Attempting to detect a lie: Do we think it through?*

Iuliia Grabova[†] Hedda Nielsen[‡] Georg Weizsäcker[§]

17 June 2023

Abstract

Our experiment measures belief hierarchies regarding a message that may be a lie. In a two-player game between a sender and a receiver, the sender knows the state of the world and has a transparent incentive to deceive the receiver. The receiver chooses a binary reaction. For a wide set of non-equilibrium beliefs, the reaction and the receiver's second-order belief should dissonate: she should follow the sender's statement if and only if she believes that the sender believes that she does not follow the statement. The opposite is true empirically, constituting a new pattern of inconsistency between actions and beliefs.

JEL-classification: D01, D83
Keywords: Strategic information transmission; lying; higher-order beliefs.

*The authors thank Dorothea Kübler and Ronald Peeters for helpful discussions, the German Science Foundation for financial support via CRC TRR 190 (project number 280092119) and the Einstein Foundation Berlin for financial support via the Einstein Visiting Fellowship of Bertil Tungodden. The data analysis was preregistered as AsPredicted #109270, available at https://aspredicted.org/SML_RN3.

[†]Humboldt-Universität zu Berlin and DIW Berlin, graboviu@hu-berlin.de
[‡]Humboldt-Universität zu Berlin, hedda.nielsen@hu-berlin.de
[§]Humboldt-Universität zu Berlin, weizsaecker@hu-berlin.de

1 Introduction

A robust finding in social psychology is that most people are bad at lie detection. Success rates of identifying a truth as a truth, and a deception as a deception, are close to the level that one can achieve by flipping a coin (Bond and DePaulo, 2006), (Vrij, 2008). In this paper, we focus on the receiver's second-order belief as a factor in this weak performance. Judging a statement's veracity requires judging the sender's incentives, i.e., the extent to which he would benefit from each of his possible statements. Yet, this judgement is about the sender's *subjective* incentives – how much does he believe to benefit? – and it thus depends on his belief about his audience's reaction. That is, in a two-person setup the receiver should aim to predict what he, the sender, believes her, the receiver, to do in response to the possible statements.[1]

For constant-sum games, where the payoff rules prescribe that one person wins if and only if the other person loses, the logic of best response generates an interesting prediction about such beliefs: the receiver chooses her reaction to the statement so that it is inconsistent with her second-order belief. If she believes that the sender believes that she trusts his statement (hence, that he can deceive her with a lie), then she should not trust it. Conversely, if she believes that he believes that she does not trust the statement, then she should trust it. We test this prediction in a two-player communication game, and find it to be violated: the empirical correlation between second-order beliefs and choices is significantly positive (Spearman coefficient of 0.2). This pattern not only contradicts the above-stated prediction but is also sub-optimal in the sense of missing that the senders, in fact, aim to exploit credulity: if they believe that the receiver is more likely to trust than not to trust, they are more likely to lie, by a difference of close to 50 percent of the average lying rate. This correlation is lost on the receivers – a player-role specific inability to take another person's perspective, and to think it through.[2]

Relative to the literature that measures beliefs in games, this highlights a new kind of inconsistency. In communication games, the receiver's relevant first-order belief, upon receipt of the statement, is about the statement's truth. This belief has a close tie to the receiver's action – to follow the statement or not – and indeed we find that the re-

[1] For an extensive classification of belief hierarchies in communication, see the book manuscript Weizsäcker (2023).

[2] The correlations that we measure among the senders' beliefs far more accurately predict the receiver.

ceiver's choice tends to be highly consistent with her first-order belief about the truth. This consistency is *larger* than in normal-form games where the first-order belief is about an opponent's action (Costa-Gomes and Weizsäcker, 2008; Rey Biel, 2009; Polonio and Coricelli, 2019). In contrast, our data show that a disconnect arises between first-order beliefs and second-order beliefs: the receiver's belief about the truth tends to be inconsistent with her belief about the sender's first-order belief. This connection between beliefs of first order and second order was not a focus of previous studies on players' beliefs in games. The more specialized literature on beliefs in communication games has, with only one exception that we are aware of, not made quantitative measurements of the receiver's second-order beliefs – perhaps due to methodic concerns about the measurement of second-order beliefs.[3] This highlights the need for an appropriate data context. We build our measurement on a game that has proven to be suitable, by Peeters et al. (2015).[4] Their game is constant sum and makes the sender's motive to deceive highly transparent to the participants. It uses a particular feature to facilitate belief measurements: the game is mirror-symmetric in the sense that a lie about one state of the world is the mirror image of a lie about the other state of the world. Therefore, and under the mild assumption of label independence of the strategies, asking very few questions per participant suffices for a full elicitation of a belief hierarchy, up to second-order beliefs. This is further explained in the next section, with details on our experimental design. Subsequent sections contain, respectively, our (pre-registered) hypotheses, the experimental results and our brief conclusion.

2 Experiment

Two players, sender and receiver, interact anonymously in a one-shot fashion. The sender knows which of two equally probable states of the world, A or B, occurs. Each state corresponds to a payoff table, reproduced in Tables 1 and 2. The sender sends a message

[3]The exception is Agranov and Schotter (2020) who study differences in second-order beliefs about truth telling in market exchange with and without competition. In games without communication, second-order beliefs were measured more widely. See, e.g., Manski and Neri (2013) and the discussion in Schotter and Trevino (2014).

[4]The authors of the original study do elicit second-order beliefs, but only for the sender (owing to the different nature of their research question). Their existing measurements provide us with benchmarks for our results regarding the variables that we duplicate, while allowing the inclusion of a new variable – the receiver's second-order belief.

indicating the state of the world: either she announces "Table A has been selected", or "Table B has been selected". The receiver reads the message and chooses a column in the (unknown) payoff table, either Option A or Option B. Payoffs are such that the receiver wants to learn the truth and the sender wants her to miss it: one and only one player wins the game – i.e., obtains the high payment – and the receiver wins if and only if she matches her choice to the identity of the table. To aid the transparency of deception incentives, the instructions are explicit in raising the possibility that the sender can send a non-truthful message at his own will. All of these procedures are equivalent to those in the original experiment by Peeters et al. (2015).[5]

Table A

	Option A	Option B
Sender	4	12
Receiver	12	4

Table B

	Option A	Option B
Sender	12	4
Receiver R	4	12

In addition to playing the game, the participants report their first order-beliefs and their second order beliefs. As they act in different player roles, these beliefs are role specific. For first-order beliefs, the sender indicates his subjective probability of the event that the receiver follows the message – chooses the option with the label that is indicated in the sender's message – and the receiver indicates her subjective probability of the event that the message corresponds to the truth. For second-order beliefs, each player reports his or her subjective expectation of the opponent's answer to her or his first-order question.[6] The payment for the belief tasks rewards accuracy via the Binarized Scoring Rule (Hossain and Okui, 2013) and the instructions indicate the rule's incentive compatibility in a way that is transparent while giving the participants the option to skip over the details.

As mentioned earlier, the fact that the game is symmetric in labels A and B is key. Due

[5]The only substantive differences between the design of Peeters et al. (2015) and ours, apart from the elicitation of the receiver's second-order belief, is that we have a mildly larger number of observations and that we use different payments for the belief variables.

[6]Full instructions are available in the Appendix. The choice of the precise belief hierarchies that are elicited corresponds to established practice in the literature, see e.g. the survey in Weizsäcker (2023). Note that the second-order belief is elicited as a point belief. For full generality of the belief hierarchy, this belief would be a distribution over the possible distributions that the first-order belief can assume. However, with only two payoffs in the game, the maintained assumptions of the next section guarantee that the mean of the distribution suffices to predict behavior.

to this feature, it suffices to elicit all responses under one out of two scenarios and impute responses for the other scenario, by using the mirrored label for all states, messages and choices.[7] Like in the experiment of Peeters et al. (2015), the instructions explain this property in simple words. The procedure is equivalent to a full elicitation of all possible scenario-contingent actions and the beliefs about them, under the assumption that strategies and beliefs are label independent.

A total of 251 participants are matched in pairs, in 12 sessions.[8] Like in Peeters et al. (2015), each participant plays the game once in each role, for a total of two games, with fixed partners but without any feedback after the first game. The chronological order of playing the two games is randomized across participants. All payoffs in the tables are in euro amounts. For each participant, one of the two games is randomly selected as payoff relevant at the conclusion of the experiment, and one of three tasks is being paid: the actual game, the first-order belief task, or the second-order belief task. The payment occurs in addition to paying a participation fee of 5 euro per person.

3 Hypotheses

Notating the two players' indexes by $\{s, r\}$, let the scenario of the sender be the state of the world, $\theta_s \in \{A, B\}$ and the scenario of the receiver be the message $\theta_r \in \{A, B\}$, with the interpretation that the sender knows the true state θ_s and the receiver hears the message "Table θ_r has been selected." The two players' families of actions are denoted as $a_i(\theta_i) \in \{0, 1\}$, for $i \in \{s, r\}$, where the action with value 1 is, in each case, the action that corresponds to the players' scenario: the sender tells the truth and the receiver follows the message. The players' first-order beliefs are $b_i^1(\theta_i) \in [0, 1]$ (a belief about a_{-i}) and their second-order beliefs are $b_i^2(\theta_i) \in [0, 1]$ (a belief about the mean of b_{-i}^1), for $i \in \{s, r\}$. The assumption of label independence, which we henceforth make, is that $a_i(A) = a_i(B) =: a_i$,

[7] We ask for the sender's message in the scenario that A is the state of the world, and impute the message for the case that the state is B. For the receiver's action we ask what option she chooses if she receives the message that Table A was selected, and impute her choice for the scenario that the message indicates B. For all belief statements, we ask for the beliefs about the other player's behavior under the scenario that is used in the instructions, and impute the beliefs for the other scenario as the correspondingly mirrored beliefs.

[8] In sessions with odd numbers of participants, one person's choices were payoff irrelevant for the other participants, although he/she obtained payments as if matched with one of the others. In the preregistration, we indicated a maximum number of 245 participants. Turnout was slightly higher than expected and we decided to work with all data before looking at them.

$b_i^1(A) = b_i^1(B) =: b_i^1$ and $b_i^2(A) = b_i^2(B) =: b_i^2$, for $i \in \{s, r\}$.[9] To denote distributions of actions, $\sigma_i(a_i)$ describes the probability of player i choosing action a_i.

The experimental observations (including beliefs) that correspond to the game's unique Weak Perfect Bayesian Equilibrium are $\sigma_i(a_i) = b_i^1 = b_i^2 = \frac{1}{2}$. The equilibrium is uninformative, prescribing that sender and receiver each randomize with equal probability in each scenario. Yet, the more interesting case for the analysis arises for non-equilibrium beliefs. We continue to maintain the assumption that the players maximize their subjective expected utility (SEU) and that this is twice-mutually known by the players, but we relax the assumption that beliefs are in equilibrium. Since the SEU assumption implies that players maximize their subjectively perceived probabilities of receiving the high payment in the game, we can straightforwardly predict the connection between actions and first-order belief:

Hypothesis 1: $\sigma_s(a_s)$ decreases in b_s^1 and $\sigma_r(a_r)$ increases in b_r^1.

A remark is in order about the fact that the formulation of Hypothesis 1 is weaker than the corner solution that SEU predicts: if $b_s^1 > \frac{1}{2}$, then $\sigma_s(a_s) = 0$, and if $b_r^1 > \frac{1}{2}$, then $\sigma_r(a_r) = 1$. The reasons for the weaker formulation of the hypothesis is (a) that it covers non-degenerate choice frequencies in the experiment, and (b) that the weaker formulation allows for possible modifications of the players' objective function. For instance, adding the assumption of a random utility perturbation of player i's two actions that is uncorrelated with b_i^1 (leading to logistic choice or similar choice models) would be covered by the hypothesis. Likewise, any utility shifters that describe a systematic preference for or against any action profile (a_s, a_r) that is independent of b_s^1 and b_r^1 is covered by the hypothesis. This includes many natural formulations of betrayal aversion, guilt aversion, or of direct preferences for or against stating lies or expressing distrust.

Under the maintained assumptions, the first-order belief is a best response to the second-order belief because each player expects the other player to act in congruence with Hypothesis 1:

Hypothesis 2: b_s^1 increases in b_s^2 and b_r^1 decreases in b_r^2.

Part (b) of the above-made remark on the "weaker" hypothesis formulation applies to

[9] Notice that the assumption's two sets of restrictions on beliefs correspond, respectively, to mutual first-order knowledge knowledge and mutual second-order knowledge of the property that $a_i(A) = a_i(B)$, for $i \in \{s, r\}$.

Hypothesis 2, too. Due to utility perturbations, a change in the a player's second-order belief may not induce a jump to the corner solution for the first-order belief, but the direction of the change applies nevertheless and is described by Hypothesis 1.

Combining the two hypothesis yields the connection between actions and second-order beliefs that we aim to test in this paper, chiefly for the receiver:

Hypothesis 3: $\sigma_s(a_s)$ decreases in b_s^2 and $\sigma_r(a_r)$ decreases in b_r^2.

4 Results

The following table shows the data averages of actions and beliefs in both player roles. For each participant, we use only the observations from one of the two player roles: the one that he or she holds in his or her first game.[10] Average responses are denoted as $\overline{a}_s$, $\overline{b}_s^1$, etc., for the empirical means of the distribution of the sender's action, her belief b_s^1, etc.

	Actions		Beliefs			
	$\overline{a}_s$	$\overline{a}_r$	$\overline{b}_s^1$	$\overline{b}_s^2$	$\overline{b}_r^1$	$\overline{b}_r^2$
Sender	0.6667		0.5038	0.4446		
	(0.0002)		(0.8160)	(0.0008)		
Receiver		0.6563			0.5477	0.5305
		(0.0004)			(0.0549)	(0.1870)

In parentheses: p-values of the Wilcoxon signed-rank tests against a value of 0.5, two-sided.

Table 1: Data averages

The table shows that about two thirds of the senders tell the truth and about two thirds of the receivers follow the message. While the best response of receiver participants to the behavior of sender participants would have been to always follow the message,[11] the average beliefs reported in the table appear to rationalize the actions fairly well: the receivers' average first-order beliefs are that sender tell the truth in half of the cases,

[10] This data restriction rules out order effects. Some such effects are detectable in our data set and our pre-plan specified, for this case, to include only the data from each participant's first game in the main analysis. All results are qualitatively robust to including all data.

[11] The observation that both senders and receivers are "truth-biased" is consistent with a large set of observations in the lying literature (Bond and DePaulo, 2006; Levine, 2014). All of our measurements are consistent with, indeed very close to, those in Peeters et al. (2015).

which makes the average receiver indifferent between following and not following. These first-order beliefs, in turn, can be justified by second-order beliefs: on average, receivers predict that the senders expect the receivers to trust the message with probability one half. Moreover, these second-order beliefs are quite accurate, on average.

Our main interest lies, however, in the correlation between the variables, in the sense of the previous section's hypotheses. The following table reports the Spearman correlation coefficients for the variables of each player role, with statistical significance of less than 0.05, 0.01 and 0.001 (two-sided) indicated by one, two or three asterisks, respectively.[12]

Sender				**Receiver**			
	a_s	b_s^1	b_s^2		a_r	b_r^1	b_r^2
a_s	1.000			a_r	1.000		
b_s^1	-0.205*	1.000		b_r^1	0.469***	1.000	
b_s^2	0.033	0.467***	1.000	b_r^2	0.203*	0.211*	1.000

The correlations show that Hypothesis 1 is supported for both player roles, that Hypothesis 2 is supported for the sender but rejected for the receiver, and that Hypothesis 3 is neither rejected nor supported for the sender but rejected for the receiver. In particular, the tables show a striking asymmetry between the belief hierarchies of the sender and those of the receiver: while the senders correctly anticipate a strongly positive relation between the receivers' first-order beliefs and their actions (correlations of about 0.47 in each case), the receivers' views are off target: the correlation between b_r^1 and b_r^2 is significantly positive whereas the correlation that it tries to predict, that of a_s and b_r^1, is significantly negative (both around 0.2, in absolute terms). The receivers fail to understand that senders tend to exploit the credulity of receivers.

To assess the magnitude of the effect, it is useful to consider only those receiver participants whose second-order belief expresses a non-zero tendency – that is, we drop the 12 percent of receivers who report a second-order belief of $b_r^2 = 0.5$ – and to ask how differently these groups act depending on the sign of their tendency. For the participants with $b_r^2 > 0.5$, the frequency of following the message is 72 percent, and for the participants

[12]Having pre-formulated and derived our hypotheses as directed hypotheses, we might have considered one-sided test as appropriate. However, the fact some test statistics deviate from zero in the opposite direction makes it arguably easier to interpret two-sided tests. In the text we describe the hypotheses as "rejected" in these cases, and as "neither confirmed nor rejected" in the case of a zero test statistic.

with $b_r^2 < 0.5$ (who should have a higher propensity to follow the message, according to Hypothesis 3) the average frequency of following is merely 60 percent. A corresponding separation of senders, however, shows that, in fact, senders lie much more often if they expect the receiver to have a positive tendency to follow: senders with $b_s^1 > 0.5$ show a lying rate of 42 percent, versus 25 percent for the senders with $b_s^1 > 0.5$ for these two groups.

One may attempt to explain the violation of Hypothesis 3 – the positive correlation between the receivers' actions and their second-order belief – by an illusion of transparency: perhaps the receivers believe (too much) that the sender can detect their inclination to follow the message or not. We point out, however, that this explanation fails to explain a key pattern: that the correlation between b_r^1 and b_r^2 violates Hypothesis 2. Any second-order belief, even an illusion-of-transparency-laden one, should generate a first-order belief that aims to exploit it. The receivers' first-order beliefs tell the opposite story. The inconsistency of b_r^1 and b_r^2 is, thus, a separate phenomenon from any illusion of transparency.

To examine the correlations of our experimental variables in a bi-variate choice model, the following table reports Probit regressions for a_s and a_r. It shows again that the correlation between the receiver's action and her second-order belief is positive, not negative. Moreover, the regressions show that if one controls for first-order beliefs, the second-order beliefs have little predictive power. This is consistent with our maintained assumptions of SEU maximization, where the only way in which second-order beliefs correlate with actions is via their role as the basis for first-order beliefs.

5 Conclusion

A striking pattern of our results is that the sender appears to be "smarter" than the receiver: the inconsistency between second-order beliefs on the one hand, and actions and first-order beliefs on the other, appears for the receiver but not for the sender. The receiver does not appear to think the situation through and fails to predict that a sender who believes in the receiver's gullibility will exploit this. In contrast, the sender, while not confirming all hypotheses, largely follows the subjectively perceived incentives that

Table 2: Probit Regressions

	a_s			a_r		
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	1.0282***	0.3170	0.7230*	-1.0048**	-0.1140	-1.3203**
	(0.2730)	(0.2954)	(0.3186)	(0.3334)	(0.2423)	(0.4312)
b_i^1	-1.1524*		-1.4688**	2.7333***		2.6142***
	(0.4746)		(0.5278)	(0.6329)		(0.6503)
b_i^2		0.2569	1.0414		1.0056*	0.7414
		(0.6122)	(0.7243)		(0.4085)	(0.4585)
Observations	123	123	123	128	128	128

Note: Standard errors in parentheses. Data include only one game per participant.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

arise under his beliefs.

The behavioral/experimental literature on biased beliefs in communication cannot, to our knowledge, explain this asymmetry. To the extent that this literature focuses on the role of second-order beliefs, it has an emphasis on lying cost and guilt aversion, as in Charness and Dufwenberg (2006), Kartik (2009) and Gneezy et al. (2018), or level-k reasoning as in Crawford (2003), but the literature does yet offer a wide set of hypotheses on other out-of-equilibrium second-order beliefs. Recent contributions include Cohen and Li (2022) and Fong et al. (2023) who extend the notion of Eyster and Rabin (2005) cursed equilibrium to extensive-form games, including communication games. However, at least the theory of Cohen and Li (2022) would, if anything, tend to predict that the receiver's beliefs are more accurate than the sender's, not vice versa. It may be useful to examine these and other theories with respect to the conditions under which the receiver may show an inconsistency that the sender does not show.[13] The level-k analyses of communication by Crawford (2003), make an observation that may be related to our result, however: if the thought process of a player starts with the instinctive level-0 behavior that the sender is truthful and the receiver is credulous, then a pair of (perhaps plausible) level-1 behaviors would be for the sender to lie and for the receiver to follow the sender's message. This illustrates that communication games, by their nature, create an asymmetry in "smartness" between the two player roles, also in the sense of level-k

[13]Certainly, there are numerous other unexplored ways in which established biases in mental models may generate hypotheses about directed effects in communication. For instance, a simple prediction appears about overoptimism or motivated beliefs: a sender overestimates how well he is understood by the receiver if the incentives are aligned, and underestimates it if they are not aligned.

analyses.[14]

References

Agranov, Marina, D. U. and A. Schotter (2020): "Trust me: Communication and Competition in Psychological Games," Unpublished manuscript.

Bond, C. F. J. and B. M. DePaulo (2006): "Accuracy of Deception Judgments," *Personality and Social Psychology Review*, 10, 214–243.

Charness, G. and M. Dufwenberg (2006): "Promises and Partnership," *Econometrica*, 74, 1570–1601.

Cohen, S. and S. Li (2022): "Sequential Cursed Equilibrium," Unpublished manuscript.

Costa-Gomes, M. A. and G. Weizsäcker (2008): "Stated Belief and Play in Normal-Form Games," *Review of Economic Studies*, 75, 729–762.

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

Crawford, V. P., M. A. Costa-Gomes, and N. Iriberri (2013): "Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications," *Journal of Economic Literature*, 51, 5–62.

Eyster, E. and M. Rabin (2005): "Cursed Equilibrium," *Econometrica*, 73, 1623–1672.

Fong, M.-J., P.-H. Lin, and T. R. Palfrey (2023): "Cursed Sequential Equilibrium," Unpublished manuscript.

Gneezy, U., A. Kajackaite, and J. Sobel (2018): "Lying Aversion and the Size of the Lie," *American Economic Review*, 108, 419–453.

Hossain, T. and R. Okui (2013): "The Binarized Scoring Rule," *Review of Economic Studies*, 80, 984–1001.

[14]A level-k analyses cannot directly apply to our game, however, as it does not make a prediction about how beliefs are stated for orders of beliefs that are higher than one's own k. For more references on level-k analysis in communication games, see Crawford et al. (2013).

Kartik, N. (2009): "Strategic Communication with Lying Costs," *Review of Economic Studies*, 76, 1359–1395.

Levine, T. R. (2014): "Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection," *Journal of Language and Social Psychology*, 33, 379–392.

Manski, C. F. and C. Neri (2013): "First- and second-order subjective expectations in strategic decision-making: Experimental evidence," *Games and Economic Behavior*, 81, 232–254.

Peeters, R., M. Vorsatz, and M. Walzl (2015): "Beliefs and truth-telling: A laboratory experiment," *Journal of Economic Behavior and Organization*, 113, 1–12.

Polonio, L. and G. Coricelli (2019): "Testing the level of consistency between choices and beliefs in games using eye-tracking," *Games and Economic Behavior*, 113, 566–586.

Rey Biel, P. (2009): "Equilibrium Play and Best Response to (Stated) Beliefs in Normal Form Games," *Games and Economic Behavior*, 65, 572–585.

Schotter, A. and I. Trevino (2014): "Belief Elicitation in the Laboratory," *Annual Review of Economics*, 6, 103–128.

Vrij, A. (2008): *Detecting Lies and Deceit: Pitfalls and Opportunities*, John Wiley & Sons.

Weizsäcker, G. (2023): "Misunderstandings: False Bliefs in Communication," Unpublished manuscript.

A Appendix: Instructions

A.1 General Instructions

Thank you for participating in this experiment. For completing the experiment you will receive 5 euros. Additionally, you will have the possibility of earning up to 12 euros depending on your decisions and answers throughout the experiment.

Participation will take around 45 minutes. The experiment consists of three parts: instructions, the experiment itself, and payment.

Please read through the instructions carefully to make sure that you have fully understood the task and the questions. At the end of the instructions you will be asked to answer a few short questions to check your understanding. All your choices and answers are anonymous and will remain confidential. They will be used for research purposes only.

We kindly ask you to respect the following rules. Please put aside your mobile phone and do not use it until the end of this experiment. Do not use your computer, laptop or other electronic devices for any purposes that are not connected with the experiment. Please do not engage in conversations with anyone apart from the experimenters. If you have any questions, please use the Zoom chat to write a personal message to the "Experimenter". If for any reason you do not complete all the questions of the experiment you will only be paid the participation fee.

A.2 Experiment Instructions

A.2.1 Procedure

In this experiment, we simulate an interaction between two individuals where Person S (*Sender*) sends a message to Person R (*Receiver*). Person S will have to choose one of two possible messages to send, and Person R will have to choose one of the two possible reactions to the message, which will later be called *Option A* and *Option B*.

At the start of the experiment, you will be randomly matched with another participant. Neither you nor the participant you are matched with will learn the identity of his/her match. Moreover, you will not learn your role until the end of the experiment. This

means that you will have to make decisions as both Person S and Person R.

Table A and Table B represent two possible scenarios of the experiment. Which of the two is the relevant scenario is uncertain at the beginning of the experiment – it is selected at random by the computer. The numbers in the tables represent payoffs of Person S and Person R, in euro amounts.

Table A

	Option A	Option B
Person S	4	12
Person R	12	4

Table B

	Option A	Option B
Person S	12	4
Person R	4	12

The payoffs depend on the table selected by the computer (*Table A* or *Table B*) and the choice made by Person R (*Option A* or *Option B*). Person R earns more money if and only if option and table coincide.

For example, if Person R chooses *Option A* and *Table A* is randomly chosen, Person R gets 12 euros and Person S gets 4 euros as a payoff. In case that Person R chooses *Option A* and *Table B* is randomly chosen, Person R gets 4 euros and Person S gets 12 euros as a payoff. And so on, for the two cases where Person R chooses *Option B*.

Not also: Following this payoff rule, Person S, who sends the message, is always better off if Person R chooses an option that does not coincide with the table selected by the computer.

The experiment proceeds as follows. The computer randomly selects one of the two tables with equal probability. Only Person S is informed about the table that has been selected by the computer. Person S chooses which of the following two messages to send to Person R:

- "Table A has been selected",
- "Table B has been selected".

Note, Person S is completely free in the choice of her message.

Person R observes the message from Person S and chooses one of the two options:

- Option A,

- Option B.

Note, also Person R is completely free to choose any of the two options.

After making choices about messages and options, both Person S and Person R are asked to answer two further questions. First, you are asked to estimate the other person's action. Second, you are asked to predict what the other person estimates your action to be. For both questions, the closer your answer is to the true value, the larger your payoff will be.

Altogether, in the course of the experiment, you will first have to answer three questions as Person S and then three questions as Person R.

A.2.2 Payment

Your payment from the experiment consists of two parts: a fixed payment (participation fee) of 5 euros and a variable payment that depends on your answers.

To determine the amount of the variable payment, the computer will randomly select your role (Person S or Person R) and one of the three questions corresponding to this role.

If the first question - the question about your choice (message for Person S, option for Person R) - is selected, the variable payment is calculated according to the payoff table. Depending on the random choice of the table (A or B) made by the computer and the choice of the option (A or B) made by Person R (either you or a participant matched with you), you will earn either 4 euros or 12 euros.

If the second question or the third question - the questions about expectations - is selected, the variable payment is calculated using a rule called the Binarized Scoring Rule. According to this rule, you can also earn 12 or 4 euros and the probability of earning a high payoff (12 euros) increases if your answer is closer to the true value; conversely, the probability of earning a low payoff (4 euros) increases if your answer is further from the true value. While you need not understand the exact algorithm of the Binarized Scoring Rule, a full description of it can be found below, in the section labelled "Supplement: ...". All you have to know is that the rule makes it optimal for all participants to state their true beliefs in response to the second question and to the third question.

A.3 Supplement: Binarized Scoring Rule

This section describes the payment rule for the second and third questions. In these two questions you are not asked to make a choice as in the first question. Instead, you need to estimate the answer of another person on a scale from 0 to 100. These estimates show how likely you consider the other person to provide a particular answer. Here, their answer is either a choice or an estimate of the other person.

The expected payment for either of two questions increases with the accuracy of your estimate. The measure of the realized error in your estimate (subsequently referred to by the letter l) is calculated as follows: $l = ((x - \theta)/100)^2$, where x is your guess about the answer of another person matched with you, and θ is the actual answer of this person. Thus, l measures the distance between your estimate and the actual answer of the other person.

To explain θ: if you are asked to estimate a *choice* of the other person (which you are in the second question), then θ is either 0 or 100, depending on the actual choice of this person. Your decision screen will clarify which of the possible choices of the other person corresponds to 0 and which to 100. If you are asked to guess an *estimate* made by the other person (in the third question), θ is the actual estimate provided by the other person.

Your payment is calculated as follows. The computer draws at random an additional integer number z between 0 and 100, with equal probability for each integer. If the error measure l is strictly less than $z/100$ ($l < z/100$), you receive 12 euros. If l is greater or equal than $z/100$ ($l \geq z/100$), you receive 4 euros.

It follows from this rule that it is optimal for you to enter a relatively high number x (close to 100) if you think that the answer θ of the other person is large. Conversely, it it optimal for you to enter a relatively small number if you think that the answer of the other person matched with you is small. In this way, you maximize the probability of receiving 12 euros for every possible realization of the integer z. Please note: how large or small your optimal x is, depends on your precise assessment of the how likely the other person chooses each of their possible answers.

With this in mind, the Binarized Scoring Rule implies that it is always

optimal for the participants to truthfully state their own estimate. This was proven by Hossain and Okui (2013).[15]

A.3.1 Understanding checks

1. What is the outcome for Person S if *Table B* is randomly selected by the computer, Person S sends message *"Table A has been selected"*, and Person R chooses *Option B*?

    - Options: 4 , 8 , 12

2. What is the outcome for Person S if *Table B* is randomly selected by the computer, Person S sends message *"Table B has been selected"*, and Person R chooses *Option B*?

    - Options: 4 , 8 , 12

3. From the perspective of which role will you need to make decisions?

    - Options: Person S, Person R, both

4. Is the probability of you earning a higher payoff from the questions about expectations decreasing, increasing or independent of you giving an answer that is closer to the true value?

    - Options: decreasing, increasing, independent

A.4 Experiment

[Text in box remains on left hand side of screen throughout experiment]

[15] Tanjim Hossain, Ryo Okui; The Binarized Scoring Rule, *The Review of Economic Studies*, Volume 80, Issue 3, 1 July 2013, Pages 984 – 1001.

Table A			Table B		
	Option A	Option B		Option A	Option B
Person S	4	12	Person S	12	4
Person R	12	4	Person R	4	12

1. Payoffs from the experiment depend on either *Table A* or *Table B*
2. The computer will randomly select one of these tables (each with equal probability)
3. Only Player S is told which table is selected
4. Player S sends one of the following messages to Player R: *"Tables A has been selected"* or *"Table B has been selected"*
5. Player R is asked to choose either *Option A* or *Option B*
6. If the option chosen coincides with the table selected by computer, Player S gets 4 euros and Player R receives 12 euros. If they do not coincide, the payoffs are reversed.

A.4.1 Person *S*

You have now been matched with another participant of the experiment.

Question 1S

Suppose you are Person S and the computer has randomly selected *Table A*.

Which message do you send to Person R?

- "Table A has been selected"
- "Table B has been selected"

NOTE: Above you only chose a message to send when Table A is selected. To simplify the experiment, we use symmtry: to determine payoffs, we assume that if you send "Table A has been selected" in the case above, then by symmetry you would send "Table B has been selected" if Table B was actually selected. If instead you send "Table B has been

selected" in the case above, then we assume that you would send "Table A has been selected" if Table B was actually selected.

You will now be asked to estimate the answers of the Person R with whom you are matched. More precisely, you will be asked to indicate how likely you consider a given event or answer by Person R to be. Remember, stating a likelihood is equivalent to stating how many out of 100 Person Rs you think would choose a certain answer.

Question 2S

Suppose you are Person S and that you sent the message "Table A has been selected".

How likely do you think it is that Person R chooses Option A?

☐ out of 100

Question 3S

Suppose you are Person S.

Person R is asked the following question: "Suppose you are Person R and that Person S sent you the message "Table A has been selected". How likely do you think it is that Table A was randomly selected by the computer?".

You, as Person S, are asked to estimate the answer given by Person R. Please read the question in the previous paragraph once again and estimate: what answer do you think does Person R gives to this question?

☐ out of 100

A.4.2 Person *R*

Remember, you are still matched with the same other participant of the experiment.

Question 1R

Now, suppose instead you are Person R and that Person S sent you the message "Table A has been selected".

Which option do you choose?

- Option A

- Option B

NOTE: Above you only chose an option for when you receive the message "Table A has been selected". Just as before, we simplify and use symmetry: to determine your payoffs, we assume that if you choose Option A in the case above, then you would choose Option B if you receive the message "Table B has been selected". If instead you choose Option B in the case above, then we assume you would choose Option A if you receive the message"Table B has been selected".

You will now be asked to estimate the answers by Person S. More precisely, you will be asked to indicate how likely (out of 100) you consider a given event or answer by Person S to be.

Question 2R

Suppose you are Person R and that Person S sent you the message "Table A has been selected".

How likely do you think it is that Table A was randomly selected by the computer?

☐ out of 100

Question 3R

Suppose you are Person R.

Person S is asked the following question: "Suppose you are Person S and that you sent the message "Table A has been selected". How likely do you think it is that Person R chooses Option A?"

You, as Person R, are asked to estimate the answer given by Person S. Please read the question in the previous paragraph once again and estimate: what answer do you think does Person S give to this question?

☐ out of 100