

THE ANATOMY OF HONESTY: LYING AVERSION VS. DECEPTION AVERSION[†]

SYNGJOO CHOI, CHANJOO LEE, AND WOORYOUNG LIM*

ABSTRACT. This paper experimentally dissects the preferences for honesty into two components: lying aversion and deception aversion. For a separate identification, we consider two reputation-building environments with a two-dimensional belief domain, where lying without deception occurs in one environment and deception without lying occurs in the other environment as a unique equilibrium phenomenon. The choice data combined with belief data collected in the lab enable us to differentiate between individuals' aversion to making statements that are literally untrue (lying aversion) and their aversion to statements intended to manipulate others' beliefs (deception aversion), excluding other plausible explanations such as erroneous inferences and misunderstanding.

1. Introduction

In many economic and social interactions, individuals often engage in dishonest behavior. Traditional economic theory assumes that people act dishonestly primarily when it offers material benefits. This prediction is crucial for standard equilibrium analysis in various situations involving adverse selection (Baron and Myerson, 1982), moral hazard

Date: May 6, 2025.

Keywords: Honesty, lying, deception, sender-receiver games, experiments.

We are grateful to Johannes Abeler, Marina Agranov, Yunus Aybas, Ian Ball, Douglas Bernheim, Andreas Blume, Yeon-Koo Che, Wonki Jo Cho, Martin Dufwenberg, Andrew Fudowsian, Evan Friedman, Amanda Friedenberg, Uri Gneezy, Aram Grigoryan, Faruk Gul, Nobuyuki Hanaki, Ilwoo Hwang, Changkuk Im, Philippe Jehiel, Navin Kartik, Jeongbin Kim, Jinwoo Kim, Fuhito Kojima, Shengwu Li, Alessandro Lizzeri, Stephen Morris, Roger Myerson, Doron Ravid, Ilya Segal, Joel Sobel, Charles Sprenger, Dmitry Taubinsky, and Rakesh Vohra for their valuable comments and suggestions. We have benefitted from comments by conference participants at the 2023 North American Summer Meeting of the Econometric Society, 2023 Asia Meeting of the Econometric Society, the 34th Stony Brook International Conference on Game Theory, the 2023 Asia-Pacific Economic Science Association Meeting, the 2024 Game Theory Society World Congress, and seminar participants at the CalTech, City University of Hong Kong, Florida State University, HKUST, KAIST College of Business, Kansai University, Kyung Hee University, Nanyang Technological University, Seoul National University, Sogang University, SUSTech, University of Arizona, University of Michigan, University of Tokyo, and Yonsei University. This study is supported by the Creative-Pioneering Researchers Program through Seoul National University, the BK21 FOUR funded by the Ministry of Education and National Research Foundation of Korea, and a grant from the Research Grants Council of Hong Kong (Grant No. GRF16506019). The IRB approval is obtained from Seoul National University (IRB No. 2205/003-012). This paper was previously circulated and presented under the title "Lying and Deception in Reputation Building." Choi is affiliated with the Department of Economics, Seoul National University, Seoul, South Korea; Lee is affiliated with Stanford University Graduate School of Business, California, United States; Lim is with the Department of Economics, the Hong Kong University of Science and Technology, Hong Kong. Email addresses: syngjooc@snu.ac.kr, chanjoo@stanford.edu, wooyoung@ust.hk.

(Holmström, 1979), tax compliance (Allingham and Sandmo, 1972), reputation building (Ely and Välimäki, 2003), and school choices (Abdulkadiroglu et al., 2006). A market or institution designer is then required to set up adequate economic incentives to restrain agents from engaging in dishonest behavior that may disrupt markets and law enforcement. However, a growing literature documents that individuals exhibit preferences for honesty despite material losses of such behaviors (e.g., Gneezy, 2005; Erat and Gneezy, 2012; Fischbacher and Föllmi-Heusi, 2013; Gneezy, Kajackaite and Sobel, 2018; Abeler, Nosenzo and Raymond, 2019; Abeler, Falk and Kosse, 2024). For the policymaker or mechanism designer to utilize agents' intrinsic motivations, it is crucial to scrutinize the nature of preferences for honesty.

In this paper, we experimentally investigate the nature of preferences for honesty. Specifically, we focus on decomposing preferences for honesty into two components: lying aversion and deception aversion. For distinct identification, we consider two reputation-building environments where lying (making a factually incorrect statement) and deception (manipulating the listener's beliefs) emerge separately as unique equilibrium phenomena. By referencing the equilibrium behavior of lying and deception, we conducted two experiments to identify individuals' aversion to deception and lying aversion.

The terms “dishonesty,” “lying,” and “deception” are often used interchangeably, both in the literature of behavioral economics and in everyday conversation. However, while these concepts are related, they are not synonymous. Sobel (2020) defines deception as a deliberate act of steering someone's beliefs in the wrong direction while lying involves conveying a message whose literal meaning is different from the true state, irrespective of the speaker's intention or the listener's beliefs. It is thus possible for deception to occur without lying, and for lies to be told without deceiving.¹ Therefore, honesty encompasses both the absence of deceit and the commitment to truthfulness (Stevenson, 2010).

The conceptual distinction between lying aversion and deception aversion has only recently gained attention in the literature. Kartik (2009) introduces a model of lying aversion, capturing the disutility that arises from the discrepancy between the sender's true type and the type stated in the message, as determined by the message's *literal interpretation*. More recently, Eilat and Neeman (2023) developed the first formal model of *deception cost*, which accounts for the moral cost of dishonesty stemming from individuals' reluctance to manipulate others' beliefs. Building on the definitions proposed by Sobel (2020),

¹Sutter (2009) and Blazquiz-Pulido et al. (2024) are two experimental studies examining whether subjects engage in “sophisticated lying” by telling the truth. In this context, “sophisticated lying” refers to what we define as deception without lying. Ettinger and Jehiel (2021) experimentally investigate reputation-building in a laboratory repeated-communication environment and find that a significant proportion of sender participants employ a deceptive tactic, where they initially tell the truth and then switch to lying at a certain point.

their model quantifies deception costs as the distance between the belief induced by the sender's message and the belief that should have been induced.

Prior to this emerging focus on the distinction between lying and deception (Sobel, 2020), empirical research primarily identified three key motives for honest behavior in environments involving the communication of private information: (1) lying aversion (Sánchez-Pagés and Vorsatz, 2007; Hurkens and Kartik, 2009), which reflects individuals' discomfort with making false statements; (2) consequentialism (Gneezy, 2005; Erat and Gneezy, 2012), which highlights the role of payoff consequences to others in guiding honest behavior; and (3) image concern (Gneezy et al., 2018; Dufwenberg and Dufwenberg, 2018; Abeler et al., 2019; Khalmetski and Sliwka, 2019), which suggests that individuals refrain from dishonesty to maintain a positive self-image. In these settings, the distinction between lying aversion and deception aversion is often either conceptually inseparable or empirically irrelevant, as lying and deception typically occur simultaneously.

The distinction between lying aversion and deception aversion is both conceptually and practically important.² First, without this distinction, it is challenging to evaluate whether economic models accurately capture the intrinsic motivations for honesty.³ The cognitive processes involved in aversion to lying and aversion to deception are fundamentally different; the former does not require higher-order reasoning, while the latter does. Second, deception aversion and lying aversion have different implications for market and mechanism design literature. If lying aversion is the primary component of preferences for honesty, then design efforts must focus on constructing the message space to amplify the moral cost of lying.⁴ However, such efforts may be of little value if deception costs are the predominant factor.⁵

²Without this distinction, it is difficult to grasp the long-standing normative debate surrounding human morality regarding lying (Mahon, 2008), which involves two contrasting views. One perspective, known as deceptionism, argues that lying is inherently morally wrong, regardless of the intentions behind it (Kant, 1797). In contrast, the non-deceptionism view posits that the intentions behind both truth-telling and lying are crucial for assessing their moral value; thus, a truth told with the intent to manipulate may be worse than a lie told with good intentions (Blake, 1790).

³The literature on the equivalence between global and local incentive compatibility (Sato, 2013; Kumar et al., 2021; Cho and Park, 2023) explores various types of network structures to account for the constraints on possible deviations in reporting individuals' types. The diversity of networks considered is well justified by the inclusion of deception costs as part of preferences for honesty.

⁴As a result, the revelation principle may not hold. See Rivera Mora (2024) for more discussion.

⁵For instance, when policymakers consider adopting a new school choice mechanism that faces a trade-off between efficiency and strategy-proofness, one could argue in favor of efficiency, expecting students to submit their true preferences even without strategy-proofness. The persuasiveness of this argument heavily relies on whether students are averse to misstating their preferences, even if their confidentially submitted statements have no intention to deceive other market participants. Perjury laws vary by country: some penalize only outright lies, others focus on deception, while some address both. This variation reflects different societal attitudes toward lying and deception.

In spite of its conceptual and practical importance, there is currently no direct empirical evidence of deception aversion in the literature. This difficulty may arise from the fact that in many well-known communication environments, lying and deception occur simultaneously. Although the literature identifies some environments in which deception occurs in the form of truth-telling (Sutter, 2009; Ettinger and Jehiel, 2021; Blazquiz-Pulido et al., 2024), these environments can only demonstrate that experimental subjects are willing to engage in deceptive communication as a non-equilibrium phenomenon; they are not suitable for identifying intrinsic aversion to deception.

We adopt a revealed preference approach and employ experimental methodology to identify deception aversion in distinction from lying aversion. As part of our identification strategy, we consider two distinct environments from the canonical reputation-building framework with repeated communication (Sobel, 1985; Benabou and Laroque, 1992; Morris, 2001), where lying without deception occurs in one environment and deception without lying occurs in the other as a unique equilibrium phenomenon. This framework encompasses a belief domain that consists of two dimensions: the preference type of the sender (referred to as “types”) capturing whether the sender’s preference is aligned or misaligned with that of the receiver, and the payoff-relevant state of nature (referred to as “states”). A sender (referred to as “she”) possesses private information about both her preference type and the state that affects payoffs. She communicates with a receiver (referred to as “he”) over two periods of interactions. The message space available to the sender is identical to the state space. The sender establishes a reputation regarding her preference type through a message regarding the state sent in the first period.⁶

Each environment involves a specific behavioral type of senders who commit to non-strategic behavior. The receiver is randomly paired with either a strategic or a behavioral sender with equal chance.⁷ The strategic sender’s preference type can be either good (aligned preference) or bad (misaligned preference). While a good sender and the receiver want the receiver’s action to match the state, a bad sender wants the receiver to choose a high action regardless of the state. In our first environment, the strategic sender is of the bad type and the behavioral sender commits to always telling the truth. In our second environment, the strategic sender is of the good type and the behavioral sender commits to always sending a high message. A state is randomly drawn in each period, which is

⁶The absence of direct means to transmit information about private characteristics or types is not only an inherent and distinct characteristic of reputation-building models, but also an assumption made without loss of generality. Even if a message were available to describe the preference type, it would not be possible to transmit credible information through this channel, thereby preserving the value of reputation building.

⁷In our experiments, strategic senders are human participants, whereas behavioral senders are simulated by computer agents adhering to predetermined information transmission rules. Prior to the experiments, participants were explicitly notified that the behavioral senders were computer-controlled entities and were informed about the specific fixed information transmission rule governing the actions of these computer senders.

the private information of the sender. Then, the sender sends a message about the state to the receiver who will take an action. The state becomes common knowledge at the end of the first period.

We assign a substantially higher weight to the second-period payoff compared to the first-period payoff to ensure that reputation-building emerges as the unique equilibrium phenomenon in each environment. In the unique equilibrium of the first environment, the strategic (bad) sender tells the truth about the state in the first period to conceal her preference type. Conversely, in the unique equilibrium of the second environment, the strategic (good) sender lies about the state in the first period to reveal her preference type. The equilibrium analysis in the two environments allows us to separate between lying (about the state of nature) and deception (regarding the preference type).⁸

In Experiment I, we bring our theoretical environments directly to the lab. To reduce the complexity of strategic interaction faced by human subjects, we simplify the sender's choices to only those contingencies that are relevant for reputation-building. Experiment I documents deviations from the reputation-building equilibrium by senders in both environments. Notably, the distributions of sender strategies reveal that the vast majority of senders are categorized into two groups: those who tell the truth with certainty and those who tell a lie with certainty. This observation suggests that deviations from reputation building primarily stem from a subgroup of subjects exhibiting extreme behavior. As traditional behavioral explanations including pure noise/mistakes, learning, and risk-aversion fail to account for these observations, we propose two potential channels: (i) aversion to lying and aversion to deception (preference channel) and (ii) erroneous inferences (inference channel) and misunderstanding.⁹ To separate the preference channel from the alternative channels, we conduct Experiment II.

In Experiment II, we make two adjustments to the setup of Experiment I. First, we replace the second-period communication with the receiver's direct reporting of beliefs about the sender's preference type. The sender's payoff directly depends on the receiver's belief. Second, we elicit the sender's second-order beliefs about the receiver's belief regarding the sender's preference type, contingent on each message. The first modification reduces potential strategic uncertainty arising from second-period interactions, and the

⁸Deception is defined separately for the sender's type space and the state space. From the receiver's perspective, uncertainty about the first-period state will be resolved by the end of the first period which allows the receiver to form his posterior belief about the preference type of the sender. Our notion of *deception with respect to the preference types* captures how a message influences the receiver's posterior belief about the preference type when the uncertainty about the state is resolved. It remains possible for a message to be deceptive regarding the state at the interim stage before the state is revealed, so we further define *deception with respect to the state*. However, the only dimension relevant to deception turns out to be that of the preference type because, in our environments, no equilibrium message can be deceptive about the state.

⁹Here, misunderstanding covers the possibility that participants do not fully understand the environments such that they do not recognize the benefit of reputation building.

second modification enables us to determine whether deviations from reputation building are driven by inference errors and/or misunderstanding.

Experiment II replicates the overall behaviors of senders and receivers observed in Experiment I. Moreover, the data on senders' second-order beliefs reveals that 16%-37% of senders believe that receivers may not make their inferences properly, resulting in the material gain from reputation-building not being significant enough to pursue. However, we also observe a substantial portion (ranging from 47% to 53%) of senders who perceive the material value of reputation-building to be sufficiently large but still do not pursue it. This deviation from reputation-building indicates an aversion to deceptive truth-telling in the first environment and an aversion to non-deceptive lying in the second environment.

Our experimental findings suggest that individuals possess an inherent aversion to deceiving others, which is distinct from their aversion to telling literal lies. We present a parsimonious model in Section 5 that incorporates the concept of deception cost. We propose a measure of deception cost based on the discrepancy between the posterior belief generated by a message and the posterior belief that is closest to the true type among all alternative messages.¹⁰ This measure serves as an intuitive way to quantify the cost associated with deception. Additionally, we discuss alternative measures that can be used to capture the deception cost and the potential problems of these alternative measures (Appendix F). Equilibrium analysis demonstrates that, depending on the relative and absolute magnitudes of lying and deception costs, agents engage in reputation building in both environments, in only one of them, or not engage in reputation building at all. We also identify the nonparametric joint distribution of lying and deception costs using our experimental data and report the result in Appendix H. Assuming that the cost parameters for each player were drawn from the same distribution across treatments, we measure the proportion of players who exhibit a greater aversion to deception than lying, and vice versa.

The rest of the paper is organized as follows. Section 2 describes the reputation-building environment and defines lying and deception in our context. Section 3 presents the experimental design, procedure, and results of Experiment I, and provides motivations for Experiment II. The design of Experiment II, its procedure, and its results are presented in Section 4. The formal equilibrium analysis that incorporates lying and deception costs is presented in Section 5. Lastly, the connection to the literature is provided in Section 6. Detailed discussions of experimental designs, additional results, statistical

¹⁰Our deception cost model is thus a psychological game (Geanakoplos et al., 1989) and related to guilt aversion (Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007). It is also in line with the modeling approach in Eilat and Neeman (2023).

tests, partial identification of lying and deception costs, discussions of alternative explanations, omitted proofs, and sample experimental instructions are relegated to the Appendices A-O.

2. Environment

Consider the following two-period interaction between two agents: an expert (she, sender) and a public (he, receiver). In each period $i = 1, 2$, the expert is privately informed about the **state of nature** (or simply “states”) $\theta_i \in \Theta = \{0, 1\}$. The state is independently drawn from the identical, uniform distribution in each period. The expert sends a message $m_i \in M = \{0, 1\}$ to the public, who decides what action $a_i \in A = [0, 1]$ to take. At the end of each period, the outcome of the stage game—the state, message, and action profile—is revealed to both players.

There are two **preference types** (or simply “types”) of the expert, denoted by $\tau \in T = \{G, B\}$, good (G) and bad (B), which is private information of the expert and time-invariant. One can interpret it as the expert’s persistent individual characteristic or intention. The preferences are perfectly aligned between the good type expert and the public as both prefer an action that is closer to the state. The bad type expert wants the public to choose the higher action regardless of the state. The total payoff of a player is the weighted average between the stage payoffs from period 1 and period 2. Formally, let U_P , U_G , and U_B denote the payoff of the public, the good type, and the bad type expert, respectively. Then

$$\begin{aligned} U_P(\theta_1, \theta_2, a_1, a_2) &= -\sum_{i=1}^2 x_i (a_i - \theta_i)^2, \\ U_G(\theta_1, \theta_2, a_1, a_2) &= -\sum_{i=1}^2 x_i (a_i - \theta_i)^2, \text{ and} \\ U_B(\theta_1, \theta_2, a_1, a_2) &= -\sum_{i=1}^2 x_i (a_i - 1)^2. \end{aligned}$$

where $x_2/x_1 > 0$ denotes the importance of period 2 relative to that of period 1. A formal specification of strategies is presented in Appendix A.

We now propose the definitions of lying and deception, leveraging the definitions from Sobel (2020). To define lying, observe that messages in our environments are framed as a report about the **state** such that each message has its literal meaning that corresponds to each state. The absence of a direct channel through which the expert can transmit information about his type is a fundamental feature of all reputation-building models including Sobel (1985), Benabou and Laroque (1992), and Morris (2001). This assumption is made without loss of generality. Even if a message describing the preference type were available, it would be impossible to convey any information through this channel, thereby maintaining the value of reputation building through the transmission of messages describing the state. We assume that there is a common understanding that $m = \theta$ means the state is θ . Then Definition 1 in Sobel (2020) boils down to the following.

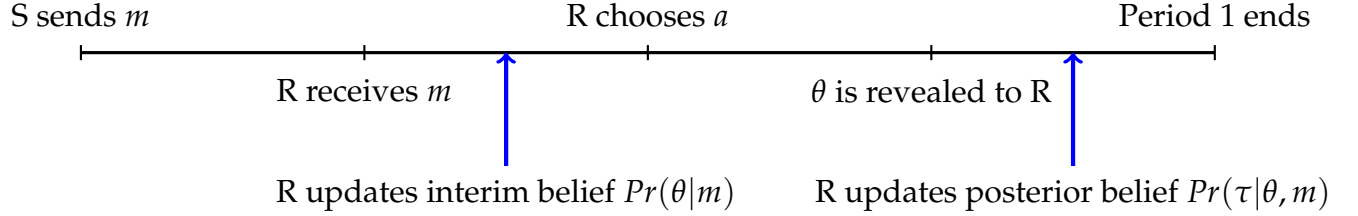


FIGURE 1. Timeline of Each Period and Belief Updating

Definition 1. (*Lying*)

- a. The message m is literally a **lie** given θ if $m \neq \theta$.
- b. The message m is literally a **truth** given θ if $m = \theta$.

Sobel (2020) points out that the definition of lying does not make any reference to how the sender's statements might influence the receiver. In our environment, however, the expert's message may influence the public's belief about the expert's preference type. As we accept the definition of deception as "a deliberate attempt by the sender to induce incorrect beliefs" (Sobel, 2020, pp 919), reputation-building may involve deception in our environment. We define deception in the domain of the *preference type* of the expert as follows. A formal definition of it is presented in Appendix B.

Definition 2. (*Deception about preference types*) A sender's message m is **deceptive with respect to the preference type** if there exists another message m' such that the receiver's posterior belief about the sender's preference type induced by m' is closer to the true preference type of the sender than that induced by m .

Our definition, based on Definition 4 in Sobel (2020), diverges from Sobel's in that it focuses on a specific aspect of deception. Rather than encompassing the entire uncertainty space $\Theta \times T$, here we first define deception with regard to the binary preference type of the expert and will subsequently define deception with respect to the state.¹¹ We shall argue that this distinction is not a matter of choice but rather an inevitable consequence of the belief-updating process inherent in the game.

Figure 4 illustrates the timeline of our game. In our framework, belief updating occurs at two points. The first is during the interim belief updating about the state when the receiver receives a message, $Pr(\theta|m)$. The second is when the outcome (θ, m, a) becomes common knowledge, resulting in sender-type belief updating, $Pr(\tau|\theta, m)$. Given that uncertainty about θ was completely resolved before the sender-type belief updating, deception at this point must be defined solely on the domain of preference type.

It remains possible for a message to be deceptive with respect to the state at the interim stage. Thus we develop a corresponding definition of deception in the domain of the state

¹¹This binary space implies that if one message is deceptive, the other message is non-deceptive.

that incorporates interim belief updating below. A formal definition of it is presented in Appendix B. However, we will subsequently clarify that in our specific environments, no equilibrium message can be deceptive regarding the states.

Definition 3. (*Deception about states*) A sender's message m is **deceptive with respect to the state** if there exists another message m' such that the receiver's interim belief about the state induced by m' is closer to the true state than that induced by m .

We now consider two reputation-building environments. In the first environment, the strategic expert is of the bad type while the behavioral type is committed to truth-telling. In the second environment, the strategic expert is of the good type while the behavioral type is committed to always sending the higher message.

2.1. Reputation Building with Bad-type Truth-telling (BT). Consider the environment in which the strategic expert is the bad type whose interests are misaligned with that of the public as in Benabou and Laroque (1992).¹² The behavioral type always reports the state truthfully. That is, the behavioral type commits to the strategy of a myopic good-type expert. The common prior is that the expert is of the behavioral type with probability $1/2$. In this environment, the equilibrium behavior of the strategic expert in period 2 is straightforward in the absence of the reputation-building motive. She will always send the higher message.

Let's consider the strategic expert in period 1. On one hand, she has incentives to pursue immediate benefits from leading the public to take action $a_1 = 1$ given her preference type. On the other hand, she has incentives to build a reputation by sending truthful messages to pursue benefits in period 2.¹³ When the relative importance of period 2 is large enough, the reputation-building incentives dominate the misguiding incentives. Thus, in the unique equilibrium, the strategic (bad type) expert tells the truth to pretend to be the good type. A formal characterization of the set of equilibria is presented in Appendix A.1.

Consider the strategic (bad type) expert's message in state $\theta_1 = 0$. According to Definition 1, $m_1 = 0$ is a truth. According to Definition 2, $m_1 = 0$ is deceptive with respect to the preference type because $m_1 = 1$ reveals the expert's preference type. That is, when $\theta_1 = 0$, $m_1 = 0$ is a deceptive truth.¹⁴ It is important to note that no equilibrium message

¹²Note that the conflict of interest between the expert and the public occurs at both states in Benabou and Laroque (1992), while the conflict of interest occurs at only one state in our environment.

¹³Note that the trade-off between the incentives to lie and to build reputation exists only when $\theta_1 = 0$.

¹⁴One caveat is that whether a message is deceptive or not in a given state sometimes depends on the specification of off-the-equilibrium path beliefs. This occurs in the reputation-building equilibrium of the BT environment in which the expert never sends $m_1 = 0$ conditional on $\theta_1 = 1$. We circumvent this issue in our experiment by fixing the expert's message to be $m_1 = 1$ conditional on $\theta_1 = 1$. As a result, when $\theta_1 = 1$, the expert has no decision to make, and thus whether a message is deceptive does not have any payoff consequences.

can be deceptive with respect to the state in this environment. This is because the equilibrium messages are truthful about the state. By definition, no other message can result in an interim belief about the state that is strictly closer to the true state. These results are summarized in the following proposition.

Proposition 1. *(Deceptive truth-telling) When the relative importance of period 2 to period 1 is large enough, in the unique equilibrium, the strategic (bad type) expert tells the truth in period 1 to pretend to be the good type. As a result, when $\theta_1 = 0$, the equilibrium message $m_1 = 0$ is*

- (a) *a truth,*
- (b) *deceptive with respect to the preference type, and*
- (c) *not deceptive with respect to the state.*

It is noteworthy that no babbling equilibrium exists in our environment because the behavioral expert is constrained to always tell the truth, so the message is informative regardless of the strategic expert's strategy. Thus, the receiver does not completely ignore the expert's message in any equilibrium, thereby ruling out the possibility of babbling equilibria.

2.2. Reputation Building with Good-type Lying (GL). Consider an environment where the strategic expert is of the good type, meaning their interests are perfectly aligned with those of the public. The behavioral type, however, always sends the higher message regardless of the state, effectively committing to the strategy of a myopic bad-type expert.¹⁵ The common prior is that the expert is of the behavioral type with probability $1/2$. In this environment, the equilibrium behavior of the strategic expert in period 2 is straightforward in the absence of the reputation-building motive. She will always report truthfully.

Let's consider the strategic expert in period 1. On one hand, she has incentives to pursue immediate benefits from leading the public to take action that matches the state. On the other hand, she has incentives to build a reputation by sending $m_1 = 0$ to perfectly reveal her preference type and pursue benefits in period 2.¹⁶ In contrast to the BT environment in which the expert's desire to **conceal** her preference type leads to reputation-building incentives, the expert's desire to **reveal** her type in the GL environment is the cause of reputation-building incentives. When the relative importance of period 2 is large enough, the reputation-building incentives dominate the truth-telling incentives. A formal characterization of the set of equilibria is presented in Appendix A.2.

¹⁵In this environment, the equilibrium behavior reflects the concept of "political correctness" introduced by Morris (2001), who does not assume any behavioral type. Introducing a behavioral type with an exogenous strategy in our setting ensures that reputation-building behavior with non-deceptive lying emerges as the unique equilibrium outcome. Additionally, introducing a behavioral type imposes the literal meaning of language on messages.

¹⁶Note that the trade-off between state-revealing incentive and reputation-building incentive exists only when $\theta_1 = 1$.

Consider the strategic (good type) expert's message in the state $\theta_1 = 1$. Definition 1 implies that $m_1 = 0$ is a lie. At the same time, $m_1 = 0$ is non-deceptive with respect to the preference type because it reveals the expert's preference type. That is, when $\theta_1 = 1$, $m_1 = 0$ is a non-deceptive lie in the GL environment. In this environment, it is not possible for a message to be deceptive with respect to the state in equilibrium. This is due to the fact that the message sent by the strategic type in one state is exactly the message sent by the behavioral type in the other state, and vice versa, resulting in both equilibrium messages being uninformative about the state. Consequently, the updated belief at the interim stage, induced by each message, remains the same as the prior belief, $1/2$. These results are summarized in the following proposition.

Proposition 2. (*Non-deceptive lying*) *When the relative importance of period 2 to period 1 is large enough, in the unique equilibrium, the strategic (good type) expert always sends message 0 in period 1 to reveal her good type. As a result, when $\theta_1 = 1$, the equilibrium message $m_1 = 0$ is*

- (a) *a lie,*
- (b) *not deceptive with respect to the preference type, and*
- (c) *not deceptive with respect to the state.*

In conclusion, in the BT environment, reputation building requires that the strategic expert **conceal** her preference type by mimicking the behavioral type and sending a **truthful** message when the state is 0 in period 1. In the GL environment, reputation building takes place when the strategic expert **reveals** her type by behaving differently from the behavioral type and sending a **non-truthful** message when the state is 1 in period 1. In short, reputation building in the first environment involves *deceptive truth-telling* while that in the second environment involves *non-deceptive lying*.

3. Experiment I

3.1. Design. We would like to experimentally investigate if the expert can successfully build her reputation using her message in period 1 depending on whether reputation-building involves a non-deceptive lie or a deceptive truth. Thus, we use the two reputation-building environments as our experimental treatments. We choose the relative importance of period 2 to be $x_2/x_1 = 20$ such that there is a unique reputation-building equilibrium in each environment.

When implementing the reputation-building games in the lab, we simplify it by restricting the expert's choices under the contingencies that are irrelevant to reputation-building. Precisely, we constrain the expert to send a truthful message conditional on $\theta_i = 1$ in the BT environment and conditional on $\theta_i = 0$ in the GL environment.

Treatment	BT	GL
Behavioral Expert	Truth-telling $\theta_1 = 0 \longrightarrow m_1 = 0$ $\theta_1 = 1 \longrightarrow m_1 = 1$	Always sending $m_1 = 1$ $\theta_1 = 0 \longrightarrow m_1 = 1$ $\theta_1 = 1 \longrightarrow m_1 = 1$
Strategic Expert	Truth-telling $\theta_1 = 0 \longrightarrow m_1 = 0$ $\theta_1 = 1 \longrightarrow m_1 = 1$	Always sending $m_1 = 0$ $\theta_1 = 0 \longrightarrow m_1 = 0$ $\theta_1 = 1 \longrightarrow m_1 = 0$

■ The expert's deliberate choices are highlighted in boldface.

TABLE 1. Equilibrium Strategy of the Expert in Period 1

Table 1 presents the experimental design and the predictions about the expert's behavior in period 1 from the reputation-building equilibrium in each environment. The expert's deliberate choices are highlighted in blue boldface.

3.2. Experimental Procedure. In our experiment, there are three roles: H (human)-sender, C (computer)-sender, and Receiver. The C-sender was programmed to play the truth-telling strategy in the BT treatment and to always send $m_i = 1$ in the GL treatment, regardless of the state. Participants were randomly assigned to the role of H-sender (one-third) or Receiver (two-thirds) at the beginning of each session, and their roles remained fixed throughout. Receivers were randomly paired with senders, without knowledge of whether the sender was an H-sender or C-sender, to play two periods of sender-receiver games. Each pair's interaction, consisting of two periods, was referred to as a round. Each subject played the game in one treatment condition, following a random matching protocol and a between-subjects design.

We further illustrate the experimental procedure based on Treatment BT. At the start of each stage, the state (Red or Gray) was randomly chosen, with an equal likelihood for both states. Before being informed about the realized state, H-senders were prompted to select their message transmission rule by choosing the relative proportions of red and gray colors in a wheel that will be spun under the contingency that the realized state is Gray. In case the realized state was Red, the whole range of the spinning wheel was automatically assigned to be red, without allowing for any deliberate choice by the senders. Once the state was randomly chosen, the corresponding wheel was spun and the message chosen was sent to the paired receiver based on the spinning wheel outcome, without disclosing the submitted spinning wheel itself (involving the relative proportion of red and gray colors) to the receiver. This design allows us to obtain the sender's complete state-contingent message plan. Receivers then received the corresponding message and made a conjecture about the state using a slider bar. Before stage 2 began, receivers were

informed of the state of stage 1. At the end of each round, both senders and receivers received information feedback.

At the end of the experiment, subjects played a dictator game. To measure each subject's other-regarding preference, we made each participant propose their share of the given amount of money (10,000 KRW; approximately 8 USD) in the position of a dictator. Then, each subject was randomly paired and assigned to either the dictator or the dictated. In each pair, the dictator claimed the share she proposed before, and the dictated received the remaining share accordingly. One pair was randomly selected in each session for the actual payment according to the outcome of the dictator game. For more details about the experimental instructions, see the sample instructions presented in Appendix O.

We conducted a total of 8 sessions, with 4 sessions for each treatment. Each session consisted of 24 or 27 participants, resulting in a total of 198 participants across both treatments (96 participants for BT treatment, 102 participants for GL treatment). The experiment took place in January 2023 at Seoul National University (SNU), and participants were recruited through SNU's online community, mainly consisting of students or recent graduates from SNU, Yonsei University, and Korea University. The experiment was programmed using oTree (Chen, Schonger and Wickens, 2016), and instructions were provided at the beginning of each session. A considerable amount of time and effort was dedicated to ensuring the subjects had a comprehensive understanding of the experimental instructions and our experimental design incorporated several measures to ensure participants fully comprehended the game. Subjects were only allowed to participate if they passed a comprehension quiz. They were provided with a practice game before the official games began. On average, 45 minutes were spent explaining the instructions, and subjects were provided with additional instruction reading time, a practice game, several Q&A sessions before the beginning of the official game, and complete feedback at the end of each official round. On average, a session lasted approximately 100 minutes, and participants received an average payment of approximately 32,000 Korean won (equivalent to roughly 26 USD), ranging from 15,000 KRW to 36,000 KRW.

3.3. Results. In this section, we present the key experimental results from Experiment I, focusing on the sender's behavior in stage 1. Additional results, including receiver strategies and welfare analysis, can be found in Appendix D.¹⁷

¹⁷All nonparametric tests were conducted using session-level data, aggregated across all rounds for each stage of the experiment. This approach is well justified because of the absence of noticeable learning among the subjects, as shown in Figure 11 in Appendix D. Detailed p-values resulting from the non-parametric tests are provided in Table 5 of Appendix E.

Figures 2(a) and 2(b) display the sender strategies in the BT and GL environments, respectively, each with two panels. The left panel shows the average truth-telling probability aggregated over all rounds and sessions for each stage. The blue diamonds represent the theoretical predictions. The right panel presents the distribution of individual truth-telling probabilities, with blue crosses indicating the theoretical predictions.

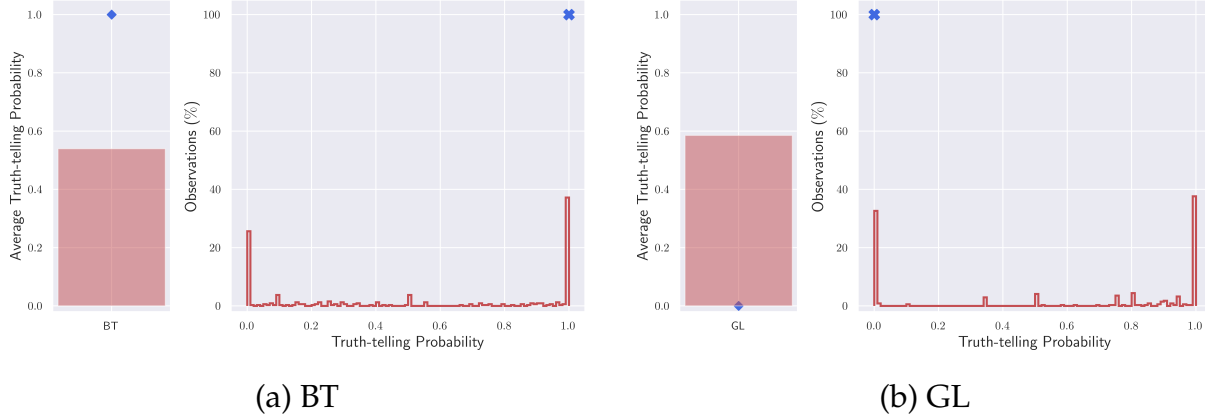


FIGURE 2. Sender Strategy (Stage 1)

Note: In each (a) and (b), the bar graph on the left panel shows the average truth-telling probability aggregated over all rounds and sessions for the first stage and the distribution on the right panel shows the distribution of the truth-telling probabilities by each individual in each round of each session. Blue squares and crosses show the theoretical predictions in each panel.

We make several notable observations from the results. Firstly, both treatments exhibit a significant degree of deviation in sender strategies. Secondly, the distribution of sender strategies shows a bimodal pattern, with two distinct peaks representing lying with certainty and truth-telling with certainty. These peaks account for the majority of observations in both treatments, suggesting that the deviations in stage 1 primarily stem from a subgroup of subjects exhibiting extreme behavior.

Result 1. *A significant proportion of subjects abstain from engaging in reputation building when it entails deceptive truth-telling in the BT treatment and non-deceptive lying in the GL treatment.*

The observed bimodality poses a challenge for standard behavioral models. Traditional explanations, such as mistakes from optimal play (e.g., quantal response equilibrium (McKelvey and Palfrey, 1998)), heterogeneity in strategic sophistication (Crawford, 2003), inequity aversion (Fehr and Schmidt, 1999), or efficiency-seeking (Charness and Rabin, 2002), do not easily account for this pattern. Learning does not change the overall behavior as indicated in Figure 9 for the sender strategies in the last 3 rounds and Figure 11 for the time-trend of the average sender strategy presented in Appendix D.1. Additionally, our analysis of the receiver strategies (see Appendix I) rules out the possibility that the observed sender behavior is an empirical best response.

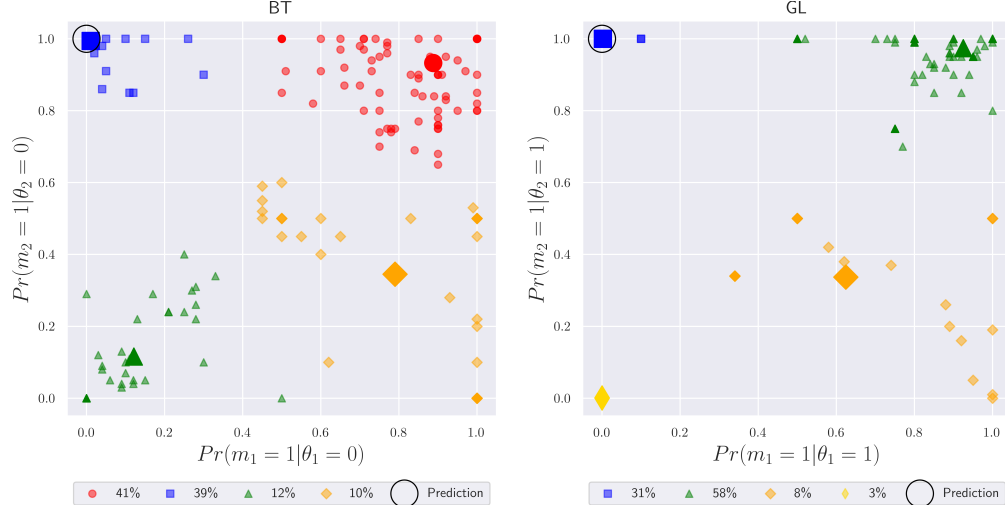


FIGURE 3. Clustering of Sender's Strategy

Note: In both panels, the horizontal axis represents the sender's strategy in stage 1, and the vertical axis represents that in stage 2. The blue square denotes the equilibrium cluster. The red circle indicates the deception aversion cluster in the BT treatment, while the green triangle represents the lying aversion cluster in both treatments. The yellow diamonds in the GL treatment indicate the lying preference cluster. The orange diamonds in both treatments correspond to noise clusters. The center of each cluster is highlighted by a larger shape.

Could the observed deviations in the sender's strategy in stage 1 be driven by her concerns on the payoff consequences for receivers in stage 2?¹⁸ Our data suggests otherwise. Figure 3 presents the joint empirical distribution of sender strategies in stages 1 and 2, along with the results of the k -means clustering analysis (MacQueen, 1967) using four clusters. The horizontal axis represents the sender's strategy in stage 1, while the vertical axis represents the strategy in stage 2. The black empty circle denotes the theoretical prediction. In the BT treatment, the equilibrium strategy is to tell the deceptive truth in stage 1 and lie in stage 2. The value on the horizontal axis and that on the vertical axis indicate departures from the equilibrium predictions that can be driven by deception aversion and preference for lying, respectively. In contrast, in the GL treatment, the equilibrium strategy is to tell a non-deceptive lie in stage 1 and tell the truth in stage 2. Both the value on the horizontal axis and that on the vertical axis can be understood as degrees of lying aversion.

The k -means clustering analysis reveals that approximately 37% of observations in the BT environment (left panel) and 31% in the GL treatment (right panel) conform to the equilibrium prediction, represented by the blue squares. However, a notable portion of

¹⁸Sobel (2020) defines "damage" as the payoff consequence of a sender's message on the listener(s), distinguishing it from lying and deception.

observations—41% in the BT treatment (indicated by red circles)—deviate from the equilibrium prediction in stage 1 in a manner that aligns with deception aversion. Despite this initial deviation, these participants opt to lie in stage 2, resulting in negative payoff consequences for the receiver. Similarly, in the GL treatment, 58% of observations (represented by green triangles) deviate in stage 1 in a manner consistent with lying aversion, even though their choices ultimately lead to negative outcomes for the receiver in stage 2. These findings suggest that the observed deviations cannot be solely attributed to concerns regarding the payoff consequences for the receivers.

It is still premature to conclude that the observed deviations are primarily due to people’s aversion to lying and deception, as there remains an alternative plausible explanation: erroneous inferences by receivers (Eyster and Rabin, 2005; Jehiel, 2005) or senders not being able to fully recognize the benefit of reputation building, which may create a situation where building a reputation in stage 1 is (subjectively) not worthwhile for senders. Without access to inference or belief data, we cannot definitively determine which explanation is more plausible or dominant. This limitation motivates our next experiment, which we present in the following section.

4. Experiment II: Preferences vs. Inferences

4.1. Experimental Design and Procedure. In this section, we introduce a new experiment designed to provide clear identification for a more dominant explanation between the preference-based and inference-based approaches for the observed deviations from the equilibrium predictions. We will focus on highlighting the key differences in the new experimental design from Experiment I.

The new experiment maintains the same Stage 1 strategic interaction as the original one, but the Stage 2 communication game is replaced with the receiver’s direct belief reporting about the sender’s preference type, which directly affects the sender’s payoff. Each receiver is asked to answer the question: “What do you think is the likelihood of the sender you are paired with being H-sender given the message you received?” They are presented with a slider to indicate their belief in any number between 0 and 1 (See Figure 29 in Appendix P). Secondly, we elicit the sender’s second-order belief about the receiver’s belief regarding the sender’s preference type for each message $m \in \{0, 1\}$, denoted by $\lambda^S(m, \hat{\theta})$, where $\hat{\theta}$ denotes the state that is our main interest of analysis in each treatment, i.e., $\hat{\theta} = 0$ in the BT treatment and $\hat{\theta} = 1$ in the GL treatment.. Each sender is asked: “What do you think is the likelihood that the receiver believes you are a H-sender when she receives message m and learns the realized state is $\hat{\theta}$?” for each message $m \in \{0, 1\}$. Two sliders are presented for them to indicate their belief between 0

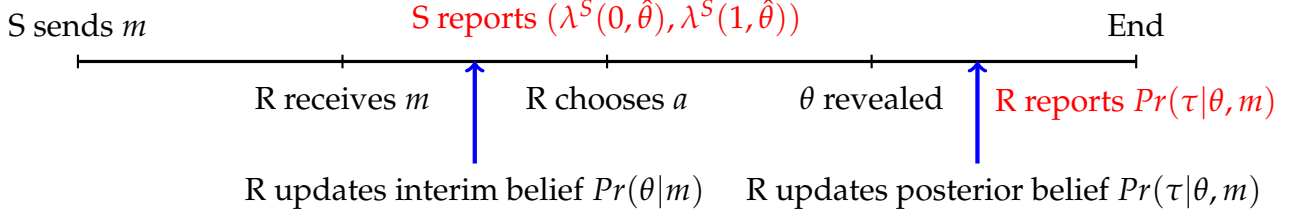


FIGURE 4. Timeline of the Modified Game

Note: $\lambda^S(m, \theta)$ denotes the sender's second-order belief for the receiver's posterior belief that the sender is good after observing the message m and state θ . $\hat{\theta}$ denotes the state that is our main interest of analysis in each treatment, i.e., $\hat{\theta} = 0$ in the BT treatment and $\hat{\theta} = 1$ in the GL treatment. We elicit the message strategy and second-order beliefs conditional on the realization of $\theta = \hat{\theta}$.

and 1 conditional on each message.¹⁹ The timeline of the modified game is illustrated in Figure 4.

Combined with the choice data, the reported second-order belief enables us to clearly identify the main drivers of deviations from equilibrium. On one hand, if the reported belief indicates that the sender perceives reputation-building as unworthy, it implies that deviations are primarily caused by inference errors and cursed beliefs. On the other hand, if the reported belief suggests that the sender views reputation-building as worthwhile, it indicates that the observed deviations are driven by preferences. Our next subsection (Section 4.2) presents a concrete argument for why the sender's second-order beliefs are sufficient to evaluate the subjective value of reputation-building.

Direct belief reporting by receivers also eliminates strategic uncertainty while preserving the incentives for reputation-building. In the environment of Experiment I, the higher the receiver's belief about the sender's preference (good) type, the more likely the receiver is to conform to the message from the sender, resulting in a higher payoff for the sender in period 2. This implies that the value of reputation building is realized through the receiver's period-2 action involving two-step belief updating, i.e. updating about the sender type and updating about the period-2 state given the period-2 message. Thus, the sender-side uncertainty over the receiver's second-period action (and complicated two-step belief updating) might have created strategic uncertainty in the first period. In the new environment, the sender's payoff is directly determined by and increases with the elicited belief of the receiver, mitigating such potential strategic uncertainty arising from the period-2 interaction while preserving the incentives for reputation-building. In converting our 2-period game to the static game, we carefully chose parameter values so

¹⁹Our belief elicitation for both sender and receiver are incentivized with the quadratic scoring rule. For more discussion about (behavioral) incentive compatibility of belief elicitation methods, see Danz et al. (2022).

that the characteristics of the game, such as the equilibrium predictions and the equilibrium characterization over the space of lying and deception costs, remain almost the same quantitatively. For the concrete design and the logic behind it, refer to Appendix K.

Both of these elicitation tasks are incentivized, and we carefully select parameters to ensure that the equilibrium behavior in the modified environment closely aligns with that of the original environment. For a detailed description of the experimental design and procedure, please refer to Appendix K. Appendix P presents sample experimental instructions.

We conducted a total of 8 sessions, with 4 sessions for each treatment. Each session consisted of 24 participants, resulting in a total of 192 participants, with 96 in each treatment. The sessions took place from September to October 2023 at Seoul National University, and participants were recruited in the same way as in Experiment I. On average, each session lasted approximately 105 minutes, and participants received an average payment of 28000 KRW (approximately 25 USD) per person.

4.2. Separating Preference Channel from Inference Channel. In this section, we concretely show that we can calculate each sender's subjective value of reputation building by using the sender's second-order beliefs on each contingency of messages. Let $EU(m|\theta)$ denote the (strategic) sender's subjective expected utility when she sends the message m given the state θ in the first period. Also, let $\lambda^S(m, \theta)$ denote the sender's second-order belief for the receiver's posterior belief that the sender is good after observing the first-period message m and state θ . Then, the condition for the subjective (net) value of reputation building to be positive is

$$\begin{aligned} BT : \quad & EU(0|0) - EU(1|0) > 0 \iff \lambda^S(0,0) - \lambda^S(1,0) > T_{BT}(a(0), a(1)), \\ GL : \quad & EU(0|1) - EU(1|1) > 0 \iff \lambda^S(0,0) - \lambda^S(1,1) > T_{GL}(a(0), a(1)), \end{aligned}$$

where $a(m)$ is the receiver's action after receiving the message m , and $T_{BT}(a(0), a(1))$ and $T_{GL}(a(0), a(1))$ are the threshold values uniquely determined in equilibrium of the respective environment. By observing the sender's second-order beliefs contingent on each message and comparing the difference between the two second-order beliefs with the threshold, we can evaluate whether the sender has an incentive to build a reputation or deviate from the equilibrium. If the sender deviates from the reputation-building equilibrium, even though she evaluates the value of reputation-building positive, then we categorize the observed strategy as arising from the preference channel. Figure 5 summarizes the classification of the pairs of observed strategies and second-order beliefs.

Note that the thresholds T_{BT} and T_{GL} depend on the sender's expectation of the receiver's action. Assuming equilibrium, we get $T_{BT} = 0.2$ and $T_{GL} = 0$. Instead, if we

	Truth-telling	Lying
$\lambda^S(0,0) - \lambda^S(1,0) > T_{BT}$	Equilibrium Prediction	Deception Aversion
$\lambda^S(0,0) - \lambda^S(1,0) \leq T_{BT}$	Lying Aversion	Inference Error

(a) BT

	Truth-telling	Lying
$\lambda^S(0,1) - \lambda^S(1,1) > T_{GL}$	Lying Aversion	Equilibrium Prediction
$\lambda^S(0,1) - \lambda^S(1,1) \leq T_{GL}$	Inference Error	Noise

(b) GL

FIGURE 5. Decomposition of Preference and Inference Channels

Note: The specification of T_{BT} and T_{GL} depends on the sender's expectation of the receiver's action. Using the receiver's action in the equilibrium, we get $T_{BT} = 0.2$ and $T_{GL} = 0$. Using the receiver's average action in Experiment I, we get $T_{BT} = 0.163$ and $T_{GL} = 0.122$. Importantly, for any choice of the receiver's action, $T_{BT} \leq 0.2$ and $T_{GL} \leq 0.4$.

take the receiver's average action empirically observed in Experiment I as a proxy of the sender's expectation, we get $T_{BT} = 0.163$ and $T_{GL} = 0.122$. Importantly, for any $(a(0), a(1))$ that satisfies $a(1) \geq a(0)$, $T_{BT} \leq 0.2$ and $T_{GL} \leq 0.4$. For our empirical analysis, we take these *most conservative* thresholds that are robust to the sender's expectation as well as the risk aversion: $T_{BT} = 0.2$ and $T_{GL} = 0.4$.²⁰

4.3. Results. In this section, we present the key experimental results from Experiment II, again focusing on the sender's behavior. Additional results, including k -means clustering for sender strategies, receiver strategies, and welfare analysis, can be found in Appendix M.

Figures 6(a) and 6(b) display the sender strategies in the BT and GL environments, respectively, each with two panels. The left panel shows the average truth-telling probability aggregated over all rounds and sessions for each stage. The blue diamonds represent the theoretical predictions. The right panel presents the distribution of individual truth-telling probabilities, with blue crosses indicating the theoretical predictions.

The internal validity of the new experimental design should be confirmed first. The main question is whether we could replicate the key result on the sender strategy we obtained from Experiment I. Figure 6, in comparison with Figure 2, provides evidence

²⁰These values are calculated under the assumption that $a(1) - a(0) = 1$, thereby maximizing the expected utility gap. This gap cannot exceed 1 under any concave utility transformation, ensuring robustness to risk aversion. See Appendix K for more details.

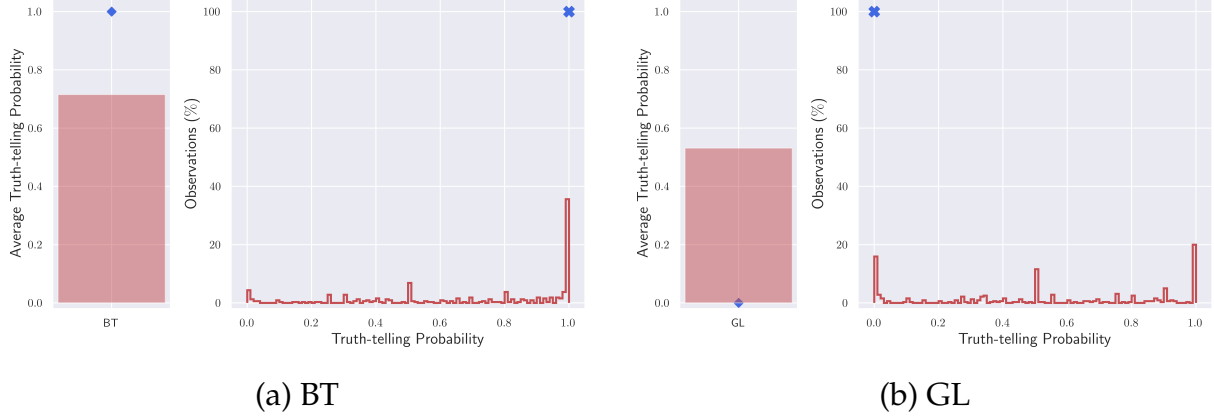


FIGURE 6. Sender Strategy (Stage 1) in Experiment II

Note: In each (a) and (b), the bar graph on the left panel shows the average truth-telling probability aggregated over all rounds and sessions for the first stage, and the distribution on the right panel shows the distribution of the truth-telling probabilities by each individual in each round of each session. Blue squares and crosses show the theoretical predictions in each panel.

supporting it. Firstly, in the BT treatment, the proportions of sender strategies consistent with the equilibrium prediction are almost the same between the two experiments. Secondly, in the GL treatment, the proportion of sender strategies consistent with the equilibrium prediction is smaller in Experiment II than in Experiment I. Although the difference is statistically significant ($p = 0.04$ in the two-sided Mann-Whitney test), its magnitude is small. These two observations imply that Experiment II, both qualitatively and quantitatively, replicates the reputation-building failures observed in both the BT and GL environments of Experiment I.

There are, however, a few notable differences in the results between Experiments I and II. First, in the BT treatment, the bimodality of the distribution disappears in Experiment II. The majority of sender strategies are now mixed, representing partial lying rather than maximal lying. Second, in the GL treatment, there are three peaks, each representing maximal lying, partial lying, and truth-telling. While the proportions of maximal lying and truth-telling strategies both decrease compared to Experiment I, the majority of sender strategies are still partial lying. Notwithstanding such differences, we continue to observe sizable deviations (over 60%) from the equilibrium predictions in each treatment.

The belief elicitation data from Experiment II enables us to identify the main sources of such deviations. Figures 7(a) and (b) describe the difference between the sender's second-order beliefs conditional on the truth-telling contingency and lying contingency in each treatment, respectively. Each figure consists of two panels. The left panel presents the average difference in the sender's second-order beliefs (= the sender's belief about the receiver's belief induced by a truthful message minus that induced by lying) aggregated

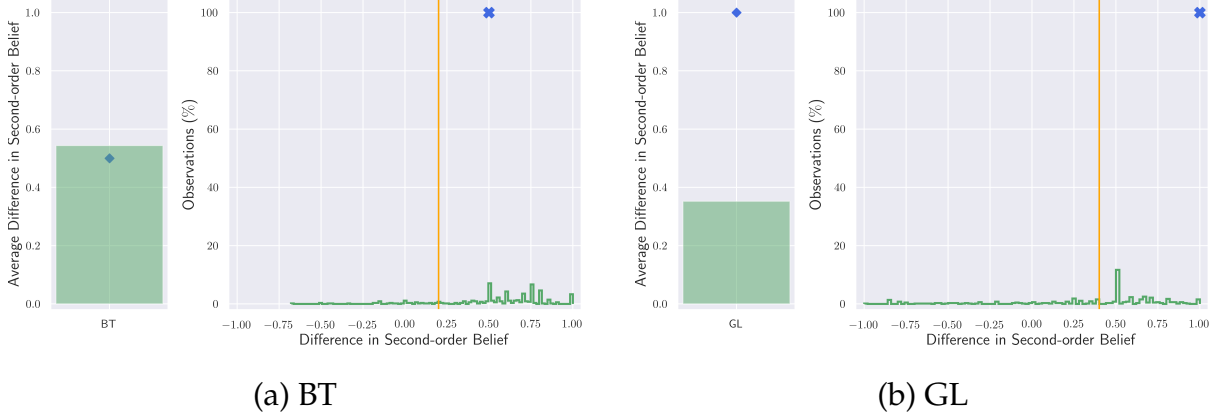


FIGURE 7. Description of Sender's Second-order Belief

Note: In each (a) and (b), the bar graph on the left panel shows the average difference in the sender's second-order beliefs aggregated over all rounds and sessions for the first stage, and the distribution on the right panel shows the distribution of the difference in the second-order beliefs of each individual sender in each round of each session. Blue squares and crosses show the theoretical predictions in each panel. The orange vertical line on the right panel describes the belief threshold determining whether the sender's subjective value of reputation building is positive or negative.

over all rounds and sessions. The blue diamonds again represent the theoretical predictions. The right panel depicts the distribution of individual differences in second-order beliefs, with blue crosses indicating the theoretical predictions. The higher the difference in the reported second-order beliefs, the larger the subjective gain from reputation building.

The orange vertical line in each treatment describes T_{BT} and T_{GL} , the most conservative belief threshold determining whether the sender's subjective value of reputation building is positive or negative. We classify that, if a reported belief difference is above (resp. below) the threshold, reputation building is subjectively (resp. not) valuable. Conditional on observing a sender's strategy departing from the theoretical prediction, the observed deviation can be attributed to either the preference channel or the inference channel, depending on the reported belief difference being above or below the line. The theoretical derivation of such thresholds is described in Appendix K and illustrated in Figure 5.

We make several notable observations from the belief elicitation. First, the sender's second-order belief difference is on average consistent with the theoretical prediction in the BT treatment but deviates substantially from the prediction in the GL treatment. Second, the distributions of the sender beliefs show that a large portion of observations (85.9% in the BT and 62.5% in the GL treatment) are above the orange line. This observation implies that a significant proportion of senders believed that the payoff gain from the reputation-building is strictly positive.

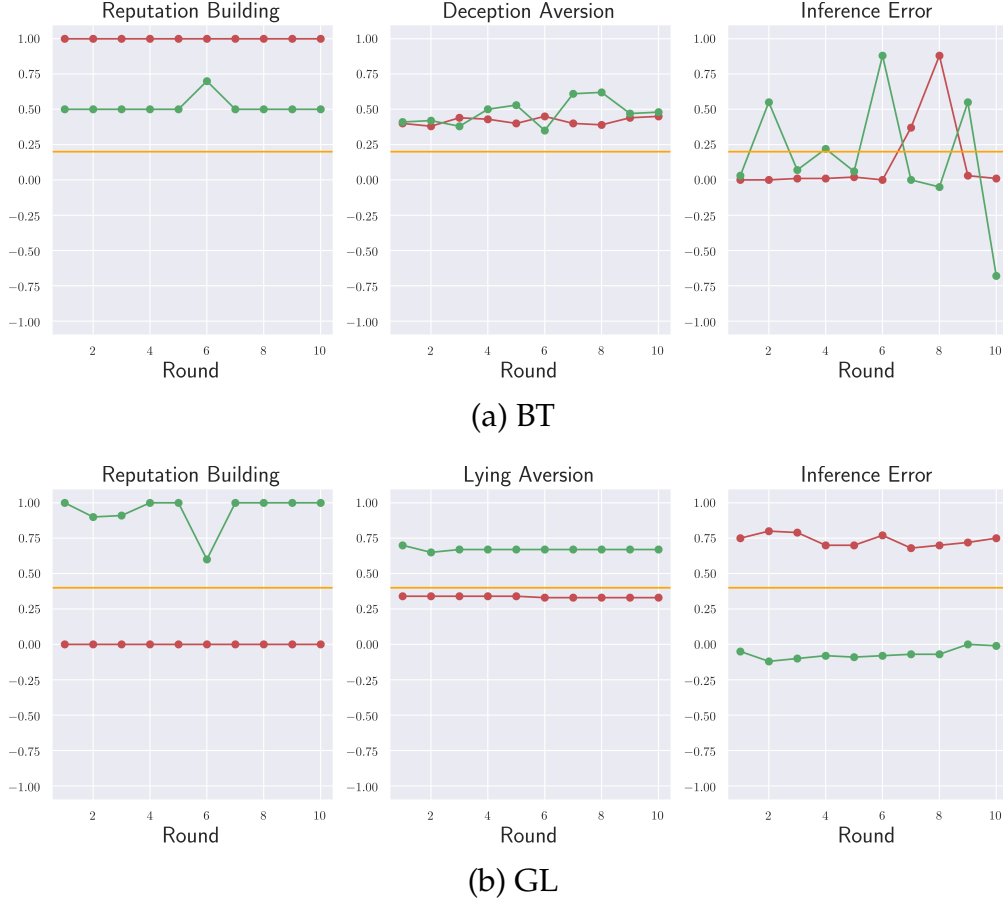


FIGURE 8. Three Representative Individual Behaviors

Note: Each graph represents the strategy and beliefs of each participant representative of each category of behavior across all rounds. The red line represents the truth-telling rate. The green line represents the belief difference. The orange horizontal line is the threshold of belief difference: if the green line is above the orange line, then the sender evaluates the value of reputation sufficiently large so that the equilibrium strategy is preferred to deviations, and vice versa.

To account for individual decision-making, we consider both the strategy and belief of each sender. Figure 8 illustrates three representative individual behaviors across rounds in each treatment. The red line represents the truth-telling rate (strategy), while the green line represents the belief difference. If the green line is above (resp. below) the orange horizontal line, it indicates that reputation-building is (resp. not) materially incentive-compatible based on the reported belief. Figure 8(a) showcases three representative behaviors observed in the BT treatment. In the leftmost panel, the participant engages in equilibrium play, finding reputation-building incentive-compatible and behaving accordingly. The middle panel represents deception aversion, where reputation-building is materially incentive-compatible but not pursued by the participant. The rightmost panel demonstrates inference error, where the participant does not find it incentive-compatible

to build a reputation and behaves accordingly. Figure 8(b) displays three representative behaviors observed in the GL treatment. The leftmost panel represents equilibrium play, the middle panel represents lying aversion, and the rightmost panel represents inference error. Notably, both individual strategy and belief remain relatively stable across rounds, except for some participants who experience inference errors.

	BT	GL
Equilibrium	31%	16%
Aversion	53%	47%
Inference Error	16%	37%

TABLE 2. Individual Classifications

Lastly, Table 2 classifies all senders into one of the following three categories: equilibrium plays, deviations due to preference (aversion), and deviations due to inference errors. For classification, we use each individual's strategy and belief difference averaged over whole rounds. In the BT treatment, 31% are reputation-builders, 53% are those with an aversion to deceiving, and 16% are those with inference errors. In the GL treatment, 16% are reputation-builders, 47% are those with an aversion to lying, and 37% are those with inference errors. The exact percentage differs depending on the cutoffs we use for the classification exercise, but the qualitative distributions remain robust.²¹ This result indicates that deception aversion and lying aversion are important impediments to reputation building.

Result 2. *A significant proportion of subjects abstain from engaging in reputation building when it entails deceptive truth-telling in the BT treatment and non-deceptive lying in the GL treatment. The departure from equilibrium behavior is influenced by both inference error and a preference to avoid lying and deception, with the latter being the primary driver.*

4.4. Other Regarding Preference. The literature on lying aversion has emphasized the role of other-regarding preferences in shaping both the occurrence and extent of lying behavior (Gneezy, 2005; Erat and Gneezy, 2012). In the GL environment, reputation-building involves non-deceptive lying that increases material payoffs for both the sender and the receiver. This suggests that other-regarding preferences are unlikely to be the main driver behind deviations from reputation-building in this setting. In contrast, in the BT environment, reputation-building relies on deceptive truth-telling, which benefits the sender at the expense of the receiver—implying that other-regarding preferences could plausibly influence behavior. To assess this, we examined sender behavior in the stage

²¹For robustness results, please refer to Table 20 in Appendix M.

2 communication game in Experiment I. We found in Figure 3 that most senders who avoided deceptive communication in stage 1 nevertheless acted selfishly in stage 2. This pattern indicates that concerns for the receiver's payoff are unlikely to be the primary explanation for the observed deviations.

To further support this conclusion, we measured individual other-regarding preferences using a dictator game conducted at the end of each session. We then performed OLS regressions, using various measures of sender behavior as dependent variables and the sender's giving share in the dictator game as the key independent variable.

Table 7 reported in Appendix M presents the OLS regression results for each treatment, assessing whether other-regarding preferences are related to the observed deviations from the equilibrium. Across all specifications, the independent variable is the giving share in the dictator game. In columns (1) and (5), the dependent variable is the probability of telling the truth averaged over the 10 rounds. In columns (2) and (6), the dependent variable is the dummy variable which indicates whether the sender deviates from the equilibrium strategy. In columns (3) and (7), we exclude those who deviate from the equilibrium strategy due to inference error and consider the same regression as in the previous columns. Finally, in columns (4) and (8), the dependent variable is the dummy variable indicating whether the sender exhibits aversion to lying/deception. All results show that the correlation between any of these measures of sender behavior and other-regarding preference is not statistically significant. These tables provide evidence that other-regarding preferences do not systematically confound our analysis.

5. A Model of Lying and Deception Costs

In this section, we present a simple model incorporating lying costs and deception costs separately.²²

Assumption 1. (*Lying cost*) *The expert incurs the lying cost $c_l \geq 0$ when telling a lie, and 0 otherwise.*

Defining the deception cost is less straightforward than defining the lying cost. According to Definition 2, given the state θ , we know that one message in our binary message space is deceptive and the other is not. Let m^n denote the non-deceptive message.

Assumption 2. (*Deception cost*) *If the expert sends a message m and the receiver's resulting posterior is $\lambda(m, \theta)$, then she incurs the deception cost $c_d |\lambda(m, \theta) - \lambda(m^n, \theta)|$ where $c_d \geq 0$.*

Intuitively, if $m = m^n$, no deception cost is incurred. In Definition 2, a message is considered deceptive if it causes the public's belief to deviate farther from the correct

²²See Sobel (2020, Section 8.G. Incorporating Costs of Lying and Deception) for more discussions. See Eilat and Neeman (2023) for a model of deception costs in the Crawford and Sobel (1982) environment.

belief than the other available message. Our measure of deception cost captures this feature of deception.²³

Assumptions 1 and 2 imply that the expected utility of the expert type θ who sends a message m is given by

$$EU^a(m|\theta) = EU(m|\theta) - c_l I\{m \neq \theta\} - c_d |\lambda(m, \theta) - \lambda(m^n, \theta)|,$$

where $EU(m|\theta)$ is the expected utility when there is no lying and deception cost.

We now conduct equilibrium characterization when lying and deception costs exist. This will explicitly reveal how the distribution of the participant's lying and deception costs explains our experimental results. When providing behavioral predictions from the model, we limit our attention to the parameter values that we selected for our experimental implementation. Since we carefully specified the parameters of Experiment I and II so that their characterization is sufficiently close to each other, we propose statements common to both experimental settings. For the full characterization statements and proofs for each experiment in detail, refer to Appendix G and L. First, we summarize the equilibrium characterization in the BT environment.

Proposition 3. (*Less Reputation Building by Deception Cost*) *Under Assumptions 1 and 2, there exists at most two thresholds $\underline{c}(c_d) < \bar{c}(c_d)$ for each $c_d \geq 0$ such that when $\theta_1 = 0$, the equilibrium message is*

- i. $m_1 = 0$ (Full reputation-building) if $c_l > \bar{c}(c_d)$;
- ii. $m_1 \in [0, 1]$ (Partial reputation-building) if $\underline{c}(c_d) < c_l < \bar{c}(c_d)$;²⁴
- iii. $m_1 = 1$ (No reputation-building) if $c_l < \underline{c}(c_d)$,

and the set of (c_d, c_l) supporting each case is nonempty.

Proof. See Appendix J. □

Proposition 3 states that the relative size between c_l and c_d decides the equilibrium strategy. Remember that the equilibrium strategy in Proposition 1 requires deceptive truth-telling. When c_l is large compared to c_d , deviation from the deceptive truth-telling is costly to the expert. Thus, the equilibrium strategy remains the same. When c_l is small compared to c_d , deviating from deceptive truth-telling becomes profitable for the expert.

²³Whether a particular message on the equilibrium path incurs a deception cost may depend on the posterior belief assigned to off-the-path messages. Following convention in the literature and in line with the assumption in Eilat and Neeman (2023), we assume that any off-path message can only induce posterior beliefs generated by an on-the-path message. This assumption ensures that: 1) off-path messages never present a tempting deviation for the Sender, and 2) an artificially created perception of deception stemming from the posterior beliefs assigned to off-the-path messages is not permitted.

²⁴By slight abuse of notation, $m_1 \in [0, 1]$ means sending $m_1 = 1$ with some probability in $[0, 1]$. Here, i.e. $\underline{c}(c_d) < c_l < \bar{c}(c_d)$, there always exists an equilibrium where $m_1 = 1$ with probability in $(0, 1)$. Additionally, there potentially exist equilibria with a pure strategy.

She begins to reduce the degree of deception by selecting $m_1 = 1$ with positive probability despite the lying cost it incurs.

Next, we summarize the equilibrium characterization in the GL environment.

Proposition 4. (*Less Reputation Building by Lying Cost*) Under Assumptions 1 and 2, there exists at most two thresholds $c_*(c_d) < c^*(c_d)$ for each $c_d \geq 0$ such that when $\theta_1 = 1$, the equilibrium message is

- i. $m_1 = 0$ (Full reputation-building) if $c_l < c_*(c_d)$;
- ii. $m_1 \in [0, 1]$ (Partial reputation-building) if $c_*(c_d) < c_l < c^*(c_d)$;²⁵
- iii. $m_1 = 1$ (No reputation-building) if $c_l > c^*(c_d)$,

and the set of (c_d, c_l) supporting each case is nonempty.

Proof. See Appendix J. □

According to Proposition 4, the relative size between c_l and c_d decides the equilibrium strategy. Remember that the equilibrium strategy in Proposition 2 requires non-deceptive lying. When c_d is large compared to c_l , deviation from the non-deceptive lying is costly to the expert. Thus, the equilibrium strategy remains the same. When c_d is small compared to c_l , deviation from the non-deceptive lying becomes profitable to the expert. She begins to reduce the degree of lying by selecting $m_1 = 1$ with positive probability despite the deception cost it incurs.

Figure 16 in Appendix G summarizes the above characterization results on the (c_d, c_l) –space. The equilibrium characterization with lying and deception costs provides a reasonable explanation of why the sizable portion of senders deviate from the equilibrium predicted in Section 3 and 4. For each sender strategy observed in the experiment, we can find the range of relative ratio between lying cost c_l and deception cost c_d by revealed preference argument in Proposition 3 and 4. Using this, we also conduct partial identification exercise in Appendix H, where we identify the nonparametric distribution of c_d and c_l that rationalizes our experimental data. The main takeaway from this exercise is that participants have a heterogeneous aversion to lying and deception. A sizable portion of participants show behavior consistent with having an aversion to either literal lying or deception exclusively, while some participants show behavior consistent with an aversion to both or neither of them.

6. Related Literature.

Sobel (2020) provides separate definitions of lying and deception in a strategic environment. Eilat and Neeman (2023) present a model of deception costs in the Crawford and

²⁵By slight abuse of notation, $m_1 \in [0, 1]$ means sending $m_1 = 1$ with some probability in $[0, 1]$. Here, i.e. $c_l < c^*(c_d)$, there always exists an equilibrium where $m_1 = 1$ with probability in $(0, 1)$. Additionally, there also exist equilibria with each pure strategy.

Sobel (1982) environment. To our knowledge, we are the first to experimentally disentangle lying aversion and deception aversion in strategic communication environments. The previous literature in economics has focused on the role of payoff consequences in lying. For instance, Gneezy (2005) discovers that participants consider the payoff consequences of lying and are concerned not only about their own payoff but also about the harm caused to the other party. Hurkens and Kartik (2009) find that Gneezy's results are compatible with a model in which lying aversion is independent of social preferences. Erat and Gneezy (2012) conduct experiments comparing black lies, white lies, and blue lies (referred to as Pareto white lies) and find that people tend to avoid even blue lies. In these papers, the terms "lie" and "deception" are used interchangeably. Sobel's notion of deception, as well as ours, is defined independently of material consequences, and our experimental results demonstrate that deception aversion plays distinct roles in shaping human behavior against the role of the associated payoff consequences.

Recently, the economics literature has begun to explore the moral considerations surrounding lying, which are believed to contribute to individuals' reluctance to deceive. In the context of the die-rolling and self-reporting game (Fischbacher and Föllmi-Heusi, 2013), Gneezy et al. (2018) and Abeler et al. (2019) observe that individuals may hesitate to lie due to concerns about how they will be perceived as liars. Theoretical models developed by Dufwenberg and Dufwenberg (2018) and Kholmetski and Sliwka (2019) incorporate this concern for social image. However, these studies do not differentiate between lying aversion and deception aversion, as this distinction is empirically irrelevant in the communication environments they consider.

A few previous studies have explored the emergence of deception in communication environments. Sutter (2009) argues that senders may choose to tell the truth while anticipating that receivers will not follow their message, leading to a "sophisticated lie." This concept aligns with the notion of deceptive truth-telling in our environment. More recently, Innes (2022) introduces an experimental design that aims to compare deceptive lying and non-deceptive lying. In their Deception Treatment, a sender communicates the color of a dot to an anonymous receiver who then takes an action. The color of the dot and the receiver's action determine the payoff for both parties. In the No-Deception Treatment, the receiver remains anonymous, but the entity taking the action is replaced by a computer, eliminating the possibility of deception. Innes (2022) finds that deception actually promotes lying. Unlike the main goal of Innes (2022), which is to understand the interplay between lying and deception, our primary objective is to identify the distinct roles of lying and deception. Our reputation-building environments involve non-deceptive lying and deceptive truth-telling, enabling us to separately examine individuals' intrinsic aversion to lying and their aversion to deception.

Similarly to our study, Ettinger and Jehiel (2010, 2021) examine a multi-period reputation-building environment with a multi-dimensional belief domain to provide evidence of deceptive communication. They find that a significant proportion of sender participants adopt a “deceptive tactic,” initially telling the truth before switching to lying at a certain point. Notably, this tactic is not part of a Perfect Bayesian Equilibrium but rather an Analogy-Based Expectation Equilibrium (Jehiel, 2005; Ettinger and Jehiel, 2010), which arises when senders are rational while receivers, reasoning coarsely, base their expectations on the aggregate lie probability of senders by type, treating it as if it were stationary across all periods. Their primary focus is to document the frequent occurrence of this deceptive strategy, despite its absence in the fully rational case, and to investigate whether receivers are deceived by it. In contrast to our study, which examines the preference aspect of deception, their work emphasizes its cognitive dimension.

Recently, we learned of a separate attempt by Elmschauser, Friedman and Jo (2025) to obtain experimental evidence for deception aversion. Their study builds on the die-rolling game introduced by Fischbacher and Föllmi-Heusi (2013) by introducing an additional dimension of private information for the sender. The sender observes both the die outcome and the result of a coin flip, where the coin outcome is payoff-relevant for the receiver. The sender sends a message about the die, which may affect the receiver’s beliefs about the coin. The authors show theoretically how to identify deception aversion with minimal assumptions on the sender’s second-order beliefs. They find that a significant fraction of participants exhibit signs of deception aversion, a result that qualitatively aligns with our findings. Both their environment and ours share the feature of a multi-dimensional belief domain combined with a uni-dimensional message space. However, unlike in our setting, their environment lacks substantive information to convey in the dimension to which the literal meaning of the message corresponds.

Deception aversion and guilt aversion share a conceptual relationship as both concepts capture the disutility generated by the discrepancy between what could have happened and what actually occurred. Guilt, as defined by Battigalli and Dufwenberg (2007), involves a sense of disappointment measured by the disparity between player i ’s belief about player j ’s payoff resulting from player i ’s strategy and the payoff player j anticipates.²⁶ In contrast, deception aversion is characterized by the disparity between the posterior belief (rather than the payoff consequence) induced by a message and the closest posterior belief to the true state among all possible beliefs that can be induced by available messages in the context of strategic transmission of private information.

²⁶Experimental evidence of guilt aversion in a trust game with pre-play communication is provided by Charness and Dufwenberg (2006).

7. Conclusion

At the heart of the inquiry into the morality of lying and deception lies the recognition that these behaviors are integral parts of human preferences and social interactions. Lying and deception shape how individuals communicate with each other in settings with private information. Moreover, the acceptance or condemnation of lying and deception by members of each society can lead to the development of norms that either allow or discourage such behaviors in certain situations, as well as legal and institutional mechanisms that either punish or tolerate them. The diverse norms and institutions further influence information transmission and communication in each society, underscoring the importance of understanding individual attitudes toward lying and deception as a crucial first step.

In this paper, we explore lying and deception in a reputation-building environment with repeated communication. We identify distinct scenarios where non-deceptive lying and deceptive truth-telling emerge as unique equilibrium phenomena. In our laboratory experiment, we find that the proportion of senders successfully employing the equilibrium strategy to build reputation is consistently lower than the theoretical predictions. Furthermore, our additional experiment allows us to demonstrate that this deviation from the equilibrium stems from inference errors and individuals' intrinsic aversion to lying and deception.

Our findings reveal that individuals possess an intrinsic aversion to deceiving others, which differs from their aversion to telling literal lies. We introduce a simple model of deception cost, utilizing the distance between posterior beliefs generated by different messages as a reasonable measure. However, it is crucial to explore alternative modeling approaches to comprehensively capture the deception cost, warranting further theoretical and experimental investigation. We leave this intriguing avenue for future research. Additionally, future studies could examine the generalizability of our results to diverse contexts and populations, as well as explore alternative communication mechanisms that may be more effective in situations where lying or deception is prevalent.

References

- Abdulkadiroglu, Atila, Parag A Pathak, Alvin E Roth, and Tayfun Sönmez**, “Changing the Boston school choice mechanism,” 2006.
- Abeler, Johannes, Armin Falk, and Fabian Kosse**, “Malleability of preferences for honesty,” *The Economic Journal*, 2024, p. ueae044.
- , **Daniele Nosenzo, and Collin Raymond**, “Preferences for Truth-telling,” *Econometrica*, 2019, 87 (4), 1115–1153.
- Allingham, Michael G and Agnar Sandmo**, “Income tax evasion: A theoretical analysis,” *Journal of public economics*, 1972, 1 (3-4), 323–338.
- Baron, David P. and Roger B. Myerson**, “Regulating a Monopolist with Unknown Costs,” *Econometrica*, 1982, 50 (4), 911–930.
- Battigalli, Pierpaolo and Martin Dufwenberg**, “Guilt in games,” *American Economic Review*, 2007, 97 (2), 170–176.
- Benabou, Roland and Guy Laroque**, “Using Privileged Information to Manipulate Markets: Insiders, Gurus, and Credibility,” *The Quarterly Journal of Economics*, 1992, 107 (3), 921–958.
- Blake, William**, “Proverbs of Hell,” In *The Marriage of Heaven and Hell* 1790. Available online at https://www.gutenberg.org/files/17976/17976-h/17976-h.htm#Page_35.
- Blazquiz-Pulido, Juan Francisco, Luca Polonio, and Ennio Bilancini**, “Who’s the deceiver? Identifying deceptive intentions in communication,” *Games and Economic Behavior*, 2024, 145, 451–466.
- Charness, Gary and Martin Dufwenberg**, “Promises and partnership,” *Econometrica*, 2006, 74 (6), 1579–1601.
- and **Matthew Rabin**, “Understanding social preferences with simple tests,” *The quarterly journal of economics*, 2002, 117 (3), 817–869.
- Chen, Daniel L, Martin Schonger, and Chris Wickens**, “oTree—An open-source platform for laboratory, online, and field experiments,” *Journal of Behavioral and Experimental Finance*, 2016, 9, 88–97.
- Cho, Wonki Jo and Changwoo Park**, “The Local-global Equivalence on General Networks,” *Available at SSRN 4551159*, 2023.
- Crawford, Vincent P**, “Lying for strategic advantage: Rational and boundedly rational misrepresentation of intentions,” *American Economic Review*, 2003, 93 (1), 133–149.
- Crawford, Vincent P. and Joel Sobel**, “Strategic Information Transmission,” *Econometrica*, 1982, 50 (6), 1431–1451.
- Danz, David, Lise Vesterlund, and Alistair J Wilson**, “Belief elicitation and behavioral incentive compatibility,” *American Economic Review*, 2022, 112 (9), 2851–2883.

- Dufwenberg, Martin and Martin A Dufwenberg**, "Lies in disguise—A theoretical analysis of cheating," *Journal of Economic Theory*, 2018, 175, 248–264.
- Eilat, Ran and Zvika Neeman**, "Communication with endogenous deception costs," *Journal of Economic Theory*, 2023, 207, 105572.
- Elmshauser, Béla, Evan Friedman, and Yoo Joo Jo**, "Deception Aversion," *Working Paper*, 2025.
- Ely, Jeffrey C. and Juuso Välimäki**, "Bad Reputation*," *The Quarterly Journal of Economics*, 08 2003, 118 (3), 785–814.
- Erat, Sanjiv and Uri Gneezy**, "White Lies," *Management Science*, 2012, 58 (4), 723–733.
- Ettinger, David and Philippe Jehiel**, "A theory of deception," *American Economic Journal: Microeconomics*, 2010, 2 (1), 1–20.
- and ———, "An experiment on deception, reputation and trust," *Experimental Economics*, September 2021, 24 (3), 821–853.
- Eyster, Erik and Matthew Rabin**, "Cursed equilibrium," *Econometrica*, 2005, 73 (5), 1623–1672.
- Fehr, Ernst and Klaus M Schmidt**, "A theory of fairness, competition, and cooperation," *The quarterly journal of economics*, 1999, 114 (3), 817–868.
- Fischbacher, Urs and Franziska Föllmi-Heusi**, "Lies in Disguise—An Experimental Study on Cheating," *Journal of the European Economic Association*, 06 2013, 11 (3), 525–547.
- Geanakoplos, John, David Pearce, and Ennio Stacchetti**, "Psychological games and sequential rationality," *Games and economic Behavior*, 1989, 1 (1), 60–79.
- Gneezy, Uri**, "Deception: The Role of Consequences," *American Economic Review*, March 2005, 95 (1), 384–394.
- , **Agne Kajackaite, and Joel Sobel**, "Lying Aversion and the Size of the Lie," *American Economic Review*, February 2018, 108 (2), 419–53.
- Holmström, Bengt**, "Moral Hazard and Observability," *The Bell Journal of Economics*, 1979, 10 (1), 74–91.
- Hurkens, Sjaak and Navin Kartik**, "Would I lie to you? On social preferences and lying aversion," *Experimental Economics*, 2009, 12, 180–192.
- Innes, Robert**, "Does deception raise or lower lie aversion? Experimental evidence," *Journal of Economic Psychology*, 2022, 90, 102525.
- Jehiel, Philippe**, "Analogy-based expectation equilibrium," *Journal of Economic theory*, 2005, 123 (2), 81–104.
- Kant, Immanuel**, "On a supposed right to lie from philanthropy," *Practical philosophy*, 1797, 612.

- Kartik, Navin**, "Strategic Communication with Lying Costs," *The Review of Economic Studies*, 10 2009, 76 (4), 1359–1395.
- Khalmetski, Kiryl and Dirk Sliwka**, "Disguising lies—Image concerns and partial lying in cheating games," *American Economic Journal: Microeconomics*, 2019, 11 (4), 79–110.
- Kumar, Ujjwal, Souvik Roy, Arunava Sen, Sonal Yadav, and Huaxia Zeng**, "Local-global equivalence in voting models: A characterization and applications," *Theoretical Economics*, 2021, 16 (4), 1195–1220.
- MacQueen, J**, "Classification and analysis of multivariate observations," in "5th Berkeley Symp. Math. Statist. Probability" University of California Los Angeles LA USA 1967, pp. 281–297.
- Mahon, James Edwin**, "The definition of lying and deception," 2008.
- McKelvey, Richard D and Thomas R Palfrey**, "Quantal response equilibria for extensive form games," *Experimental economics*, 1998, 1, 9–41.
- Mora, Ernesto Rivera**, "Mechanism design with belief-dependent preferences," *Journal of Economic Theory*, 2024, 216, 105782.
- Morris, Stephen**, "Political Correctness," *Journal of Political Economy*, 2001, 109 (2), 231–265.
- Sánchez-Pagés, Santiago and Marc Vorsatz**, "An experimental study of truth-telling in a sender–receiver game," *Games and Economic Behavior*, 2007, 61 (1), 86–112.
- Sato, Shin**, "A sufficient condition for the equivalence of strategy-proofness and nonmanipulability by preferences adjacent to the sincere one," *Journal of Economic Theory*, 2013, 148 (1), 259–278.
- Sobel, Joel**, "A Theory of Credibility," *The Review of Economic Studies*, 1985, 52 (4), 557–573.
- , "Lying and Deception in Games," *Journal of Political Economy*, 2020, 128 (3), 907–947.
- Stevenson, Angus**, *Oxford dictionary of English*, Oxford University Press, 2010.
- Sutter, Matthias**, "Deception Through Telling the Truth?! Experimental Evidence From Individuals and Teams," *The Economic Journal*, 2009, 119 (534), 47–60.

Appendices

Appendix A. Equilibrium Analysis

In this section, we conduct the full equilibrium analysis of each environment. In both environments, the strategic type expert's strategy consists of two parts. The first part is her state-contingent message plan in period 1, denoted by $\sigma_1 : \Theta \rightarrow M$. $\sigma_1(m_1|\theta_1)$ specifies the probability that the expert sends message m_1 given θ_1 in period 1. The second part is her message plan contingent upon the outcome of the period 1 interaction as well as the state in period 2, denoted by $\sigma_2 : \Theta \times \Theta \times M \times A \rightarrow M$. $\sigma_2(m_2|\theta_2; \theta_1, m_1, a_1)$ specifies the probability that the expert sends message m_2 given θ_2 in period 2 after the history (θ_1, m_1, a_1) . The strategy of the public also consists of two parts. The first part is his message contingent action plan in period 1, denoted by $a_1 : M \rightarrow A$.²⁷ $a_1(m_1)$ specifies the action the public takes after receiving message m_1 in period 1. The second part is his action plan in period 2 that is contingent upon the outcome of the period 1 interaction as well as the message he receives in period 2, denoted by $a_2 : M \times M \times A \times \Theta \rightarrow A$. $a_2(m_2; \theta_1, m_1, a_1)$ specifies the action that the public takes upon receiving message m_2 in period 2 after the history (θ_1, m_1, a_1) .

A.1. Reputation Building with Bad-type Truth-telling (BT). Note that the trade-off between misguiding incentive and reputation-building incentive exists only for the expert of state 0 in period 1. The expert of state 1 can conceal her preference type and at the same time lead the public to match the state by sending $m_1 = 1$. It implies that the expert of state 1 sends message 1 with probability 1 in any informative equilibrium. We thus focus on the equilibrium in which the expert of state 1 sends message 1 with probability 1 (i.e., $\sigma_1(1|1) = 1$) while the expert of state 0 sends message 1 with probability v (i.e., $\sigma_1(1|0) = v$). Given the expert's strategy, the public's best response is to choose $a_1(0) = 0$ and $a_1(1) = \frac{1}{1+v/2}$ in period 1.

Let $\lambda(m_1, \theta_1) := \Pr(\tau = G|m_1, \theta_1)$ denote the public's belief that the expert is the good type at the beginning of period 2 given that m_1 and θ_1 are realized in period 1. The public's best response in period 2 depends on the belief. Given the expert's strategy in period 1, we have

$$\begin{aligned}\lambda(0,0) &= \frac{1/2}{1/2+(1/2)(1-v)} = \frac{1}{2-v} \\ \lambda(1,0) &= 0, \text{ and} \\ \lambda(1,1) &= \frac{1}{2}\end{aligned}$$

²⁷It is without loss of generality to consider the pure strategies of the public given the quadratic loss and the continuous action space.

by Bayes rule. $\lambda(0, 1)$ can be any number in $[0, 1]$ because it is off the equilibrium path. Given the arbitrary belief $\lambda(\cdot, 0) = \lambda$ at the beginning of period 2, the public chooses

$$a_2 = \frac{\frac{1}{2}}{\frac{1}{2} + \frac{(1-\lambda)}{2}} = \frac{1}{2-\lambda}$$

upon receiving $m_2 = 1$ while choosing $a_2 = 0$ upon receiving $m_2 = 0$.

Now, we can derive the expert's optimal choice of v . Given $\theta_1 = 0$, the expert's payoffs from sending $m_1 = 0$ and $m_1 = 1$ in period 1 are $-(0-1)^2$ and $-\left(\frac{1}{1+v/2} - 1\right)^2$, respectively. Furthermore, given that the only message sent in period 2 is $m_2 = 1$, the two messages in period 1 induce updated beliefs $\lambda(0, 0)$ and $\lambda(1, 0)$ which lead to the action $a_2 = \frac{1}{2-\lambda(0,0)} = \frac{2-v}{3-2v}$ and $a_2 = \frac{1}{2-\lambda(1,0)} = \frac{1}{2}$, respectively. Then the expected payoff of the expert from sending each message given $\theta_1 = 0$ becomes

$$\begin{aligned} EU_B(m_1 = 0 | \theta_1 = 0) &= -x_1(0-1)^2 - x_2\left(\frac{2-v}{3-2v} - 1\right)^2 \text{ and} \\ EU_B(m_1 = 1 | \theta_1 = 0) &= -x_1\left(\frac{1}{1+v/2} - 1\right)^2 - x_2\left(\frac{1}{2} - 1\right)^2 \end{aligned}$$

The two curves can cross at most once because $EU_B(m_1 = 1 | \theta_1 = 0)$ is strictly decreasing in v and $EU_B(m_1 = 0 | \theta_1 = 0)$ is strictly increasing in v . If they intersect in the domain of v , $v \in [0, 1]$ obtains from the indifference condition between the two expected payoffs. Otherwise, one of $EU_B(m_1 = 1 | \theta_1 = 0)$ and $EU_B(m_1 = 0 | \theta_1 = 0)$ dominates the other for all $v \in [0, 1]$ and thus the expert uses a pure strategy in the equilibrium.

Proposition 5. *In the unique equilibrium of the game, $\sigma_1(1|1) = 1$ and $\sigma_1(1|0) = v \in [0, 1]$ while $a_1(0) = 0$ and $a_1(1) = \frac{1}{1+v/2}$, where $v = 1$ if $x_2/x_1 \leq 32/9$, $v \in (0, 1)$ if $32/9 < x_2/x_1 < 36/5$, and $v = 0$ if $x_2/x_1 \geq 36/5$.*

Proof. See Appendix J. □

A.2. Reputation Building with Good-type Lying (GL). Note that the trade-off between type-revealing incentive and reputation-building incentive exists only for the expert of state 1 in period 1. The expert of state 0 can reveal her preference type and at the same time lead the public to match the state by sending $m_1 = 0$. It implies that the expert of state 0 sends message 0 with probability 1 in any informative equilibrium. We thus focus on the equilibrium in which the expert of state 0 sends message 0 with probability 1 (i.e. $\sigma_1(0|0) = 1$) while the expert of state 1 sends message 0 with probability w (i.e. $\sigma_1(0|1) = w$). Given the expert's strategy, the public's best response is to choose $a_1(0) = \frac{w}{1+w}$ and $a_1(1) = \frac{2-w}{(2-w)+1} = \frac{2-w}{3-w}$ in period 1.

As in the previous section, let $\lambda(m_1, \theta_1) := \Pr(\tau = G | m_1, \theta_1)$ denote the public's belief that the expert is the good type at the beginning of period 2 given that m_1 and θ_1 are realized in period 1. The public best response in period 2 depends on the belief. Given

the expert's strategy in period 1, we have

$$\begin{aligned}\lambda(0,0) &= 1, \\ \lambda(0,1) &= 1, \\ \lambda(1,0) &= 0, \text{ and} \\ \lambda(1,1) &= \frac{(1/2)(1-w)}{(1/2)(1-w)+1/2} = \frac{1-w}{2-w}\end{aligned}$$

by Bayes rule. Given the arbitrary belief $\lambda(\cdot, 1) = \lambda$ at the beginning of period 2, the public chooses

$$a_2 = \frac{\frac{1}{2}}{\frac{1}{2} + \frac{(1-\lambda)}{2}} = \frac{1}{2-\lambda}$$

upon receiving $m_2 = 1$ while choosing $a_2 = 0$ upon receiving $m_2 = 0$.

Now, we can derive the expert's optimal choice of w . Given $\theta_1 = 1$, the expert's payoffs from sending $m_1 = 0$ and $m_1 = 1$ in period 1 are $-\left(\frac{w}{1+w} - 1\right)^2$ and $-\left(\frac{2-w}{3-w} - 1\right)^2$, respectively. Furthermore, the two messages in period 1 induce updated beliefs $\lambda(0,1)$ and $\lambda(1,1)$. If $m_2 = 1$, this results in the action $a_2 = \frac{1}{2-\lambda(0,1)} = 1$ and $a_2 = \frac{1}{2-\lambda(1,1)} = \frac{2-w}{3-w}$, respectively. If $m_2 = 0$, then the public further updates his belief and realizes that the expert is the good type, leading to $a_2 = 0$. Then the expected payoff of the expert from sending each message given $\theta_1 = 1$ becomes

$$\begin{aligned}EU_G(m_1 = 0|\theta_1 = 1) &= -x_1\left(\frac{w}{1+w} - 1\right)^2 - x_2\left[\frac{1}{2}(0-0)^2 + \frac{1}{2}(1-1)^2\right] = -\frac{x_1}{(1+w)^2}, \text{ and} \\ EU_G(m_1 = 1|\theta_1 = 1) &= -x_1\left(\frac{2-w}{3-w} - 1\right)^2 - x_2\left[\frac{1}{2}(0-0)^2 + \frac{1}{2}\left(\frac{2-w}{3-w} - 1\right)^2\right] = -\frac{2x_1+x_2}{2(w-3)^2}.\end{aligned}$$

At $w = 1$ it is always the case that $EU_G(m_1 = 1|\theta_1 = 1) < EU_G(m_1 = 0|\theta_1 = 1)$. Given that $EU_G(m_1 = 1|\theta_1 = 1)$ is strictly decreasing in w and $EU_G(m_1 = 0|\theta_1 = 1)$ is strictly increasing in w , whether or not the two curves intersect in the domain of w depends on the ranking between the two expected payoffs at $w = 0$. If they don't intersect, then $EU_G(m_1 = 0|\theta_1 = 1)$ dominates $EU_G(m_1 = 1|\theta_1 = 1)$ for all $w \in [0, 1]$ such that the expert uses a pure strategy ($w = 1$) in equilibrium. If they intersect in the interior domain of w , the mixed-strategy equilibrium with $w \in (0, 1)$ obtains from the indifference condition between the two expected payoffs. In this case, in contrast to the BT environment, the incentive compatibility condition for $\theta_1 = 1$ is automatically satisfied at $w = 0$ and $w = 1$ such that there are two additional pure-strategy equilibria. This result is summarized in the following proposition.

Proposition 6. *In any informative equilibrium of the game, $\sigma_1(0|0) = 1$ and $\sigma_1(0|1) = w \in [0, 1]$ while $a_1(0) = \frac{w}{1+w}$ and $a_1(1) = \frac{2-w}{3-w}$. If $x_2/x_1 \geq 16$, equilibrium is unique and $w = 1$. If $x_2/x_1 < 16$, then there exist three equilibria, each with $w = 0$, $w = 1$, and $w \in (0, 1)$.*

Proof. See Appendix J. □

The proposition says that the equilibrium with $w = 1$ always exists. It is the unique informative equilibrium when the relative importance of period 2 is large enough. There are two additional informative equilibria with $w < 1$ when the relative importance of period 2 is small.

It is noteworthy that there is an uninformative equilibrium in which the strategic expert mimics the behavioral type such that the public completely ignores the expert's message. However, this equilibrium is supported by the implausible off-path belief: when receiving $m_1 = 0$, the public keeps his prior belief without realizing that the message must come from the good type expert.

Appendix B. Formal Definition of Deception in Reputation Building Environment

Formally, we define deception with respect to the preference type as the following to make our definition consistent with Sobel (2020).

Definition 4. (*Deception about the preference type*) Fix the state θ . The good type expert's message m is **deceptive with respect to the preference type** if there exists a message m' such that $\lambda(m', \theta) > 0$ and a number $p \in [0, 1)$ such that

$$\lambda(m, \theta) = p\lambda(m', \theta).$$

Similarly, the bad type expert's message m is **deceptive with respect to the preference type** if there exists a message m' such that $1 - \lambda(m', \theta) > 0$ and a number $p \in [0, 1)$ such that

$$1 - \lambda(m, \theta) = p(1 - \lambda(m', \theta)).$$

Consider the strategic (bad) type expert's message in state $\theta_1 = 0$ in the reputation-building equilibrium characterized in Proposition 5. According to Definition 1, $m_1 = 1$ is a lie while $m_1 = 0$ is not. Moreover, because $\lambda(1, 0) < \lambda(0, 0)$, there exists $p \in [0, 1)$ such that $1 - \lambda(0, 0) = p(1 - \lambda(1, 0))$. According to Definition 2, $m_1 = 0$ is deceptive while $m_1 = 1$ is not. That is, when $\theta_1 = 0$, $m_1 = 0$ is a deceptive truth in the BT environment.²⁸

Consider the strategic (good) type expert's message in the state $\theta_1 = 1$ in the reputation-building equilibrium characterized in Proposition 6. Definition 1 implies that $m_1 = 0$ is a lie while $m_1 = 1$ is not. Because $\lambda(0, 1) > \lambda(1, 1)$, there exists $p \in [0, 1)$ such that $\lambda(1, 1) = p\lambda(0, 1)$. Thus, $m_1 = 1$ is deceptive while $m_1 = 0$ is not. That is, when $\theta_1 = 1$, $m_1 = 0$ is a non-deceptive lie in the GL environment.

²⁸One caveat is that whether a message is deceptive or not in a given state sometimes depends on the specification of off-the-equilibrium path beliefs. This occurs in the reputation-building equilibrium of the BT environment in which the expert never sends $m_1 = 0$ conditional on $\theta_1 = 1$. We circumvent this issue in our experiment by fixing the expert's message to be $m_1 = 1$ conditional on $\theta_1 = 1$. As a result, when $\theta_1 = 1$, the expert has no decision to make, and thus whether a message is deceptive does not have any payoff consequences.

Let $Pr(\theta|m)$ denote the interim belief of the receiver which is formed after a message m is received but before the state θ is revealed. We now define deception with respect to the state using the interim belief.

Definition 5. (*Deception about the state*) Given the true state θ , an expert's message m is **deceptive with respect to the state** if there exists a message m' such that the interim belief $Pr(\theta|m') > 0$ satisfies

$$Pr(\theta|m) = p \times Pr(\theta|m').$$

with a number $p \in [0, 1)$.

Appendix C. Experiment I: Design

Payoff Detail. The exact payoff (unit: KRW) used in Experiment I is

$$\begin{aligned} \text{Receiver} &: 1000[1 - (a_1 - \theta_1)^2] + 20,000[1 - (a_2 - \theta_2)^2], \\ \text{Sender(BT)} &: 1000[1 - (a_1 - 1)^2] + 20,000[1 - (a_2 - 1)^2], \\ \text{Sender(GL)} &: 1000[1 - (a_1 - \theta_1)^2] + 20,000[1 - (a_2 - \theta_2)^2], \end{aligned}$$

Hypothesis. By choosing $x_2/x_1 = 20$, Proposition 5 and 6 imply that the strategic (bad) type expert tells the truth in period 1 in the BT environment and the strategic (good) type expert sends $m_1 = 0$ regardless of the state in the GL environment. That is, if individuals do not incur any intrinsic lying and/or deception costs in building reputation, then we expect that experts pursue reputation-building via her period 1 messages equally in the two environments. We thus have our first null hypothesis as follows:

Hypothesis 1. [*Expert's Messages in Period 1*] Expert sends $m_1 = 0$ with probability 1 conditional on the state $\theta_1 = 0$ in the Bad-type Pooling (BT) treatment. Expert sends $m_1 = 0$ with probability 1 conditional on $\theta_1 = 1$ in the Good-type Separating (GL) treatment.

Reputation building is essentially a costly act to influence the public's belief about the preference type of the expert to pursue the gain that comes in period 2. If reputation building is successful, then the influenced belief of the public should lead the public to take favorable action for a given message. In the BT environment, successful reputation building ensures that the public's post period 1 belief stays the same as the prior when $\theta_1 = 0$ and $m_1 = 0$, i.e., $\lambda(0,0) = 1/2$. As a result, the public partially separates his actions depending on the messages given by the expert in period 2. In the GL environment, successful reputation building allows the public to fully identify the (good) preference type of the strategic expert when $\theta_1 = 1$ and $m_1 = 0$, i.e., $\lambda(0,1) = 1$. As a result, the public fully separates his actions depending on the messages given by the expert in period 2. These predictions are summarized in Table 3 below. Our second hypothesis is

set to test if the partial and full separations of the public's actions in period 2 driven by reputation building is observed in the laboratory.

Hypothesis 2. *[Public's Actions in Period 2] (i) In the BT treatment, conditional on $(m_1, \theta_1) = (0, 0)$ and $(1, 1)$, the public's action induced by $m_2 = 0$ is substantially lower than that induced by $m_2 = 1$. (ii) In the GL treatment, conditional on $(m_1, \theta_1) = (0, 0)$ and $(0, 1)$, the public's action induced by $m_2 = 0$ is substantially lower than that induced by $m_2 = 1$. (iii) The degree of separation measured by the distance between the two actions in (i) is lower than that in (ii).*

m_1	θ_1	BT	GL
0	0	$m_2 = 0 \longrightarrow a_2 = 0$ $m_2 = 1 \longrightarrow a_2 = 2/3$	$m_2 = 0 \longrightarrow a_2 = 0$ $m_2 = 1 \longrightarrow a_2 = 1$
0	1	Off-path	$m_2 = 0 \longrightarrow a_2 = 0$ $m_2 = 1 \longrightarrow a_2 = 1$
1	0	Off-path	$m_2 = 0 \longrightarrow \text{Off-path}$ $m_2 = 1 \longrightarrow a_2 = 1/2$
1	1	$m_2 = 0 \longrightarrow a_2 = 0$ $m_2 = 1 \longrightarrow a_2 = 2/3$	$m_2 = 0 \longrightarrow \text{Off-path}$ $m_2 = 1 \longrightarrow a_2 = 1/2$

■ The public's actions contingent on period 2 messages that are directly affected by the expert's deliberate choices for reputation building are highlighted in boldface.

TABLE 3. Equilibrium Strategy of the Public in Period 2

Recall that, from the perspective of the expert, reputation building is a costly act in period 1 to influence the public's belief about her preference type to pursue the gain that comes in period 2. It thus entails an intertemporal tradeoff with respect to the degree of information being transmitted to the public in each period. In the BT treatment, the bad type expert sacrifices instantaneous gain by telling the truth in period 1 to deceive the public, thereby inducing the public to follow the expert's recommendation in period 2. Successful reputation building in the BT treatment results in a higher degree of information transmission (and thus a higher expected payoff for the public) in period 1 than in period 2. In contrast, in the GL treatment, the good type expert sacrifices spontaneous gain by lying in period 1 to reveal her preference type, thereby inducing the public to fully trust the expert in period 2. Successful reputation building in the GL treatment thus implies a lower degree of information transmission (and thus a lower expected payoff for the public) in period 1 than in period 2. We use π_i to denote the receiver's expected payoff in period i , and $\Delta\pi_{1,2} = \pi_1 - \pi_2$ to denote the payoff difference. Table 4 presents the intertemporal tradeoff of reputation building in each environment by reporting the expected payoff of the public in each period. If there were no reputation building, then

each player would behave as if each period is a one-shot game, resulting in $\Delta\pi_{1,2} = 0$. Our next hypothesis summarizes this result.

	BT	GL
π_1	0	$-1/4$
π_2	$-1/6$	$-1/8$
$\Delta\pi_{1,2}$	$1/6$	$-1/8$

TABLE 4. Intertemporal Tradeoff of Reputation Building

Hypothesis 3. *[Intertemporal Tradeoff of Reputation Building] $\Delta\pi_{1,2} > 0$ in the BT treatment while $\Delta\pi_{1,2} < 0$ in the GL treatment.*

Appendix D. Experiment I: Additional Results

In this section, we present additional results from Experiment I that complement Section 3.

D.1. Sender Strategies in the Last 3 Rounds. Figure 9 below presents the sender strategy only in the last 3 rounds, showing that our main result is robust to learning.

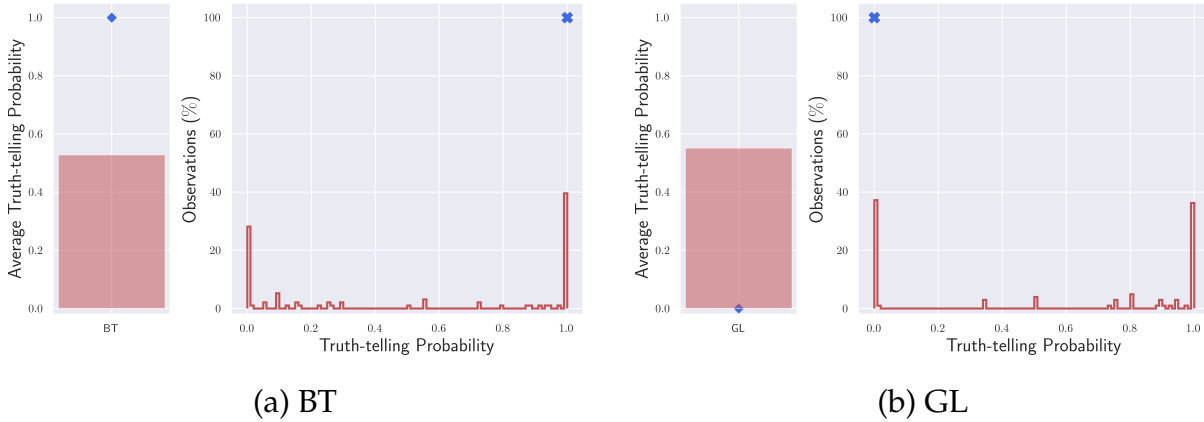


FIGURE 9. Sender Strategy (Stage 1, Last 3 Rounds)

Note: In each (a) and (b), the bar graph on the left panel shows the average truth-telling probability aggregated over all rounds and sessions for the first stage and the distribution on the right panel shows the distribution of the truth-telling probabilities by each individual in the last 3 rounds of each session. Blue squares and crosses show the theoretical predictions in each panel.

D.2. K-means Clustering of Sender Strategy: Additional Results. Figure 10 displays the (stages 1 and 2) joint empirical distribution of the sender strategy, with the sender's strategy in stage 1 on the horizontal axis and that in stage 2 on the vertical axis, as in Figure 3. The difference is that we used each observation as data points in Figure 3, whereas we used the average sender strategies across rounds for each sender as data points in Figure 10. Figure 10 yields the following observations: (i) In the BT treatment, we observe the reputation builders (blue square), the deception haters (red circle), and the always truth-teller (green triangle) as in the disaggregated clustering. The only distinction is the existence of individuals who employ strategies denoted by the pink star (25% of total players). This indicates that a quarter of individuals found it arduous to comprehend the game at the outset and then learned to play strategically as the game progressed. Indeed, when we plot the same figure using only the data from the last 5 rounds, we observe that the proportion of pink stars drops to 12%, while the proportions of reputation builders and deception haters increase. (ii) In the GL treatment, we observe the reputation builders (blue square), the lying haters (green triangle), and the irrational players (orange diamond) as in the disaggregated clustering. The only distinction is the existence of individuals who employ strategies denoted by the cyan hexagon (15% of total players). They experience a small degree of lying cost. If we restrict the data to the last 5 rounds, the proportion of cyan hexagons remains the same, while the proportion of reputation builders increases by 9% and the proportion of lying haters decreases by 9%. Thus, a small fraction of those who seemed to exhibit perfect or partial lying aversion actually had trouble determining the optimal strategy, while most deviators are still accounted for by intrinsic lying aversion.

D.3. Receiver Behavior and Low Compliance in GL. Figure 12(c) presents the average action taken in stage 2 conditional on each stage 1 history ($\theta_1 = 0/m_1 = 0$ and $\theta_1 = 1/m_1 = 1$ are on-the-path, but $\theta_1 = 0/m_1 = 1$ is off-the-path) and stage 2 message in the BT treatment. Figure 12(d) presents the average action taken in stage 2 conditional on each stage 1 history ($\theta_1 = 1/m_1 = 0$ and $\theta_1 = 0/m_1 = 0$ are on-the-path, and rest two histories are off-the-path) and stage 2 message in the GL treatment. The blue diamonds indicate the theoretical predictions. In both treatments, consistent with the theoretical predictions, the average receiver actions induced by $m_2 = 1$ in any on-the-path history are substantially higher than those induced by $m_2 = 0$ ($p = 0.015$, Mann-Whitney U Test). We thus accept both Hypotheses 2(i) and 2(ii). However, the differences in the induced actions are substantially smaller in the GL treatment (conditional on $\theta_1 = 1/m_1 = 0$) than that in the BT treatment ($p = 0.015$, Mann-Whitney U Test), allowing us to reject Hypothesis 2(iii).

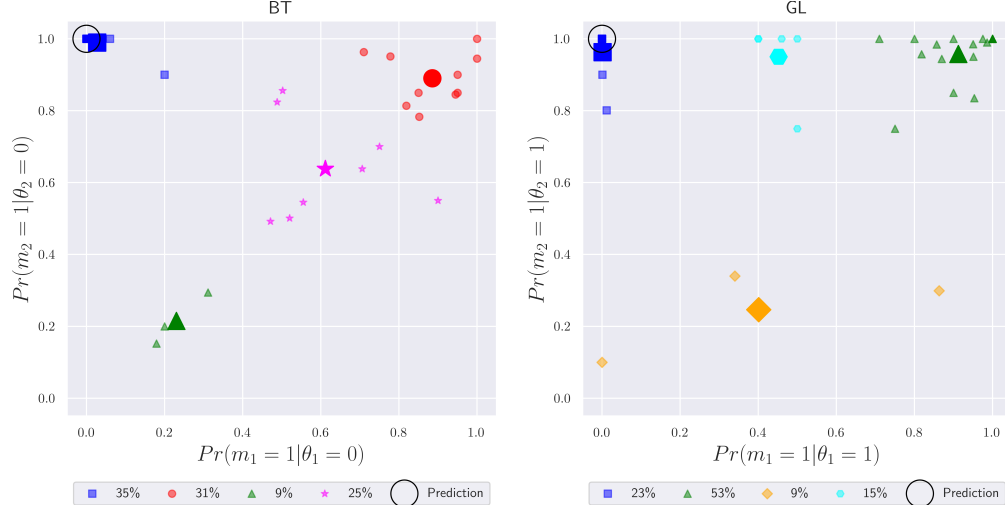


FIGURE 10. Clustering of Sender Strategy (Individual Average)

Note: Aggregated version of Figure 3. Each observation indicates the individual average across all 10 rounds. Each marker means the same cluster as in Figure 3, except for magenta stars and cyan hexagons in each treatment.

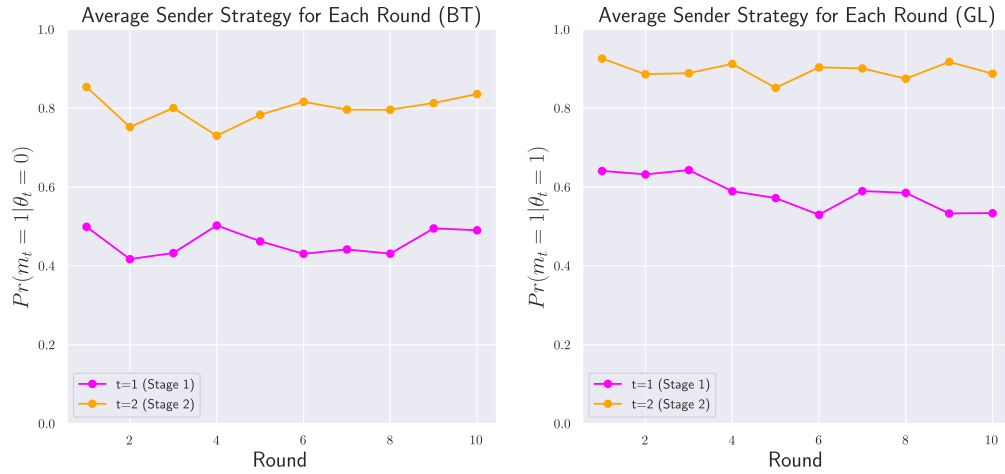


FIGURE 11. Sender Strategy - Time Trend

Note: Sender strategies averaged over all players at each round. This time trend shows that there is no noticeable trend across rounds in each stage of each treatment.

This discrepancy from the theory is largely observed in the GL treatment. In theory, the sender's message $m_1 = 0$ in each history perfectly reveals that the sender is of a good type and hence the receiver should comply with the recommendation from the sender. However, receivers in our experiment comply substantially less (85% vs. 65%) with the sender's recommendation in stage 2 when reputation building requires lying (i.e., when $\theta_1 = 1$ and $m_1 = 0$) compared to when reputation building is achieved via truth-telling (i.e., when $\theta_1 = 0$ and $m_1 = 0$). Figure 13 highlights that each individual do respond

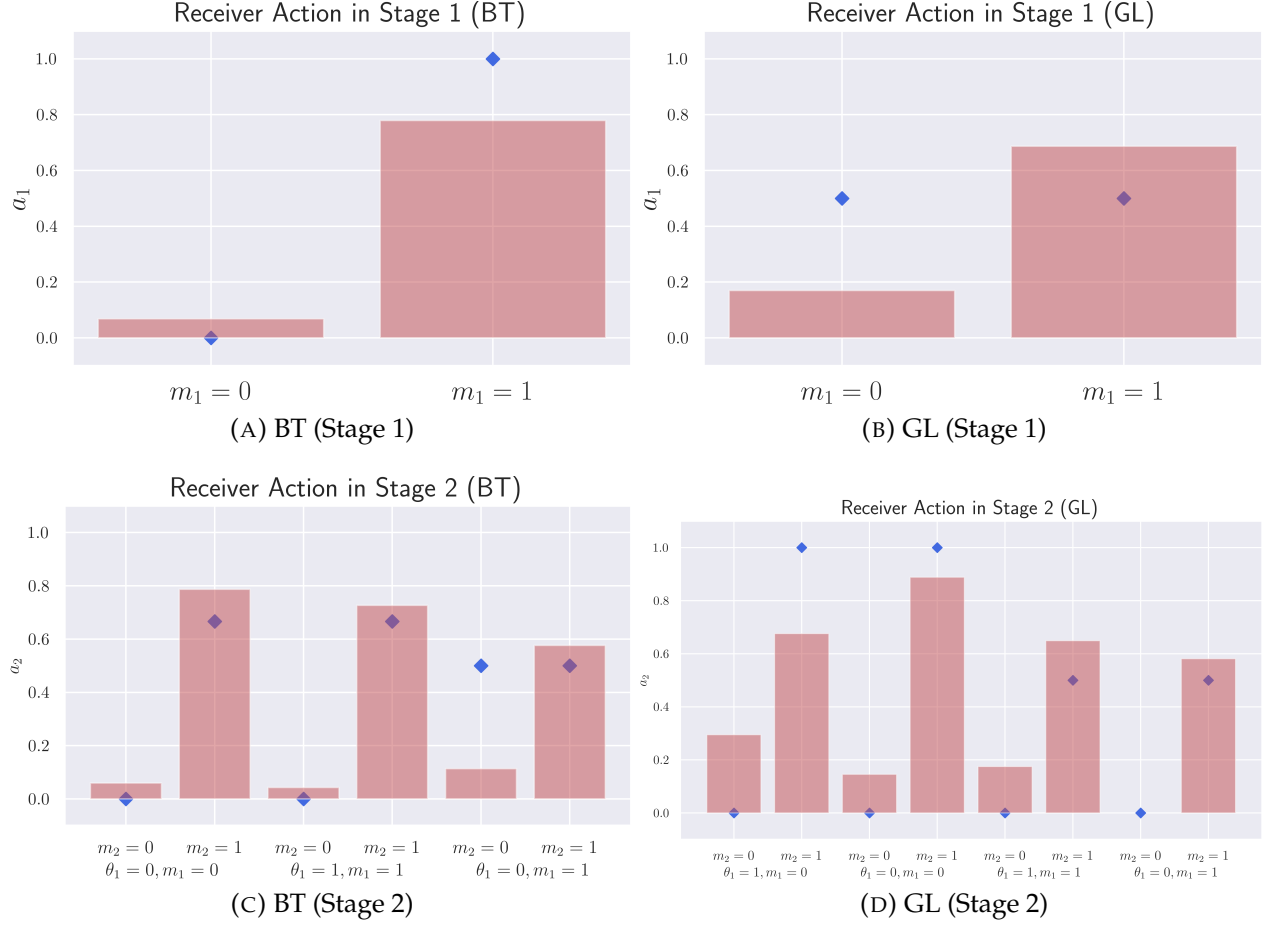


FIGURE 12. Receiver Strategy in Each Stage

Note: Averaged over all players and rounds. Blue diamonds indicate equilibrium predictions. Each panel considers all possible history.

to lying and truth-telling history differently. This figure illustrates the distribution of receivers' "compliance gap," which measures the difference between average compliance to the sender's message after lying history ($\theta_1 = 1, m_1 = 0$) and that after truth-telling history ($\theta_1 = 0, m_1 = 0$). Each receiver's average compliance after a given history is calculated by averaging the differences between the sender's message and the receiver's action in stage 2 of the rounds where a particular history occurred in stage 1. Thus, the compliance gap is well-defined only for those who experienced each history at least once. In the GL treatment, we can calculate the compliance gap for 37 out of 68 receivers. In Figure 13, most deviations from the theoretical prediction (blue cross) show positive compliance gaps, meaning a sizable portion of receivers complied with the sender's stage 2 less after the lying history than after the truth-telling history.

This observation can be rationalized in two ways: inference or preference. First, inferring that the sender is a good type from the lie may be harder than inferring the same

fact from the truth-telling. Second, preference over being told the truth and a lie may play a role in the receiver's decision-making: receivers may not like to receive lies. In this viewpoint, the receivers who observed a lie in stage 1 appear to punish the sender regardless of the intention behind the lie. That is, lying aversion occurs bilaterally while deception aversion occurs unilaterally. In Experiment II, we explicitly observe receivers' beliefs and confirm that low compliance should be attributed to the "punishing a lie" argument (Figure 25), rather than the inference error argument.

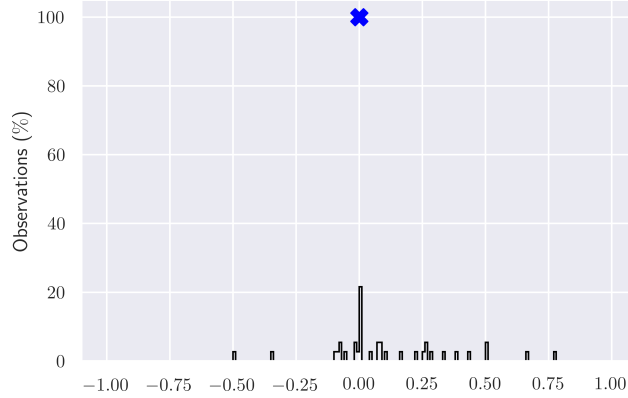


FIGURE 13. Distribution of Individual Receiver's Compliance Gap in GL

Note: The horizontal axis represents the compliance gap, which is the difference in a receiver's compliance to a sender's message after on-the-path lying history ($\theta_1 = 1, m_1 = 0$) and on-the-path truth-telling history ($\theta_1 = 0, m_1 = 0$). There should be no compliance gap according to the equilibrium prediction (blue cross). Among 37 receivers, almost 80 percent deviated from the prediction, and most of them showed positive compliance gaps, meaning that they complied less with the sender's message after the sender's previous lie, even though the sender lied to reveal that she is a good type.

D.4. Intertemporal Tradeoff of Reputation Building. Figure 14 depicts the relative degrees of information transmission between stage 1 and stage 2, where information transmission is measured by the receiver's empirical stage payoff. These results are in line with Hypothesis 3. In the BT treatment, the predicted positive $\Delta\pi_{1,2}$ is observed ($p = 0.034$, Wilcoxon signed-rank test), indicating that the bad type expert sacrifices immediate gain by being truthful in period 1, in order to deceive the public and lead them to follow the expert's recommendation in period 2. In the GL treatment, the predicted negative $\Delta\pi_{1,2}$ is observed ($p = 0.034$, Wilcoxon signed-rank test), indicating that the good type expert sacrifices immediate gain by lying in period 1 in order to reveal her preference type and gain the full trust of the public in period 2. However, the actual difference in information transmission between stages is smaller than the theoretical predictions in both treatments (0.051 vs 0.167 in BT, and -0.046 vs -0.125 in GL). This observation is consistent with our main finding reported in the previous section that only a third of senders successfully built their reputations.

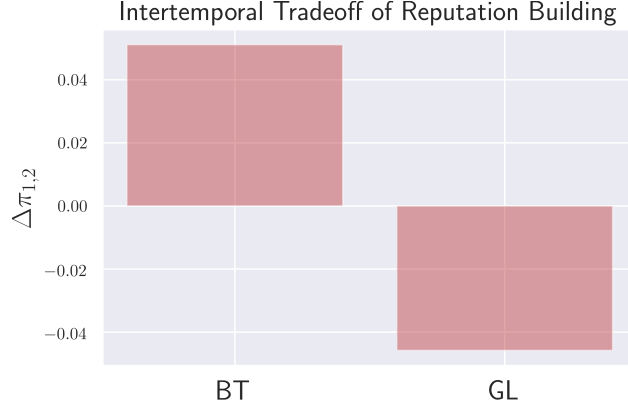


FIGURE 14. Intertemporal Tradeoff of Reputation Building

Note: The height of each bar represents the average difference between senders' payoffs in the first stage and those in the second stage. Positive difference means that senders get on average a higher payoff in the first stage than in the second stage, and vice versa.

Appendix E. Experiment I: Non-parametric Tests

Table 5 presents the results from the non-parametric tests regarding the sender and receiver strategies. The statistical test results confirm the conclusions presented in Section D.

Appendix F. Discussion on the Models of Deception Costs

Our model of deception cost has some benefits over alternative model specifications. There are two other natural specifications of the deception cost, but they turn out to be problematic for the following reasons.

Alternative specification 1. One possibility is to replace $\lambda(m^n, \theta)$ with λ_0 , which is defined as the expert's belief in her own preference type. Specifically, $\lambda_0 = 1$ if she is a good type and $\lambda_0 = 0$ if she is a bad type, so it is equivalent to the correct belief. If the expert sends a message that perfectly reveals her preference type, then $\lambda(m, \theta) = \lambda_0$ and she incurs no deception cost. However, one issue with this measure is that a non-deceptive message could have a strictly positive deception cost. For instance, in the BT environment, sending $m_1 = 1$ given $\theta_1 = 1$ incurs the deception cost of $c_d/2$, even though $m_1 = 1$ is non-deceptive.

Alternative specification 2. Another alternative possibility is to use the distance between the prior and the posterior as a measure of belief distortion. However, this definition can lead to unacceptable consequences in some cases. For instance, in the BT environment, sending $m_1 = 1$ given $\theta_1 = 0$ leads to a negative deception cost because $\lambda(m_1 = 1 | \theta_1 = 0) - 1/2 = -1/2$.

One caveat is that any definition of deception cost depends on beliefs, so deception cost on the off-the-equilibrium-path would differ by the equilibrium refinement. However,

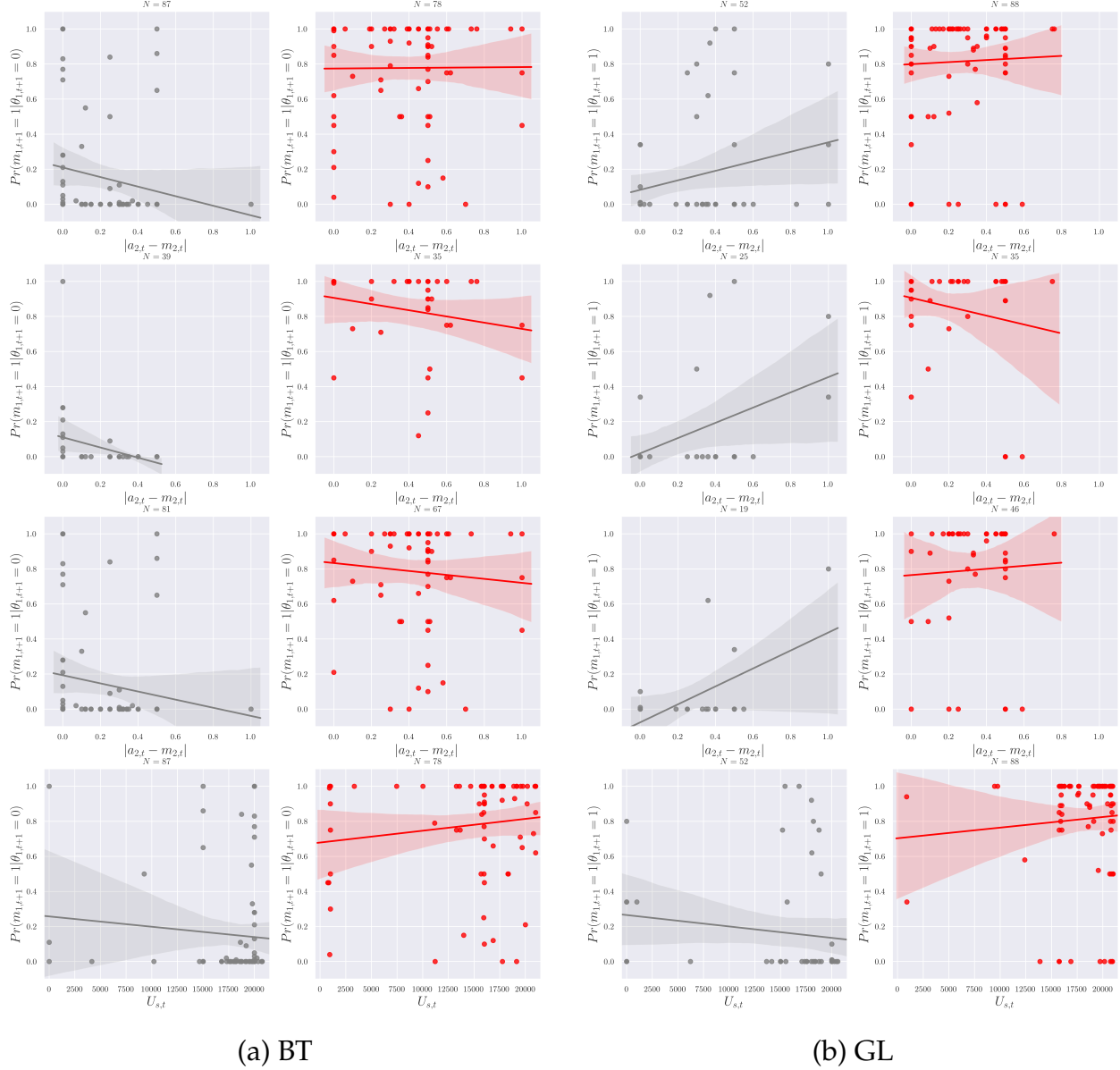


FIGURE 15. Sender Response to Previous Round

Note: In the BT treatment (Panel (a)), graphs in first column (gray) are when $\theta_1 = 0, m_1 = 0$ and those in second column (red) are when $\theta_1 = 0, m_1 = 1$. In the GL treatment (Panel (b)), graphs in first column (gray) are when $\theta_1 = 1, m_1 = 0$ and those in second column (red) are when $\theta_1 = 1, m_1 = 1$. In each panel, first three rows represent Senders' Stage 1 strategy in Round $t + 1$ in response to Receivers' compliance (defined as the absolute difference between sender's message and receiver's action in Stage 2) in Round t . First row uses data from whole rounds, second row uses data from last 4 rounds, and third row uses data from cases where Sender's Stage 2 message in Round t was 1, which is the only circumstance that Sender reaps the benefit of reputation. Lastly, fourth row represents Senders' Stage 1 strategy in Round $t + 1$ in response to her round payoff in Round t . Mostly flat regression lines in these figures imply that most of senders did not change their strategies in response to the receiver's response in the previous period. Therefore, most of sender strategies cannot be attributed to the empirical best response to receivers' strategies.

Who?	Test	Two-sided?	Null Hypothesis	p-values
Sender	MWU	One-sided (<)	Sender strategy is the same across BT and GL in stage 1.	0.235
	MWU	Yes	Sender strategy is the same across BT and GL in stage 1.	0.470
	MWU	One-sided (<)	Sender strategy is the same across BT and GL in stage 2.	0.097
	MWU	Yes	Sender strategy is the same across BT and GL in stage 2.	0.194
	Wilc	One-sided (>)	Sender strategy is zero in stage 1 of BT.	0.034
	Wilc	One-sided (>)	Sender strategy is zero in stage 1 of GL.	0.034
Receiver	MWU	Yes	Receiver strategy given $m_2 = 0$ in BT is the same conditional on $\theta_1 = 0, m_1 = 0$ and conditional on $\theta_1 = 1, m_1 = 1$.	0.665
	MWU	Yes	Receiver strategy given $m_2 = 1$ in BT is the same conditional on $\theta_1 = 0, m_1 = 0$ and conditional on $\theta_1 = 1, m_1 = 1$.	0.312
	MWU	Yes	Receiver strategy given $m_2 = 0$ in GL is the same conditional on $\theta_1 = 1, m_1 = 0$ and conditional on $\theta_1 = 0, m_1 = 0$.	0.061
	MWU	Yes	Receiver strategy given $m_2 = 1$ in GL is the same conditional on $\theta_1 = 1, m_1 = 0$ and conditional on $\theta_1 = 0, m_1 = 0$.	0.030
	MWU	One-sided (<)	Receiver strategy conditional on $\theta_1 = 0, m_1 = 0$ in BT is the same given $m_2 = 0$ and given $m_2 = 1$.	0.015
	MWU	One-sided (<)	Receiver strategy conditional on $\theta_1 = 1, m_1 = 1$ in BT is the same given $m_2 = 0$ and given $m_2 = 1$.	0.015
	MWU	One-sided (<)	Receiver strategy conditional on $\theta_1 = 1, m_1 = 0$ in GL is the same given $m_2 = 0$ and given $m_2 = 1$.	0.015
	MWU	One-sided (<)	Receiver strategy conditional on $\theta_1 = 0, m_1 = 0$ in GL is the same given $m_2 = 0$ and given $m_2 = 1$.	0.015
	MWU	Yes	Receiver strategy given $m_2 = 0$ is the same conditional on $\theta_1 = 0, m_1 = 0$ in BT and conditional on $\theta_1 = 1, m_1 = 0$ in GL.	0.030
	MWU	Yes	Receiver strategy given $m_2 = 1$ is the same conditional on $\theta_1 = 0, m_1 = 0$ in BT and conditional on $\theta_1 = 1, m_1 = 0$ in GL.	0.470
	MWU	Yes	Receiver strategy given $m_2 = 0$ is the same conditional on $\theta_1 = 0, m_1 = 0$ in BT and conditional on $\theta_1 = 0, m_1 = 0$ in GL.	0.030
	MWU	Yes	Receiver strategy given $m_2 = 1$ is the same conditional on $\theta_1 = 0, m_1 = 0$ in BT and conditional on $\theta_1 = 0, m_1 = 0$ in GL.	0.030
	MWU	Yes	Differences in the induced actions are the same in BT ($\theta_1 = 0, m_1 = 0$) and BT ($\theta_1 = 1, m_1 = 1$).	0.194
	MWU	One-sided (<)	Differences in the induced actions are the same in GL ($\theta_1 = 1, m_1 = 0$) and GL ($\theta_1 = 0, m_1 = 0$).	0.015
	MWU	One-sided (>)	Differences in the induced actions are the same in BT ($\theta_1 = 0, m_1 = 0$) and GL ($\theta_1 = 1, m_1 = 0$).	0.015
	MWU	One-sided (>)	Differences in the induced actions are the same in BT ($\theta_1 = 1, m_1 = 1$) and GL ($\theta_1 = 1, m_1 = 0$).	0.015
	MWU	Yes	Differences in the induced actions are the same in BT ($\theta_1 = 0, m_1 = 0$) and GL ($\theta_1 = 0, m_1 = 0$).	0.885
	MWU	One-sided (<)	Differences in the induced actions are the same in BT ($\theta_1 = 1, m_1 = 1$) and GL ($\theta_1 = 0, m_1 = 0$).	0.015
Welfare	Wilc	One-sided (>)	$\Delta\tau_{1,2} = 0$ in BT.	0.034
	Wilc	One-sided (<)	$\Delta\tau_{1,2} = 0$ in GL.	0.034

■ MWU and Wilc refer to the Mann-Whitney U (rank-sum) test and one-sample Wilcoxon (signed rank) test, respectively.

■ Sender strategy in this table denotes the probability to send message $m = 1$ given the state θ (BT: $\theta = 0$, GL: $\theta = 1$).

■ Receiver strategy in this table only considers the stage 2 action.

■ One-sided test is implemented only when the direction of the alternative hypothesis is clear.

TABLE 5. Non-parametric Tests Results

this is not a concern in our setting. To see this, recall that in the BT environment, the deception cost incurred in the off-equilibrium message $\lambda(0, 1)$ can be chosen to have any value in the range of $[0, 1]$. However, this is not a concern in our game as the equilibrium message $m_1 = 1$ is not deceptive given $\theta_1 = 1$, and therefore there is no deviation to $m_1 = 0$ regardless of the value of $\lambda(0, 1)$.

Appendix G. Equilibrium Characterization with Lying and Deception Costs

We now turn to the equilibrium analysis. In period 2, the deception cost does not affect the expert's strategy because the public's belief about the expert's preference type after receiving m_2 plays no role in our game. In contrast, the lying cost can alter the expert's strategy in period 2 depending on the environment. In the GL environment, the strategic expert tells the truth in period 2 and therefore experiences no lying cost. However, in

the BT environment, the strategic expert always sends $m_2 = 1$. Thus, she experiences the lying cost in the state $\theta_2 = 0$. If the lying cost is small, she does not change her strategy. If it is large enough, she modifies her strategy. The following proposition summarizes this argument.

Proposition 7. *Consider period 2 of the BT environment. Suppose that $\theta_2 = 0$. If $c_l \leq 3/4$, then the (strategic) expert sends a message $m_2 = 1$ with the probability 1. If $c_l > 3/4$, then she sends $m_2 = 1$ with the probability (weakly) less than 1.*

Proof. See Appendix J. □

Proposition 7 implies that we could empirically identify the individuals with their intrinsic lying cost $c_l > 3/4$ by observing their period 2 strategy in the BT environment. From now on, we focus on the case of $c_l \leq 3/4$ and turn to the first-period behavior of the (strategic) expert.

G.1. Lying and Deception in the BT Environment. Consider period 1 of the BT environment. When $\theta_1 = 1$, the expert incurs no lying and deception costs by sending a message $m_1 = 1$ as in the equilibrium of Proposition 5. When $\theta_1 = 0$, the expected utility of the expert from sending each message becomes

$$\begin{aligned} EU_B^a(m_1 = 0|\theta_1 = 0) &= -x_1 - x_2\left(\frac{v-1}{3-2v}\right)^2 - c_d(\lambda(0,0) - \lambda(1,0)) = -\frac{1}{20} - \left(\frac{v-1}{3-2v}\right)^2 - c_d\frac{1}{2-v}; \\ EU_B^a(m_1 = 1|\theta_1 = 0) &= -x_1\left(\frac{v}{v+2}\right)^2 - x_2\frac{1}{4} - c_l = -\frac{1}{20}\left(\frac{v}{v+2}\right)^2 - \frac{1}{4} - c_l. \end{aligned}$$

As in Proposition 5, $EU_B^a(m_1 = 1|\theta_1 = 0)$ is strictly decreasing in v . However, whether $EU_B^a(m_1 = 0|\theta_1 = 0)$ is strictly increasing in v or not depends on the size of c_d . This opens the possibility of multiple equilibria.

Proposition 8. *Under Assumptions 1 and 2, in any equilibrium of the game, $\sigma_1(1|1) = 1$, $\sigma_1(1|0) = v \in [0,1]$, $a_1(0) = 0$, and $a_1(1) = \frac{1}{1+v/2}$. There exists two curves $c_l = \bar{c}(c_d)$ and $c_l = \underline{c}(c_d)$ over the (c_d, c_l) -space such that $0 \leq \underline{c}(c_d) < \bar{c}(c_d) \leq 3/4$ for all c_d and the following holds:*

- i. *if $c_l > \bar{c}(c_d)$, there exists the unique equilibrium with $v = 0$;*
- ii. *if $\underline{c}(c_d) < c_l < \bar{c}(c_d)$, there exists an equilibrium with $v \in (0,1)$, and multiple equilibria with $v \in [0,1]$ may arise;*
- iii. *if $c_l < \underline{c}(c_d)$, there exists the unique equilibrium with $v = 1$.*

Proof. See Appendix J. □

Proposition 8 states that the relative size between c_l and c_d decides the equilibrium strategy. Remember that the equilibrium strategy in Proposition 5 requires deceptive truth-telling. When c_l is large compared to c_d , deviation from the deceptive truth-telling is costly to the expert. Thus, the equilibrium strategy remains the same. When c_l is small

compared to c_d , deviating from deceptive truth-telling becomes profitable for the expert. She begins to reduce the degree of deception by selecting $v > 0$, despite the lying cost it incurs.

G.2. Lying and Deception in the GL Environment. Consider period 1 of the GL environment. When $\theta_1 = 0$, the expert incurs no lying and deception costs by sending a message $m_1 = 0$ as in the equilibrium of Proposition 6. When $\theta_1 = 1$, the expected utility of the expert from sending each message becomes

$$\begin{aligned} EU_G^a(m_1 = 0|\theta_1 = 1) &= -\frac{x_1}{(1+w)^2} - c_l = -\frac{1}{20} \frac{1}{(1+w)^2} - c_l; \\ EU_G^a(m_1 = 1|\theta_1 = 1) &= -\frac{2x_1+x_2}{2(w-3)^2} - c_d(\lambda(0,1) - \lambda(1,1)) = -\frac{11}{20} \left(\frac{1}{3-w}\right)^2 - c_d \frac{1}{2-w}. \end{aligned}$$

As in Proposition 6, $EU_G^a(m_1 = 1|\theta_1 = 1)$ is strictly decreasing in w and $EU_G^a(m_1 = 0|\theta_1 = 1)$ is strictly increasing in w .

Proposition 9. *Under Assumptions 1 and 2, in any equilibrium of the game, $\sigma_1(0|0) = 1$, $\sigma_1(0|1) = w \in [0, 1]$, $a_1(0) = \frac{w}{1+w}$, and $a_1(1) = \frac{2-w}{3-w}$. Moreover,*

- a. *in the unique equilibrium, $w = 0$ if $c_l > c_d + \frac{1}{8}$;*
- b. *there are three equilibria, each with $w = 0$, $w \in (0, 1)$, and $w = 1$ if $\frac{c_d}{2} + \frac{1}{90} < c_l < c_d + \frac{1}{8}$;*
- c. *in the unique equilibrium, $w = 1$ if $c_l < \frac{c_d}{2} + \frac{1}{90}$.*

Proof. See Appendix J. □

According to Proposition 9, the relative size between c_l and c_d decides the equilibrium strategy. Remember that the equilibrium strategy in Proposition 6 requires non-deceptive lying. When c_d is large compared to c_l , deviation from the non-deceptive lying is costly to the expert. Thus, the equilibrium strategy remains the same. When c_d is small compared to c_l , deviation from the non-deceptive lying becomes profitable to the expert. She begins to reduce the degree of lying by selecting $w < 1$, despite the deception cost it incurs.

Figure 16 illustrates the findings presented in Propositions 7, 8, and 9, with the deception cost c_d on the horizontal axis and the lying cost c_l on the vertical axis. The light green region represents the equilibrium in which senders tell the truth even in the second stage of the BT environment. The blue region represents the equilibrium in which senders employ the reputation-building strategy, as in the original equilibrium in Section 2. The red region represents the equilibrium in which senders maximally deviate from the reputation-building strategy only in the first stage due to deception aversion. Similarly, the purple region represents the equilibrium in which senders partially deviate from the reputation-building strategy by employing a mixed strategy, due to deception aversion. The green region represents the equilibrium in which senders maximally deviate from the reputation-building strategy only in the first stage due to lying aversion. Similarly,

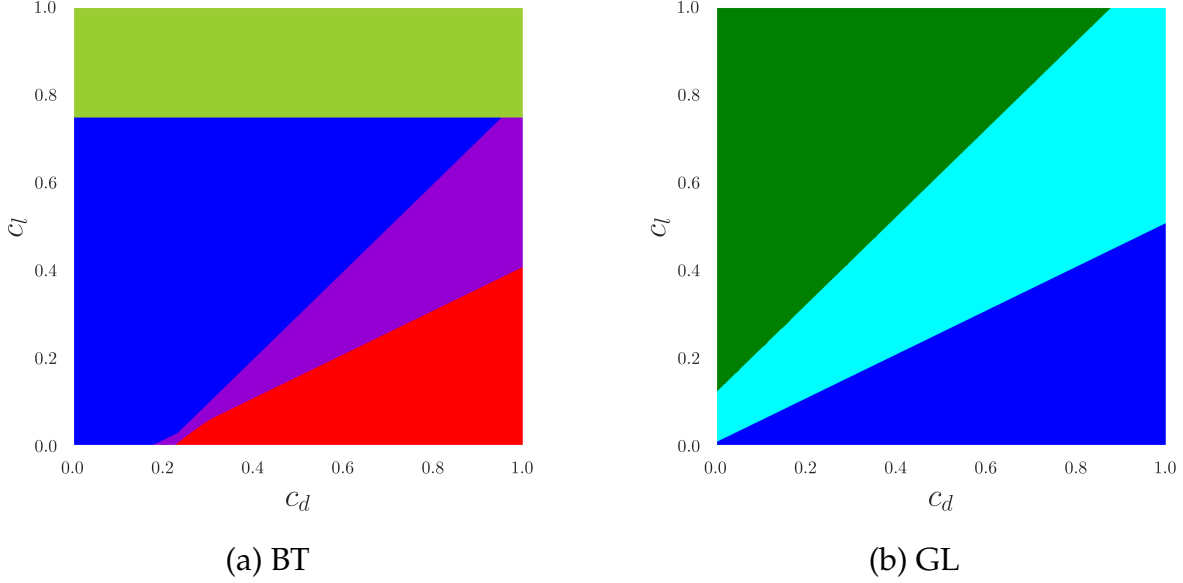


FIGURE 16. Equilibrium under Lying and Deception Costs

Note: The horizontal axis is the size of the deception cost. The vertical axis is the size of the lying cost. The maximal size of each cost is normalized to 1. The equilibrium strategy of a sender remains the same within the region of cost parameters that is described by the same color.

the cyan region represents the equilibrium in which senders partially deviate from the reputation-building strategy by employing a mixed strategy, due to lying aversion.

The equilibrium characterization with lying and deception costs provides one reasonable explanation of why the sizable portion of senders deviate from the equilibrium predicted in Section 3. The color of each area in Figure 16 and that in Figure 3 mean the same type of equilibrium.²⁹ That is, for each sender strategy cluster in Figure 3, we can find the range of relative ratio between lying cost c_l and deception cost c_d .

Appendix H. Partial Identification of the Distribution of Lying and Deception Costs

In this section, we investigate how we can identify the empirical distributions of lying cost c_l and deception cost c_d among our experiment participants.³⁰ To achieve this, we assume that there is a joint distribution over the (c_l, c_d) -space, and each participant's cost vector (c_l, c_d) is drawn randomly from this distribution. Overlapping the parameter space

²⁹Figure 3 does not illustrate mixed-strategy equilibria separately from the pure-strategy equilibria, while Figure 16 does. Thus, we denote a mixed-strategy equilibrium region in Figure 16 by the color mixing of two adjacent pure-strategy equilibrium regions. Also, light green ($c_l > 3/4$ in BT) region includes both “always truth-tellers (green in Figure 3)” and “noise (orange in Figure 3).”

³⁰To our knowledge, we are the first who provides a systematic answer to the question raised by Sobel (2020, pp.943-944) “It is an empirical question to describe these costs. There is a smaller experimental literature on deception, but again my model provides a way to include costs of deception in a strategic model.”

of each treatment in Figure 16 generates eight distinct regions, as shown in Figure 17. We denote the probability of the cost vector being realized in region k as $\tilde{\zeta}_k$.

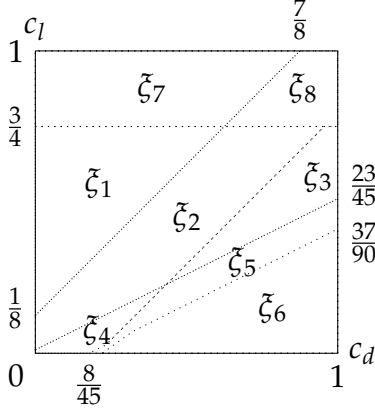


FIGURE 17. Eight regions in the (c_l, c_d) -space

Region	BT	GL
1	$\theta_1 = 0 \rightarrow m_1 = 0$	$\theta_1 = 1 \rightarrow m_1 = 1$
2	$\theta_1 = 0 \rightarrow m_1 = 0$	$\theta_1 = 1 \rightarrow m_1 \in [0, 1]$
3	$\theta_1 = 0 \rightarrow m_1 \in [0, 1]$	$\theta_1 = 1 \rightarrow m_1 \in [0, 1]$
4	$\theta_1 = 0 \rightarrow m_1 = 0$	$\theta_1 = 1 \rightarrow m_1 = 0$
5	$\theta_1 = 0 \rightarrow m_1 \in [0, 1]$	$\theta_1 = 1 \rightarrow m_1 = 0$
6	$\theta_1 = 0 \rightarrow m_1 = 1$	$\theta_1 = 1 \rightarrow m_1 = 0$
7	$\theta_2 = 0 \rightarrow m_2 \in [0, 1]$	$\theta_1 = 1 \rightarrow m_1 = 1$
8	$\theta_2 = 0 \rightarrow m_2 \in [0, 1]$	$\theta_1 = 1 \rightarrow m_1 \in [0, 1]$

FIGURE 18. Equilibrium strategy in each region

In region 1, the sender selects the reputation-building strategy in the BT treatment and chooses the maximally lying-averse strategy in the GL treatment. In region 2, she selects the reputation-building strategy in the BT treatment and chooses the partially lying-averse strategy in the GL treatment. In region 3, she selects the partially deception-averse strategy in the BT treatment and chooses the partially lying-averse strategy in the GL treatment. In region 4, she selects the reputation-building strategy in both treatments. In region 5, she selects the partially deception-averse strategy in the BT treatment and chooses the reputation-building strategy in the GL treatment. In region 6, she selects the maximally deception-averse strategy in the BT treatment and chooses the reputation-building strategy in the GL treatment. In region 7, she deviates from always selecting $m_2 = 1$ in stage 2 of the BT treatment and chooses the maximally lying-averse strategy in the GL treatment. Finally, in region 8, she deviates from always selecting $m_2 = 1$ in stage 2 of the BT treatment and chooses the partially lying-averse strategy in the GL treatment. These are summarized in Figure 18.

We aim to identify $(\hat{\zeta}_1, \hat{\zeta}_2, \dots, \hat{\zeta}_8)$ that match the empirical moments observed from the observed sender strategies, without imposing any parametric assumptions regarding the distribution over the (c_l, c_d) -space. However, because the number of parameters is greater than the number of moments, the model is under-identified. Therefore, we identify the set of parameters $(\tilde{\zeta}_1, \tilde{\zeta}_2, \dots, \tilde{\zeta}_8)$ that are consistent with our experimental data. Although $(\tilde{\zeta}_1, \tilde{\zeta}_2, \dots, \tilde{\zeta}_8)$ are not point-estimated, this exercise still provides meaningful ranges of

estimates for the distribution of lying and deception costs.³¹ The moment conditions derived from the empirical sender strategies are as follows.

$$\begin{aligned}
\hat{\xi}_1 + \hat{\xi}_2 + \hat{\xi}_4 &= 0.325, \\
\hat{\xi}_3 + \hat{\xi}_5 &= 0.053, \\
\hat{\xi}_4 + \hat{\xi}_5 + \hat{\xi}_6 &= 0.398, \\
\hat{\xi}_6 &= 0.191, \\
\hat{\xi}_1 + \hat{\xi}_7 &= 0.482, \\
\hat{\xi}_2 + \hat{\xi}_3 + \hat{\xi}_8 &= 0.120, \\
\hat{\xi}_7 + \hat{\xi}_8 &= 0.431, \\
\sum_{k=1}^8 \hat{\xi}_k &= 1.
\end{aligned}$$

Since the last three equations are redundant, the first five equations are enough to capture all restrictions imposed on the estimators.

Several features of the moment estimators are noteworthy.

Observation. The moment estimators $(\hat{\xi}_1, \hat{\xi}_2, \dots, \hat{\xi}_8)$, which characterizes the distribution of c_l and c_d , satisfy the following:

- The proportion of those who are averse only to deception $\hat{\xi}_6 = 0.19$.
- The proportion of those who are not averse to either lying or deception $\hat{\xi}_4 \in [0.15, 0.21]$.
- For each fixed $\hat{\xi}_4$, $(\hat{\xi}_3, \hat{\xi}_5)$ are determined uniquely. As $\hat{\xi}_4$ increases, $\hat{\xi}_3$ increases and $\hat{\xi}_5$ decreases.
- For each fixed $(\hat{\xi}_3, \hat{\xi}_4, \hat{\xi}_5)$, there are multiple solutions for $(\hat{\xi}_1, \hat{\xi}_2, \hat{\xi}_7, \hat{\xi}_8)$. As $\hat{\xi}_1$ increases, $\hat{\xi}_2$ and $\hat{\xi}_7$ decrease while $\hat{\xi}_8$ increases.

The above observation implies that the shape of the distribution over the (c_l, c_d) -space is jointly determined by $\hat{\xi}_1$ and $\hat{\xi}_4$. This guides us to fully describe the range of the distributions compatible with our experimental data. Figure 19 depicts each distribution in the four extreme cases. Panel (a) illustrates the distribution with maximal $\hat{\xi}_4$ and maximal $\hat{\xi}_1$. Panel (b) illustrates the distribution with maximal $\hat{\xi}_4$ and minimal $\hat{\xi}_1$. Panel (c) illustrates the distribution with minimal $\hat{\xi}_4$ and maximal $\hat{\xi}_1$. Panel (d) illustrates the distribution with minimal $\hat{\xi}_4$ and minimal $\hat{\xi}_1$. Higher $\hat{\xi}_4$ implies the higher proportion of senders who build a reputation in both treatments. Higher $\hat{\xi}_1$ implies the higher proportion of senders who build a reputation only in BT treatment and shows maximal lying aversion in GL treatment. All other distributions that are compatible with our experimental data

³¹About 26.7% of the senders in the GL treatment deviate from the equilibrium prediction in stage 2. This deviation cannot be explained by any lying and deception costs. We drop these observations in the identification analysis. Including these observations does not change the qualitative feature of the identified distribution.

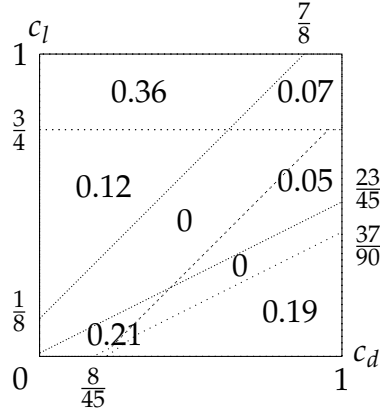
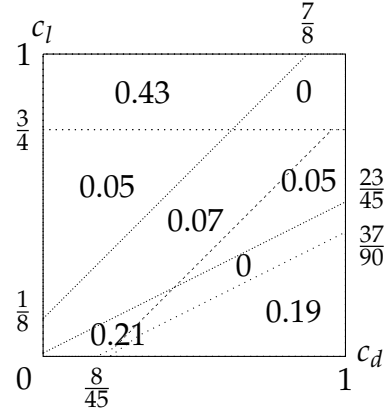
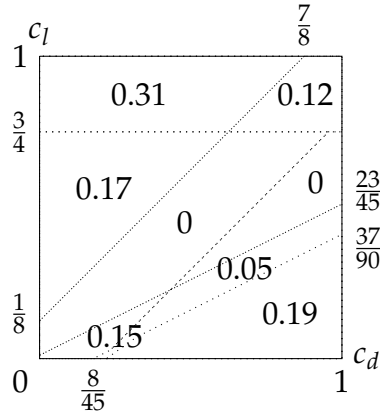
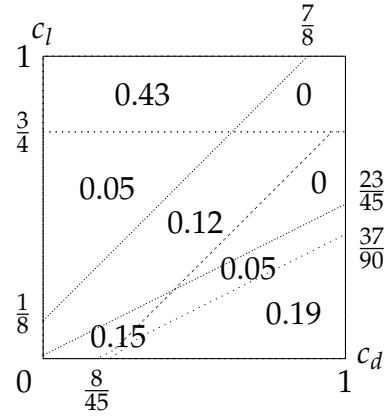
(a) Maximal ξ_4 , Maximal ξ_1 (b) Maximal ξ_4 , Minimal ξ_1 (c) Minimal ξ_4 , Maximal ξ_1 (d) Minimal ξ_4 , Minimal ξ_1

FIGURE 19. Partially Identified Distributions

are convex combinations of some of these four distributions. The distributions reveal that those who have high c_l and high c_d constitute the largest portion, followed by those with low c_l and high c_d and those with low c_l and low c_d . Other areas where senders play mixed strategies in at least one of the two treatments constitute a relatively small portion in most cases. Quantitatively, we identify that $\hat{\xi}_1 \in [0.05, 0.12]$, $\hat{\xi}_2 \in [0, 0.12]$, $\hat{\xi}_3 \in [0, 0.05]$, $\hat{\xi}_5 \in [0, 0.05]$, $\hat{\xi}_7 \in [0.31, 0.43]$, and $\hat{\xi}_8 \in [0, 0.12]$.

Appendix I. Alternative Explanations

Level-k. Following Crawford (2003), we assume that L0 sender is naive. L0 receiver is best responding to L0 sender. L1 sender is best responding to L0 receiver. L-k receiver is best responding to L-k sender. L-k sender is best responding to L-(k-1) receiver. The most plausible assumption for the naive (L0) sender in our environment is that the bad type always sends the high message, while the good type always tells the truth. If we accept this assumption, then Receiver's best response is to take action 0 upon receiving

the message “0” and action 2/3 upon the message “1” in both environments. Then, the higher-level strategic sender’s best response is always the same as that of L0. That is, the unique prediction for all levels is the reputation-building failure. In conclusion, this level-k model (which is natural and comparable to the conventional model in the literature) cannot explain our data at all. Table 6 presents level-k analysis in our Experiment I environment.

BT Environment	Strategic Sender		Behavioral Sender		Receiver	
State/Message	0	1	0	1	“0”	“1”
Level 0	“1”	“1”	“0”	“1”	0	2/3
Level 1 or higher	“1”	“1”	“0”	“1”	0	2/3

GL Environment	Strategic Sender		Behavioral Sender		Receiver	
State/Message	0	1	0	1	“0”	“1”
Level 0	“0”	“1”	“1”	“1”	0	2/3
Level 1 or higher	“0”	“1”	“1”	“1”	0	2/3

TABLE 6. Level-k Analysis

Efficiency Seeking. In each treatment, efficiency-seeking always incentivizes reputation building. Thus, this argument cannot account for our experimental data.

Empirical Response. Here, we examine whether (i) the observed sender strategy is the best response to the average empirical receiver strategy and vice versa, (ii) the previous history leads to deviation from or convergence to the theoretical prediction. We start by examining the receiver’s response to the sender’s off-the-path behavior.

In the BT treatment, the receiver strategy is closely consistent with the equilibrium prediction. Figure 12 (a) illustrates the average receiver strategy in stage 1. A small departure from the prediction in stage 1 increases the reputation-building incentive because low receiver compliance to $m_1 = 1$ makes $m_1 = 0$ relatively attractive. Figure 12 (c) illustrates the average receiver strategy in stage 2. The second and last bar imply that the sender’s empirical best response to the average receiver strategy still remains to use reputation-building strategy ($x_2/x_1 > 9$ is the parsimonious condition).

In the GL treatment, the receiver strategy is not the best response to the sender’s average strategy. In stage 1, deviations from the reputation-building equilibrium by some senders make the message $m_1 = 1$ less credible. However, Figure 12 (b) shows that the average receiver action after observing $m_1 = 1$ is higher than the equilibrium prediction. In stage 2, less compliance after the history $\theta_1 = 1, m_1 = 0$ cannot be the best response to any sender strategy.

On the contrary, the sender strategy in the GL treatment can be explained as the empirical best response to the average receiver strategy. Figure 12 (b) and Figure 12 (d) show

that the average receiver actions after the history $\theta_1 = 1, m_1 = 0$ and $\theta_1 = 1, m_1 = 1$ are similar with each other, while receivers are credulous in stage 1. Thus, the sender can benefit from deviating from the reputation-building equilibrium. Our experiment II design in K rules out this possibility.

Appendix J. Omitted Proofs

Proof of Proposition 5. Note that $EU_B(m_1 = 1|\theta_1 = 0)$ is strictly decreasing in v and $EU_B(m_1 = 0|\theta_1 = 0)$ is strictly increasing in v . Thus, $EU_B(m_1 = 1|\theta_1 = 0)$ and $EU_B(m_1 = 0|\theta_1 = 0)$ intersect at at most one point. To examine whether the two functions intersect, it is enough to consider two endpoints of each function at $v = 0$ and $v = 1$. These endpoints are

$$EU_B(m_1 = 0|\theta_1 = 0) = \begin{cases} -x_1 - \frac{x_2}{9} & \text{if } v = 0, \\ -x_1 & \text{if } v = 1, \end{cases} \quad EU_B(m_1 = 1|\theta_1 = 0) = \begin{cases} -\frac{x_2}{4} & \text{if } v = 0, \\ -\frac{x_1}{9} - \frac{x_2}{4} & \text{if } v = 1. \end{cases}$$

- **Case 1:** $\frac{x_2}{x_1} \leq \frac{32}{9}$. Then, $EU_B(m_1 = 0|\theta_1 = 0) < EU_B(m_1 = 1|\theta_1 = 0)$ for all $v \in [0, 1]$. Thus, only $v = 1$ is incentive compatible for the expert. Knowing this, the public chooses $(a_1(v = 1), a_2(v = 1))$, which gives us the unique equilibrium.
- **Case 2:** $\frac{x_2}{x_1} \in (\frac{32}{9}, \frac{36}{5})$. Then, $EU_B(m_1 = 0|\theta_1 = 0)$ crosses $EU_B(m_1 = 1|\theta_1 = 0)$ from below only once. At this intersection v , the expert is indifferent between the two messages given the public's belief about v , thus this gives the unique incentive compatible mixed strategy v . Given the expert strategy, the public's belief about v is consistent. This is the unique equilibrium because all pure strategies are not incentive-compatible. At the public's belief that $v = 0$, $EU_B(m_1 = 0|\theta_1 = 0) < EU_B(m_1 = 1|\theta_1 = 0)$, so any $v > 0$ is a profitable deviation. Similarly, at the public's belief that $v = 1$, $EU_B(m_1 = 0|\theta_1 = 0) > EU_B(m_1 = 1|\theta_1 = 0)$, so any $v < 1$ is a profitable deviation.
- **Case 3:** $\frac{x_2}{x_1} \geq \frac{36}{5}$. Then, $EU_B(m_1 = 0|\theta_1 = 0) > EU_B(m_1 = 1|\theta_1 = 0)$ for all $v \in [0, 1]$. Thus, only $v = 0$ is incentive compatible for the expert. Knowing this, the public chooses $(a_1(v = 0), a_1(v = 0))$, which gives us the unique equilibrium.

Proof of Proposition 6. Note that $EU_G(m_1 = 1|\theta_1 = 1)$ is strictly decreasing in w and $EU_G(m_1 = 0|\theta_1 = 1)$ is strictly increasing in w . Thus, $EU_G(m_1 = 1|\theta_1 = 1)$ and $EU_G(m_1 = 0|\theta_1 = 1)$ intersect at at most one point. To examine whether the two functions intersect, it is enough to consider two endpoints of each function at $v = 0$ and $v = 1$. These endpoints are

$$EU_G(m_1 = 0|\theta_1 = 1) = \begin{cases} -x_1 & \text{if } w = 0, \\ -\frac{x_1}{4} & \text{if } w = 1, \end{cases} \quad EU_G(m_1 = 1|\theta_1 = 1) = \begin{cases} -\frac{x_1}{9} - \frac{x_2}{18} & \text{if } w = 0, \\ -\frac{x_1}{4} - \frac{x_2}{8} & \text{if } w = 1. \end{cases}$$

- **Case 1:** $\frac{x_2}{x_1} \geq 16$. Then, $EU_G(m_1 = 0|\theta_1 = 1) > EU_G(m_1 = 1|\theta_1 = 1)$ for all $w \in [0, 1]$. Thus, only $w = 1$ is incentive compatible for the expert. Knowing this, the public chooses $(a_1(w = 1), a_2(w = 1))$, which gives us the unique equilibrium.
- **Case 2:** $\frac{x_2}{x_1} < 16$. Then, $EU_G(m_1 = 0|\theta_1 = 1)$ crosses $EU_G(m_1 = 1|\theta_1 = 1)$ from below only once. At this intersection w , the expert is indifferent between the two messages given the public's belief about w , thus this gives the incentive-compatible mixed strategy w . Given the expert strategy, the public's belief about w is consistent. However, this is not the unique equilibrium because all pure strategies are also incentive-compatible. At the public's belief that $w = 0$, $EU_G(m_1 = 0|\theta_1 = 1) < EU_G(m_1 = 1|\theta_1 = 1)$, so all $w > 0$ are not profitable deviations for the expert. Given this, the public's belief is consistent with the expert's strategy. Similarly, at the public's belief that $w = 1$, $EU_G(m_1 = 0|\theta_1 = 1) > EU_G(m_1 = 1|\theta_1 = 1)$, so all $w < 1$ are not profitable deviations for the expert. Given this, the public's belief is consistent with the expert's strategy.

Proof of Proposition 7. We prove the general version of Proposition 7.

Proposition 10. (General Version of Proposition 7) Consider period 2 of the BT treatment. Suppose that $\theta_2 = 0$. There exists the increasing function (in λ) $\underline{c}_l(\lambda) \geq 3/4$ so that the (strategic) expert sends a message $m_2 = 1$ with probability 1 if $c_l \leq \underline{c}_l(\lambda)$, with probability $v_2 \in (0, 1)$ if $c_l \in (\underline{c}_l(\lambda), 1)$, and with probability 0 if $c_l \geq 1$.

Proof. In period 2 of the BT environment, the lying aversion potentially arises only when $\theta_2 = 0$. Thus, $\sigma_2(1|1) = 1$ and $\sigma_2(1|0) = v_2$.³² Given the expert strategy and belief λ , the public chooses

$$a_2 = \frac{\lambda/2 + (1 - \lambda)/2}{\lambda/2 + (1 - \lambda)(1 + v_2)/2} = \frac{1}{\lambda + (1 - \lambda)(1 + v_2)}$$

upon receiving $m_2 = 1$ and chooses $a_2 = 0$ upon receiving $m_2 = 0$. Then, the expected payoff of the expert from sending each message given $\theta_2 = 0$ becomes

$$\begin{aligned} EU_B(m_2 = 0|\theta_2 = 0) &= -(0 - 1)^2 \text{ and} \\ EU_B(m_2 = 1|\theta_2 = 0) &= -\left(\frac{1}{\lambda + (1 - \lambda)(1 + v_2)} - 1\right)^2 \end{aligned}$$

Observe that $EU_B(m_2 = 0|\theta_2 = 0)$ is constant and $EU_B(m_2 = 1|\theta_2 = 0)$ is strictly decreasing in v_2 . Also, $EU_B^a(m_2 = 1|\theta_2 = 0)$ is lower-translation of $EU_B(m_2 = 1|\theta_2 = 0)$ by c_l , while $EU_B^a(m_2 = 0|\theta_2 = 0) = EU_B(m_2 = 0|\theta_2 = 0)$. Thus, $EU_B^a(m_2 = 1|\theta_2 = 0)$ crosses $EU_B^a(m_2 = 0|\theta_2 = 0)$ at most once (from above) depending on the value of c_l .

³²In our experiment, the babbling equilibrium cannot arise in period 2 even at the public's belief $\lambda = 0$. This is because $\sigma_2(1|1) = 1$ is fixed by the experimental design, so $m_2 = 0$ becomes a perfect signal of $\theta_2 = 0$.

- **Case 1:** $c_l \leq 1 - \left(\frac{1}{2-\lambda} - 1\right)^2$. In this case,

$$EU_B^a(m_2 = 1|\theta_2 = 0) = -\left(\frac{1}{\lambda + (1-\lambda)(1+v_2)} - 1\right)^2 - c_l \geq -1 = EU_B^a(m_2 = 0|\theta_2 = 0)$$

for all v_2 , where the equality holds only when $c_l = 1 - \left(\frac{1}{2-\lambda} - 1\right)^2$ and $v_2 = 1$. Thus, $v_2 = 1$ in the equilibrium.

- **Case 2:** $c_l \in \left(1 - \left(\frac{1}{2-\lambda} - 1\right)^2, 1\right)$. In this case, $EU_B^a(m_2 = 1|\theta_2 = 0)$ intersects $EU_B^a(m_2 = 0|\theta_2 = 0)$ only once from above, at $v_2 = \frac{\sqrt{1-c_l}}{(1-\lambda)(1-\sqrt{1-c_l})}$. At this v_2 , the expert is indifferent between sending each message, so there exists the unique mixed-strategy equilibrium with $v_2 = \frac{\sqrt{1-c_l}}{(1-\lambda)(1-\sqrt{1-c_l})} \in (0, 1)$.

It remains to show there is no pure-strategy equilibrium. At $v_2 = 1$, $EU_B^a(m_2 = 1|\theta_2 = 0) < EU_B^a(m_2 = 0|\theta_2 = 0)$, so the expert deviates to $v_2 < 1$. At $v_2 = 0$, $EU_B^a(m_2 = 1|\theta_2 = 0) > EU_B^a(m_2 = 0|\theta_2 = 0)$, so the expert deviates to $v_2 > 0$.

- **Case 3:** $c_l \geq 1$. In this case,

$$EU_B^a(m_2 = 1|\theta_2 = 0) = -\left(\frac{1}{\lambda + (1-\lambda)(1+v_2)} - 1\right)^2 - c_l \leq -1 = EU_B^a(m_2 = 0|\theta_2 = 0)$$

for all v_2 , where the equality holds only when $c_l = 1$ and $v_2 = 0$. Thus, $v_2 = 0$ in the equilibrium.

Letting $\underline{c}_l(\lambda) \equiv 1 - \left(\frac{1}{2-\lambda} - 1\right)^2$ completes the proof of Proposition 10. \square

Therefore, the expert chooses $v_2 = 1$ if and only if $c_l \leq 1 - \left(\frac{1}{2-\lambda} - 1\right)^2$. Since $\lambda(0, 0) = \frac{1}{2-v_1} \geq \frac{1}{2}$ and $\lambda(1, 0) = 0$, we get $1 - \left(\frac{1}{2-\lambda(0,0)} - 1\right)^2 \geq \frac{8}{9}$ and $1 - \left(\frac{1}{2-\lambda(1,0)} - 1\right)^2 = \frac{3}{4}$. Thus, if $c_l \leq \frac{3}{4}$, we can guarantee that the expert sends a message $m_2 = 1$ with probability 1.³³ This completes the proof of Proposition 7.

Proof of Proposition 8. We require that (c_l, c_d) is the common knowledge in the equilibrium with lying and deception costs. First, note that $EU_B^a(m_1 = 1|\theta_1 = 0)$ is the lower-translation of $EU_B(m_1 = 1|\theta_1 = 0)$ by c_l and therefore downward-sloping. Second, $EU_B^a(m_1 = 0|\theta_1 = 0)$ is not a translation of $EU_B(m_1 = 0|\theta_1 = 0)$. It is an inverted U shape when $c_d < 8/27$ and downward-sloping otherwise.

Fix $c_d \geq 0$. Since $EU_B^a(m_1 = 0|\theta_1 = 0)$ and $EU_B^a(m_1 = 1|\theta_1 = 0)$ are continuous and changing c_l does not change the shape of $EU_B^a(m_1 = 1|\theta_1 = 0)$, we can find $\bar{c} \in \mathcal{R}$ such that $c_l > \bar{c}$ if and only if $EU_B^a(m_1 = 0|\theta_1 = 0) > EU_B^a(m_1 = 1|\theta_1 = 0)$ for all $v \in [0, 1]$.

³³If $c_l > \frac{3}{4}$, then the expert deviates from $v_2 = 1$ when $m_1 = 1, \theta_1 = 0$ (this history occurs with positive probability unless $v_1 = 0$). Likewise, if $c_l > 1 - \left(\frac{1}{2-\lambda(0,0)} - 1\right)^2$, then the expert can deviate from $v_2 = 1$ in any period 1 history, depending on the period 1 behavior of players.

Then, only $v = 0$ is incentive compatible for the expert. Knowing this, the public chooses $(a_1(v = 0), a_1(v = 0))$, which gives us the unique equilibrium.

Likewise, we can find $\underline{c} \in \mathcal{R}$ such that $c_l < \underline{c}$ if and only if $EU_B^a(m_1 = 0|\theta_1 = 0) < EU_B^a(m_1 = 1|\theta_1 = 0)$ for all $v \in [0, 1]$. Then, only $v = 1$ is incentive compatible for the expert. Knowing this, the public chooses $(a_1(v = 1), a_2(v = 1))$, which gives us the unique equilibrium. Define $\bar{c}(c_d) = \min\{\bar{c}, 3/4\}$ and $\underline{c}(c_d) = \max\{\underline{c}, 0\}$. Then, we get Cases i and iii in Proposition 8.

Finally, consider the case $c_l \in (\underline{c}(c_d), \bar{c}(c_d))$. Then, $EU_B^a(m_1 = 0|\theta_1 = 0)$ and $EU_B^a(m_1 = 1|\theta_1 = 0)$ cross in $v \in (0, 1)$ at least once. At this intersection, the expert is indifferent between the two messages given the public's belief about v , thus this gives the incentive compatible mixed strategy v . Given the expert strategy, the public's belief about v is consistent. Then, we get Case ii in Proposition 8. In addition, there are two possibilities of equilibrium multiplicity. First, $EU_B^a(m_1 = 0|\theta_1 = 0)$ and $EU_B^a(m_1 = 1|\theta_1 = 0)$ can cross at two points. This gives rise to two mixed-strategy equilibria. Second, if either $EU_B^a(m_1 = 0|\theta_1 = 0) > EU_B^a(m_1 = 1|\theta_1 = 0)$ at $v = 0$ or $EU_B^a(m_1 = 0|\theta_1 = 0) < EU_B^a(m_1 = 1|\theta_1 = 0)$ at $v = 1$, a pure-strategy equilibrium also exists.

Proof of Proposition 9. Consider $EU_G^a(m_1 = 0|\theta_1 = 1)$ and $EU_G^a(m_1 = 1|\theta_1 = 1)$, which are functions of w . First, note that $EU_G^a(m_1 = 0|\theta_1 = 1)$ is the lower-translation of $EU_G(m_1 = 0|\theta_1 = 1)$ by c_l and therefore upward-sloping. Second, the gap $EU_G(m_1 = 1|\theta_1 = 1) - EU_G^a(m_1 = 1|\theta_1 = 1)$ is increasing in w and $EU_G(m_1 = 1|\theta_1 = 1)$ is downward-sloping, so $EU_G^a(m_1 = 1|\theta_1 = 1)$ is also downward-sloping. Thus, $EU_G^a(m_1 = 0|\theta_1 = 1)$ and $EU_G^a(m_1 = 1|\theta_1 = 1)$ intersect at at most one point. To examine whether the two functions intersect, it is enough to consider two endpoints of each function at $w = 0$ and $w = 1$. These endpoints are

$$EU_G^a(m_1 = 0|\theta_1 = 1) = \begin{cases} -\frac{1}{20} - c_l & \text{if } w = 0, \\ -\frac{1}{80} - c_l & \text{if } w = 1, \end{cases} \quad EU_G^a(m_1 = 1|\theta_1 = 1) = \begin{cases} -\frac{11}{180} - \frac{c_d}{2} & \text{if } w = 0, \\ -\frac{11}{80} - c_d & \text{if } w = 1. \end{cases}$$

- **Case 1:** $c_l > c_d + \frac{1}{8}$. Then, $EU_G^a(m_1 = 0|\theta_1 = 1) < EU_G^a(m_1 = 1|\theta_1 = 1)$ for all $w \in [0, 1]$. Thus, only $w = 0$ is incentive compatible for the expert. Knowing this, the public chooses $(a_1(w = 0), a_2(w = 0))$, which gives us the unique equilibrium.
- **Case 2:** $c_l \in \left[\frac{c_d}{2} + \frac{1}{90}, c_d + \frac{1}{8}\right]$. Then, $EU_G^a(m_1 = 0|\theta_1 = 1)$ crosses $EU_G^a(m_1 = 1|\theta_1 = 1)$ only once. At this intersection, the expert is indifferent between the two messages given the public's belief about w , thus this gives the incentive compatible mixed strategy w . Given the expert strategy, the public's belief about w is consistent. This is not the unique equilibrium strategy because all pure strategies are incentive-compatible for the expert given the consistent belief of the public. At the public's belief that $w = 0$, $EU_B^a(m_1 = 1|\theta_1 = 1) > EU_B^a(m_1 = 0|\theta_1 = 1)$, so any

$w > 0$ are not profitable deviations for the expert. Given this, the public's belief that $w = 0$ is consistent with the expert's strategy. Similarly, at the public's belief that $w = 1$, $EU_B^a(m_1 = 1|\theta_1 = 1) < EU_B^a(m_1 = 0|\theta_1 = 1)$, so any $w < 1$ are not profitable deviations for the expert. Given this, the public's belief that $w = 1$ is consistent with the expert's strategy.

- **Case 3:** $c_l < \frac{c_d}{2} + \frac{1}{90}$. Then, $EU_G^a(m_1 = 0|\theta_1 = 1) > EU_G^a(m_1 = 1|\theta_1 = 1)$ for all $w \in [0, 1]$. Thus, only $w = 1$ is incentive compatible for the expert. Knowing this, the public chooses $(a_1(w = 1), a_1(w = 1))$, which gives us the unique equilibrium.

Appendix K. Experiment II: Design

In this section, we illustrate how we specify the design of Experiment II.

Timeline. Contingent on message m and state θ , let $\lambda^R(m, \theta)$ denote the receiver's belief about the sender's preference type, and let $\lambda^S(m, \theta)$ denote the sender's second-order belief about $\lambda^R(m, \theta)$. The timeline of the new design is as follows.

- (1) Senders choose their message transmission rule using the real-time spinning wheel.
- (2) Senders report their conjecture on the receiver's sender-type beliefs after observing each message ($\lambda^S(m = 1, \theta = 0)$ and $\lambda^S(m = 0, \theta = 0)$ in BT, $\lambda^S(m = 1, \theta = 1)$ and $\lambda^S(m = 0, \theta = 1)$ in GL).³⁴
- (3) Senders observe the state θ and a message m is sent to the paired receiver according to the spinning wheel.
- (4) Receivers see their paired sender's message m and make a conjecture about the state using the slider bar.
- (5) Receivers see the state and report their conjecture about the sender's type, $\lambda^R(m, \theta)$.
- (6) The round ends and new sender-receiver pairs are randomly formed.

Specification of Payoff Functional Form. The payoffs are specified as the following.

$$\begin{aligned}
 U_P(\theta, a, \lambda^R) &= -(a - \theta)^2, \\
 U_G(\theta, a, \lambda^R) &= -x_1(a - \theta)^2 + \frac{x_2}{4}\lambda^R, \text{ and} \\
 U_B(\theta, a, \lambda^R) &= -x_1(a - 1)^2 + \frac{x_2}{8}\lambda^R.
 \end{aligned}
 \tag{1}$$

³⁴In the BT treatment, $m_1 = 1$ is automatically sent to the receiver given $\theta_1 = 1$. Thus, $\lambda^S(m_1|\theta_1 = 1)$ has no role when the sender decides $Pr(m_1 = 1|\theta_1 = 0)$. Similarly, in the GL treatment, $m_1 = 0$ is automatically sent to the receiver given $\theta_1 = 0$. Thus, $\lambda^S(m_1|\theta_1 = 0)$ has no role when the sender decides $Pr(m_1 = 1|\theta_1 = 1)$. Therefore, we do not elicit these beliefs.

The linear reputational payoff is derived from the first-order Taylor approximation of the expected Stage 2 sender payoff. To see this, recall that the expected Stage 2 sender payoffs of Experiment I are

$$BT : 1 - \left(\frac{1}{2 - \lambda^R} - 1 \right)^2, \quad GL : 1 - \frac{1}{2} \left(\frac{1}{2 - \lambda^R} - 1 \right)^2.$$

The linear approximations in each environment are

$$BT : \frac{\lambda^R + 3}{4}, \quad GL : \frac{\lambda^R + 7}{8}.$$

The purpose of linear approximations is to alleviate the computational complexity in calculating the complicated rational payoff functions. Note that linear approximations perform better than quadratic ones for all $\lambda^R \in (0, 1)$ in each environment.³⁵ Since the constant term does not affect the sender's message choice, we drop the constant term.

The weight on λ^R is two times greater in the BT treatment than in the GL treatment. However, this does not mean that the reputation incentive is stronger in the BT treatment than in the GL treatment because the ranges of λ^R on the equilibrium path are different across treatments. That is, the range of equilibrium λ^R is $[0, 1/2]$ in the BT treatment and $[0, 1]$ in the GL treatment. Thus, the ranges of the equilibrium reputational payoffs are the same across the two treatments. Lastly, $x_1 = 1000$ KRW and $x_2 = 20,000$ KRW as in Experiment I.

Belief Elicitation. Each belief elicitation is incentivized so that truthful reporting becomes the dominant strategy. That is, each player gets an additional quadratic-loss payoff from belief reporting. The bliss point of the receiver's quadratic payoff is given by λ^* , the true probability that the sender is the good type. The bliss point of the sender's quadratic payoff is given by λ^R .

Round Payoff. In summary, the round payoff we used in Experiment II is given as the following:

$$\begin{aligned} \text{Receiver}(BT) &: 1000[1 - (a - \theta)^2] + 5000[1 - (\lambda^R - \lambda^*)^2], \\ \text{Sender}(BT) &: 1000[1 - (a - 1)^2] + 5000\lambda^R + 500[1 - (\lambda^S - \lambda^R)^2], \\ \text{Receiver}(GL) &: 1000[1 - (a - \theta)^2] + 2500[1 - (\lambda^R - \lambda^*)^2], \\ \text{Sender}(GL) &: 1000[1 - (a - \theta)^2] + 2500\lambda^R + 500[1 - (\lambda^S - \lambda^R)^2], \end{aligned}$$

³⁵The quadratic approximations are given by $1 - \frac{(\lambda^R - 1)^2}{4}$ in the BT treatment and $1 - \frac{(\lambda^R - 1)^2}{8}$ in the GL treatment.

Since the receiver makes separate decisions regarding the first and second terms of his payoff, the relative weights of the receiver's action payoff and belief payoff do not affect the theoretical predictions of our experiment. Therefore, to simplify instructions, we made the receiver's payoff weights equal to the sender's payoff weights. Also, we used a small weight (500 KRW) on the sender's incentives for second-order belief elicitation.

The participation fee was 15,000 KRW as in Experiment I. Lastly, recall that we dropped the constant term in the linear approximation of reputation payoffs (1). To make the total average payment similar between Experiment I and Experiment II, participants in Experiment II received an additional constant compensation of 10,000 KRW. We separated this final compensation from the participation fee and let participants know that they receive additional constant compensation at the end of experiments to make sure that the size of the fixed compensation would not affect participants' incentives in unexpected ways.

Discussion of Experiment II Design. The simple abstraction of Stage 2 has several benefits. First, the receiver's two-step inference (about the sender's type, then the best action response) boils down to simple belief reporting about the sender's type. This not only simplifies the receiver's decision-making but also alleviates the strategic uncertainty arising from Stage 2. For example, senders who tell the truth in Stage 2 of the BT treatment and receivers who are less responsive in Stage 2 after observing a lying history in the GL treatment are interesting but unexpected observations in our main experiment. By abstracting from Stage 2, we rule out such anomalies. Second, the one-shot design shortens the experiment time significantly, which enables us to implement belief elicitation.³⁶

Separating Preference Channel from Inference Channel. Belief elicitation of the sender's second-order belief lets us classify the observed deviation from reputation-building equilibrium into the two underlying channels: preference and inference. The classification rule is summarized in Figure 5.

The classification rule is characterized by a threshold on the difference between two second-order beliefs. This rule is derived from the following arguments. Suppose that the preference channel does not work at all, i.e. $c_l = c_d = 0$. Then, in the BT treatment, the sender prefers sending $m = 0$ to sending $m = 1$ if and only if

$$\begin{aligned} EU_B(m = 0|\theta = 0) - EU_B(m = 1|\theta = 0) &> 0 \\ \iff \lambda^S(0,0) - \lambda^S(1,0) &> \frac{4x_1}{x_2}[(a(0) - 1)^2 - (a(1) - 1)^2] \equiv T_{BT}, \end{aligned}$$

³⁶To separate the preference channel from the inference channel, we need to implement the strategy method in eliciting the sender's second-order belief. Such belief elicitation in the 2-period design is burdensome for participants because our main experiment already took about 90-110 minutes per session.

where $T_{BT} \leq \frac{4x_1}{x_2} = 0.2$ for any $(a(0), a(1))$ satisfying $a(1) \geq a(0)$. Similarly, in the GL treatment, the sender prefers sending $m = 0$ to sending $m = 1$ if and only if

$$\begin{aligned} EU_G(m = 0|\theta = 1) - EU_G(m = 1|\theta = 1) &> 0 \\ \iff \lambda^S(0,0) - \lambda^S(1,0) &> \frac{8x_1}{x_2}[(a(0) - 1)^2 - (a(1) - 1)^2] \equiv T_{GL}, \end{aligned}$$

where $T_{GL} \leq \frac{8x_1}{x_2} = 0.4$ for any $(a(0), a(1))$ satisfying $a(1) \geq a(0)$. If the elicited pair of second-order beliefs does not satisfy the above threshold rule and the strategy shows deviation from the reputation-building equilibrium strategy, then we classify such strategy-belief pair as the preference channel: deception aversion (BT) or lying aversion (GL).

Note that the specification of T_{BT} and T_{GL} depends on the sender's expectation of the receiver's action. In the (reputation-building) equilibrium, $T_{BT} = 0.2$ and $T_{GL} = 0$. Instead, if we use the receiver's average Stage 1 action in Experiment I, $T_{BT} = 0.163$ and $T_{GL} = 0.122$. Note that $T_{BT} = 0.2$ and $T_{GL} = 0.4$ are the most conservative thresholds of the belief gaps in that these require the strongest conditions for a sender's strategy-belief pair to be attributed to the preference channel. Thus, in our analysis, we use $T_{BT} = 0.2$ and $T_{GL} = 0.4$ for robustness.

Explaining Receiver's Low Compliance in Experiment I (GL). Belief elicitation of the receiver enables us to separate the channel behind the receiver's low compliance after the lying history in the GL treatment. If $\lambda^R(0,1) = \lambda^R(0,0)$, then the low-compliance issue reported in Section D.3 is attributed to the punishment argument. Otherwise, low compliance is attributed to the inference error argument. Indeed, Experiment II results reported in Figure 25 of Appendix 25 documents $\lambda^R(0,1) = \lambda^R(0,0)$, providing strong evidence in favor of the punishment argument.

Appendix L. Equilibrium Analysis of the Psychological Game

In this section, we provide theoretical results about Experiment II in Section 4. Let $y_2/y_1 > 0$ denote the importance of period 2 relative to period 1. Recall from Appendix K that we use $y_1 = x_1$ and $y_2 = x_2/4$ in the BT environment, while $y_1 = x_1$ and $y_2 = x_2/8$ in the GL environment. Here, we derive theoretical predictions for arbitrary y_1 and y_2 . We drop subscripts since there is no second period. We first suppose that $c_l = c_d = 0$ and characterize equilibria of the modified game, which corresponds to the modified version of Proposition 5 and 6.

Proposition 11. *In the unique equilibrium of the game, $\sigma(1|1) = 1$ and $\sigma(1|0) = v \in [0,1]$ while $a(0) = 0$ and $a(1) = \frac{1}{1+v/2}$, where $v = 1$ if $y_2/y_1 \leq 8/9$, $v \in (0,1)$ if $8/9 < y_2/y_1 < 2$, and $v = 0$ if $y_2/y_1 \geq 2$.*

Proof. In the modified game, the expert's second-period payoff is substituted with the reputation utility. Thus, the expert's expected payoff from sending each message given $\theta = 0$ becomes

$$\begin{aligned} EU_B(m = 0|\theta = 0) &= -y_1(0-1)^2 + y_2\lambda(0,0) = -y_1 + \frac{y_2}{2-v}, \text{ and} \\ EU_B(m = 1|\theta = 0) &= -y_1(\frac{1}{1+v/2} - 1)^2 + y_2\lambda(1,0) = -y_1(\frac{v}{v+2})^2. \end{aligned}$$

The rest of the proof exactly follows that of Proposition 5. \square

Proposition 12. *In any informative equilibrium of the game, $\sigma(0|0) = 1$ and $\sigma(0|1) = w \in [0, 1]$ while $a(0) = \frac{w}{1+w}$ and $a(1) = \frac{2-w}{3-w}$. If $y_2/y_1 \geq 16/9$, equilibrium is unique and $w = 1$. If $y_2/y_1 < 16/9$, then there exist three equilibria, each with $w = 0$, $w = 1$, and $w \in (0, 1)$.*

Proof. In the modified game, the expert's second-period payoff is substituted with the reputation utility. Thus, the expert's expected payoff from sending each message given $\theta = 1$ becomes

$$\begin{aligned} EU_G(m = 0|\theta = 1) &= -y_1(\frac{w}{1+w} - 1)^2 + y_2\lambda(0,1) = -\frac{y_1}{(1+w)^2} + y_2, \text{ and} \\ EU_G(m = 1|\theta = 1) &= -y_1(\frac{2-w}{3-w} - 1)^2 + y_2\lambda(1,1) = -\frac{y_1}{(w-3)^2} + y_2\frac{1-w}{2-w}. \end{aligned}$$

The rest of the proof exactly follows that of Proposition 6. \square

We now introduce lying cost c_l and deception cost c_d . Since the proofs follow the exactly same steps as in the proofs of Proposition 8 and 9, we omit the proofs. The only notable difference from our original game is that there is no analogous version of Proposition 7. This is because second period does not exist in our modified game.

Proposition 13. *Under Assumptions 1 and 2, in any equilibrium of the game, $\sigma(1|1) = 1$, $\sigma(1|0) = v \in [0, 1]$, $a(0) = 0$, and $a(1) = \frac{1}{1+v/2}$.*

a. When $c_d < y_2$, in the unique equilibrium

- i. $v = 0$ if $c_l > \frac{c_d}{2} + (y_1 - \frac{y_2}{2})$;
- ii. $v \in (0, 1)$ if $\frac{c_d}{2} + (y_1 - \frac{y_2}{2}) < c_l < c_d + (\frac{8}{9}y_1 - y_2)$;
- iii. $v = 1$ if $c_l < c_d + (\frac{8}{9}y_1 - y_2)$.

b. When $c_d \geq y_2$,

- i. equilibrium is unique and $v = 0$ if $c_l > \frac{c_d}{2} + (y_1 - \frac{y_2}{2})$ and $c_l > c_d + (\frac{8}{9}y_1 - y_2)$;
- ii. there are three equilibria, each with $v = 0$, $v \in (0, 1)$, and $v = 1$ if $c_l > \frac{c_d}{2} + (y_1 - \frac{y_2}{2})$ and $c_l < c_d + (\frac{8}{9}y_1 - y_2)$;
- iii. equilibrium is unique and $v \in (0, 1)$ if $c_l < \frac{c_d}{2} + (y_1 - \frac{y_2}{2})$ and $c_l > c_d + (\frac{8}{9}y_1 - y_2)$;
- iv. equilibrium is unique and $v = 1$ if $c_l < \frac{c_d}{2} + (y_1 - \frac{y_2}{2})$ and $c_l < c_d + (\frac{8}{9}y_1 - y_2)$.

Proposition 14. *Under Assumptions 1 and 2, in any equilibrium of the game, $\sigma(0|0) = 1$, $\sigma(0|1) = w \in [0, 1]$, $a(0) = \frac{w}{1+w}$, and $a(1) = \frac{2-w}{3-w}$. Moreover,*

- a. in the unique equilibrium, $w = 0$ if $c_l \geq c_d + y_2$;
- b. there are three equilibria, each with $w = 0$, $w \in (0, 1)$, and $w = 1$ if $\frac{c_d}{2} + \left(\frac{y_2}{2} - \frac{8y_1}{9}\right) < c_l < c_d + y_2$;
- c. in the unique equilibrium, $w = 1$ if $c_l \leq \frac{c_d}{2} + \left(\frac{y_2}{2} - \frac{8y_1}{9}\right)$.

Appendix M. Experiment II: Additional Results

	BT	GL
Equilibrium	44%	16%
Aversion	47%	47%
Inference Error	9%	37%

(a) threshold= 0.1, last 5 rounds

	BT	GL
Equilibrium	47%	19%
Aversion	37%	44%
Inference Error	16%	37%

(b) threshold= 0.2, 10 rounds

	BT	GL
Equilibrium	47%	28%
Aversion	44%	34%
Inference Error	9%	38%

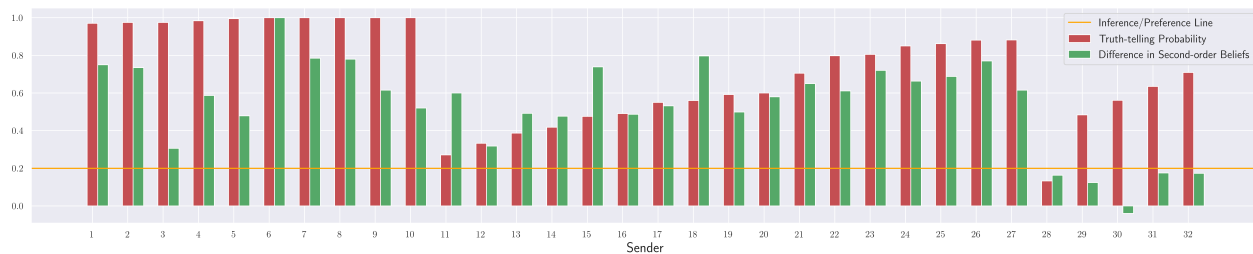
(c) threshold= 0.2, last 5 rounds

FIGURE 20. Individual Classifications

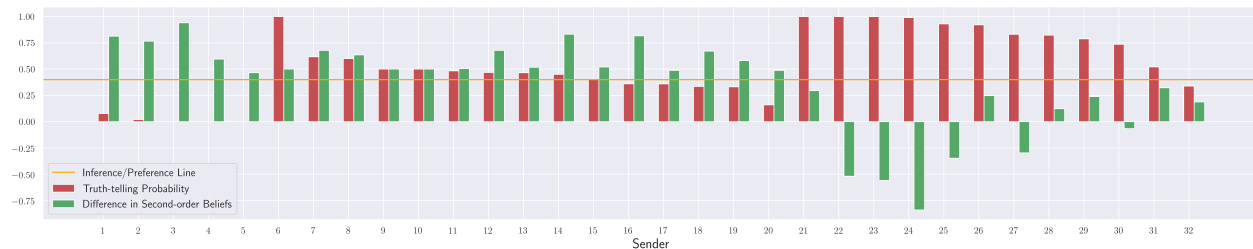
TABLE 7. The Role of Other Regarding Preference

	BT				GL			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Giving Share	-0.361 (0.239)	0.372 (0.453)	0.079 (0.548)	-0.469 (0.486)	0.229 (0.307)	0.449 (0.337)	0.486 (0.553)	-0.258 (0.475)
Constant	0.772*** (0.058)	0.629*** (0.110)	0.620*** (0.118)	0.605*** (0.118)	0.494*** (0.077)	0.770*** (0.085)	0.690*** (0.121)	0.511*** (0.120)
Observations	32	32	27	32	32	32	20	32

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Standard errors in parentheses. The first column in each treatment (i.e., (1) and (5)) uses individual truth-telling probability averaged over all 10 rounds as the dependent variable. The second column (i.e., (2) and (6)) uses the dummy variable which equals 1 if a sender deviated from the equilibrium strategy and 0 otherwise. The third column (i.e., (3) and (7)) excludes senders who are categorized as indicating inference errors from the regression specification for the second column. The fourth column (i.e., (4) and (8)) uses the dummy variable which equals 1 if a sender is categorized as indicating deception aversion and 0 otherwise. In each column, the independent variable is the measure of other-regarding preference, which is the percentage share of money each player proposed in a dictator game played at the end of each session. All variables are normalized to range from 0 to 1.



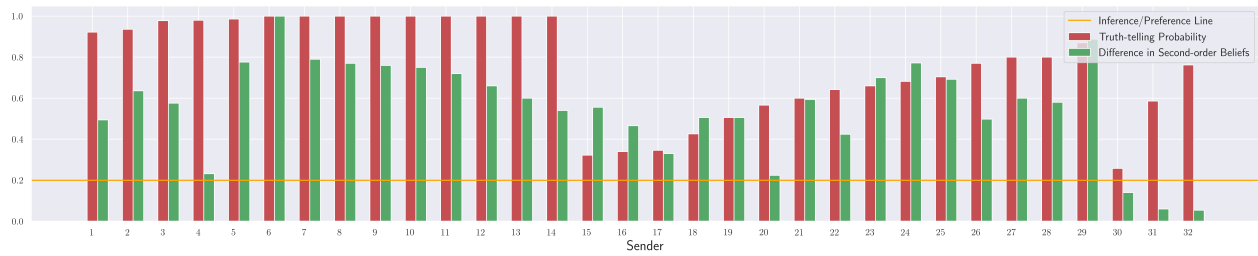
(a) BT



(b) GL

FIGURE 21. Description of Individual Average Play

Each red (resp. green) bar represents each sender's strategy (resp. belief difference) aggregated across all 10 rounds. In the BT treatment, players 1-10 (31%) are reputation-builders, players 11-27 (53%) are deception-haters, and players 28-32 (16%) are subject to inference errors. In the GL treatment, players 1-5 (16%) are reputation-builders, players 6-20 (47%) are lying-haters, and players 21-32 (37%) are subject to inference errors.



(a) BT



(b) GL

FIGURE 22. Description of Individual Average Play (Last 5 Rounds)

This figure replicates Figure 21 using only the observations in the last 5 rounds. Overall, each three groups remains robust in each treatment.

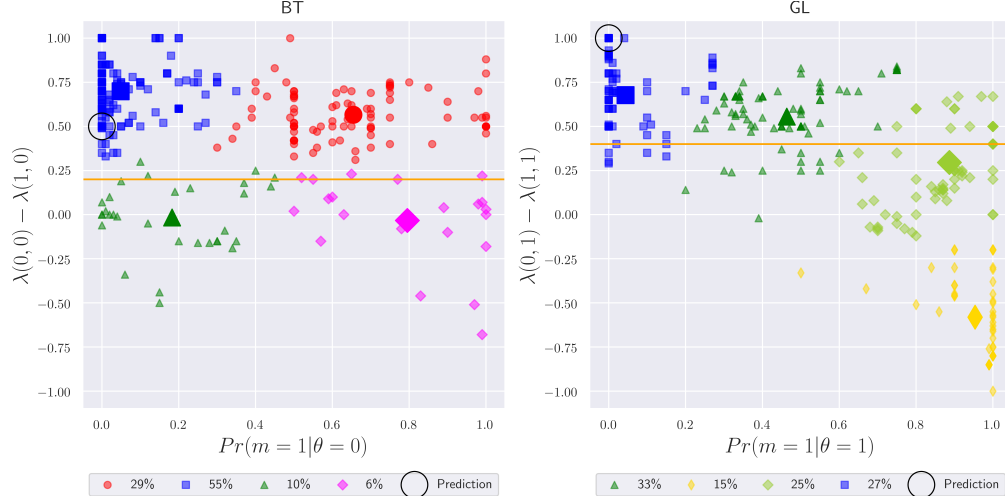


FIGURE 23. Clustering of Sender's Strategy

Note: In both panels, the horizontal axis represents the sender's strategy, and the vertical axis represents the difference in the relevant second-order beliefs. The blue square denotes the equilibrium cluster. The red circle indicates the deception aversion cluster in the BT treatment, while the green triangle represents the lying aversion cluster in both treatments. The magenta and yellow-green diamonds represent the cursed equilibrium cluster in each treatment. The yellow diamonds in the GL treatment indicate the noise. The center of each cluster is highlighted by a larger shape. The orange line is the threshold dividing the two channels. If the strategy (horizontal axis) deviates from the equilibrium prediction, it is attributed to the preference channel if its corresponding belief is above the orange line, whereas it is attributed to the inference channel if its corresponding belief is below the line. The height of the line in the BT treatment is $T_{BT} = 0.2$ and that in the GL treatment is $T_{GL} = 0.4$, which are parsimonious in the sense that they minimize the set of observations attributed to the preference channel.

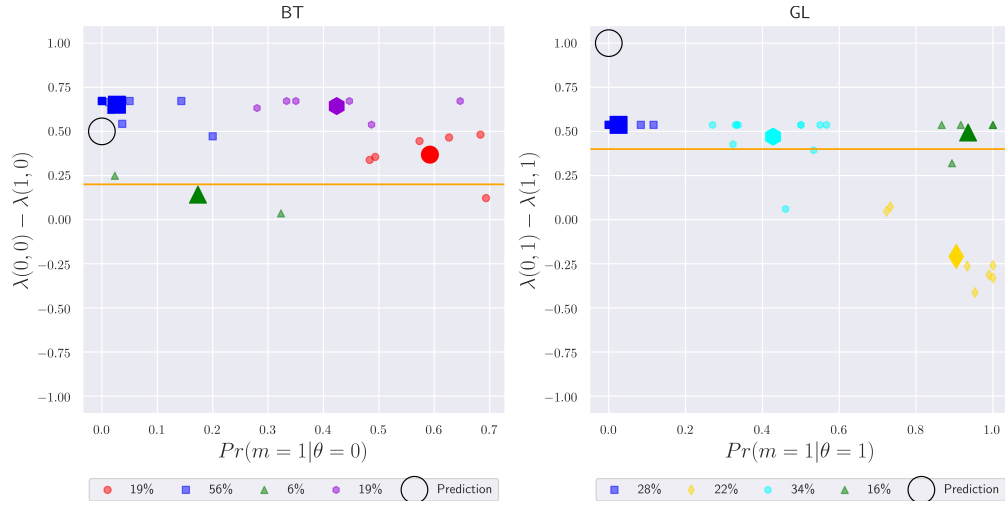


FIGURE 24. Clustering of Sender Strategy (Individual Average)

Note: Aggregated version of Figure 23. Each observation indicates the individual average across all 10 rounds. Each marker means the same cluster as in Figure 23, except for violet and cyan hexagons in each treatment. The violet hexagons represent the deception aversion cluster with a weaker extent compared to the red circles. Similarly, the cyan hexagons represent the lying aversion cluster with mixed strategies. These two clusters are attributed to the preference channel because they are located above the orange line. The observation that inference error clusters disappear in each treatment implies that each individual sender infers well on average over the entire rounds.



FIGURE 25. Receiver Strategy in Each Stage

Note: The red bars (Panel (a) and (b) for each treatment) describe the receiver's action given each message averaged over all individuals and rounds. The green bars (Panel (c) and (d) for each treatment) describe the receiver's belief about the sender type given each history. The blue diamond indicates the theoretical predictions. Panel (a) and (b) show that the average receiver strategy remains very similar to that in Stage 1 of Experiment I. Panel (c) shows that in the BT treatment, the average receiver belief is consistent with the theoretical predictions, with the exception that receivers become slightly more optimistic after observing $\theta = m = 0$ history than $\theta = m = 1$ history. Lastly, Panel (d) implies that the average receiver belief is consistent with the theoretical predictions, except for sizable optimism about the sender type after observing $\theta = m = 1$ history. Importantly, the first bar in Panel (d) indicates that almost all receivers did not experience inference error after observing $\theta = 1, m = 0$ history, providing very strong evidence that receivers' under-compliance to senders' message, reported in Section D.3, is mainly attributed to "punishing a lie" argument.

Appendix N. Experiment II: Non-parametric Tests

Table 8 presents the results from the non-parametric tests regarding the sender and receiver strategies. The statistical test results confirm the conclusions presented in Section M.

Who?	Test	Two-sided?	Null Hypothesis	p -values
Sender	MWU	Yes	Sender strategy is the same across BT and GL.	0.030
	MWU	Yes	Sender belief difference is the same across BT and GL.	0.061
	Wilc	One-sided ($>$)	Sender strategy is zero in BT.	0.034
	Wilc	One-sided ($>$)	Sender strategy is zero in GL.	0.034
	Wilc	Yes	Sender belief difference is 1/2 in BT.	0.465
	Wilc	One-sided ($<$)	Sender belief difference is 1 in GL.	0.034

■ MWU and Wilc refer to the Mann-Whitney U (rank-sum) test and one-sample Wilcoxon (signed rank) test, respectively.

■ Sender strategy in this table denotes the probability to send message $m = 1$ given the state θ (BT: $\theta = 0$, GL: $\theta = 1$).

■ One-sided test is implemented only when the direction of the alternative hypothesis is clear.

TABLE 8. Non-parametric Tests Results (Experiment II)

Appendix O. Experimental Instructions - Experiment I (Treatment BT)

In this section, we present a translated version of a sample instruction. In the actual experiments, the instruction was in Korean.

[Page 1]

Welcome! This experiment is about strategic decision-making within a group consisting of two participants. You will participate in 1 practice round and 10 official rounds going forward. The reward you will receive will be determined by your decisions and the decisions of other participants in the official rounds.

Your Role and Decision Group

Roles within the group are divided into two categories: the **Sender**, who sends messages, and the **Receiver**, who receives messages and takes actions.

Once the experiment begins, one-third of you will be randomly assigned the role of Sender, while the remaining two-thirds will be assigned the role of Receiver. Once your roles are determined, they will **not change** throughout the experiment. Additionally, an equal number of Senders to those assigned to be Senders among the participants will be played by the computer. The Senders played by experiment participants are referred to as **H (human)-senders**, while those played by the computer are referred to as **C (computer)-senders**.

At the start of each round, Senders and Receivers will form random Sender-Receiver pairs to play that round. Since the number of H-senders and C-senders is equal, the probability of each Receiver being paired with an H-sender or a C-sender is the same. However, you will not know whether the Sender you are paired with is an H-sender or a C-sender.

[Page 2]

Now let's consider a situation with two balls inside a box. One ball is **Red**, and the other ball is **Gray**.

Each round consists of two Stages.

STAGE 1

At the beginning of Stage 1, for each Sender-Receiver pair, the computer randomly selects one of the two balls from the box. The color of the ball selected is determined independently for each pair, each stage, and each round, so there is no correlation between them.

The color of the ball chosen by the computer is revealed only to the **Sender**. The Receiver cannot see the color of the selected ball.

Now the decision-making process for the Sender and Receiver begins. Initially, Receivers will wait until all Senders have made their decisions.

[Page 3]

H-sender's Decision

Before seeing the color of the ball selected by the computer, each H-sender must decide whether to send the message "Red" or "Gray" to the Receiver they are paired with.

- To simplify the decision-making process, if the selected ball is Red, the "Red" message will be sent to the Receiver with a 100% probability.
- However, if the selected ball is Gray, the H-sender will send the "Red" or "Gray" message to the Receiver based on the probability rules they have determined as follows:
 - (1) First, the Sender will see a spinning wheel on their screen (Figure 26) to determine the probability rule for message transmission.
 - (2) While clicking on the small circle on the spinning wheel, you can rotate it clockwise (or counterclockwise) along the circumference of the large circle to determine the relative size of the Red area and the Gray area. The relative sizes of each color will be displayed as a percentage on the right side of the spinning wheel.
 - (3) After deciding on the relative sizes of Red and Gray as you wish, simply press the "Submit" button located at the bottom of the spinning wheel.

Please note that if you press the "Submit" button without operating the spinning wheel, an error message saying "Please fix the errors in the form" will appear at the top of the page. In this case, you should operate with the spinning wheel again and then press the "Submit" button.

- (4) Press the "Submit" button located at the bottom of the spinning wheel.

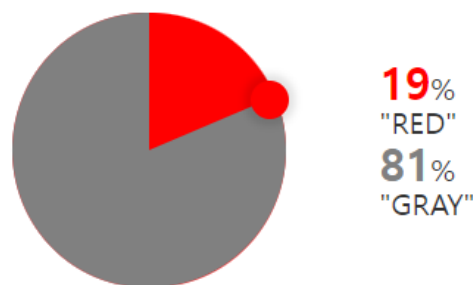


FIGURE 26. Real-time Spinning Wheel

[Page 4]

H-sender's Decision (Continued)



FIGURE 27. Submitted Spinning Wheel

- (1) Now, the H-sender will see the color of the ball selected by the computer, and the corresponding spinning wheel will also appear on the screen.
 - If the selected ball is **Red**, you will see a spinning wheel that is 100% **Red**.
 - If the selected ball is **Gray**, you will see the spinning wheel you submitted (Figure 27).
- (2) In the center of the spinning wheel, you will see a spin button and a needle pointing at 12 o'clock. Pressing the spin button will rotate the spinning wheel.
- (3) When the spinning wheel stops (the stopping point is randomly determined), the color of the area indicated by the white needle will be the message delivered to the Receiver paired with you.

In summary, Receivers paired with H-senders receive messages as follows:

- When the computer selects a **Red** ball: A 100% probability of receiving a "**Red**" message.
- When the computer selects a **Gray** ball: Messages with probabilities based on the colors you set on the submitted spinning wheel, either "**Red**" or "**Gray**."

After the wheel stops spinning, a "Continue" button will appear at the bottom of the spinning wheel. Don't forget to click this button to proceed to the next screen.

[Page 5]

C-sender's Decision

C-senders will send a message that is the same as the color of the selected ball. That is,

- When the computer selects a **Red** ball: "**Red**" message with a 100% probability
- When the computer selects a **Gray** ball: "**Gray**" message with a 100% probability

Once all H-senders and C-senders have made their decisions, Receivers will move from the waiting screen to the decision screen **altogether**. This means that even if your paired

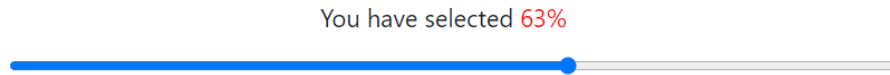


FIGURE 28. Action Slider Bar

Sender has made a decision, you will still have to wait on the waiting screen until all other Senders have completed their decisions.

Similarly, Senders will be on the waiting screen until all Receivers have completed their decisions.

[Page 6]

Receiver's Decision

Receivers receive messages from their paired Sender. However, they do not know whether their paired Sender is an H-sender or a C-sender, and they also **do not have information about which spinning wheel the Sender used.**

Receiver's Action: given the message, use the following slider (Figure 28) to estimate the probability that the selected ball is **Red**.

After making your decision, press the "Submit" button located at the bottom of the slider.

After all Receivers have completed their selections, on the next screen, **each Receiver will see the actual color of the ball that was selected in their group.**

[Page 7]

STAGE 2

In Stage 2, the Sender-Receiver pairs remain the same as in Stage 1, meaning Receivers receive messages from the same Senders as in Stage 1.

At the start of Stage 2, the computer randomly selects one of two new balls placed in a box for each Sender-Receiver pair. This is independent of the ball drawn in Stage 1, and the color of the ball drawn in Stage 2 is only revealed to the Sender, not the Receiver.

The rest of Stage 2 follows the same process as Stage 1, with H-senders and Receivers making new decisions.

[Page 8]

Your Earning in Each STAGE

- Receivers' earnings for each stage are determined based on the difference between the actual color of the selected ball and the Receivers' color prediction (the probability with which the Receivers think that the selected ball is **Red**). Specifically, the Receivers' stage earnings are as follows:

- When the selected ball is **Red**: $1 - (1 - \text{Receiver's prediction})^2$
- When the selected ball is **Gray**: $1 - (0 - \text{Receiver's prediction})^2$
- When the selected ball is **Red**, Receivers earn higher rewards the closer their prediction is to 1. The farther their prediction is from 1, the lower the reward.
- When the selected ball is **Gray**, Receivers earn higher rewards the closer their prediction is to 0. The farther their prediction is from 0, the lower the reward.
- As Receivers' predictions deviate from the actual color of the selected ball, their earnings decrease, and the rate of decrease also becomes larger.
- H-sender's earnings are higher (regardless of the color of the selected ball) if the Receiver predicts the higher probability of the selected ball being **Red**. Specifically, the H-sender's earnings are as follows:
 - $1 - (1 - \text{the paired Receiver's prediction})^2$

Your Earning in Each ROUND

Your earning for each round is determined as follows:

ROUND Earnings = **1,000 KRW** × [Stage 1 Earnings] + **20,000 KRW** × [Stage 2 Earnings]

That is, Stage 2 payoff carries **20 times** the weight of Stage 1 payoff.

[Page 9]

Information Feedback

At the end of each **round**, you will see a table consisting of two columns. Each column contains the following information for each stage:

The table at the end of each round will include the following information:

- (1) The color of the selected ball.
- (2) The message sent by the Sender.
- (3) The action chosen by the Receiver.
- (4) Your earnings.
- (5) (Receiver only) Whether the paired Sender was an H-sender or a C-sender.

[Page 10]

Your Final Cash Reward

Out of the 10 official rounds, one will be randomly selected, and the earnings you receive from that round will be your final cash reward. Therefore, it is recommended that you approach all official rounds with equal seriousness. Your earnings in this experiment will be the sum of the earnings from the selected round and a participation fee of 15,000 won.

1 Practice Round

Before starting the experiment in earnest, you will participate in 1 practice round. The practice round is part of the instructions and does not affect your earnings. The purpose of the practice round is to help you become familiar with the interface and the decision-making process. After the practice round ends, you will see the message "official round now begins!".

[Page 11]

Summary

In summary, each round proceeds in the following sequence:

- (1) At the start of the round, Senders and Receivers form random Sender-Receiver pairs.
 - Half of the Senders are H-senders, while the other half are C-senders.
 - H-senders earn higher payoffs as the Receiver predicts higher the probability of the selected ball being **Red**.
 - C-senders always send a message that is the same as the color of the selected ball.
- (2) Stage 1 begins. The computer randomly selects one of the two balls inside the box.
- (3) Senders decide the probabilities with which they will send the messages "**Red**" and "**Gray**" when the selected ball is **Gray** using a spinning wheel.
 - Receivers cannot know the specific spinning wheel chosen by Senders.
- (4) Senders, upon seeing the color of the selected ball, send the corresponding message to the Receiver.
 - If a **Red** ball is selected, a "**Red**" message is sent with 100% probability.
 - If a **Gray** ball is selected, the message is determined based on the spinning wheel.
- (5) Receivers receive the Sender's message and make a prediction about the probability that the selected ball is **Red**.
- (6) Receivers observe the color of the ball selected in Stage 1.
- (7) Stage 2 begins. The computer randomly selects one of the two balls inside the new box.
- (8) Senders and Receivers make decisions following the same procedure as in Stage 1.
- (9) Information feedback is provided at the end of the round.
 - Stage 2 payoff carries 20 times the weight of Stage 1 payoff.
- (10) The next round begins, and participants are grouped randomly again.

Appendix P. Experimental Instructions - Experiment II (Treatment GL)

[Page 1]

Same as in [Page 1] of Appendix O.

[Page 2]

Now let's consider a situation with two balls inside a box. One ball is **Red**, and the other ball is **Gray**.

At the beginning of each round, for each Sender-Receiver pair, the computer randomly selects one of the two balls from the box. The color of the ball selected is determined independently for each pair and each round, so there is no correlation between them.

The color of the ball chosen by the computer is revealed only to the **Sender**. The Receiver cannot see the color of the selected ball.

Now the decision-making process for the Sender and Receiver begins. Initially, Receivers will wait until all Senders have made their decisions.

[Page 3]

H-sender's Decision

Before seeing the color of the ball selected by the computer, each H-sender must decide whether to send the message "**Red**" or "**Gray**" to the Receiver they are paired with.

- To simplify the decision-making process, if the selected ball is **Gray**, the "**Gray**" message will be sent to the Receiver with a 100% probability.
- However, if the selected ball is **Red**, the H-sender will send the "**Red**" or "**Gray**" message to the Receiver based on the probability rules they have determined as follows:

The rest of [Page 3] is the same as in [Page 3] of Appendix O.

[Page 4]

[Bonus] H-sender's Belief Reporting

After submitting the spinning wheel, just before seeing the color of the ball selected by the computer, the Sender will be required to submit their predictions for the following. **The decisions made by the Sender on this screen will not affect any other part of the experiment and are solely used for additional compensation purposes.**

Let's consider the case where the selected ball is **Red**. You should submit predictions for the following two scenarios:

- First, you will use the following slider (Figure 29) to estimate the probability with which your paired Receiver thinks he is paired with an **H-sender** after the Receiver receives the "**Red**" message and sees that the selected ball is **Red**.
- Next, you will use the following slider to estimate the probability with which your paired Receiver thinks he is paired with an **H-sender** after the Receiver receives the "**Gray**" message and sees that the selected ball is **Red**.

You have selected 84%




FIGURE 29. Slider Bar for Belief Elicitation

After making your predictions, press the "Submit" button located at the bottom of the slider.

[Page 5]

H-sender's Decision (Continued)

- (1) Now, the H-sender will see the color of the ball selected by the computer, and the corresponding spinning wheel will also appear on the screen.
 - If the selected ball is **Gray**, you will see a spinning wheel that is 100% **Gray**.
 - If the selected ball is **Red**, you will see the spinning wheel you submitted (Figure 27).
- (2) In the center of the spinning wheel, you will see a spin button and a needle pointing at 12 o'clock. Pressing the spin button will rotate the spinning wheel.
- (3) When the spinning wheel stops (the stopping point is randomly determined), the color of the area indicated by the white needle will be the message delivered to the Receiver paired with you.

In summary, Receivers paired with H-senders receive messages as follows:

- When the computer selects a **Gray** ball: A 100% probability of receiving a "**Gray**" message.
- When the computer selects a **Red** ball: Messages with probabilities based on the colors you set on the submitted spinning wheel, either "**Red**" or "**Gray**."

After the wheel stops spinning, a "Continue" button will appear at the bottom of the spinning wheel. Don't forget to click this button to proceed to the next screen.

[Page 6]

C-sender's Decision

C-senders will transmit a "**Red**" message with a 100% probability regardless of the color of the ball selected by the computer.

Once all H-senders and C-senders have made their decisions, Receivers will move from the waiting screen to the decision screen **altogether**. This means that even if your paired Sender has made a decision, you will still have to wait on the waiting screen until all other Senders have completed their decisions.

Similarly, Senders will be on the waiting screen until all Receivers have completed their decisions.

[Page 7]

Receiver's Decision

Same as in [Page 6] of Appendix O.

[Page 8]

Receiver's Belief Reporting

Now, based on the messages received from the Sender and the color of the selected ball, the Receiver will use the following slider (Figure 29) to estimate the probability of their paired Sender being an H-sender.

After making your decision, press the "Submit" button located at the bottom of the slider.

Once all Receivers have completed their submissions, the round will come to an end.

[Page 9]

Receiver's Earning

- Receiver's earnings for the round are determined as follows:

$$\text{ROUND Earnings} = 1,000 \text{ KRW} \times [\text{Part 1 Earnings}] + 2,500 \text{ KRW} \times [\text{Part 2 Earnings}]$$

- Receivers' [Part 1 Earnings] are determined based on the difference between the actual color of the selected ball and their predicted color (the probability of them thinking the selected ball is **Red**). Specifically, Receivers' [Part 1 Earnings] are as follows:
 - When the selected ball is **Red**: $1 - (1 - \text{Receiver's prediction})^2$
 - When the selected ball is **Gray**: $1 - (0 - \text{Receiver's prediction})^2$
 - When the selected ball is **Red**, Receivers earn higher rewards the closer their prediction is to 1. The farther their prediction is from 1, the lower the reward.
 - When the selected ball is **Gray**, Receivers earn higher rewards the closer their prediction is to 0. The farther their prediction is from 0, the lower the reward.
 - As Receivers' predictions deviate from the actual color of the selected ball, their earnings decrease, and the rate of decrease also becomes larger.
- Receivers' [Part 2 Earnings] are determined based on how accurate their predictions are regarding the probability that their paired Sender is an H-sender in that round. Specifically,
 - When the Sender is an H-sender: $1 - (1 - \text{Receiver's prediction})^2$

- When the Sender is a C-sender: $1 - (0 - \text{Receiver's prediction})^2$
- Receivers earn higher rewards when their predictions are closer to 1 if the actual Sender is an H-sender.
- Receivers earn higher rewards when their predictions are closer to 0 if the actual Sender is a C-sender.

[Page 10]

H-sender's Earning

- H-senders' earnings for the round are determined as follows:

$$\text{ROUND Earnings} = 1,000 \text{ KRW} \times [\text{Part 1 Earnings}] + 2,500 \text{ KRW} \times [\text{Part 2 Earnings}]$$

- H-senders' [Part 1 Earnings] are determined based on the difference between the color predictions submitted by their paired Receiver and the actual color of the selected ball. Specifically, H-senders' [Part 1 Earnings] are as follows:
 - When the selected ball is **Red**: $1 - (1 - \text{the paired Receiver's prediction})^2$
 - When the selected ball is **Gray**: $1 - (0 - \text{the paired Receiver's prediction})^2$
 - In other words, H-sender's [Part 1 Earnings] are identical to Receiver's [Part 1 Earnings].
- H-senders' [Part 2 Earnings] are **increasing** in their paired Receiver's prediction on the probability of being paired with an H-sender. Specifically,
 - Part 2 Earnings = the paired Receiver's prediction
 - This means that [Part 2 Earnings] can also range from 0 to 1.
- Bonus: In addition to the round earnings mentioned above, if the selected ball is **Red**, H-sender receives a bonus based on how close their prediction regarding "Receiver's prediction of the probability that the Sender is an H-sender" was to the actual Receiver's prediction. The bonus is calculated as:
 - $500 \text{ KRW} \times [1 - (\text{the paired Receiver's prediction} - \text{the Sender's prediction regarding the paired Receiver's prediction})^2]$

[Page 11]

Information Feedback

At the end of each **round**, you will see a table containing the following information for that round:

The table at the end of each round will include the following information:

- (1) The color of the selected ball.
- (2) The message sent by the Sender.

- (3) The action chosen by the Receiver.
- (4) The Receiver's prediction for their paired Sender.
- (5) Your earnings.
- (6) (Receiver only) Whether the paired Sender was an H-sender or a C-sender.
- (7) (Sender only, [Bonus]) The Sender's prediction for "Receiver's prediction of the probability that the Sender is an H-sender."

[Page 12]

Same as in [Page 10] of Appendix O.

[Page 13]

Summary

In summary, each round proceeds in the following sequence:

- (1) At the start of the round, Senders and Receivers form random Sender-Receiver pairs.
 - Half of the Senders are H-senders, while the other half are C-senders.
 - H-senders earn higher payoffs the more accurately the Receiver predicts the color of the selected ball.
 - C-senders always send a fixed "Red" message.
- (2) The computer randomly selects one of the two balls inside the box.
- (3) Senders decide the probabilities with which they will send the messages "Red" and "Gray" when the selected ball is Red using a spinning wheel.
 - Receivers cannot know the specific spinning wheel chosen by Senders.
- (4) Senders, upon seeing the color of the selected ball, send the corresponding message to the Receiver.
 - If a Gray ball is selected, a "Gray" message is sent with 100% probability.
 - If a Red ball is selected, the message is determined based on the spinning wheel.
- (5) Receivers receive the Sender's message and make a prediction about the probability that the selected ball is Red.
- (6) Receivers observe the color of the selected ball.
- (7) Receivers predict the probability that the paired Sender is an H-sender.
- (8) Information feedback is provided at the end of the round.
- (9) The next round begins, and participants are grouped randomly again.