Bayesian persuasion without full commitment as a situation in which the sender can commit only to the message distribution induced by an information structure, rather than to the structure itself. While the details are provided in Lin and Liu (2024), whether the outcome replicates the BP equilibrium depends on the payoff structure. Specifically, under the payoff structure we call "ALG," the BP equilibrium still arises, whereas under "MIS" it does not.

The intuition is as follows. In "MIS," suppose the receiver follows the sender's message as in the BP equilibrium. Since $u_S(red, \text{Blue}) > u_S(red, \text{Red})$, the sender then has an incentive to deviate by increasing the probability of sending $m = r$ when the state is Blue. Since, in practice, such deviations are possible without changing the message distribution, the presence of both the lack of full commitment and the misaligned payoff structure leads the equilibrium to collapse into a babbling equilibrium. In this equilibrium, the sender transmits no meaningful information to the receiver, and the receiver simply plays $a = blue$ according to the prior.

3 Lab Implementation

We tested six treatments (AF, MF, AP, MP, AN and MN) in the laboratory by combining two payoff structures (ALG and MIS) with three commitment conditions (Full, Partial, and No). As discussed earlier, the two payoff structures are presented in Table 1. The commitment conditions, on the other hand, determine the structure of the game played by subjects in the laboratory.

Here, we describe the most complex case, Partial Commitment. The game proceeds as follows. First, the computer randomly draws one ball from the box. Without observing the ball color, the sender then designs an information structure $\tilde{\pi}$ and announces it to the receiver. After this announcement, the sender may revise $\tilde{\pi}$ to another information structure π, but only as long as the message distribution specified by $\tilde{\pi}$ remains unchanged. The receiver observes the announced information structure $\tilde{\pi}$, the message distribution that is common to both $\tilde{\pi}$ and π, and the realized message m drawn from π. Based on this information, the receiver finally chooses an action $a \in A$ by guessing the ball color.

In the Full and No Commitment conditions, the sender designs only an information structure π. In the Full Commitment condition, the receiver directly observes π, whereas in the No Commitment condition, the receiver observes only the message distribution induced by π. The equilibrium predictions for each treatment are summarized in Table 2.

We employed a between-subject design, conducting one pilot session for each treatment. Each session involved 16 to 24 participants and consisted of 12 rounds. Subjects were assigned fixed roles throughout the session, while matching was randomized in every round. Participants received the earnings from one randomly selected round out of the 12 in addition to a participation fee.

Table 2: Equilibrium predictions by treatment.

	Commitment		
	Full	**Partial**	**No**
ALG	**AF** BP	**AP** BP	**AN** BP
MIS	**MF** BP	**MP** Babbling	**MN** Babbling

Table 3: Average $\phi(\theta, a)$: (a) Theoretical predictions vs. (b) Experimental data.

	Commitment				
	Full		**Partial**		**No**
ALG	0.53	$\approx$	0.53	$\approx$	0.53
	$\wr$		$\vee$		$\vee$
MIS	0.53	$>$	0.00	$\approx$	0.00

	Commitment				
	Full		**Partial**		**No**
ALG	0.247	$>$	0.138	$\approx$	0.101
	$\vee$		$\vee$		$\vee$
MIS	0.064	$>$	0.017	$>$	-0.021

Notes: ">" indicates $p < 0.05$ in M-W test. Green: as predicted. Red: not as predicted.

4 Results

To quantify the overall informativeness of the interaction between the sender and the receiver, we define $\phi(\theta, a) := \text{Corr}_{(\pi,a)}(\theta, a)$. This measure captures the extent to which information about the state is actually transmitted through the information structure π designed by the sender and the actions a chosen by the receiver. Overall informativeness is a standard metric in experimental studies of communication games, including Bayesian persuasion (Fréchette et al., 2022).

Table 3 presents the theoretical predictions and the average values of $\phi(\theta, a)$ across treatments. The theory predicts that under the BP equilibrium, $\phi(\theta, a) = 0.53$, indicating a moderate level of information transmission, whereas under the babbling equilibrium, $\phi(\theta, a) = 0$.

Looking at the averages from the experimental data, we find that the presence of both the lack of commitment and conflicting interests reduces the overall informativeness to nearly zero, which is consistent with the babbling equilibrium (MP, MN). By contrast, the data also show that lack of commitment (AF $\rightarrow$ AP) and conflicting interests (AF $\rightarrow$ MF) each independently reduces the informativeness, which is inconsistent with the theoretical prediction. Even when the theory predicts a moderate level of informativeness under the BP equilibrium, the average values obtained from the experimental data are generally lower.

5 Discussion

Due to space constraints, we omit the details here and instead summarize some of our results. We find that the deviations from theory observed in $\phi(\theta, a)$ are driven by undercommunication by the sender. Figure 1 reports the average values of $\pi(r|\cdot)$ across treatments, which also reveals undercommunication relative to the theoretical predictions. By contrast, although some deviations are observed, receivers respond to empirical senders in a manner largely consistent with the theoretical predictions.

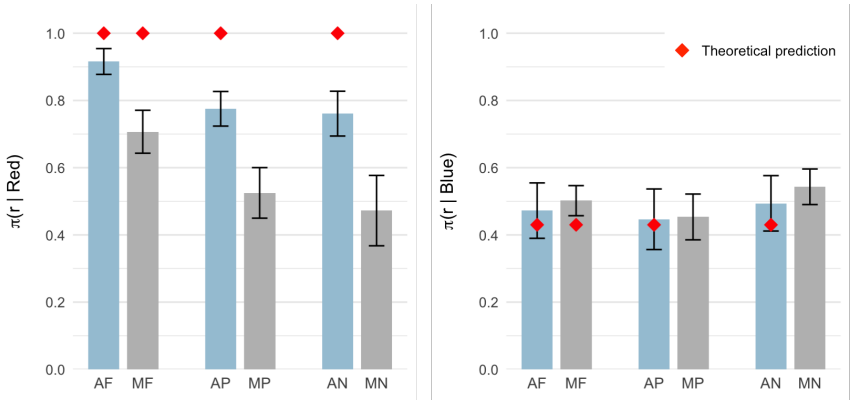

Figure 1: Sender's information structure: (A) $\pi(r|\text{Red})$. (B) $\pi(r|\text{Blue})$.

References

Fréchette, G. R., A. Lizzeri and J. Perego, 2022. Rules and commitment in communication: an experimental analysis. Econometrica 90, 2283–2318.

Kamenica, E. and M. Gentzkow, 2011. Bayesian persuasion. American Economic Review 101, 2590–2615.

Lin, X. and C. Liu, 2024. Credible persuasion. Journal of Political Economy 132, 2228–2273.