Pinocchio's Pupil: Using Eyetracking and Pupil Dilation

To Understand Truth-telling and Deception in Sender-Receiver Games

Joseph Tao-yi Wang, Michael Spezio and Colin F. Camerer*

Abstract

We report experiments on sender-receiver games with an incentive for senders to exaggerate. Subjects "overcommunicate"— messages are more informative of the true state than they should be, in equilibrium. Eyetracking shows that senders look at payoffs in a way that is consistent with a level-k model. A combination of sender messages and lookup patterns predicts the true state about twice as often as predicted by equilibrium. Using these measures to infer the state would enable receiver subjects to hypothetically earn 16-21 percent more than they actually do, an economic value of 60 percent of the maximum increment.

Keywords: *Cheap talk, Truth-bias, Lie detection, behavioral game theory, eyetracking, experimental economics, behavioral economics*

JEL codes: *C72*, *C92*, *D82*

Wang: Department of Economics, National Taiwan University, 21 Hsu-Chow Road, Taipei 100, Taiwan (email: josephw@ntu.edu.tw); Spezio: Division of Humanities and Social Sciences, California Institute of Technology, MC228-77, 1200 E California Blvd, Pasadena, CA 91125 and Department of Psychology, Scripps College, 1030 Columbia Avenue, Claremont California, 91711 (e-mail: mlspezio@hss.caltech.edu); Camerer: Division of Humanities and Social Sciences, California Institute of Technology, MC228-77, 1200 E California Blvd, Pasadena, CA 91125 (e-mail: camerer@hss.caltech.edu). Research support was provided by grants from the Betty and Gordon Moore Foundation, and the Human Frontiers of Society Program (HFSP). Alex Brown, Noah Myung, Eileen Chou, Min Jeong Kang, Dan Knoepfle and Chun-Ting Chen provided excellent research assistance. Thanks to comments from Vincent Crawford, Moran Cerf, Robert Östling, Navin Kartik, David Eil, Carlos R. Lever, two anonymous referees, and the audience of the ESA 2005 Tucson Meeting, the 10th BDRM Conference, the 2007 AEA Annual Meeting , the Yale SOM Whitebox Student Conference on Behavioral Science, and the 26th Arne Ryde Symposium of Communication in Games and Experiment.

"Why do almost all people tell the truth in ordinary everyday life? —Certainly not because a god has forbidden them to lie. The reason is, firstly because it is easier; for lying demands invention, dissimulation and a good memory."

– Friedrich Nietzsche, Human, All Too Human, II.54, 1878/1996

During the tech-stock bubble, Wall Street security analysts were alleged to inflate recommendations about the future earnings prospects of firms, in order to win investment banking relationships with those firms.[1] Specifically, analysts in Merrill Lynch used a five-point rating system (1=Buy to 5=Sell) to predict how the stock would perform. They usually gave two 1-5 ratings for short run (0-12 months) and long run (more than 12 months) performance separately. Henry Blodget, Merrill Lynch's famously optimistic analyst, "did not rate any Internet stock a 4 or 5" during the bubble period (1999 to 2001). In one case, the online direct marketing firm LifeMinders, Inc. (LFMN), Blodget first reported a rating of 2-1 (short run "accumulate"—long run "buy") when Merrill Lynch was pursuing an investment banking relationship with LFMN. Then, the stock price gradually fell from $22.69 to the $3-$5 range. While publicly maintaining his initial 2-1 rating, Blodget privately emailed fellow analysts that "LFMN is at $4. I can't believe what a POS [piece of shit] that thing is." He was later banned from the security industry for life and fined millions of dollars.[2]

---

[1] For a detailed description of the tech-stock bubble, see Michael J. Brennan (2004). For evidence regarding analyst recommendations affected by conflicts of interest, see Hsiou-wei Lin and Maureen F. McNichols (1998) and Roni Michaely and Kent L. Womack (1999).

[2] See Complaint in Securities and Exchange Commission v. Henry M. Blodget, 03 CV 2947 (WHP) (S.D.N.Y.) (2003), paragraph 11-12 and 70-72, Securities and Exchange Commission Order Against Henry M. Blodget (2003), and United States District Court Final Judgement on Securities and Exchange Commission v. Henry M. Blodget 03 Civ. 2947 (WHP) (S.D.N.Y.) (2003).

This case is an example of a sender-receiver game with divergent preferences (sometimes called a "cheap talk" or "strategic information transmission" game). Sender-receiver games are simple models of economic situations in which one agent has an incentive to exaggerate the truth to another agent. The central issues in these games are how well uninformed players infer the private information from the actions of players who are better-informed, and what informed players do, anticipating the behavioral inference of the uninformed players. Vincent P. Crawford and Joel Sobel (1982) showed that in such games, the sender's incentive to exaggerate when his preferences differ from the receiver's precludes equilibria in which communication is perfectly informative. Instead, all equilibria are noisy, and the larger the difference between the sender's and receiver's preferences, the noisier is the most informative equilibrium. But as in any equilibrium model, there is no systematic deception: The receiver's beliefs conditional on the sender's message are unbiased estimates of the true state. Although previous experimental work starting with John Dickhaut et al. (1995), Andreas Blume et al. (1998, 2001), and cumulating with Hongbin Cai and Joseph Tao-yi Wang (2006) general confirms Crawford and Sobel's comparative statics prediction regarding the noisiness of communication, experiments also tend to disconfirm their model's prediction that with divergent preferences, senders never tell the truth except by accident, and that receivers are never systematically deceived. Crawford (2003) and Navin Kartik, Macro Ottaviani and Francesco Squintani (2007) show that both "overcommunication" (senders telling the truth more than equilibrium predicts) and systematic deception can be explained by a class of non-equilibrium models of strategic thinking called level-k models. Cai and Wang (2006) show that such a model can describe the data from their sender-receiver experiments and could help to explain the overcommunication their subjects exhibits. The present paper builds on these results and attempts to investigate the cause behind the behavior patterns in such games. Understanding these behavioral patterns better should aid in the design of institutions to foster more accurate transmission of information when preferences diverge.

Incentives for exaggeration are common. Besides the Blodget case mentioned above, similar dramatic accounting frauds in the last few years, such as Enron, Worldcom, and Tyco, might have been caused by the incentives of managers (and perhaps their accounting firms) to inflate earnings prospects (Brennan, 2004, and Brian J. Hall and Kevin J. Murphy, 2003). For instance, Enron executives told shareholders at meetings that earnings prospects were rosy, at the same time as the executives were selling their own shares, leading to indictments and trials in 2006.[3] In universities, grade inflation and well-polished recommendation letters help schools promote their graduates (Henry Rosovsky and Matthew Hartley, 2002). Other examples of incentives for strategic information transmission include government-expert relationships in policy making, doctor-patient relationships in health care choices, teacher cheating on student tests (Brian A. Jacob and Steven D. Levitt, 2003) and the floor-committee relationship in Congress.

This paper reports experiments on a sender-receiver game. In the game, a sender learns the true state (a number S) and sends a costless message M to a receiver who then chooses an action A. Payoffs only depend on S and A so the message M is “cheap talk.” The receiver prefers to choose an action that matches the state, but the sender wants the receiver to choose an

[3] According to an SEC complaint filed in court, Kenneth Lay, Enron’s then chairman and CEO, said “We will hit our numbers” and “My personal belief is that Enron stock is an incredible bargain at current prices” in an employee online forum on September 26, 2001. However, in the prior two months he was actually making net sales of over $20 million in Enron stock (back to Enron). See Second Amended Complaint in Securities and Exchange Commission v. Richard A. Causey, Jeffrey K. Skilling and Kenneth L. Lay, Civil Action No. H-04-0284 (Harmon) (S.D. Tx.) (2004), paragraph 81-82.

action closer to S+b, where b is a known bias parameter. The value of b is varied across rounds. When b=0 senders prefer that receivers choose *S*, so they almost always just announce S (i.e., M=S), and receivers believe them and choose A=M. When b>0 senders would prefer to exaggerate and announce M>S if they thought receivers would believe them. Since subjects choose 1-5, the numbers in our game are coincidentally the same as those used by Merrill Lynch. Indeed, when b>0, we find that our subjects hardly ever report the number 1 (in only 8 percent of 208 rounds), much as Blodget never rated a stock 4 or 5 (the equivalent of 1-2 in our game). However, our game is presented in abstract terms without reference to stock analysts or deception. This could make subjects feel less guilty when 'deceiving' others in the experiment.

Besides measuring choices in these games, our experiment uses video-based "eyetracking" to measure what payoffs or game parameters sender subjects are looking at (see Appendix: Methods). Eyetracking software records where players are looking on a computer screen every 4 milliseconds. These data are a useful supplement to econometric analysis of choices, when decision rules which produce similar choices make distinctive predictions about what information is needed to execute these rules. Previous "eyetracking" studies used a "Mouselab" system in which moving a cursor into a box opens the box's contents and are more accurately described as "mouse-tracking." See Camerer et al. (1993), Miguel Costa-Gomes et al. (2001), Eric Johnson et al. (2002), Costa-Gomes and Crawford (2006), Xavier Gabaix et al. (2006); and Crawford (2008).[4]

[4] One small handicap of the Mouselab system is that the experimenter cannot be certain the subject is actually looking at (and processing) the contents of the open box. Our system measures the eye fixation so we can tell if the subject's eye is wandering, and pupil dilation is measured at the same time (which Mouselab cannot do). Nevertheless, Mouselab systems can be installed cheaply in many computers to measure lookups of many agents at the same time, which could prove useful in running efficient subjects and studying attention simultaneously in complex markets with many agents.

The eyetracking apparatus also measures how much subjects' pupils "dilate" (expand in width and area). Pupils dilate under stress, cognitive difficulty, arousal and pain.[5] Pupillary responses have also been measured in the lie-detection literature for many years (See for example, F. K. Berrien and G. H. Huntington, 1942, I. Heilveil, 1976, Michel P. Janisse, 1973, M. T. Bradley and Janisse, 1979, 1981, Janisse and Bradley, 1980, R. E. Lubow and Ofer Fein, 1996, and Daphne P. Dionisio et al., 2001). These studies suggest that pupil dilation might be used to infer deceptive behavior because senders find deception stressful or cognitively difficult.

The choices, eyetracking, and pupil dilation measures generate four basic findings:

1. Overcommunication in sender-receiver game is consistent with L0, L1, L2, and equilibrium (EQ) sender behavior produced by a level-k model of the sender-receiver game in which L0 sender behavior is anchored at truth-telling.
2. Eyetracking data provide the following support for the level-k model:
    a. *Attention to structure and own payoffs:* Sender subjects pay attention to important parameters (state and bias) of the sender-receiver game. This indicates subjects are thinking carefully about the basic structure of the game, even if they are not following equilibrium theory. Sender subjects also look at their own payoffs more than their opponents'.

[5] For pupillary responses to arousal, see R. A. Hicks et al. (1967), R. Bull and G. Shead (1979), and Darren C. Aboyounand James N. Dabbs (1998). For pupillary responses to cognitive difficulty, see Jackson Beatty (1982) and B. C. Goldwater (1972). For pupillary responses to pain, see C. Richard Chapman et al. (1999) and Shunichi Oka et al. (2000). Min Jeong Kang et al (2008) show that pupils dilate in anticipation of finding out the answers to trivia questions about which they are curious. (Their self-reported curiosity is also shown by fMRI to activate the ventral striatum, a brain region involved in anticipated reward or "prediction error" and learning; and curiosity also enhances later memory for mistaken answers.)

b. *Truth bias*: Sender subjects focus too much on the true state payoff row. This bias is consistent with a failure to "think in the opponent's shoes" as in Meghana Bhatt and Colin F. Camerer (2005).

c. *Individual level-k lookup patterns*: Sender subjects focus on the payoffs corresponding to the action A = S (L0 reasoning), A=S+b (L1 reasoning), …, up to the corresponding level-k reasoning for each individual subject based on his or her level-k type. This indicates particular level-k type subjects do generally exhibit the corresponding lookup patterns.

3. Right before and after the message is sent, senders' pupils dilate more when their deception is larger in magnitude. This suggests subjects feel guilty for deceiving (as in Uri Gneezy, 2005), or deception is cognitively difficult (as the level-k model assumes).[6]
4. Prediction: Based on the eyetracking results, we can try to predict the true state observed by the sender using lookup data and messages. This prediction exercise suggests it is possible to increase the receiver's payoff (beyond what was earned in the experiments) by 16-21 percent, resulting in an economic value of 60 percent of the maximum achievable increase.

Finally, this study shows the possible relevance of psychology and neuroscience to economics. Douglas Bernheim (2008) suggests that neuroeconomics will be successful if it can show how new non-choice data can solve a prediction or normative problem that could not be solved by standard choice data. One such problem is how to extract private information from

[6] Note that although the pupil dilation results are consistent with both the guilt and cognitive difficulty explanations, the lookup results are more consistent with the cognitive difficulty story of overcommunication, since different lookup patterns each suggest a specific (level-k) reasoning process that has a particular level of cognitive difficulty. It is not obvious how guilt alone (or variations in guilt) can produce this result.

choices. In the standard model, private information is, by definition, not directly observable to outsiders (such as receivers in our game); it can only be inferred assuming a particular model of behavior (e.g., inferring private values from auction bids). If eyetracking, pupil dilation, fMRI, or other biological measures enable one to infer more about private information than by using only choices, those "new" data—new to economists, that is—have some added value for something economists care about. Our data satisfy this criterion because lookups and pupil dilation enhance prediction of the true state beyond the predictions derived simply from observed messages (choice) and equilibrium theory.

This is the first study in experimental economics to use a combination of video-based eyetracking and pupil dilation, and is, of course, exploratory and is therefore hardly conclusive. But the eyetracking and pupil dilation results by themselves suggest that the implicit assumption in equilibrium theories of "cheap talk" in games with communication—namely, that deception has no (cognitive or emotional) cost—is not completely right. This provides the foundation for alternative theories such as costly talk (as in Kartik, Ottaviani and Squintani, 2007, Ying Chen, 2007, and Kartik, 2008) or the level-k model (as in Crawford, 2003, Cai and Wang, 2006). The Nietzsche passage quoted above describes the cognitive load of deception, and is explored in Jennifer Maria Nuñeza et al. (2005). Mark Twain also famously quipped, "*If you tell the truth, you don't have to remember anything*," indicating the memory cost of deception.[7] The corollary principle is that if subjects want to misrepresent the state to fool receivers, they have to figure out precisely how to do so (and whether receivers will be fooled). This process is not simple and seems to leave a psychological signature in looking patterns and pupil dilation. Future theories

[7] Quotation taken from Mark Twain's Notebook, 1894. In fact, Daniel Kahneman and Beatty (1966) showed how more difficult memory tasks induced larger pupillary response. Hence, memory load could also be a channel for deception to affect pupil dilation.

could build in an implicit cost to lying (which might also vary across subjects and with experience) and construct richer economic theories about when deception is expected to be widespread or rare.

I. The Sender-Receiver Game

In each round of the experiments, subjects play a game of strategic information transmission, involving cheap talk (Crawford and Sobel, 1982). One player always acts as the sender, and the other as the receiver. The sender's eye movements and pupil dilation are measured with a head-mounted Eyelink II eyetracker (see Appendix: Methods). At the beginning of the round, the sender is informed about the true state of the world, which is described as a "secret" number S uniformly drawn from the state space **S** = {1, 2, 3, 4, 5}, and is informed about the bias b, which is either 0, 1, or 2 with known probabilities. The receiver knows the bias b, but not the realization of the state S. Both players are informed in instructions about the basic structure of the game.

The sender then sends a message to the receiver, from the set of messages **M** = {1, 2, 3, 4, 5}.[8] After receiving a message from the sender, the receiver chooses an action from the action space **A** = {1, 2, 3, 4, 5}. The true state and the receiver's action determine the two players' payoffs in points according to $u_R = 110 - 20 \cdot |S - A|^{1.4}$, and $u_S = 110 - 20 \cdot |S + b - A|^{1.4}$, where u_R and u_S are the payoffs for the receiver and the sender, respectively. Note that the receiver earns the most money if her action matches the true state (since her payoff falls with the absolute difference between A and S). The sender prefers the receiver to choose an action equal to S+b.

---

[8] Following Cai and Wang (2006), we use the message, "The number I received is X" to eliminate possible misinterpretation of the message (which contributes to the multiple equilibria problem typical in these types of games resulting form the need to assign meaning to messages).

This payoff structure is made known to both senders and receivers. Figure S1 shows the screen display for b=1 and S=4.

When the bias is large (b=2), the most informative equilibrium has the sender send an uninformative message, while the receiver ignores the message and chooses A=3 based on her prior beliefs (babbling equilibrium). When b=1, the most informative equilibrium requires the senders to send messages {1} when S=1, and send {2,3,4,5} when S is 2-5. When b=1 the receivers should choose action A=1 when seeing M={1}, and A=3 or 4 when seeing M={2,3,4,5}.[9] When b=0, truth-telling by choosing M=S (and receivers choosing A=M) is the most informative equilibrium.

On the other hand, following Crawford (2003) and Cai and Wang (2006), the level-k model for the sender-receiver game starts with L0 senders (who has the lowest level of sophistication) would simply tell the truth, and L0 receivers best responding to L0 senders by following the message. Moving up the hierarchy, L1 senders best respond to the L0 receivers by inflating the message (stating their preferred states), and L1 receivers best respond to L1 senders by discounting the message. Such procedure is continued until we reach the most informative equilibrium prediction.[10] In addition, we include a sophisticated type (SOPH) which best

---

[9] Thanks to David Eil for clarifying the equilibrium analysis. Due to discreteness, there is another knife-edge equilibrium with b=1 that produces higher information transmission: Senders send messages M={1,2} and {3,4,5}, while receivers choose A=2 and 4. However, this equilibrium is not robust since senders who see S=2 are exactly indifferent between sending M={1,2} and M={3,4,5}. Moreover, the main results of the paper do not change even if we consider this equilibrium (then Corr(S,A) = 0.791, and u_R= 94.56).

[10] Note that the level-k model itself provides an equilibrium selection criterion—it selects the most informative equilibrium where senders report the upper bound of the interval of true states. This pins down both the amount of information transmitted *and* the language used in the message sent. In general, level-k models will provide *more* precision (given a particular parameter value specification) than equilibrium concepts when there are multiple equilibria. Also note that, due to signal jamming, higher level types do not simply add (or subtract) multiples of the bias. This is particularly true when approaching the upper bound of the state space. For example, when b=1, L2

responds to the empirical distribution of opponent's behavior. This represents the highest level of strategic thinking, knowing the exact heterogeneity of opponent types and behavior. Table 1 provides the list of different level-k types for b=0, 1, and 2.[11] Note that in our data, SOPH senders act like L2 senders when b=1 and act like EQ(=L3) senders when b=2, both a best response to L1 receivers.

Under both the equilibrium and level-k models, the comparative statics are similar: Information transmission decreases as the bias increases, though the level-k model still allows transmission even when the bias is so big that the equilibrium model predicts babbling (zero transmission). Informativeness is measured by the correlation between actions and the true states, and by receiver payoffs (more informative equilibria have higher expected payoffs). In addition, we assume a literal interpretation of messages, and measure the "informativeness" of senders' messages by the correlation between the true states and the messages M. How "trusting" the receivers are can be measured by the correlation between the messages M they receive and the actions A they take.[12] These comparative statics predictions were tested by Dickhaut et al. (1995), Blume et al. (1998, 2001), and Cai and Wang (2006). Overcommunication—messages are more informative of the true state than they should be, in equilibrium—are typically found in

senders who see S=3-5 will all send M=5 (since higher messages are not feasible), and L2 receivers, knowing the true state is equally likely to be 3, 4 or 5, would choose A=4, instead of 3 (=5-1*2).

[11] Cai and Wang (2006) constructed a level-k model for the case where the most informative equilibrium is babbling. Here we extend it to other biases. Also, we use the econometric methods developed by Costa-Gomes and Crawford (2006) to estimate individual types.

[12] Such a natural language interpretation is justified by Blume et al. (2001) findings that equilibrium messages tend to be consistent with their natural language meanings, and is used in Cai and Wang (2006). Moreover, many behavioral theories of lying, such as Crawford (2003) and Kartik, Ottaviani and Squintani (2007), also lead to this sort of natural language interpretation since naïve receivers would take the message at face value.

these studies, and Cai and Wang (2006) suggest two bounded rationality explanations: the level-k model and quantal response equilibrium.

II. The Experiment

To be sure that subjects learn, and to collect a lot of trials to pool across, the same game is played 45 times among the two players with random choices of bias b (and random states) in each round. Because we could only eyetrack one or two subjects at a time, only the senders were hooked up to the mobile Eyelink eyetracker (although studying receivers' eye fixations would be useful in future work). We randomly matched six subjects into pairs using a stranger-matching protocol, with different receivers in each round (with no immediate re-matching with the same receiver), and eyetracked two of the senders in each group. Values of b=0, 1, 2 were used with known probabilities (0.2, 0.4, 0.4) since we are less interested in the no-bias (b=0) case than in the bias (b>0) cases. We also added some noise (integers -4 to +4 with equal probability, i.i.d. across payoff cells) to each payoff to minimize memory effects. Since the noise is small, the equilibrium remains the same. To further eliminate any memory effect, the bias parameter was not shown to the eyetracked senders on the screen, although it was mentioned in the instructions. Instead, subjects were forced to look at the payoff table to infer it. Thus, this set of experiments is called the "hidden bias-stranger" design. The results reported below focus entirely on the eye fixations and pupil dilation of the eyetracked senders, and the message choices of all senders and action choices of receivers.[13]

[13] Two of the twelve eyetracked subjects experienced technical difficulty during the experiment and their data were excluded (along with the corresponding receiver subjects' choices).

Subjects' choices are compared to the most informative equilibrium in the one-shot game.[14] We also use predictions from a level-k model (Table 1 to estimate individual sender types with a quantal response-like "spike-logit" error structure, using the econometric analysis developed by Costa-Gomes and Crawford (2006). In particular, we assume each sender subject exactly follows a certain level-*k* type and plays t^k (the "spike" of probability) with probability $(1-\varepsilon)$. With probability ε, they make mistakes following a logit error density

$d^k(m,\lambda \mid s) = \dfrac{\exp\left[\lambda \cdot \Pi^k(m \mid s)\right]}{\sum_{\mu \neq t^k} \exp\left[\lambda \cdot \Pi^k(\mu \mid s)\right]}$, in which $\Pi^k(m \mid s) = \sum_{\alpha=1}^{5} \pi(s,\alpha) f^k(\alpha \mid m)$ is the expected payoff of sending message *m* when the true state is *s*. $\pi(s,a)$ is the payoff for true state *s* and receiver action *a*, and $f^k(\alpha \mid m)$ is level-*k* sender's belief about receiver's actions (seeing each message). The likelihood for observing a level-*k* subject *i* play $m^i = \left\{m_g^i\right\}_{g \in G}$ in the set of games *G* (making mistakes in subset $N^{ik} \subset G$, $n^{ik} = |N^{ik}|$) is therefore

$d^k(m^i,\varepsilon,\lambda) = (1-\varepsilon)^{|G|-n^{ik}} \cdot \varepsilon^{n^{ik}} \prod_{g \in N^{ik}} d^k\left(m_g^i,\lambda \middle| s(g)\right)$. The level-*k* type distribution is $p = (p^1,\ldots,p^K)$.

For each individual subject, we estimate the parameters (p,ε,λ) that maximizes empirical log-likelihood $L^i(p,\varepsilon,\lambda \mid m^i) = \ln\left[\sum_{k=1}^{K} p^k d^k(m^i,\varepsilon,\lambda)\right]$. Note that *p* will be estimated to have $p^k = 1$ for some *k*, so estimation results for a subject could be written as (k,ε,λ).

[14] We do not consider a possible dynamic equilibrium that might sustain higher information transmission levels. This is not a problem for b = 0 or 2. When b = 2, babbling is the only equilibrium in the one shot game and backward induction yields the babbling equilibrium for all finitely repeated games; when b = 0, the one shot game equilibrium already has full information transmission and there is no room for improvement. Also note that overcommunication is the most striking when b = 2. Random rematching also limits repeated-game effects.

We also ran an earlier set of experiments using a partner protocol in which a pair of subjects played repeatedly in a fixed-role protocol where b=0, 1, 2 with known equal probability. The bias parameters were always revealed (sender subjects indeed look at them), and there was no payoff perturbation. This is a simpler design to implement logistically, requiring only one eyetracked subject and his/her (open box) opponent, but creates potential repeated-game effects. We refer to this as the "display bias-partner" design. Results of this design are briefly discussed in comparison to that of the "hidden bias-stranger" design in Section III.D. Corresponding tables and figures are in the appendix.

Subjects were 60 Caltech students recruited from the Social Science Experimental Laboratory subject pool. Six sessions of six subjects were randomly matched in the "hidden bias-stranger" design, and twelve pairs were run in the "display bias-partner" design. They earned between $12 and $27 in addition to a $5-15 show-up fee. For the "hidden bias-stranger" sessions, we used different randomly pre-drawn parameters for each of the six sessions. But in the "display bias-partner" design we used the same set of randomly drawn biases and states for 9 of the 12 pairs, and used two other sets of parameters for the remaining 3 pairs to see if there were any effects for using the same parameters.

While 60 subjects might appear to be a small sample size,[15] most experimental studies with larger samples have many fewer choices per subject. The eyetracked subjects played 45 games, and made a very large number of eye fixations; so we recorded a lot of data for each subject and could often draw confident statistical conclusions from these sample sizes.

[15] We successfully eyetracked 22 of the 60 subjects, which is considered a *large* sample size for psychophysical studies involving eyetracking.

III. Results

III.A Comparative Statics and Behavior

What do players choose? Figures 1-3 display the three dimensions of the raw choice data—states, messages and actions—for the three bias levels b = 0, 1 and 2. Each Figure is a 5-by-5 display. The true states 1-5 correspond to the five rows. The sender messages 1-5 correspond to the five columns. Within each stage-message cell, there is a pie chart. The area of the pie-chart in each cell is scaled by the number of occurrences for the corresponding state and message ; i.e., the most common state-message pairs have the largest pies. Hence, the rows indicate senders' behavior with respect to different states and the columns represents the "informativeness" of each message, determined by the distribution of states conditional on each particular message. Several diagonal lines connect predicted messages for various level-k types. Each pie chart also shows the distribution of actions chosen by the receiver for that state and message, using a gray-scale ranging from white (action 1) to black (action 5). The average receiver action is the number inside the pie.

For example, when b=0, and there is no conflict of interest, large pie-charts are concentrated on the diagonal (L0/Eq sender behavior), which is a visual way of showing that the senders almost always send a message corresponding to the true state. Moreover, these pie-charts mostly contain the same color ranging from light (lower actions) to dark (higher actions) as the message number increases across columns, showing that the receivers' actions are typically equal to the message. The distribution of state frequencies conditional on each message (i.e., down each column) almost degenerates into mass points of the true states, indicating nearly full information transmission. This corresponds to the (most informative) truth-telling equilibrium predicted by equilibrium theory, as well as the L0/EQ type in the level-k model.

When b=1, and there is an incentive to bias the message upward, the results are different. There is a large tendency for deception, which is evident from having some large pie charts off the diagonal. This departure is lopsided—only the upper diagonal of Figure 2 is populated with large pie charts.[16] That is, for a given state, the most common messages are the state itself or higher messages (not lower messages). Furthermore, the largest pie charts of each row are mainly on the line one column or two columns to the right of the diagonal (i.e., states S+1 and S+2), consistent with L1 and L2 sender behavior. Within the upper diagonal, the pie-chart gets darker and darker going down and right, showing how the receivers correctly increase their actions as the state and message increase. Since the conditional distribution of states (columns in Figure 2) shift from a mass point on the true state (as in Figure 1) to a distribution skewed toward state 3 to 5, some information is transmitted. However, this distribution is not consistent with the {1}-{2, 3, 4, 5} partition equilibrium, which requires that distributions of messages and actions for the bottom four rows (states 2-5) should all look the same.[17]

Finally, when b=2, equilibrium theory predicts a babbling equilibrium. If subjects were playing this equilibrium, the pie-chart distributions in each column would be roughly the same (up to random sampling error of state frequencies) and the shading distributions on each pie-chart would be the same. In fact, there is still a substantial amount of information transmitted, since the columns in Figure 3 do not show the same uniform distribution of state frequencies. However, many senders still sent message 5, especially for states 2-5. And a substantial amount

[16] Note that this one-sided deception can potentially backfire since if seeing a message 1 indicates the true state is 1, the state is less likely to be 1 when other messages were sent.

[17] If subjects were playing according to the partition equilibrium, column 1 should have probability 1 on state 1, and zero probability elsewhere, indicating the state being in partition {1}, while column 2 to 5 should all have equal probability distributions (say, a mass point at 5 or 1/4 each) on states 2 through 5, and zero elsewhere (indicating the state being in partition {2,3,4,5}).

of receivers did choose action 3, as predicted in the babbling equilibrium. This is consistent with the level-k model, since L1, L2, and EQ senders all send message 5 for states 3 to 5.

<u>What are the comparative static results?</u> Table 2 shows that the actual information transmitted, measured by the correlation between states S, actions A, and messages M. The key comparative static prediction of Crawford and Sobel (1982) holds in the data: As the bias b increases, the information transmitted decreases, measured either by the correlation r(S,A) between state S and action A, or by receiver payoffs. But note that even when the bias is so large (b=2) that theory predicts babbling (i.e., no correlation between S, A and M), the correlations are still around 0.3. There are also very small learning effects: correlations and payoffs rise across trials for b=0 and fall for b>0 reflecting (weak convergence toward equilibrium (see supplementary Appendix, Table S1). Payoffs also decline with bias b, as predicted by theory (Table 3). Data from both senders who are eyetracked and senders with "open boxes" (no eyetracking), are reported separately as a check on whether eyetracking, per se, changes behavior. There is no discernible effect of being eyetracked versus seeing all parameters ("open boxes").

When the bias b is large, information transmission is higher (measured by correlations among S, M and A), and payoffs are higher, than predicted by equilibrium theory. These data replicate the "overcommunication" (too much truth-telling) reported in Cai and Wang (2006).

<u>Can individual players be classified as level-k types?</u> Based on all behavioral data, we classify individual sender subjects into various types according to Table 1, assuming subjects remain the same type across different biases using the "spike-logit" estimation as in Costa-Gomes and Crawford (2006). The results are shown in Table 4. Subjects are classified as types (percentages) L0-L2 (18 percent, 25 percent, and 25 percent), SOPH (14 percent) and EQ (18 percent), with good compliance (above 60 percent, except for one).[18] Individual level classifications therefore confirm that subjects are mostly choosing according to stable level-k types, as hinted by the aggregate choice data. Comparing the classification results with that of Cai and Wang (2006), we see a similar pattern (having few L0, mostly L1 beyond), although they use a more primitive way to conduct the classification.

III.B Lookup Patterns

There are several goals in observing lookup patterns: First, we want to know what the aggregate lookup patterns are during the decision process. This indicates the subjects' attention to different information, and provides the basis for theorizing about subjects' decision-making process. Moreover, since the level-k model relaxes the assumption that people hold consistent beliefs about others, beliefs about other's beliefs, and so on, we expect the lookup patterns to indicate this. Finally, since the level-k model predictions explain individual behavior, it is natural to ask whether additional lookup data can provide more direct evidence supporting the level-k model than choices alone. In particular, we would like to ask whether individual subjects who are classified into different level-k types produce different lookup patterns matching their types.

The lookup results are organized according to the above goals as follows:

[18] Using only trials such that b=1, 2 yield the exact same classification. Using a logit structure (instead of spike-logit) on b=1, 2 also yields a similar distribution, in which only two subjects are classified differently: Subject #3 (2-1) is classified as SOPH instead of L1, and subject #5 (3-2) is classified as L2 instead of SOPH. See Table S12. Note that SOPH and L2 are almost identical, and from the lookup results below (Table 7), subject #3 has a lookup score more similar to SOPH than L1. Finally, using a logit structure on all data adds three more SOPH types (2-2, 4-3 and 5-2), all from "neighboring" types which often coincide with SOPH (EQ, EQ, and L2, respectively).

1. *Attention to structure:* In reporting aggregate lookup counts and time spent on different parts of the screen, we expect to see different level-k subjects paying differential attention to important parameters of the sender-receiver game, such as state, bias, and payoffs.

2. *Truth bias:* The level-k model assumes subjects best respond to perceived beliefs about their opponents' behavior, which are inconsistent with what opponent's actually do.[19] If senders cannot think like receivers (who do not know the true state), they would put too much attention on the payoff row corresponding to the true state, instead of treating all states equally. Hence, excessive attention to payoffs corresponding to the true state demonstrates a "curse of knowledge" and could be an attentional marker of these incorrect beliefs.

3. *Individual level-k type lookup patterns:* The level-k model assumes an anchoring L0 behavior of truth-telling. Higher types go through beliefs about lower types until they reach their own level-k type. If this decision process is reflected in the lookup patterns, attention should be paid to payoffs corresponding to the action A=S (L0), A=S+b (L1), and so on, up to the corresponding level-k type for each individual subject. For example, when bias is 2, a L2 sender under state 2 would look at the payoffs corresponding to state 2 and action 2 (the L0 outcome if the message is taken literally), action 4 (the L1 outcome if the message is taken literally), and action 5 (the L2 outcome if the message is taken literally). In Table 1, this corresponds to the first three elements (L0~L2 Senders) of the second column (S=2) in the bottom panel (b=2). Thus, in addition to the lookups required to figure out the bias parameter,[20] a level-*k* type sender

[19] If *all* subjects are SOPH who correctly best respond to others, SOPH behavior should coincide with equilibrium (EQ) behavior.

[20] In the hidden bias-stranger design, subjects must at least look at two payoffs to determine the bias: A=S and A=S+b. Potentially, this S could be any state, but should correspond to the true state due to the truth bias. In this case, the lookups would coincide with lookups linked to L0 and L1 thinking.

(with truth bias) would follow the prediction of the level-*k* model (first *(k+1)* elements of column *S* in Table 1) up to his own level.

What are senders paying attention to? Table 5 shows the average lookup time (excluding fixations shorter than 50msec) for various numbers on the screen which are parameters of the game.[21] Senders clearly are thinking carefully about the game because they look up the state for 0.86 seconds total (which is 3.2 fixations, about 270msec per fixation). The low time per lookup is a reminder that the eye glances around very rapidly, making frequent quick fixations, as is typical of other tasks including reading text passages.

Senders also look at their own payoffs longer. In particular, subjects look at their own (sender) payoffs at least 40% more than receiver payoffs. This difference is surprising since senders need to look carefully at receiver payoffs in order to determine the bias. Note that the ratio of lookup time for sender and receiver payoffs is the same for a small bias (b=1) and large bias (b=2). For b=2, which creates the most scope for guilt to constrain deception, we divide senders into those who looked more often at receiver payoffs, and those who looked less often (relative to the median sender-receiver looking ratio). Importantly, the high receiver-lookup group is actually more deceptive than the low group, which is inconsistent with the guilt hypothesis that the more one cares about other's payoffs and looks at them, the less one should deceive. For the high group, the correlation between states and messages is 0.55, and the average LIE_SIZE (|M-S|) is 0.88; for the low group, the correlation is 0.69, and the average LIE_SIZE is 0.71.

[21] The number of separate fixations is very highly correlated with the lookup time—in no cell is the average time per fixation less than 250msec or greater than 300msec—so the number of fixations can be approximated well by dividing the Table 5 figures by 270 msec. Both fixations and lookup time are reported in the supplemental Appendix (Table S10 and S11).

Note that there *is* a reduction in total looking times across trials, about 35 percent less in later periods (31-45) than in earlier periods (1-15) (see Table 5), and this reduction is similar across bias levels and treatments (Table S7.)

Do senders have a "curse of knowledge"? Table 6 shows that subjects look about five times longer at payoffs in the rows corresponding to the true state than they look at payoffs in rows corresponding to each of the four other states. When the bias is 0 this fixation on the actual state is understandable (and subjects typically choose message M=S), but the disproportionate attention to actual state payoffs is comparable when there is a bias of b=1 or 2. This result indicates that subjects do not "think in others' shoes", and cannot fully think like a receiver (who does not know the true state). Note that Table 6 suggests lookups might have statistical power to detect the actual state (i.e., to detect lies in which the message M deviates from the true state S). That is, a receiver who had online sender looking statistics might be able to predict what the actual state was rather reliably. This possibility is explored in Section IV.

Do senders follow level-*k* predictions of lookups? Tables 5 and 6 show there is a strong bias for senders to look more at payoffs from the state they know to be the true one. More detailed information about looking patterns across state-action pairs is conveyed by the icon graph in Figures 4-7 (developed by Johnson et al., 2002). For brevity we show only data from trials with positive biases for subjects classified as L1 and L2 (aggregate data are in the supplemental appendix, Figures S6-S9).

Each box in Figures 4-7 represents the attention paid to the payoff corresponding to a different state-action combination. Parts (a) and (b) represent attention to the sender payoff boxes and the receiver payoff boxes, respectively. The width of the box is a linear function of the average number of fixations on that box. The height of the box is a linear function of the average total looking time in that box. Boxes which are wide and tall were looked at repeatedly (wide) and for a longer time (tall). The vertical bars in the first columns represent the sum of looking time across each row. Longer bars represent longer time for all state-action boxes in that state. The "ruler" in the upper right shows the scale of looking time and number of fixations that can be used to "measure" each box.

Figure 4 shows the icon graph for bias b=1 when the subject is classified as L1.[22] The first thing to notice is that subjects spend much more time looking at their own payoffs (Figure 4a) than the payoffs of receivers (Figure 4b), as the Table 5 statistics show. Subjects' lookups are also more frequent and longer for actions that are equal to the actual state S or S+1. This corresponds to the first two rows (L0 and L1 senders) of the top panel (b=1) in Table 1, as well as the lookups to determine the bias.

Figure 5 shows the lookup icon graphs for bias b=1 when the subject is classified as L2 (again when subjects are L1). Senders again look at their own payoffs more often than their opponents' payoffs. When the state S is 1-3 they tend to look at their payoffs from actions corresponding to S, S+1 and S+2. This is consistent with the first three rows (L0, L1, and L2 senders) of the top panel (b=1) in Table 1. However, when the state is 4 or 5 this pattern crumbles as states S+2 and S+1 do not exist; then lookup patterns resemble L1 lookups.

Similar patterns arise when b=2 as well. Figure 6 and 7 show the lookup icon graphs for bias b=2 when subjects are classified as L1 and L2, respectively. As the level-k model predicts,

[22] When the bias b=0 the looking data are very clear: Subjects look almost exclusively at their own payoffs corresponding to the actual state S and corresponding receiver action A, and they look at the receiver payoffs from the same S-A pair less often than they look at their own payoffs (1/2 in the display bias-partner design and 2/3 in the hidden bias-stranger). See Figure S5.

subjects look at payoffs corresponding to the first two or three rows (L0-L1 or L0-L2 senders) of the bottom panel (b=2) in Table 1.

If we calculate the linear measure of predictive success (Reinhard Selten, 1991), a subject who is classified as a certain level-k type almost always has the highest score for the corresponding lookups of the same type. In particular, let x equal the "hit rate", the proportion of lookups in a period that fall in the target cells, and let a equal the proportional area of the target cells. Then the linear measure (LM score) equals x-a, the proportional hit rate minus the proportional area. This measure controls for the size of the predicted lookup area, and takes a value of zero when subjects randomly scan the entire screen. Table 7 presents each subject's LM score for various types. Among all the six subjects classified as L1 and L2 subjects, only one (subject #8) has another type's LM score slightly higher (0.268 vs. 0.259, less than 0.01) than the score corresponding to their classification based on choices. Moreover, this subject would be classified as SOPH under the logit specification. Regarding SOPH subjects, it is not clear theoretically what their lookup patterns would be. But, the low LM scores do show that they do not look like L1, L2 or EQ.

III.C Pupil Dilation

As noted in the introduction, pupils dilate when people are aroused or make cognitively difficult decisions. Our first exploratory step is to treat pupil dilation as a dependent variable and see whether the degree of behavioral deception by the sender is correlated with pupil dilation. It may be that pupil dilation is so poorly measured, or so weakly linked to deception, that there is no reliable correlation. However, we see that deception *is* reliably correlated with pupil dilation.

To correlate pupil dilation with senders' messages, average pupil sizes are calculated for various time periods before and after the sender's message decision. Then, we try to predict averaged pupil dilation using the bias b and the amount of deception (measured by the absolute distance between states and messages, |M-S|).

To record their message M, senders are instructed to look at a series of decision boxes on the right side of the screen, which contain the numbers 1 to 5 (corresponding to the possible numerical messages). The software records a decision after the subject has fixated on a single decision box for 0.8 seconds—that is, the subjects choose by using their eyes, not their hands.[23]

Since there is a time lag of at least 0.8 second between the instant subjects "made up their minds" and the recording of this decision,[24] the *decision time* is defined as the first time subjects view any of the boxes in the decision boxes area, (provided they continue to look at the decision box area for more than 98 percent of the time until the software records a decision).

Average pupil sizes are regressed on the amount of deception for different biases, the absolute size of the deception (LIE_SIZE = |M-S|), and bias and state dummies, controlling for subject random effects and individual learning trends (picked up by round number and squared round number variables interacted with individual dummies). All standard errors are robust standard errors clustered at the individual level. The specification is:

$$\text{(1)} \qquad \mathrm{PUPIL}_i = \alpha + \sum_{b=0}^{2} \beta_{1b} \cdot LIE\_SIZE \cdot BIAS_b + \sum_{b \neq 2} \beta_{2b} \cdot BIAS_b + \sum_{s \neq 3} \beta_{3s} \cdot STATE_s$$

$$+ \sum_{k=1}^{K} \left(\gamma_{k,1} ROUND \cdot SUBJ_k + \gamma_{k,2} ROUND^2 \cdot SUBJ_k \right) + \varepsilon$$

[23] Allowing eye fixations to determine actual choices is widely used in research with monkeys. For humans, making choice hands-free is an advantage if psychophysiological measurements are being recorded simultaneously (e.g., galvanic skin conductance on the palms, heart rate) since even small hand movements add noise to those measurements.

[24] This time lag can be longer if the subject is not perfectly calibrated, and hence, needs extra time to perform the required fixation. Another possible situation is when the subject "changed her mind" and looked at different decision boxes.

where the error term ε has elements $\varepsilon_{kt} = u_k + \eta_{kt}$ (subject random effects), and

$PUPIL_i$ = Average pupil (area) size at time frame *i*: 1.2 to 0.8 seconds, 0.8 to 0.4 seconds, 0.4 to 0 seconds before, and 0 to 0.4 seconds, 0.4 to 0.8 seconds after the decision time.[25] Here, we normalize each individual's average pupil size to 100.[26]

LIE_SIZE = The "size" of the lie or the amount of deception, measured by the absolute distance between states and messages, (|M-S|).

$BIAS_b$, $STATE_s$, $SUBJ_k$ = Dummy variables for the bias b, true state s, and subject k

ROUND = Round number t

The parameter α is the average pupil size. The β_1 coefficients give us the effect of deviating from reporting the true state (deceiving more) under different bias levels. The coefficients β_{2b} and β_{3s} give us the pure effects of different biases b (relative to b=2) and states (relative to S=3) which might influence dilation, and $\gamma_{k,1}$,$\gamma_{k,2}$ capture (individual) linear and quadratic learning effects.

Look first at the coefficients on the amount of deception in Table 8, interacted with bias (denoted β_{1b} where b is the bias parameter). Immediately after the decision is made (0 seconds to 0.4 seconds and 0.4 to 0.8 seconds later), the coefficients are significantly higher at about 2 percent for all biases. Sending less accurate messages is therefore correlated with pupil dilation

[25] Hence, we are aggregating 100 observations into one data point when averaging for each 400 milliseconds interval. Rounds with very short response time are discarded if $PUPIL_i$ cannot be calculated.

[26] Pupil sizes are measured by area, in relative terms. Absolute pixel counts have little meaning since they vary by camera positions, contrast cutoffs, etc., which depend on individual calibrations. Hence, the eyetracker scales it to a pupil size measurement between 800-2000. Here, we normalize all observations by the average pupil size of each subject throughout the entire experiment, and present all results in percentage terms. (To avoid potential bias created by eyetracker adjustments, all between-round adjustment stages were excluded when doing this normalization.) Therefore, "100" means 100 percent of an individual subject's typical pupil size.

when b=1 or b=2. Before the decision is made, the pupil dilation difference is still at 1.5-2 percent (though less significant) when b=2.

Note that the bias condition by itself does not generate pupil dilation (i.e., nearly all the coefficients β_{2s} are insignificant and are omitted from Table 8). This finding implies arousal or cognitive difficulty is created by sending deceptive messages in bias conditions, not by bias per se. Furthermore, these basic patterns are reproduced when we divide the samples into two halves and compare them, which provide some assurance of statistical reliability.[27]

III.D Results of the Display Bias-Partner Design

The supplemental appendix reports results analogous to those in Table 2-6, Table 8 and Figure 2-4 for the display bias-partner condition (Tables S2-S6 and S8, Figures S2-S4). Compared to the hidden bias-stranger condition there is more overcommunication (correlations of M and S around 0.5 even when b=2) and more low-type classification (one third L0 types). These differences are probably due to the repeated game effects created by the partner matching. Subjects do also look at the bias parameter when it is available, but they look less often at receiver payoffs (which they need not look at to figure out what the bias b is).

The pupil dilation results are much stronger than in the hidden bias-stranger design. The coefficients on pupil dilation predicting the amount of deception are 2.8-4.5 percent, and are significant in all 400 millisecond intervals from -1200 milliseconds to +800 milliseconds (where zero is the decision time). It is likely that the display bias-partner design is less demanding

[27] Because we measured eyetracking and pupil dilation from ten senders, it is useful to check how reliable these results are in two subsamples of five subjects each. The 400-msec intervals from +0.4 to +0.8 secs after decision time gives the highest R^2's so we compare those. The β_{1b} coefficients across bias levels (b=0, 1, 2) are the most important. They are 6.35*, 2.40, 2.11 for the first five subjects and 6.11**, 4.14**, and 3.00*** for the second five

cognitively, and lowered baseline pupil dilation. In fact, the increase in predictive power here could be construed as consistent with the cognitive difficulty story because showing the bias parameter and eliminating noise from the payoffs make the display bias-partner design easier in general. This simplification could decrease the baseline pupil dilation of truth-telling in all conditions, which makes any additional dilation from deception easier to detect. Running similar regressions show that using criteria of 99, 95, or 90 percent all yield similar results, though slightly weaker.

IV. Lie-detection and Prediction

As noted, one goal of measuring eyetracking is to see whether these behavioral measures enable us to improve upon predictions of theory. This section reports whether using eyetracking data helps predict deception and uncover the underlying true states. The ability to detect private information in this way could eventually have many practical applications. And since private information often undermines efficiency, the ability to detect private information could be Pareto-improving in some settings.

Here, we ask how well receivers could predict the true state using *only* messages and lookup patterns (and how much they could earn by using those predictions). That is, we pretend we don't know the true state for predictive purposes, forecast it from observables, then use knowledge of the true state to evaluate predictive accuracy. We focus only on b=1 and b=2 since truth-telling is so prominent when b=0.

For the dependent variable STATE *j*, from 1-5, we ran an ordered logit regression

$$\log\left[\Pr(STATE \geq j)\right] = \theta_j + \sum_{b=1,2}(\beta_{1b} \cdot MESSAGE + \beta_{2b} \cdot ROW_{self} + \beta_{3b} \cdot ROW_{other}) \cdot BIAS_b + \varepsilon$$

---

subjects. For other intervals, as predictive power (R^2) falls the reliability across the two subsamples falls too, but the where lookups are consolidated into two integer variables, ROW_{self} and ROW_{other}, which are the states corresponding to the own (or opponent) payoff row which has the longest total lookup time of all rows.

The coefficients β_{1b} represents the information about the state contained in the message the coefficient, β_{2b} measures the effects of the "most viewed row" of one's own payoffs (i.e., the state number corresponding to the row that is viewed for the longest time), and β_{3b} represents the effects of the "most viewed row" of the opponent's payoffs. The θ_j are state-specific constants.

To evaluate how well these specifications could predict new data, out-of-sample validation is used. Each observation is used with probability 2/3 to estimate the model, then the model forecasts on a holdout sample of the remaining 1/3 of the data. For each holdout observation, the estimated logit probabilities are used to calculate the expected state, which is rounded to the nearest integer to make a precise single-state prediction. This partial estimation-prediction procedure is performed for 100 random samples of the data. Average βs and (bootstrap) standard errors across the 100 resamplings are reported in Table 9.

The significance of β_{1b} in Table 9 indicates that the messages are informative about the states (as analyses reported above established). A smaller message indicates a smaller true state, even though standard game theory predicts that little information should be transmitted by the message (β_{1b} should be zero when b=2).

The lookup data are significantly correlated with states as well. The coefficients β_{2b}, on the most-viewed *own* row variables, and the coefficients β_{3b}, on the most-viewed *other* row variables, are all positive and significant. Thus, lookup data improve predictability *even when controlling for the message.* For example, if the message is 4, but the lookup data indicate the

coefficient signs are almost always the same in the two subsamples and magnitudes are typically reasonably close.

subject was looking most often at the payoffs in row corresponding to state 2, then the model could predict that the true state is 2, not 4. This is to be expected, since Table 6 indicates subjects look at the payoff rows corresponding to the true state five times more than other rows. However, note that this sort of prediction can only come from a setting in which attention is measured. In addition, if senders knew their eye movements were being used to infer the state, they could of course change their lookups and undermine the predictions.

The error rates in predicting states in the holdout sample are never greater than 40 percent. (Keep in mind that the error rates in equilibrium would be 60 percent and 80 percent.) Most of the wrong predictions from the logit model (70 percent) miss the state by one. The model accuracy is also substantially better than the actual performance of the receiver subjects in our experiments: Subjects “missed” (chose $A \neq S$) 58.5 percent of the time when b=1, and missed 77.9 percent for b=2.

An interesting calculation is how much these predictions could potentially add to the receiver payoffs (cf. “economic value” in Camerer et al., 2004). For biases b=1 and b=2, the average actual payoffs earned by receivers who faced eyetracked senders in the random sample were 87.5 and 80.9. If receivers had based their predictions on the models estimated in Table 8, and chose an action equal to the model predicted state (for the holdout sample), their expected payoffs would be 101.7 for b=1 and 98.0 for b=2. Since the maximum payoff possible is 110, this is a large economic value of about 60 percent of the increment between actual and maximum payoffs.[28] In fact, these payoffs are already close to what subjects actually earn when b=0 and

[28] For b=1, economic value = (101.7-87.5)/(110-87.5) = 63%. For b=2, economic value = (98-80.9)/(110-80.9) = 59%. Analogous out-of-sample prediction results for the display bias-partner design are reported in Table S9. Results are weaker than that of the hidden bias-stranger design, having a modest economic value of 44 and 24 percent.

there is no bias (100.85 in Table 3).[29] These economic value statistics suggest that it could be possible to almost erase the cost to receivers of not knowing the true state just by looking at attention along with messages.

An important caveat to these analyses is that we do not know what would happen if the senders knew that their pupil dilation and lookups were being used to predict the true state. Senders would try to signal-jam by looking at the payoffs corresponding to their message more often (a kind of faked sincerity), but it is possible that excessive pupil dilation or more detailed lookup patterns could distinguish such signal-jamming. Putting senders under time pressure might also make it difficult for them use such a deliberately misleading strategy. In any case, such experiments are natural follow-ups and could be easily done.

V. Conclusion

This paper reports experiments on sender-receiver games with an incentive gap between senders and receivers, such as managers or security analysts painting a rosy picture about a firm's earnings prospects. Senders observe a state S, an integer 1-5, and choose a message M. Receivers observe M (but not S) and choose an action A. The sender prefers that the receiver choose an action A=S+b, which is b units higher than the true state, where b=0 (truth-telling is optimal), or b=1 or b=2. The bias number b is the size of the incentive gap. Receivers know the payoff structure, so they should be suspicious of inflated messages M.

Our experimental results show "overcommunication"—messages are more informative of the state than they should be, in equilibrium. This result is consistent with a level-k model of communication anchored at level-0 truth-telling. To explore the cognitive foundations of

[29] Such gains in the hidden bias-stranger design are not surprising since subjects are forced to look at the payoff table to discover the bias parameter, and they focus disproportionally on the "true state" row along the way.

overcommunication, eyetracking was used to record what payoffs the sender subjects are looking at, and how widely their pupils dilate (expand) when they send message.[30]

The lookup data show that senders look disproportionally at the payoffs corresponding to the true state. They do not appear to be thinking strategically enough by putting themselves "in other's shoes," looking and choice are roughly consistent with a cognitive hierarchy specified by the level-k model, starting from truth-telling.

Senders' pupils also dilate when they send deceptive messages ($M \neq S$), and dilate more when the deception $|M-S|$ is larger in magnitude. In a simpler pilot design that is prone to memory and repeated game effects (the display bias-partner design), these behavioral results are also present. Together, these data are consistent with the underlying assumptions of the level-k model, and that figuring out how much to deceive another player is cognitively difficult. Gneezy (2005) and Sjaak Hurkens and Kartik (2008) found that changing the known costs to others from deception lowers deception by subjects, suggesting that guilt plays a role in limiting deception. Complementing this finding, we find that guilt does not appear to be the sole driver of overcommunication, because senders who look at receiver payoffs more often are also more deceptive. In fact, Santiago Sánchez-Pagés and Marc Vorsatz (2007) show that overcommunication is caused by the tension between normative social behavior and incentives for lying.

Furthermore, combining sender messages and lookup patterns, one can predict the true state and lower the miss rate of subjects by one half. Those predictions increase receiver payoffs

[30] The sender-receiver paradigm also expands the quality of research on lie-detection in general: Deception in these games is spontaneous and voluntary (most studies use instructed lying); and both players have a clear and measurable financial incentive to deceive and to detect deception (most studies lack one or both types of incentives).

up to 16-21 percent, which is an economic value of more than half of the maximum increase above what subjects actually earn in the experiment.

There are many directions for future research.

Within this paradigm, eyetracking receivers would be useful for establishing their degree of strategic sophistication in making inferences from messages. More generally, economic theories often talk vaguely about the costs of decision making or difficulty of tradeoffs. Pupil dilation gives us one way to start measuring these costs.

Many economic models also specify a cognitive algorithm that maps acquired information into choices (e.g, dynamic programming applications which require looking ahead). The idea of allocating attention has itself gotten attention in economics (Della Vigna, 2008) and in macroeconomic studies of “rational inattention” (e.g., Christopher Sims, 2006). In both cases, measuring attention directly through (now video-based) eyetracking could improve tests of theories which make predictions about both attention and choice, and how they interact, as in previous mouse-tracking studies, such as Costa-Gomes et al. (2001), Johnson et al. (2002), and Costa-Gomes and Crawford (2006). Given the novelty of using these two methods in studying games, the results should be considered exploratory and simply show that such studies can be done and can yield surprises (e.g., the predictive power of lookups and pupil dilation for inferring private state information).

In the realm of deception, two obvious questions for future research are: Are there substantial individual differences in the capacity or willingness to deceive others for a benefit? And, can experience teach people to be better at deception, and at detecting deception? Both questions are important for extrapolating these results to domains in which there is self-selection and possibly large effects of experience (e.g., used-car sales or politics). In other domains of economic interest, the combination of eyetracking and pupil dilation could be applied to study any situation in which the search for information and cognitive difficulty are both useful to measure, such as “directed cognition” (Xavier Gabaix et al., 2006), perceptions of advertising and resulting purchase, and attention to trading screens with multiple markets (e.g., with possible arbitrage relationships).

References

Complaint in Securities and Exchange Commission V. Henry M. Blodget, 03 Cv 2947 (Whp) (S.D.N.Y.), Securities and Exchange Commission Litigation Release No. 18115, April 23, 2003. Washington, DC: Securities and Exchange Commission, 2003.

Order against Henry M. Blodget, Securities and Exchange Commission Administrative Proceedings, File No.3-11322, October 31, 2003. Washington, DC: Securities and Exchange Commission, 2003.

Second Amended Complaint in Securities and Exchange Commission V. Richard A. Causey, Jeffrey K. Skilling and Kenneth L. Lay, Civil Action No. H-04-0284 (Harmon) (S.D. Tx.), Securities and Exchange Commission Litigation Release No. 18776, July 8, 2004. Washington, DC: Securities and Exchange Commission, 2004.

United States District Court Final Judgement on Securities and Exchange Commission V. Henry M. Blodget 03 Civ. 2947 (Whp) (S.D.N.Y.), Securities and Exchange Commission Litigation Release No. 18115, Washington, DC: Securities and Exchange Commission, 2003.

Aboyoun, Darren C. and Dabbs, James N. "The Hess Pupil Dilation Findings: Sex or Novelty?" *Social Behavior and Personality*, 1998, *26*(4), pp. 415-19.

Beatty, Jackson. "Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources." *Psychological Bulletin*, 1982, *91*(2), pp. 276-92.

Bernheim, Douglas. The psychology and neurobiology of judgment and decision making. In P. Glimcher, C. Camerer, E. Fehr and R. Poldrack (Eds.), Handbook of Neuroeconomics, Elsevier, 2008.

Berrien, F. K. and Huntington, G. H. "An Exploratory Study of Pupillary Responses During

Deception." *Journal of Experimental Psychology*, 1943, *32*(5), pp. 443-49.

Bhatt, Meghana and Camerer, Colin F. "Self-Referential Thinking and Equilibrium as States of Mind in Games: fMRI Evidence." *Games and Economic Behavior*, 2005, *52*(2), pp. 424-59.

Blume, Andreas; DeJong, Douglas V.; Kim, Yong-Gwan and Sprinkle, Geoffrey B. "Experimental Evidence on the Evolution of Meaning of Messages in Sender-Receiver Games." *American Economic Review*, 1998, *88*(5), pp. 1323-40.

Blume, Andreas; DeJong, Douglas V.; Kim, Yong-Gwan and Sprinkle, Geoffrey B. "Evolution of Communication with Partial Common Interest." *Games and Economic Behavior*, 2001, *37*(1), pp. 79-120.

Bradley, M. T. and Janisse, Michel P. "Accuracy Demonstrations, Threat, and the Detection of Deception - Cardiovascular, Electrodermal, and Pupillary Measures." *Psychophysiology*, 1981, *18*(3), pp. 307-15.

Bradley, M. T. and Janisse, Michel P. "Pupil Size and Lie Detection - the Effect of Certainty on Detection." *Psychology*, 1979, *16*(4), pp. 33-39.

Brennan, Michael J. "How Did It Happen?" *Economic Notes*, 2004, *33*(1), pp. 3-22.

Bull, R. and Shead, G. "Pupil-Dilation, Sex of Stimulus, and Age and Sex of Observer." *Perceptual and Motor Skills*, 1979, *49*(1), pp. 27-30.

Cai, Hongbin and Wang, Joseph T. "Overcommunication in Strategic Information Transmission Games." *Games and Economic Behavior*, 2006, *56(1)*, pp.7-36.

Camerer, Colin F.; Ho, Teck-Hua and Chong, Juin-Kuan. "A Cognitive Hierarchy Model of Games." *Quarterly Journal of Economics*, 2004, *119*(3), pp. 861-98.

Camerer, Colin F.; Johnson, Eric J.; Rymon, Talia and Sen, Sankar. "Cognition and

Framing in Sequential Bargaining for Gains and Losses," K. G. Binmore, A. P. Kirman and P. Tani, *Frontiers of Game Theory*. Cambridge: MIT Press, 1993, 27-47.

Chapman, C. Richard; Oka, Shunichi; Bradshaw, David H.; Jacobson, Robert C. and Donaldson, Gary W. "Phasic Pupil Dilation Response to Noxious Stimulation in Normal Volunteers: Relationship to Brain Evoked Potentials and Pain Report." *Psychophysiology*, 1999, *36*(1), pp. 44-52.

Chen, Ying. "Perturbed Communication Games with Honest Senders and Naive Receivers," Unpublished paper, 2007.

Costa-Gomes, Miguel A. and Crawford, Vincent P. "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study." *The American Economic Review*, 2006, *96*, pp. 1737-68.

Costa-Gomes, Miguel; Crawford, Vincent P. and Broseta, Bruno. "Cognition and Behavior in Normal-Form Games: An Experimental Study." *Econometrica*, 2001, *69*(5), pp. 1193-235.

Crawford, Vincent P. "Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions." *American Economic Review*, 2003, *93*(1), pp. 133-49.

Crawford, Vincent P. "Look-ups as the Windows of the Strategic Soul: Studying Cognition via Information Search in Game Experiments." *Perspectives on the Future of Economics: Positive and Normative Foundations*, Volume 1, *Handbooks of Economic Methodologies*, ed. by Andrew Caplin and Andrew Schotter, Oxford University Press, 2008.

Crawford, Vincent P. and Sobel, Joel. "Strategic Information Transmission." *Econometrica*, 1982, *50*(6), pp. 1431-51.

Dickhaut, John; McCabe, Kevin and Mukherji, Arijit. "An Experimental Study of Strategic

Information Transmission." *Economic Theory*, 1995, *6*, pp. 389-403.

Della Vigna, Stefano. "Psychology and Economics: Evidence from the Field." *Journal of Economic Literature*, 2008, *forthcoming*.

Dionisio, Daphne P.; Granholm, Eric; Hillix, William A. and Perrine, William F. "Differentiation of Deception Using Pupillary Responses as an Index of Cognitive Processing." *Psychophysiology*, 2001, *38*(2), pp. 205-11.

Gabaix, Xavier; Laibson, David; Moloche, Guillermo and Weinberg, Stephen. " Information Acquisition: Experimental Analysis of a Boundedly Rational Model." *American Economic Review*, 2006, *96(4)*, pp. 1043-1068.

Gneezy, Uri. "Deception: The Role of Consequences." *American Economic Review*, 2005, *95*(1), pp. 384-94.

Goldwater, B. C. "Psychological Significance of Pupillary Movements." *Psychological Bulletin*, 1972, *77*(5), pp. 340-55.

Hall, Brian J. and Murphy, Kevin J. "The Trouble with Stock Options." *Journal of Economic Perspectives*, 2003, *17*(3), pp. 49-70.

Heilveil, I. "Deception and Pupil Size." *Journal of Clinical Psychology*, 1976, *32*(3), pp. 675-76.

Hicks, R. A.; Reaney, T. and Hill, L. "Effects of Pupil Size and Facial Angle on Preference for Photographs of a Young Woman." *Perceptual and Motor Skills*, 1967, *24*(2), pp. 388-&.

Hurkens, Sjaak and Kartik, Navin. "Would I Lie to You? On Social Preferences and Lying Aversion." *Experimental Econoimcs*, 2008, *forthcoming*.

Jacob, Brian A. and Levitt, Steven D. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics*, 2003, *118*(3), pp. 843-77.

Janisse, Michel P. "Pupil Size and Affect - Critical Review of Literature since 1960." *Canadian Psychologist*, 1973, *14*(4), pp. 311-29.

Janisse, Michel P. and Bradley, M. T. "Deception, Information and the Pupillary Response." *Perceptual and Motor Skills*, 1980, *50*(3), pp. 748-50.

Johnson, Eric J.; Camerer, Colin; Sen, Sankar and Rymon, Talia. "Detecting Failures of Backward Induction: Monitoring Information Search in Sequential Bargaining." *Journal of Economic Theory*, 2002, *104*(1), pp. 16-47.

Kahneman, Daniel and Beatty, Jackson. "Pupil Diameter and Load on Memory." *Science*, 1966, *154*(3756), pp. 1583-85.

Kang, Min Jeong; Hsu, Ming; Krajbich, Ian M.; Loewenstein, George F.; McClure, Samuel M. ; Wang, Joseph Tao-yi and Camerer, Colin F. "The Wick in the Candle of Learning: Epistemic Curiosity Activates Reward Circuitry and Enhances Memory." *Psychological Science*, 2008, *forthcoming*.

Kartik, Navin; Ottaviani, Macro and Squintani, Francesco. "Credulity, Lies, and Costly Talk." *Journal of Economic Theory*, 2007, *136 pp.* (1), pp. 749-58.

Kartik, Navin. "Strategic Communication with Lying Costs." *Review of Economic Studies*, 2008, *forthcoming*.

Lin, Hsiou-wei and McNichols, Maureen F. "Underwriting Relationships, Analysts' Earnings Forecasts and Investment Recommendations." *Journal of Accounting and Economics*, 1998, *25*(1), pp. 101-27.

Lubow, R. E. and Fein, Ofer. "Pupillary Size in Response to a Visual Guilty Knowledge Test: New Technique for the Detection of Deception." *Journal of Experimental Psychology-Applied*, 1996, *2*(2), pp. 164-77.

Michaely, Roni and Womack, Kent L. "Conflict of Interest and the Credibility of Underwriter Analyst Recommendations." *Review of Financial Studies*, 1999, *12*(4), pp. 653-86.

Nietzsche, Friedrich. *Human, All Too Human: A Book for Free Spirits*, 2nd edition. Translated by R. J. Hollingdale, Cambridge University Press, 1878, 1996.

Nuñeza, Jennifer Maria; B.J. Caseyb; Tobias Egnerc; Todd Harea and Joy Hirsch. "Intentional false responding shares neural substrates with response conflict and cognitive control." *NeuroImage*, 2005, 25(1), pp. 267-277.

Oka, Shunichi; Chapman, C. Richard and Jacobson, Robert C. "Phasic Pupil Dilation Response to Noxious Stimulation: Effects of Conduction Distance. Sex, and Age." *Journal of Psychophysiology*, 2000, *14*(2), pp. 97-105.

Rosovsky, Henry and Hartley, Matthew. "Evaluation and the Academy: Are We Doing the Right Thing? Grade Inflation and Letters of Recommendation," Cambridge, MA: American Academy of Arts and Sciences, 2002.

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.

Selten, Reinhard. "Properties of a measure of predictive success." *Mathematical Social Sciences*, 1991, 21, 153-167.

Sims, Christopher. "Rational Inattention: Beyond the Linear-Quadratic Case." *American Economic Review*, 2006, 96, 158-163.

Table 1: Behavioral Predictions of the Level-k Model

Sender Message (condition on State)						Receiver Action (condition on Message)					
State	1	2	3	4	5	Message	1	2	3	4	5
b=0											
L0/Eq Sender	1	2	3	4	5	L0/Eq Receiver	1	2	3	4	5
b=1											
L0 Sender	1	2	3	4	5	L0 Receiver	1	2	3	4	5
L1 Sender	2	3	4	5	5	L1 Receiver	1	1	2	3	4
L2 Sender	3	4	5	5	5	L2 Receiver	1	1	1	2	4
Eq Sender	4	5	5	5	5	Eq Receiver	1	1	1	1	4
SOPH Sender	3	4	5	5	5	SOPH Receiver	1	2	2	2	4
b=2											
L0 Sender	1	2	3	4	5	L0 Receiver	1	2	3	4	5
L1 Sender	3	4	5	5	5	L1 Receiver	1	1	1	2	4
L2 Sender	4	5	5	5	5	L2 Receiver	1	1	1	1	4
Eq Sender	5	5	5	5	5	Eq Receiver	1	1	1	1	3
SOPH Sender	5	5	5	5	5	SOPH Receiver	2	2	2	2	3

Note: L0 senders are truthful and L0 receivers best respond to L0 senders by following the message. L1 senders best respond to L0 receivers, while L1 receivers best respond to L1 senders, and so on. Note that when b=2, due to discreteness both L2 and Eq(=L3) senders best respond to L1 receivers.

Table 2: Information Transmission: Correlations between states S, messages M, and actions A

Bias	Eyetracked?	r(S, M)		r(M, A)		r(S, A)		Predicted r(S, A)
0	Yes	.92	.93	.90	.92	.86	.86	1.00
	No	.94		.94		.88		
1	Yes	.68	.64	.73	.71	.53	.49	.65
	No	.51		.61		.35		
2	Yes	.41	.34	.52	.58	.34	.32	.00
	No	.23		.63		.28		

Note: In the hidden bias-stranger design, some senders' eye movements were recorded ("eyetracked") and others were not ("open box"). This comparison provides a useful test of whether obtrusively tracking a subject's eye fixations affects their behavior.

Table 3: Sender and Receiver's Payoffs

Bias	Eyetracked?	u_S (std)	(combined)	u_R (std)	(combined)	Pred. u_R (std)
0	Yes	101.13 (18.68)	101.30 [a] (17.28)	100.85 (19.28)	101.27 [a] (17.69)	110.00 (0.00)
	No	101.89 (14.89)		102.07 (15.23)		
1	Yes	71.81 (39.56)	73.28 (37.46)	87.88 (28.63)	86.88 (27.59)	91.40 (19.39)
	No	75.44 (35.11)		84.44 (25.62)		
2	Yes	43.39 (52.17)	43.31 (52.79)	80.78 (27.17)	80.55 (27.57)	80.80 (20.76)
	No	43.21 (53.37)		80.21 (29.11)		

Note: [a] Payoffs are not exactly the same due to the random noise added and certain groups excluded.

Table 4: Level-k Classification Results

Session	ID	log L	*k*	Exact	lambda	Treatment
1	1	-46.23	SOPH	0.64	0.06	eyetracked subject #1
1	2	-25.99	L1	0.87	0.00	eyetracked subject #2
1	3	-15.98	L2	0.91	0.44	open box
2	1	-37.32	L1	0.60	0.52	eyetracked subject #3
2	2	-37.34	EQ	0.73	0.52	open box (eyetracked to round 20)
2	3	-25.70	SOPH	0.83	0.07	open box
3	1	-68.84	n/a	0.13	0.01	eyetracked subject #4
3	2	-17.71	SOPH	0.89	0.12	eyetracked subject #5
3	3	-54.73	EQ	0.60	0.03	open box
4	1	-50.86	L1	0.51	0.04	eyetracked subject #6
4	3	-25.22	EQ	0.82	0.48	open box
5	1	-22.26	L1	0.89	0.02	eyetracked subject #7
5	2	-35.77	L2	0.78	0.03	eyetracked subject #8
5	3	-25.17	EQ	0.87	0.04	open box
6	1	-16.27	L2	0.91	0.43	eyetracked subject #9
6	2	-42.02	SOPH	0.62	0.13	eyetracked subject #10
6	3	-52.17	L0	0.62	0.01	open box

Table 5: Average Sender Lookup Times (in seconds) Across Game Parameters

Bias b	Response Time		State	Sender Payoffs	Receiver Payoffs	Sender-to-Receiver Ratio
	Periods 1-15	Periods 31-45				
0	9.78	7.24	0.83	2.93	1.71	1.72
1	11.77	8.76	0.81	3.80	2.66	1.43
2	16.84	8.99	0.91	4.67	3.26	1.43
all	13.47	8.52	0.86	3.99	2.72	1.47

Table 6: Average Lookup Time per Row Depending on the State

Bias b	True State Rows	Other State Rows	True-to-Other Ratio
0	2.76	0.47	5.89
1	3.88	0.64	6.02
2	4.29	0.91	4.70
overall	3.83	0.72	5.33

Table 7: Individual Lookup Linear Measure Scores for Various Level-k Types

Type	Subject ID	L1	L2	L3/EQ
L1	#2 (1-2)	<u>**0.24**</u>	0.22	0.19
	#3 (2-1)	<u>**0.16**</u>	0.15	0.14
	#6 (4-1)	<u>**0.26**</u>	0.24	0.18
	#7 (5-1)	<u>**0.41**</u>	0.33	0.28
	Average	<u>**0.27**</u>	0.23**	0.19***
L2	#8 (5-2)	<u>0.27</u>	**0.26**	0.21
	#9 (6-1)	0.22	<u>**0.24**</u>	0.19
	Average	0.24	<u>**0.25**</u>	0.20*
SOPH	#1 (1-1)	<u>0.17</u>	0.16	0.13
	#5 (3-2)	<u>0.16</u>	0.15	0.11
	#10 (6-2)	<u>0.21</u>	0.13	0.07
	Average	<u>0.18</u>	0.15	0.10

Note: Highest lookups scores <u>underlined</u>. Lookup scores if choice classifications correspond to lookups **boldfaced**. Note that they almost always coincide for L1 and L2 types.
*, ** and *** denotes p<0.05, p<0.01, p<0.0001 for signed rank sum test using both own and other cells for each state, each bias, and each subject (of that type) with total lookup time > 1sec.

Table 8: Pupil Size Regressions for 400 msec Intervals

Y	$PUPIL_i$	-1.2~ -0.8sec	-0.8~ -0.4sec	-0.4~ 0.0sec	0.0~ 0.4sec	0.4~ 0.8sec
constant	α	107.27	108.03	106.19	109.56	108.67
		(2.81)	(2.55)	(2.57)	(2.05)	(2.16)
LIE_SIZE * $BIAS_b$ interactions	β_{10}	2.83	2.36	3.07	5.35**	5.57*
		(1.85)	(2.23)	(2.46)	(1.76)	(2.19)
	β_{11}	-1.02	-0.46	-0.36	2.16^	2.64*
		(1.26)	(1.31)	(1.28)	(1.21)	(1.15)
	β_{12}	2.06*	1.52^	1.47*	1.83*	2.00**
		(0.86)	(0.79)	(0.75)	(0.75)	(0.74)
	N	414	415	414	415	414
	χ^2	323.86	235.43	194.40	258.49	352.49
	R^2	0.291	0.299	0.263	0.365	0.438

Note: Robust standard error in parentheses; t-Test p-values lower than ^10 percent, *5 percent, ** 1 percent, and *** 0.1 percent. (Dummies for biases, states, individual subjects and individual learning trends are included in the regression, but results are omitted.)

Table 9: Predicting True States (Resampling 100 times) (s. e. in parentheses)

X		Hidden Bias-Stranger	
MESSAGE * BIAS = 1	β_{11}	0.46**	(0.12)
MESSAGE * BIAS = 2	β_{12}	0.42**	(0.09)
ROW_{self} * BIAS=1	β_{21}	1.07**	(0.24)
ROW_{self} * BIAS=2	β_{22}	1.72**	(0.20)
ROW_{other} * BIAS=1	β_{31}	1.27**	(0.22)
ROW_{other} * BIAS=2	β_{32}	0.44**	(0.15)
total observations N [a]		357	
N used in estimation		238.3	
N used to predict		118.7	
		Actual Data	Hold-out Sample
Percent of wrong prediction (b=1)		58.5	28.9
Percent of errors of size (1,2,3+) (b=1)		(61, 28, 11)	(79, 19, 2)
Average predicted payoff (b=1) [b]		87.5 (28.8)	101.7** (2.1)
Percent of wrong prediction (b=2)		77.9	37.9
Percent of errors of size (1,2,3+) (b=2)		(60, 30, 10)	(72, 24, 4)
Average predicted payoff (b=2) [b]		80.9 (26.9)	98.0** (2.2)

Note: * and ** Denotes p<0.05 and p<0.001 (t-test)

[a] Observation with less than 0.5 seconds lookup time and without the needed pupil size measures are excluded.

[b] Two sample t-test conducted against the actual payoffs of receivers in the experiment who are paired with eyetracked senders.

Figure 1: Raw Data Pie Charts (b=0)
(Hidden Bias-Stranger)

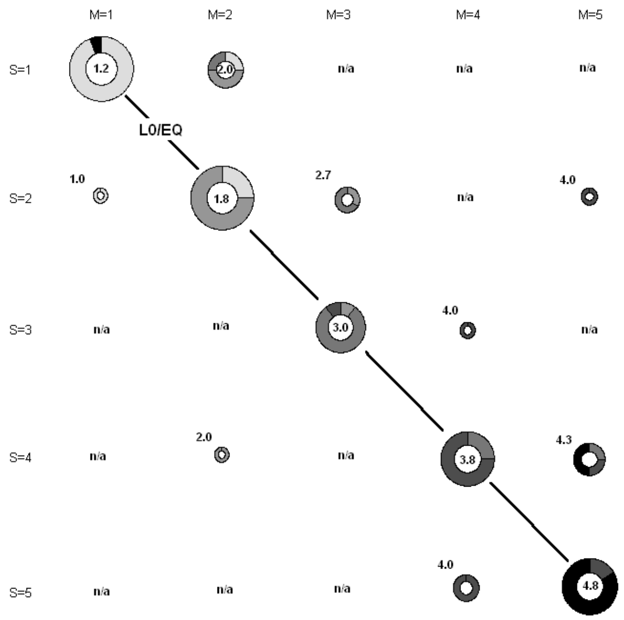


Figure 2: Raw Data Pie Chart (b=1)
(Hidden Bias-Stranger)

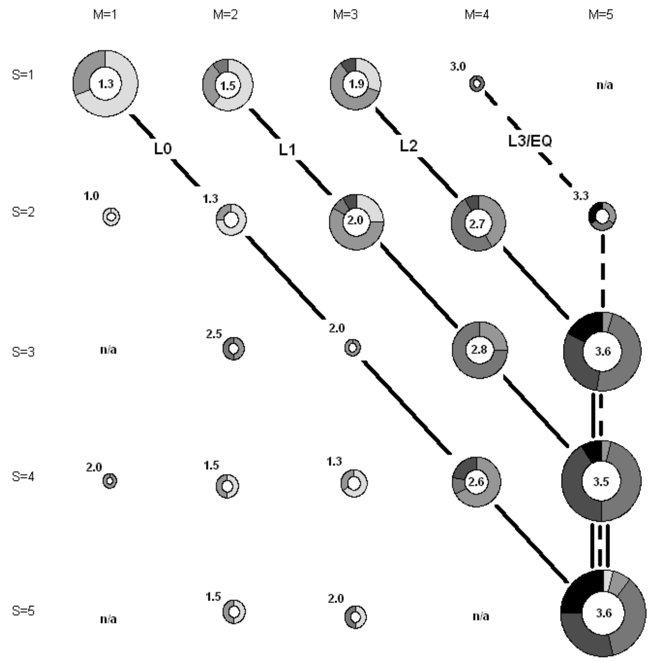


Figure 3: Raw Data Pie Chart (b=2)
(Hidden Bias-Stranger)

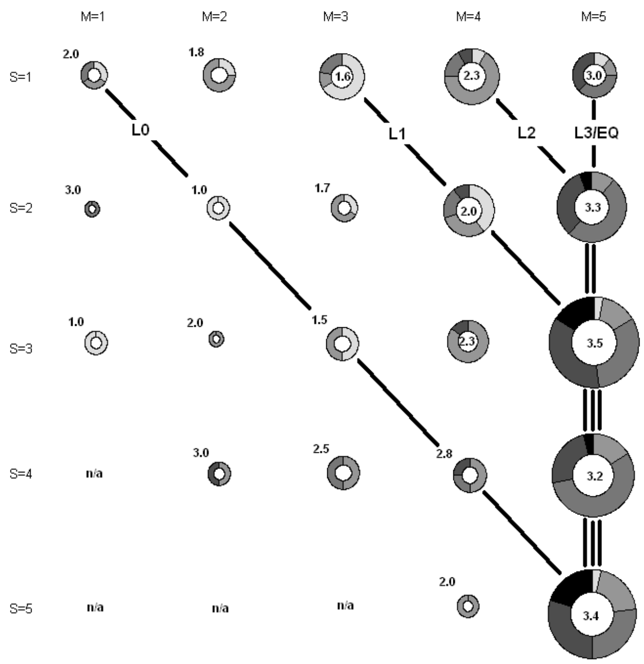
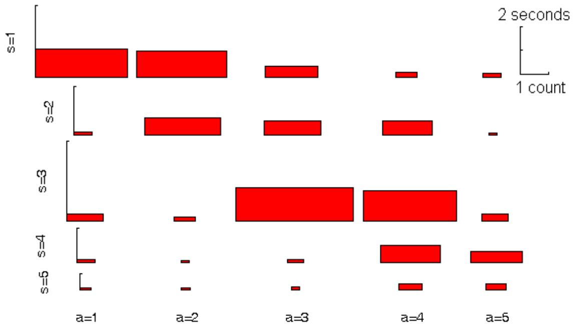


Figure 4: Lookup Icon Graph for b=1, Hidden Bias-Stranger, Type = L1

Part (a): Sender Payoffs



Part (b): Receiver Payoffs

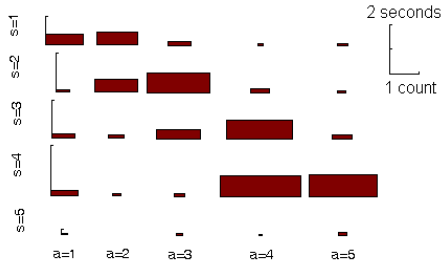
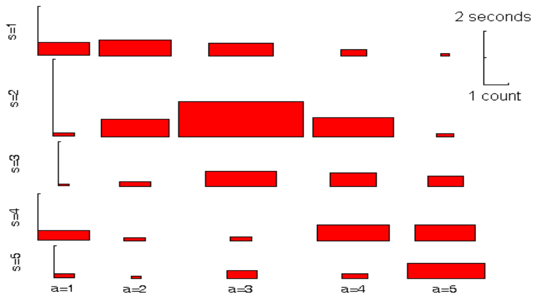
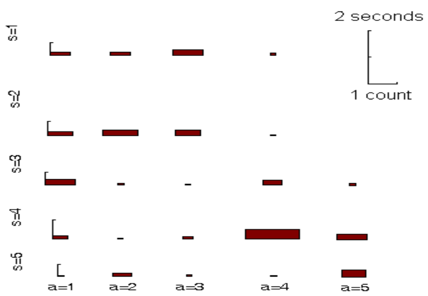


Figure 5: Lookup Icon Graph for b=1, Hidden Bias-Stranger, Type = L2

Part (a): Sender Payoffs



Part (b): Receiver Payoffs



Each row reports the lookup counts and time for the "true state row" corresponding to the given true state. The width of each box is scaled by the number of lookups and the height by the length of lookups (scaled by the little "ruler" in the upper right corner). The vertical bar on the first column icon represents the total lookup time summed across each row.

Figure 6: Lookup Icon Graph for b=2, Hidden Bias-Stranger, Type = L1

Part (a): Sender Payoffs **Part (b): Receiver Payoffs**

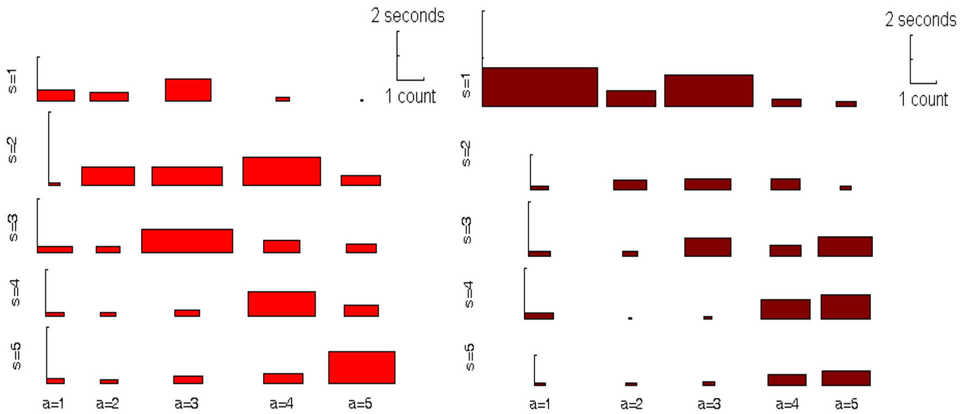
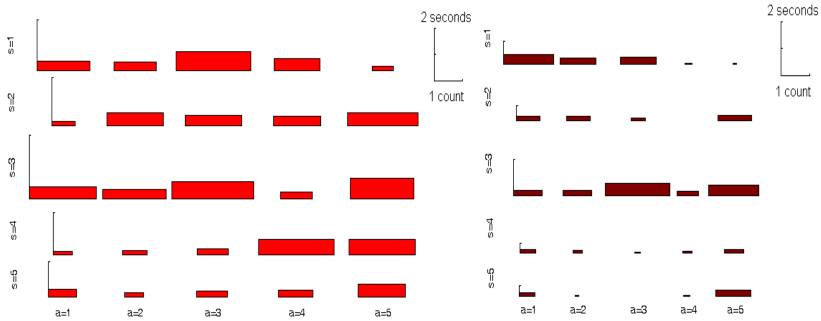


Figure 7: Lookup Icon Graph for b=2, Hidden Bias-Stranger,, Type = L2

Part (a): Sender Payoffs **Part (b): Receiver Payoffs**



Each row reports the lookup counts and time for the "true state row" corresponding to the given true state. The width of each box is scaled by the number of lookups and the height by the length of lookups (scaled by the little "ruler" in the upper right corner). The vertical bar on the first column icon represents the total lookup time summed across each row.

Appendix for Online Access [NOT INTENDED FOR PUBLICATION]

Appendix: Methods

Since this paper incorporates economics experiments in the laboratory, eyetracking devices, and studies the issue of deception, we expect to have readers who come from various backgrounds, such as economic theory, experimental economics, psychophysiology, and lie-detection. Therefore, we use this methodology appendix to address issues that might already be very familiar to some readers, but not to the rest. In particular, section 1 introduces video-based eyetracking to economists who are interested in learning about methods to study information acquisition, and section 2 demonstrates the relevance of eyetracking in economic experiments. Section 3 provides an argument for adding yet another paradigm (sender-receiver games) to study lie-detection, instead of adopting previous tasks such as CQT, GKT, etc. Section 4 provides the technical details of the equipment and software programs used in this study for those who are interested in replicating our results or applying this technique in future research.

A.I What is Eyetracking?

There are several ways to track a person's eyes. One of the most reliable and non-invasive way is video-based. Video-based eyetracking works by placing cameras in front of subject's eyes to capture eye images and corneal reflection of infrared sensors, and record changes up to 50-250Hz. Using eye movement images when subjects are told to fixate on certain positions on the screen, a procedure called "calibration," the experimenter can trace eye fixations and saccades on the screen and infer subject information acquisition patterns. In addition to information lookups, the eyetracker also records pupil dilation, which is correlated with arousal, pain, and cognitive difficulty. Therefore, eyetracking provides additional data about one's decision making process, uncovering previously unobservable parameters.[1]

A.II What Does Eyetracking Tell Us About the "Real World"?

Since economists are used to judging theories only by whether they predict choices accurately, it is useful to ask what direct measurement of eye fixations and pupil dilation can add. One possible inferential strategy from eyetracking is to separate competing theories that explain the same behavior. Previous studies compared offers and lookups in three-period alternating-offer bargaining (Camerer et al., 1993; Johnson et al., 2002), and in initial responses to normal-form games and two-person guessing games (Costa-Gomes et al., 2001; Costa-Gomes and Crawford, 2006). In those experiments, the same choices could be caused by different decision rules, such as L1 (optimize against perceived random play) and D1 (optimize against perceived random play excluding dominated strategies) in Costa-Gomes et al. (2001), but are separated by different lookup generated by these rules.[2] These studies illustrate the potential for using cognitive data, besides choices, for distinguishing between competing theories or inspiring new theory.[3]

---

[1] One potential concern of adopting eyetracking is scrutiny. For example, in our experiments senders could have been more truthful simply because they were watched. Indeed, we do find many L0 and L1 types (seven out of twelve) in the display bias-partner design. But subjects could be more truthful due to the repeated game effect. Hence, such concerns should be dealt with empirically by comparing eyetracked and open box subjects. In our experiment, the hidden bias-stranger adopts random matching and contains both eye-tracked and open boxed subjects. Overall type classification results are similar to Cai and Wang (2006). Although the sub-samples of eyetracked and open box subjects do show some interesting differences, the average level of strategic thinking is comparable: None of the eyetracked subjects were EQ (L3), but there were many SOPH; none of the open box subjects were L1, but the only L0 subject was an open box. This results in lower correlation between state and message for the open box subjects, but there is still little difference in payoffs. Hence, we conclude that there is no striking difference between the two, though the sample size is small.

[2] For example, in the three-stage bargaining game of Camerer et al. (1993) and Johnson et al. (2002), opening offers typically fell between an equal split of the first-period surplus and the subgame perfect equilibrium prediction (assuming self-interest). These offers could be caused by limited strategic thinking (i.e., players do not always look ahead to the second and third round payoffs of the game), or by computing an equilibrium by looking ahead, adjusting for fairness concerns of other players. The failure to look at payoffs in future periods showed that the deviation of offers from equilibrium was (at least partly) due to limited strategic thinking, rather than entirely due to equilibrium adjustment for fairness (unless "fairness" means not at all responding to advantages conferred by the strategic structure). Furthermore, comparing across rounds, when players do look ahead at future round payoffs their resulting offer are

Lookup patterns and pupil dilation could be useful in the sender-receiver games, because it could potentially be used to distinguish between competing theories for overcommunication. Although our experiments are not designed to separate these theories, overcommunication of the true state is consistent with two rough accounts: guilt and cognitive difficulty. Senders may feel guilty about deceiving the receivers and potentially costing the receivers money. This is the direct cost of lying. According to this theory, senders will look at the receiver payoffs (since seeing those payoffs is the basis of guilt) and their pupils will dilate when they misrepresent the state (i.e., choose M different from S) due to emotional arousal from guilt. In this story, the guilt springs from the senders' realization that their actions cost the receivers money, which depends on seeing the receiver payoffs.

A different story is that senders find it cognitively difficult to figure out how much to misrepresent the state. For example, senders might believe that some other senders always tell the truth, and receivers might therefore believe messages are truthful. Then strategic senders have to think hard about how much to misrepresent the state to take advantage of the receivers' naïveté (as in Crawford, 2003, Kartik, Ottaviani and Squintani, 2007, Chen, 2007, and Kartik, 2008). In this story, senders do not have to pay much attention to receiver payoffs but their pupils will dilate because of the cognitive difficulty of figuring out precisely how much to exaggerate.

Ultimately, the goal is to open up the black box of human brain, and model the decision process of human behavior, which is similar to what has been done to the firm. Instead of dwelling

---

closer to the self-interested equilibrium prediction (see Johnson and Camerer, 2004). Thus, the lookup data can actually be used to predict choices, to some degree.

[3] Another example comes from the accounting literature: James E. Hunton and McEwen (1997) asked analysts under hypothetical incentive schemes to make earnings forecast based on real firm data, and investigated factors that affect the accuracy of these forecasts. Using an eye-movement computer technology (Integrated Retinal Imaging System, IRIS), they find that analysts who employ a "directive information search strategy" make more accurate forecasts, both in the lab and in the field, even after controlling for years of experience. This indicates that eyetracking may provide an alternative measure of experience or expertise that is not simply captured by seniority. Had they not observed the eye movements, they could not have measured the difference in information search which is linked to accuracy.

on the neoclassical theory of the firm, which is merely a production function, modern economics has opened up the black box of the firm, and explicitly modeled its internal structure, such as the command hierarchy, principle-agent issues, and team production. Though there is still much to be done before we come close to what has been achieved in industrial organization, eyetracking provides a window to the soul and gives us a hint of the decision-making process inside the brain. Just as we may infer a factory's technology level by observing its inputs and wastes, we may also infer a person's reasoning process by observing the information he or she acquires (inputs) and how hard does he think (indexed by pupillary response).

A.III What Does Economics Have to Offer Regarding Lie-detection?

This study introduces an economic framework that is missing in most previous psychophysical studies on deception and lie detection. An advantage of the strategic information transmission game for studying deception is that game theory makes equilibrium predictions about how much informed agents will exaggerate what they know, when they know that other agents are fully-informed about the game's structure and the incentives to exaggerate. Even when equilibrium predictions fail, there are various behavioral models, such as level-k reasoning and quantal response equilibrium, which provide precise predictions that are testable in the lab. And while in most other deception studies,[4] subjects are instructed to lie or give weak or poorly controlled incentives,[5] subjects in experiments like ours choose voluntarily whether to deceive others or not (see also John

[4] For a survey of studies on (skin-conductance) polygraph, see Theodore R. Bashore and Paul E. Rapp (1993). For lie-detection studies in psychology, see the reviews of Robert E. Kraut (1980) and Aldert Vrij (2000). For a comprehensive discussion of different cues used to detect lies, see Bella M. DePaulo et al. (2003). For individual differences in lie-detection (Secret Service, CIA and sheriffs do better), see Paul Ekman and Maureen O'Sullivan (1991) and Ekman et al. (1999). More recently studies in neuroscience using functional magnetic resonance imaging (fMRI) include Sean A. Spence et al. (2001), D. D. Langleben et al. (2002) and F. Andrew Kozel et al. (2004).

[5] One exception is Samantha Mann et al. (2004) which used footage of real world suspect interrogation to test lie-detecting abilities of ordinary police. However, a lot of experimental control is lost in this setting. One interesting findings in this study is that counter to conventional wisdom, the more subjects relied on stereotypical cues such as gaze aversion to detect lies, the *less* accurate they were.

Dickhaut et al., 1995, Andreas Blume et al., 1998, 2001 and Cai and Wang, 2006).[6] Senders and receivers also have clear measurable economic incentives to deceive and to detect deception.[7]

A.IV Technological Details

Eyetracking data and button responses are recorded using the mobile Eyelink II head-mounted eyetracking system (SR Research, Osgoode, Ontario). Eyetracking data are recorded at 250 Hz. The mobile Eyelink II is a pair of tiny cameras mounted on a lightweight rack facing toward the subjects' eyes, and is supported by comfortable head straps. Subjects can move their heads and a period of calibration adjusts for head movement to infer accurately where the subject is looking. Nine-point calibrations and validations are performed prior to the start of each experiment in a participant's session. Accuracy in the validations typically is better than 0.5º of visual angle. Experiments are run under Windows XP (Microsoft, Inc.) in Matlab (Mathworks, Inc., Natick, MA) using the Psychophysics Toolbox (David H. Brainard, 1997; Denis G. Pelli, 1997) and the Eyelink Toolbox (Frans W. Cornelissen et al., 2002).

Eyetracking data are analyzed for fixations using the Eyelink Data Viewer (SR Research, Hamilton, Ontario). In discriminating fixations, we set saccade velocity, acceleration, and motion

---

[6] In fact, when the senders were asked after the experiment whether they considered sending a number different from the true state deception, 8 of the subjects said yes, while another 3 said no, but gave excuses such as "it's part of the game" or "the other player knows my preference difference." Only 1 subject said no without any explanation. These debriefing results also suggest that guilt has played little role in the experiment.

[7] Most lie-detection studies have three drawbacks: (1) They do not use naturally-occurring lies (because it is then difficult to know whether people are actually lying or not). Instead, most studies create artificial lies by giving subjects true and false statements (or creating a "crime scenario") and instructing them to either lie or tell the truth, sometimes to fool a lie-detecting algorithm or subject. However, instructed deception can be different than naturally-occurring voluntary deception, and the ability to detect instructed deception might be different than detecting voluntary deception. (2) The incentives to deceive in these studies are typically weak or poorly controlled (e.g., in Spence et al. (2001) all subjects were told that they successfully fooled the investigators who tried to detect them; in Mark G. Frank and Ekman (1997), subjects were threatened with "sitting on a cold, metal chair inside a cramped, darkened room labeled ominously XXX, where they would have to endure anywhere from 10 to 40 randomly sequenced, 110-decibel startling blasts of white noise over the course of 1 hr" but never actually enforcing it.). (3) Subjects are typically not economically motivated to detect deception. Experiments using the strategic-transmission paradigm from game theory address all these drawbacks.

thresholds to 30º/sec, 9500º/sec^2, and 0.15º, respectively. Regions of interest (ROIs), or the boxes subject look up, are drawn on each task image using the drawing functions within the Data Viewer. Measures of gaze include Fixation Number (i.e., the total number of fixations within an ROI) and Fractional Dwell Time (i.e., the time during a given round spent on fixating a given ROI divided by the total time between image onset and response). Only those fixations beginning between 50ms following the onset of a task image and offset of the task image are considered for analysis.

All task images are presented on a CRT monitor (15.9 in x 11.9 in) operating at 85 or 100 Hz vertical refresh rate with a resolution of 1600 pixels x 1200 pixels, and at an eye-to-screen distance of approximately 24 inches, thus subtending ~36 degrees of visual angle.

References

Bashore, Theodore R. and Rapp, Paul E. "Are There Alternatives to Traditional Polygraph Procedures." *Psychological Bulletin*, 1993, *113*(1), pp. 3-22.

Brainard, David H. "The Psychophysics Toolbox." *Spatial Vision*, 1997, *10*, pp. 433-36.

Camerer, Colin F.; Johnson, Eric J.; Rymon, Talia and Sen, Sankar. "Cognition and Framing in Sequential Bargaining for Gains and Losses," K. G. Binmore, A. P. Kirman and P. Tani, *Frontiers of Game Theory.* Cambridge: MIT Press, 1993, 27-47.

Chen, Ying. "Perturbed Communication Games with Honest Senders and Naive Receivers," Unpublished paper, 2007.

Cornelissen, Frans W.; Peters, Enno M. and Palmer, John. "The Eyelink Toolbox: Eye Tracking with Matlab and the Psychophysics Toolbox." *Behavior Research Methods, Instruments & Computers*, 2002, *34*, pp. 613-17.

Costa-Gomes, Miguel; Crawford, Vincent P. and Broseta, Bruno. "Cognition and Behavior in

Normal-Form Games: An Experimental Study." *Econometrica*, 2001, *69*(5), pp. 1193-235.

Crawford, Vincent P. "Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions." *American Economic Review*, 2003, *93*(1), pp. 133-49.

DePaulo, Bella M.; Lindsay, James J.; Malone, Brian E.; Muhlenbruck, Laura; Charlton, Kelly and Cooper, Harris. "Cues to Deception." *Psychological Bulletin*, 2003, *129*(1), pp. 74-118.

Ekman, Paul and O'Sullivan, Maureen. "Who Can Catch a Liar?" *American Psychologist*, 1991, *46*, pp. 913-20.

Ekman, Paul; O'Sullivan, Maureen and Frank, Mark G. "A Few Can Catch a Liar." *Psychological Science*, 1999, *10*, pp. 263-66.

Frank, Mark G. and Ekman, Paul. "The Ability to Detect Deceit Generalizes Across Different Types of High-Stake Lies." *Journal of Personality and Social Psychology*, 1997, *72*(6), pp. 1429-39.

Hunton, James E. and McEwen, Ruth A. "An Assessment of the Relation between Analysts' Earnings Forecast Accuracy, Motivational Incentives and Cognitive Information Search Strategy." *Accounting Review*, 1997, *72*(4), pp. 497-515.

Johnson, Eric J. and Camerer, Colin F. "Thinking Backward and Forward in Games," I. Brocas and J. Castillo, *The Psychology of Economic Decisions, Vol.2: Reasons and Choices.* Oxford University Press, 2004.

Kartik, Navin; Ottaviani, Macro and Squintani, Francesco. "Credulity, Lies, and Costly Talk." *Journal of Economic Theory*, 2007, *136 pp.* (1), pp. 749-58.

Kartik, Navin. "Strategic Communication with Lying Costs." *Review of Economic Studies*, 2008, *forthcoming*.

Kozel, F. Andrew; Revell, Letty J.; Lorberbaum, Jeffrey P.; Shastri, Ananda; Elhai, Jon D.; Horner, Michael David; Smith, Adam; Nahas, Ziad; Bohning, Daryl E. and George, Mark S. "A Pilot Study of Functional Magnetic Resonance Imaging Brain Correlates of Deception in Healthy Young Men." *Journal of Neuropsychiatry and Clinical Neurosciences*, 2004, *16*, pp. 295-305.

Kraut, Robert E. "Humans as Lie Detectors: Some Second Thoughts." *Journal of Communication*, 1980, *30*, pp. 209-16.

Langleben, D. D.; Schoroeder, L.; Maldjian, J. A.; Gur, R. C.; McDonald, S.; Ragland, J. D.; O'Brien, C. P. and Childress, A. R. "Brain Activity During Simulated Deception: An Event-Related Functional Magnetic Resonance Study." *NeuroImage*, 2002, *15*(3), pp. 727-32.

Mann, Samantha; Vrij, Aldert and Bull, Ray. Detecting True Lies: Police Officers' Ability to Detect Suspects' Lies, *Journal of applied psychology*, 2004, *89*(1), pp. 137-49.

Pelli, Denis G. "The Videotoolbox Software for Visual Psychophysics: Transforming Numbers into Movies." *Spatial Vision*, 1997, *10*, pp. 437-42.

Spence, Sean A.; Farrow, Tom F. D.; Herford, Amy E.; Wilkinson, Iain D.; Zheng, Ying and Woodruff, Peter W. R. "Behavioural and Functional Anatomical Correlates of Deception in Humans." *NeuroReport*, 2001, *12*(13), pp. 2849-53.

Vrij, Aldert. *Detecting Lies and Deceit: The Psychology of Lying and the Implications for Professional Practice*. Chichester: Wiley and Sons, 2000.

Appendix: Experiment Instructions

The experiment you are participating in consists of 1 session, having 45 rounds. At the end of the last session, you will be asked to fill out a questionnaire and paid the total amount you have accumulated during the course of the sessions in addition to a $5 show-up fee. Everybody will be paid in private after showing the record sheet. You are under no obligation to tell others how much you earned.

During the experiment all the earnings are denominated in FRANCS. Your dollar earnings are determined by the FRANC/$ exchange rate: 200 FRANCS = $1.

In each round, the computer program generates a secret number that is randomly drawn from the set {1,2,3,4,5}. The computer will display this secret number on member A's screen. After receiving the number, member A will send the message "The number I received is XX," to member B by staring at box XX. Hearing the message from member A, member B will then choose an action. In particular, member B can choose action 1, 2, 3, 4, or 5, using the game pad. Earnings of both members depend on the secret number and member B's action.

Member B's earnings is higher when member B's action is closer to the secret number, while member A's earnings is higher when member B's action is closer to the secret number **plus the preference difference**. The preference difference is either 0, 1 or 2, with equal chance, and will also be displayed and announced at the beginning of each round.

For example, if the preference difference is 2 and the secret number is 3, member B's earnings are higher if his or her action is closer to 3. However, member A's earnings is higher when member B's action is closer to 3 + 2 = 5. The earning tables are provided to you for convenience.

To summarize, in each round, the computer will display the preference difference and the secret number on member A's screen. Then, member A stares at a box (on the right) containing the desired message. Member B will hear the preference difference and the message "The number I received is XX," and then choose an action. The secret number is revealed after this choice, and earnings are determined accordingly.

Practice Session: 3 Rounds

Session 1: 45 Rounds

Member B: Please make sure you record the earnings in your record sheet. **Your payments will be rounded up.** Thank you for your participation.

Appendix: Supplemental Figures and Tables

Figure S1: Sender Screen for b=1 and S=4 without payoff perturbation

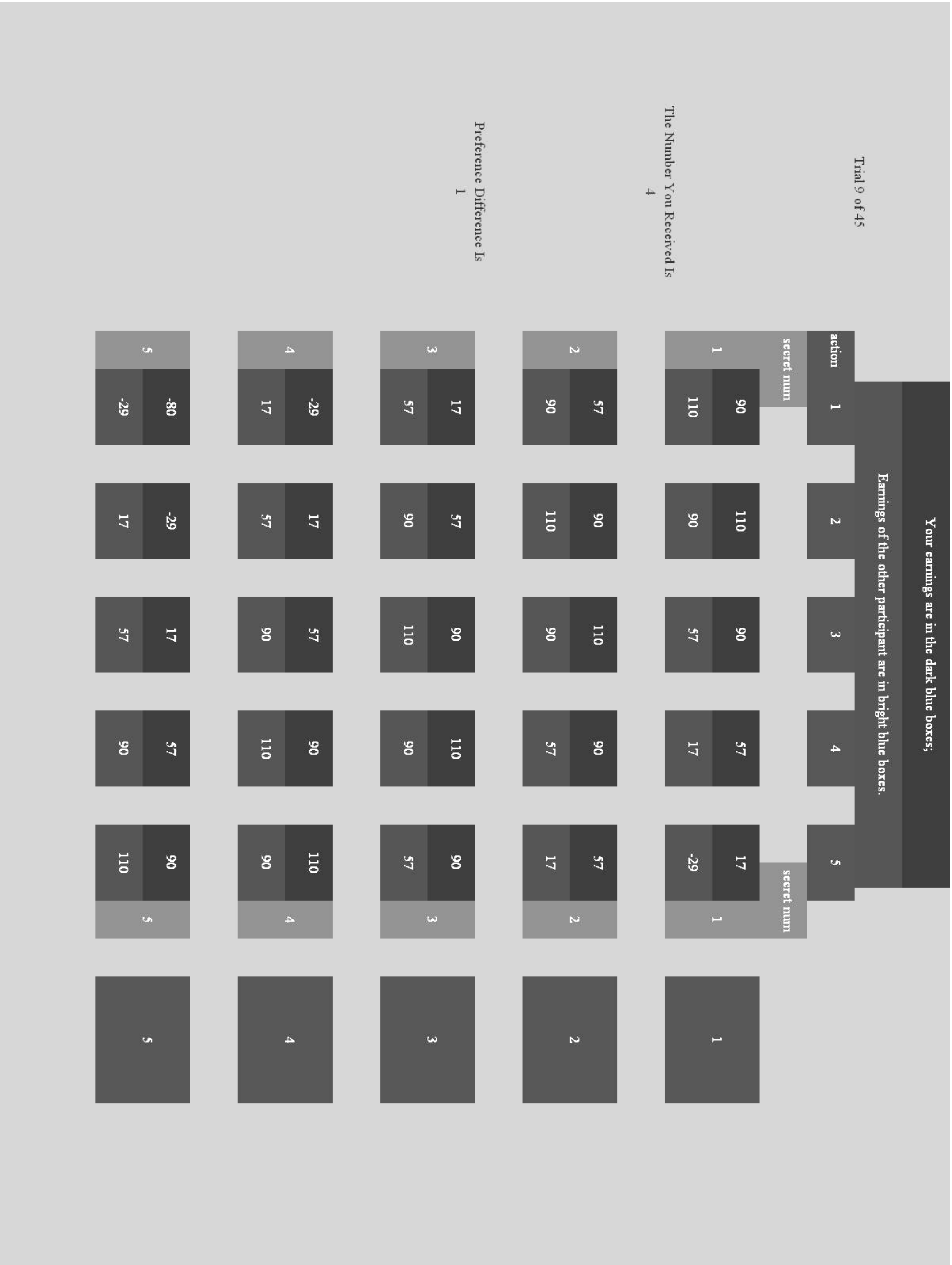


Figure S2: Raw Data Pie Charts (b=0), (Display Bias-Partner)

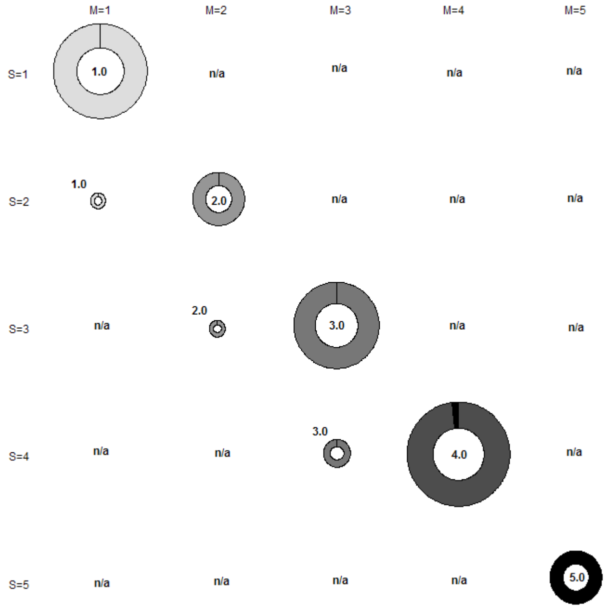


Figure S3: Raw Data Pie Chart (b=1), (Display Bias-Partner)

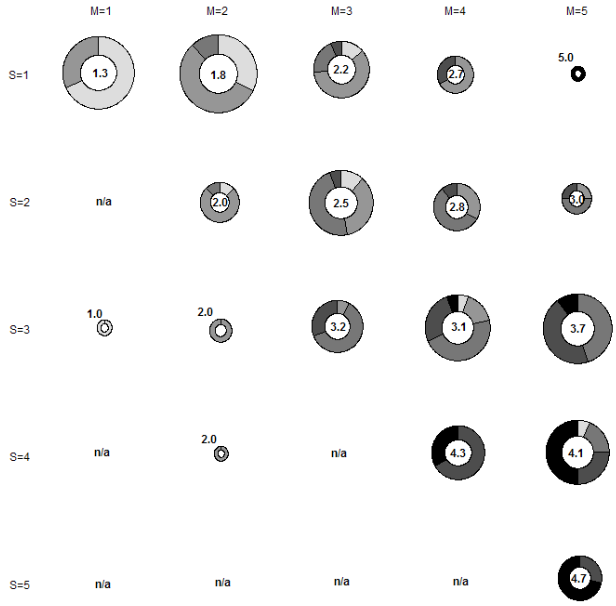
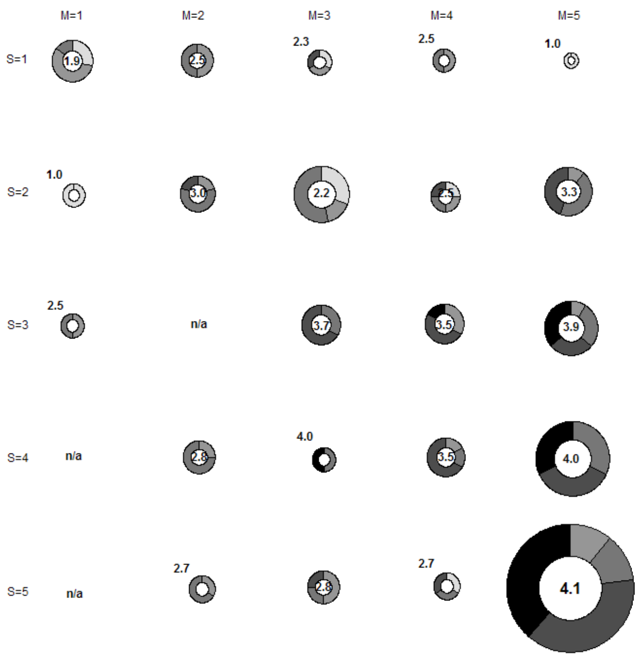


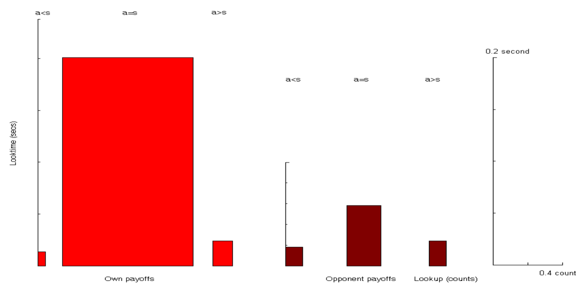
Figure S4: Raw Data Pie Chart (b=2), (Display Bias-Partner)



The true states are in rows, and senders' messages are in columns. Each cell contains the average action taken by the receivers and a pie chart break down of the actions. Actions are presented in a gray scale, ranging from white (action 1) to black (action 5). The size of the pie chart is proportional to the number of occurrences for the corresponding state and message.

Figure S5: Lookup Icon Graph for b=0, Type = all

Part (a): Display Bias-Partner



Part (b): Hidden Bias-Stranger

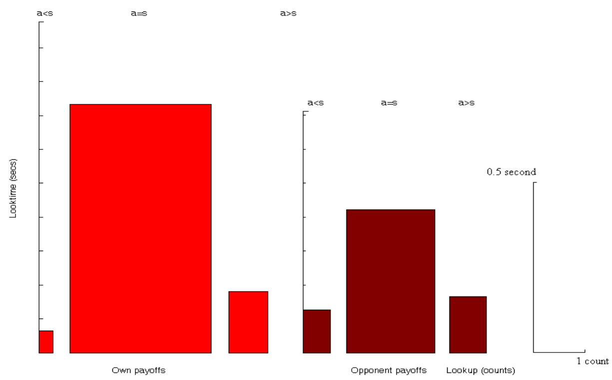
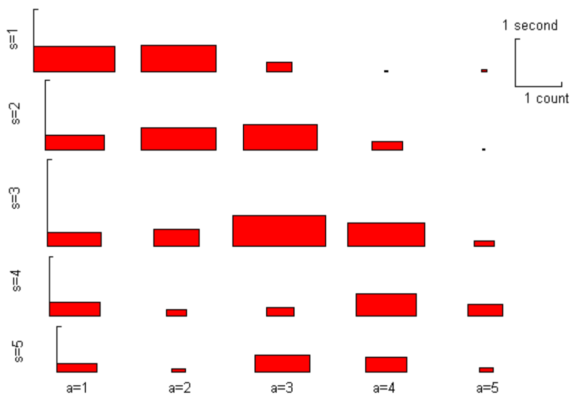


Figure S6: Lookup Icon Graph for b=1, Display Bias-Partner, Type = all

Part (a): Sender Payoffs



Part (b): Receiver Payoffs

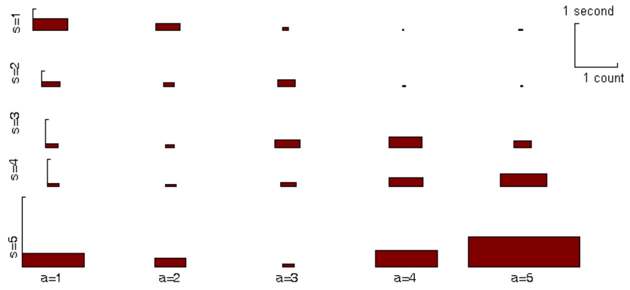
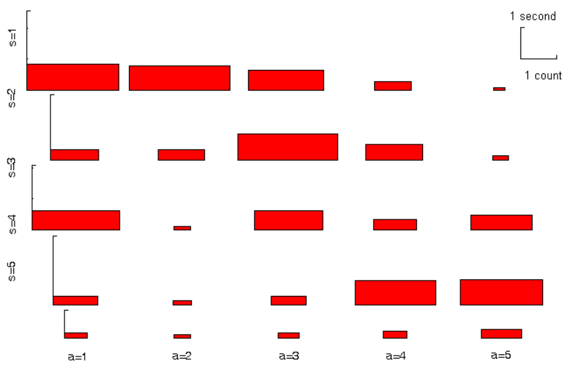
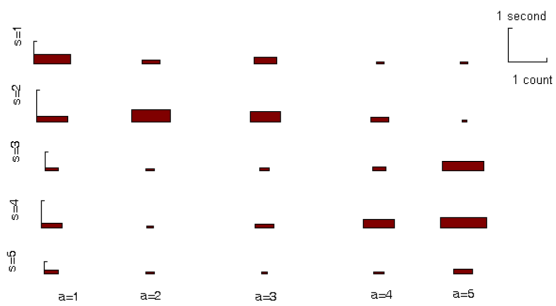


Figure S7: Lookup Icon Graph for b=2, Display Bias-Partner, Type = all

Part (a): Sender Payoffs



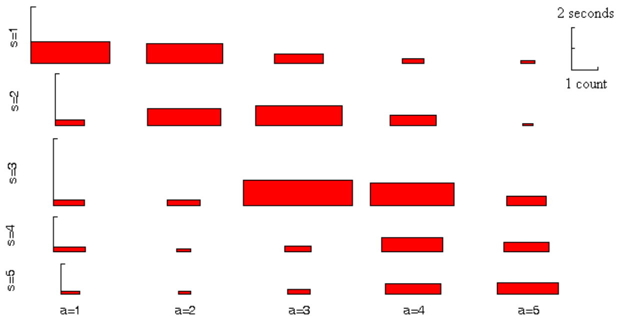
Part (b): Receiver Payoffs



Each row reports the lookup counts and time for the "true state row" corresponding to the given true state. The width of each box is scaled by the number of lookups and the height by the length of lookups (scaled by the little "ruler" in the upper right corner). The vertical bar on the first column icon represents the total lookup time summed across each row.

Figure S8: Lookup Icon Graph for b=1, Hidden Bias-Stranger, Type = all

Part (a): Sender Payoffs



Part (b): Receiver Payoffs

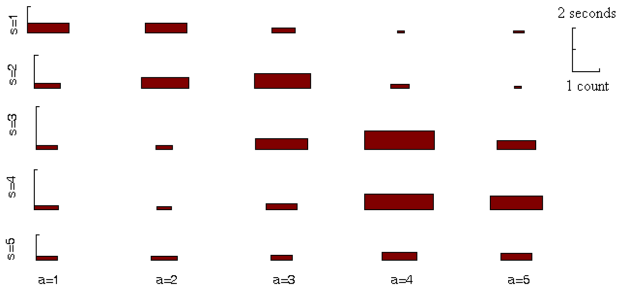
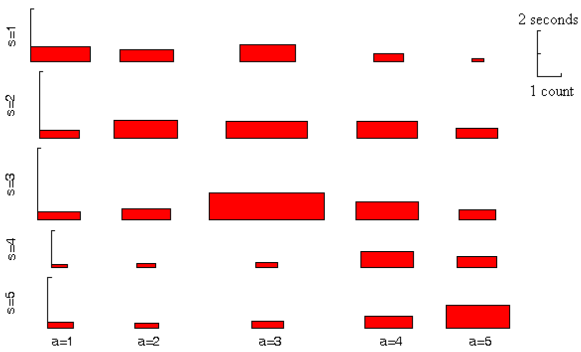
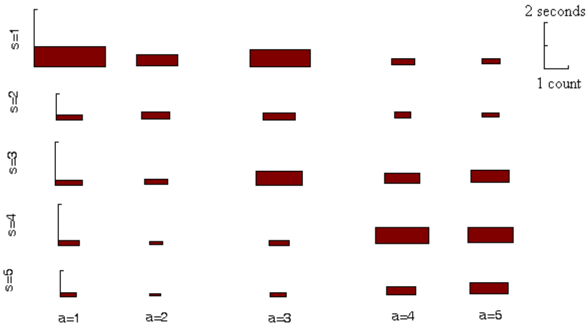


Figure S9: Lookup Icon Graph for b=2, Hidden Bias-Stranger, Type = all

Part (a): Sender Payoffs



Part (b): Receiver Payoffs



Each row reports the lookup counts and time for the "true state row" corresponding to the given true state. The width of each box is scaled by the number of lookups and the height by the length of lookups (scaled by the little "ruler" in the upper right corner). The vertical bar on the first column icon represents the total lookup time summed across each row.

Table S1A: Learning – Actual Information Transmission

Display Bias-Partner					
BIAS	Rounds	Corr(S, M)	Corr(M, A)	Corr(S, A)	Predicted Corr(S, A)
	1-15	0.880	0.833	0.732	
0	16-30	0.976	0.949	0.925	1.000
	31-45	0.937	0.942	0.919	
	1-15	0.620	0.730	0.477	
1	16-30	0.685	0.724	0.577	0.645
	31-45	0.598	0.713	0.415	
	1-15	0.384	0.584	0.372	
2	16-30	0.327	0.526	0.306	0.000
	31-45	0.279	0.643	0.291	

Hidden Bias-Stranger					
BIAS	Rounds	Corr(S, M)	Corr(M, A)	Corr(S, A)	Predicted Corr(S, A)
	1-15	0.887	0.816	0.716	
0	16-30	0.941	0.951	0.885	1.000
	31-45	0.888	0.944	0.866	
	1-15	0.602	0.730	0.436	
1	16-30	0.660	0.727	0.561	0.645
	31-45	0.555	0.714	0.393	
	1-15	0.380	0.592	0.372	
2	16-30	0.347	0.540	0.313	0.000
	31-45	0.232	0.636	0.288	

Table S1B: Learning Sender and Receiver's Payoffs

Display Bias-Partner				
BIAS	Rounds	u_S (std)	u_R (std)	Predicted u_R (std)
	1-15	96.36 (23.47)	96.48 (24.37)	
0	16-30	104.63 (11.65)	104.78 (12.01)	110.00 (0.00)
	31-45	103.50 (12.46)	103.19 (12.18)	
	1-15	79.38 (31.83)	87.04 (26.78)	
1	16-30	69.19 (40.15)	87.98 (28.94)	91.40 (19.39)
	31-45	71.83 (39.05)	85.52 (27.09)	
	1-15	46.06 (50.91)	80.63 (25.93)	
2	16-30	46.74 (51.11)	81.20 (27.63)	80.80 (20.76)
	31-45	35.87 (55.73)	79.70 (29.65)	
Hidden Bias-Stranger				
BIAS	Rounds	u_S (std)	u_R (std)	Predicted u_R (std)
	1-15	95.38 (23.56)	95.72 (24.15)	
0	16-30	102.40 (15.18)	102.52 (15.53)	110.00 (0.00)
	31-45	102.00 (16.89)	101.69 (17.30)	
	1-15	78.76 (35.63)	85.88 (28.92)	
1	16-30	69.18 (39.40)	87.45 (28.61)	91.40 (19.39)
	31-45	71.40 (38.82)	84.73 (26.87)	
	1-15	46.76 (49.84)	81.06 (26.36)	
2	16-30	46.75 (50.19)	81.81 (27.15)	80.80 (20.76)
	31-45	36.22 (55.94)	79.29 (29.10)	

Table S2: Information Transmission: Correlations between S, M and A, Display Bias-Partner

Bias	r(S, M)	r(M, A)	r(S, A)	Predicted r(S, A)
0	.99	1.00	.99	1.00
1	.73	.74	.72	.65
2	.63	.57	.50	.00

Note: In the display bias-partner design, all senders' eye movements were recorded ("eyetracked").

Table S3: Sender and Receiver's Payoffs, Display Bias-Partner

Bias	u_S (std)	u_R (std)	Pred. u_R (std)
0	109.14 (4.07) [a]	109.14 (4.07) [a]	110.00 (0.00)
1	93. 35 (20.75)	94.01 (19.86)	91.40 (19.39)
2	41.52 (49.98)	85.52 (25.60)	80.80 (20.76)

Note: [a] Payoffs are exactly the same for senders and receivers due to the symmetry of the payoffs when b=0.

Table S4: Level-k Classification Results, Display Bias-Partner

Session	ID	log L	*k*	Exact	lambda
1	1	-36.33	L0	0.71	0.06
2	2	-51.47	L0	0.64	0.00
3	3	-33.01	L0	0.78	0.03
4	4	-19.81	L1	0.82	0.49
5	5	-38.93	SOPH	0.76	0.04
6	6	-45.05	EQ	0.69	0.05
7	7	-34.89	L0	0.80	0.00
8	8	-27.36	L2	0.84	0.04
9	9	-31.80	L1	0.80	0.04
10	10	-24.30	L1	0.84	0.48
11	11	-22.35	L2	0.87	0.45
12	12	-31.07	L2	0.73	1.00

Table S5: Average Sender Lookup Times (in sec.) across Game Parameters, Display Bias-Partner

Bias b	Response Time Periods 1-15	Response Time Periods 31-45	State	Bias	Sender Payoffs	Receiver Payoffs	Sender-to-Receiver Ratio
0	5.42	2.39	0.65	0.41	0.73	0.27	2.70
1	7.92	5.44	1.47	0.99	2.29	1.05	2.18
2	9.73	8.12	1.72	1.52	3.03	1.50	2.02
all	8.07	5.25	1.34	1.02	2.14	1.00	2.14

Table S6: Average Lookup Time per Row Depending on the State, Display Bias-Partner

Bias b	True State Rows	Other State Rows	True-to-Other Ratio
0	0.54	0.11	4.91
1	2.06	0.32	6.44
2	2.24	0.57	4.28
overall	1.71	0.36	4.75

Table S7A: Average response time change for different biases, Display Bias-Partner

Bias	N	Average for first 15 rounds	N	Average for middle 15 rounds	N	Average for last 15 rounds
0	38	5.42	47	2.91	55	2.39
1	73	7.92	60	5.44	59	5.44
2	67	9.73	68	8.96	51	8.12
overall	178	8.07	175	6.13	165	5.25

* The numbers of observations are slightly different because we exclude 10 rounds where subjects had to use the keyboard to make their decision. Also, subject #4 had severe pain and the experimenter was forced to stop the experiment at the end of round 33.

Note: Since the bias was randomly determined each round, and subject #4 stopped at round 33 (due to excess pain wearing the eyetracker), numbers of observations are not equal. Dropping subject #4 does not change the results.

Table S7B: Average response time change for different biases, Hidden Bias-Stranger

Bias	N	Average for first 15 rounds	N	Average for middle 15 rounds	N	Average for last 15 rounds
0	30	9.78	24	5.54	29	7.24
1	56	11.77	58	10.78	59	8.76
2	61	16.84	65	10.23	49	8.99
overall	147	13.47	147	9.68	137	8.52

* The numbers of observations are slightly different because we exclude 12 rounds where subjects had to use the keyboard to make their decision. Also, subject #3 had calibration issues and the experimenter was forced to stop eyetracking at the end of round 40.

Note: Since the bias was randomly determined each round, and subject #4 stopped at round 40 wearing the eyetracker), numbers of observations are not equal.

Table S8: Pupil Size Regressions for 400 msec Intervals, Display Bias-Partner

Y	$PUPIL_i$	-1.2~ -0.8sec	-0.8~ -0.4sec	-0.4~ 0.0sec	0.0~ 0.4sec	0.4~ 0.8sec
constant	α	99.59	99.78	104.62	111.81	109.95
		(2.45)	(2.41)	(2.19)	(1.84)	(2.07)
LIE_SIZE * $BIAS_b$ interactions	β_{10}	1.20	6.41	3.92	-3.91	0.58
		(3.21)	(6.38)	(3.06)	(2.76)	(7.36)
	β_{11}	2.79*	3.40**	3.28**	4.55***	4.20***
		(1.19)	(1.17)	(0.97)	(0.86)	(0.73)
	β_{12}	3.49***	3.71***	3.04***	2.90**	3.28**
		(0.99)	(0.98)	(0.84)	(0.87)	(0.90)
	N	499	497	499	508	503
	χ^2	224.54	337.22	500.93	785.32	631.21
	R^2	0.271	0.346	0.455	0.539	0.557

Note: Robust standard error in parentheses; t-Test p-values lower than ^10 percent, *5 percent, ** 1 percent, and *** 0.1 percent. (Dummies for biases, states, individual subjects and individual learning trends are included in the regression, but results are omitted.)

Table S9: Predicting True States (Resampling 100 times, s.e. in parentheses), Display Bias-Partner

X		Display Bias-Partner	
MESSAGE * BIAS = 1	β_{11}	0.64*	(0.22)
MESSAGE * BIAS = 2	β_{12}	0.91**	(0.23)
ROW_{self} * BIAS=1	β_{21}	0.98**	(0.21)
ROW_{self} * BIAS=2	β_{22}	1.00**	(0.27)
ROW_{other} * BIAS=1	β_{31}	0.25	(0.16)
ROW_{other} * BIAS=2	β_{32}	0.39*	(0.17)
total observations N [a]		208	
N used in estimation		139.3	
N used to predict		68.7	

	Actual Data	Hold-out Sample
Percent of wrong prediction (b=1)	56.2	29.2
Percent of errors of size (1,2,3+) (b=1)	(80, 15, 5)	(74, 19, 7)
Average predicted payoff (b=1) [b]	93.4 (22.3)	100.7* (2.4)
Percent of wrong prediction (b=2)	70.9	58.7
Percent of errors of size (1,2,3+) (b=2)	(67, 26, 7)	(73, 22, 5)
Average predicted payoff (b=2) [b]	86.2 (23.8)	91.8* (3.4)

Note: * and ** Denotes p<0.05 and p<0.001 (t-test)

[a] Observation with less than 0.5 seconds lookup time and without the needed pupil size measures are excluded.

[b] Two sample t-test conducted against the actual payoffs of receivers in the experiment who are paired with eyetracked senders.

Table S10: Average Sender Fixation Counts and Lookup Time across Game Parameters

Treat-ment	Bias b	Res-ponse time (sec.)	State		Bias		Sender Payoffs		Receiver Payoffs	
			Fixation (count)	Lookup (sec.)	Fixation (count)	Lookup (sec.)	Fixation (count)	Lookup (sec.)	Fixation (count)	Lookup (sec.)
Displayed	0	3.59	2.6	0.65	2.1	0.41	3.0	0.73	1.4	0.27
Bias	1	6.86	5.0	1.47	3.9	0.99	8.1	2.29	3.9	1.05
- Partner	2	9.68	6.2	1.72	5.5	1.52	10.6	3.03	5.4	1.50
	overall	7.00	4.8	1.34	4.0	1.02	7.6	2.14	3.7	1.00
Hidden	0	7.65	3.0	0.83	-	-	12.0	2.93	7.5	1.71
Bias	1	10.95	3.1	0.81	-	-	14.2	3.80	10.7	2.66
- Stranger	2	12.91	3.4	0.91	-	-	17.5	4.67	12.4	3.26
	overall	11.12	3.2	0.86	-	-	15.1	3.99	10.8	2.72

Table S11: Average Fixation Counts and Lookup Time per Row

Treatment	Bias b	True State Rows		Other Rows	
		Fixation Counts (counts per row)	Lookup Time (sec. per row)	Fixation Counts (counts per row)	Lookup Time (sec. per row)
Displayed	0	2.2	0.54	0.5	0.11
Bias	1	6.8	2.06	1.3	0.32
- Partner	2	7.8	2.24	2.0	0.57
	overal l	5.9	1.71	1.3	0.36
Hidden	0	11.4	2.76	2.0	0.47
Bias	1	14.4	3.88	2.6	0.64
- Stranger	2	15.7	4.29	3.6	0.91
	overall	14.3	3.83	2.9	0.72

Table S12: Individual Types and Log Likelihood under Spike-logit and Logit Specification

Ses-	Sub-	Spike-logit (baseline)					Spike-logit (without bias=0)					Logit					Logit (without bias=0)				
sion	ject	L0	L1	L2	L3	SOPH	L0	L1	L2	L3	SOPH	L0	L1	L2	L3	SOPH	L0	L1	L2	L3	SOPH
1	1	-60.20	-55.68	-46.36	-53.16	**-46.23**	-50.47	-43.08	-36.68	-41.48	**-35.28**	-66.92	-52.59	-50.83	-54.65	**-48.72**	-54.44	-42.12	-40.07	-43.31	**-38.50**
1	2	-67.54	**-25.99**	-55.16	-56.98	-55.82	-55.14	**-24.72**	-50.15	-51.96	-50.80	-66.85	**-36.95**	-49.41	-51.79	-48.10	-57.46	**-33.19**	-42.58	-44.18	-42.21
1	3	-72.16	-50.97	**-15.98**	-40.06	-22.60	-56.92	-42.76	**-8.82**	-33.29	-17.74	-72.21	-46.26	**-16.26**	-31.31	-19.94	-57.94	-39.55	**-10.24**	-24.95	-16.32
2	1	-55.43	**-37.32**	-43.27	-43.29	-41.45	-46.88	**-30.94**	-36.70	-36.82	-35.30	-56.20	-35.57	-36.33	-37.68	-32.92	-47.12	-29.33	-29.29	-30.01	-26.56
2	2	-49.08	-47.07	-45.17	**-37.34**	-43.01	-41.28	-39.95	-38.68	**-32.57**	-37.13	-54.41	-48.18	-44.00	-40.05	-39.73	-42.03	-37.90	-34.10	**-30.40**	-31.55
2	3	-63.73	-49.05	-33.23	-31.65	**-25.70**	-49.74	-40.08	-25.05	-23.07	**-17.26**	-63.97	-43.32	-28.04	-27.62	**-24.89**	-49.89	-35.66	-20.95	-20.18	**-19.66**
3	1	**-68.32**	-68.84	-71.93	-71.16	-71.29	-56.34	-54.93	-57.48	-56.92	-57.48	**-69.40**	-71.94	-72.43	-72.43	-72.42	**-56.72**	-57.93	-57.94	-57.94	-57.94
3	2	-71.84	-47.10	-22.95	-30.78	**-17.71**	-62.49	-43.98	-21.86	-28.76	**-16.95**	-71.79	-41.49	-18.26	-27.31	-21.02	-62.77	-38.52	-16.86	-24.02	-19.27
3	3	-72.35	-71.84	-59.83	**-54.73**	-55.24	-64.40	-65.98	-57.14	**-52.56**	-53.07	-72.43	-71.80	-63.97	**-61.77**	-62.83	-65.99	-65.85	-59.46	**-57.33**	-58.77
4	1	-54.83	**-50.86**	-57.41	-62.43	-58.71	-48.26	**-43.88**	-49.87	-54.04	-50.51	-54.81	**-49.71**	-57.41	-61.08	-56.59	-47.41	**-42.74**	-48.53	-51.20	-48.16
4	3	-69.49	-43.38	-29.43	**-25.22**	-27.41	-56.24	-36.20	-22.70	**-18.81**	-21.88	-69.77	-38.12	-23.20	-22.61	-20.73	-56.15	-31.29	-16.80	**-15.57**	-15.89
5	1	-68.90	**-22.26**	-44.60	-42.75	-40.74	-61.32	**-21.50**	-41.94	-40.52	-38.51	-67.29	**-23.01**	-33.07	-35.16	-29.89	-60.41	**-21.50**	-29.46	-30.98	-27.38
5	2	-69.84	-54.26	**-35.77**	-48.07	-40.75	-54.31	-42.78	**-21.10**	-37.72	-30.23	-69.44	-48.58	-40.71	-45.07	-39.44	-54.20	-37.41	**-29.60**	-33.32	-30.16
5	3	-70.23	-44.73	-30.63	**-25.17**	-29.33	-61.00	-40.19	-26.93	**-21.44**	-26.26	-71.66	-41.34	-21.23	-19.50	-17.81	-61.16	-36.49	-17.38	-15.43	-15.35
6	1	-70.88	-46.20	**-16.27**	-35.62	-22.96	-57.94	-39.17	**-9.11**	-29.26	-17.88	-70.51	-38.41	**-14.12**	-23.23	-15.98	-57.89	-32.36	**-8.53**	-17.17	-12.80
6	2	-65.57	-49.32	-43.38	-47.52	**-42.02**	-56.82	-44.38	-38.05	-43.33	**-37.08**	-70.22	-47.91	-48.39	-52.75	**-45.64**	-57.70	-41.36	-40.60	-43.83	**-38.75**
6	3	**-53.12**	-68.57	-70.88	-71.41	-70.87	**-46.26**	-59.73	-62.40	-62.66	-62.35	**-56.49**	-67.30	-71.21	-71.31	-70.36	**-48.23**	-58.70	-62.09	-62.15	-61.50

Note: Maximum likelihood for each specification underlined. Classification results that are consistent with the baseline specification (spike-logit) are in **bold**.
Subject 3-1 has compliance rates less than 20 percent for all types under both spike-logit specifications, and hence, is deemed as unclassified.