Eliciting Private Information with Noise: The Case of Randomized Response*

Andreas Blume† Ernest K. Lai‡ Wooyoung Lim§

July 23, 2013

Abstract

The paper formalizes Warner's (1965) randomized response technique (RRT) as a game and implements it experimentally, thus linking game theoretic approaches to randomness in communication with survey practice in the field and a novel implementation in the lab. As predicted by our model and in line with Warner, the frequency of truthful responses is significantly higher with randomization than without. The model predicts that randomization weakly improves information elicitation, as measured in terms of *mutual information*, although, surprisingly, not always by RRT inducing truth-telling. Contrary to this prediction, randomization significantly reduces the elicited information in our experiment.

Keywords: Randomized Response; Lying Aversion; Stigmatization Aversion; Mutual Information; Laboratory Experiments

JEL classification: C72; C92; D82; D83

*We are grateful to Joel Sobel for valuable comments and suggestions. For helpful comments and discussions, we thank seminar participants at Lehigh University (Economics and Psychology Departments), HKUST, Rutgers University, Sungkyunkwan University, Shanghai University of Finance and Economics, Özyeğin University, Korea University, National Taiwan University, University of Arizona, IUPUI, The Chinese University of HK, City University of HK, Yeungnam University, Sogang University and conference participants at the Deception, Incentives and Behavior Conference, 2012 International ESA Conference, the 87th WEAI Annual Conference, Korean Econometric Society International Conference, the 4th World Congress of the Game Theory Society, Fall 2012 Midwest Economic Theory Meeting, the 47th Annual Conference of the Canadian Economic Association, and the 24th International Conference on Game Theory. This study is supported by a grant from the Research Grants Council of Hong Kong (Grant No. GRF-643511). Lai gratefully acknowledges financial support from the Office of the Vice President and Associate Provost for Research and Graduate Studies at Lehigh University. The paper was previously circulated and presented under the title "A Game Theoretic Approach to Randomized Response: Theory and Experiments."

†Department of Economics, The University of Arizona. ablume@email.arizona.edu

‡Department of Economics, Lehigh University. kwl409@lehigh.edu.

§Department of Economics, The Hong Kong University of Science and Technology. wooyoung@ust.hk.

1 Introduction

Recent theoretical work on noisy communication channels (Blume, Board and Kawamura, 2007), non-strategic mediators (Goltsman, Hörner, Pavlov and Squintani, 2009), strategic mediators (Ivanov, 2010), and survey methods (Ljungqvist, 1993) has explored the efficiency effects of making parts of the communication environment stochastic. It suggests that when there is conflict between those who have and those who might benefit from information, introducing randomness can help improve their communication.

Randomness moderates the inferences made from messages. When, for example, messages are sometimes lost, their non-arrival cannot entirely be attributed to those who would not send a message. Similarly, a "yes" answer is less revealing when the listener has only imperfect knowledge of the question in response to which the answer is given. In both cases randomness causes posterior beliefs of listeners and their responses to those beliefs to vary less across messages. This makes it easier for speakers to provide some but not all of their information. In short, randomness encourages information transmission by providing a cover.

This potential of noise providing a cover was recognized by Warner (1965), who proposed the *randomized response technique* (RRT) to elicit information about sensitive issues, like sexual behavior or drug use. In one version of RRT, a potential drug user is questioned by being asked to provide a yes/no answer in response to either the statement "I have used illegal drugs yesterday" or "I have not used illegal drugs yesterday." The interviewer knows the probability with which each question is asked, but in any given instance not the question itself. On one hand, this provides privacy protection for the survey respondent; a "yes" is not clear-cut evidence for drug use, even if the respondent is always truthful. On the other, it permits the interviewer to make inferences at the population level if the privacy protection is sufficient to induce truth-telling by respondents.[1]

RRT has been used to gather information about a large variety of sensitive issues, including drug use and doping (Striegel, Ulrich, and Simon, 2010), tax evasion (Houston and Tran, 2001), employee theft (Wimbush and Dalton, 1997), poaching (St John et al., 2012), regulatory non-compliance (Elffers, van der Heijden, and Hezemans, 2003) and the integrity of certified public accountants (Buchman and Tracy, 1982). Thus, at least in the domain of survey methods, there is some faith in the theoretical prediction that introducing randomness aids communication.

Is this faith justified? The use of RRT is premised on randomization inducing truth-telling. For this to be the case, a number of conditions have to hold jointly, not all of which have received close scrutiny in the literature. First, it must be the case that there is at least some preference

[1]The second question can be replaced by an unrelated question such as "Have you ever visited a local library?" with the *innocuous question* technique. Another version of the randomized response technique widely used in the survey statistics literature (e.g., St John et al., 2012) is the *forced response* technique proposed by Boruch (1971). With this approach, depending on the dice number they roll, respondents are instructed to either answer a sensitive question or to give a prescribed response irrespective of the truth.

for truth-telling so as to outweigh privacy concerns; second, respondents must appreciate and process the inference-moderating effect of randomness; and third, if the game that is induced between interviewer and respondent has multiple equilibria, then a truth-telling one must be selected.

Thus far, efforts to answer the question of whether RRT works as predicted, rather than examining the above conditions, have focused on two empirical validation methods, *individual validation* and *comparative validation*. The former relies on the rare instances when there is direct evidence on the question of interest that can be contrasted with the results from a randomized response study. The latter compares data from randomized response studies with those from alternative survey methods (self-administered questionnaires, telephone interviews, face-to-face interviews, and computer-assisted interviews). Examples of comparative validation studies are Beldt, Daniel and Garcha (1982), Wimbush and Dalton (1997), and Lensvelt-Mulders, Hox, van der Heijden and Maas (2005). Their results suggest that RRT improves on direct questioning according to the *more-is-better* criterion, where a higher population estimate of the stigmatizing trait is interpreted as being more valid.

A third avenue for validation, through controlled laboratory experiments, has thus far been largely unexplored. This is despite the fact that laboratory experiments have a number of advantages over other validation methods. As in individual validation studies, the exact population proportions are known; there is at least partial control over truth-telling preferences and stigmatization aversion; it becomes possible to study respondents' beliefs about the inferences made about them; and, an experiment can be designed to closely match a fully specified game that can be analyzed completely.

Ljungqvist (1993) has treated RRT from the perspective of utility maximizing agents, but has stopped short of a complete game-theoretic analysis, focusing instead on the conditions for truth-telling to be incentive compatible. Following in his footsteps, in this paper, we use RRT as a vehicle for illustrating the potential benefits of randomness on information transmission both theoretically and experimentally. We first formally analyze RRT with a game theoretic model and then conduct an experiment to explore whether the theoretical prediction of the model matches the behavior of subjects in a laboratory setting.

Fully specifying a communication game helps make explicit the above mentioned conditions for RRT to induce truth-telling. Following Ljungqvist (1993), we use a payoff function for the respondent that trades off lying aversion against stigmatization aversion and, in line with psychological game theory, makes the respondent's payoffs directly dependent on the interviewer's beliefs. The game formulation draws attention to the role of prior beliefs and the updating of those beliefs in equilibrium. It raises the issue of how much information can be transmitted without the use of RRT, which we refer to as the deterministic response technique (DRT), and how the informativeness of RRT equilibria compares with that of DRT equilibria. Finally, it

brings into focus the possibility of multiple equilibria, especially of equilibria that are informative without being truthful.[2]

We provide a full characterization of the equilibrium set of the proposed communication game for all relevant configurations of lying aversion and stigmatization aversion parameters. Importantly, the equilibrium analysis reveals that for some nontrivial parameter sets equilibrium responses may provide useful information even without using RRT; when there are truthful equilibria under RRT, they need not be optimal; and, when truthful RRT equilibria are not optimal, DRT strictly dominates RRT inducing truth-telling. Essentially, when truth-telling preferences are relatively strong in comparison to stigmatization aversion, while it is possible to induce truth-telling via RRT, it may be counterproductive to do so.

Our experimental results show that there are some truthful responses by stigmatized types under DRT, even for values of the lying aversion and stigmatization aversion parameters for which theory rules out communication; the frequency of truthful responses by stigmatized types under DRT responds positively to an induced increase in lying aversion; RRT does raise the incidence of truthful responses by stigmatized types relative to DRT; stigmatized types are not approximately truthful under RRT for induced preferences that admit truthful equilibria (despite the fact that lying aversion, because of prior truth-telling preferences brought to the lab, is likely to be stronger than what we induce with monetary incentives); and, the theoretical prediction that the informativeness of RRT in terms of an entropy measure ("mutual information") dominates that of DRT is rejected by the data.

In summary, we formalize randomized response as a game of information transmission and implement the game experimentally, thus linking the theory on randomness in information transmission both with practice in the field and a novel implementation in the lab. As in most findings in the experimental literature on communication games, we observe over-transmission of information (under DRT). As predicted by the theory, the frequency of truthful responses is significantly higher under RRT. Contrary to the prediction of theory, DRT proves to have a higher information-eliciting performance than RRT. Regarding validation of RRT, both our theoretical and experimental results show that a shifted distribution of responses in the direction of increased truth-telling in line with the more-is-better criterion is no guarantee that the responses came from truthful responding. Finally, our results suggest a cautionary note for optimal survey design: while Ljungqvist (1993) is right to ground RRT in utility maximization/game theory, it is important to keep in mind that the game in question has other than only truthful equilib-

[2]Recently and independently, John, Loewenstein, Acquisti and Vosgerau (2013) have relied in part on laboratory experiments to evaluate RRT. Unlike ours, their approach is not guided by a formal model, there is no attempt directly to control lying aversion and stigmatization aversion preferences, and the stigmatizing trait rather than being induced and known to be known to the experimenter, as in our experiment, is observed unbeknownst to subjects. Their setup is closer to field experiments and more naturalistic. A key interesting finding in their paper will also be relevant for the interpretation of our experimental results: RRT may fail to perform as promised if using this method leads respondents who lack the stigmatizing trait to engage in protective behaviors, avoiding responses that would jeopardize the perception of not having the trait.

ria and that sometimes the price of inducing truthfulness – the direct informational loss from randomization – may be too steep.

In the next section we discuss some of the related literature. Section 3 sets up and motivates our model. In Section 4 we fully characterize the equilibrium sets of both the RRT and the DRT versions of our model. The information-eliciting performance of RRT and DRT is compared in Section 5. Section 6 lays out our experimental design and formulates hypotheses based on our characterization of equilibrium sets in the theoretical model. In Section 7 we report our experimental findings. We conclude in section 8 with a review of how the rationale for RRT relates to the theory of information transmission through noisy channels, how this is reflected in the model we put forward, and an assessment of our theoretical and experimental findings.

2 Related Literature

Communication through a noisy channel is an example of mediated communication (e.g., Forges, 1986; Myerson, 1986). Myerson (1991, p.285-288) was the first to emphasize that introducing noise can improve transmission of private information. He considers a simple sender-receiver game and shows that communication is possible when messages are sent via a carrier pigeon that arrives only half the time, while there is no communicative equilibrium in the same game when messages are transmitted without error. Blume at al. (2007) generalize this insight of noise improving communication in the framework of Crawford and Sobel (1982). For Crawford and Sobel's leading example, with quadratic payoff functions and a uniform type distribution, they find that for almost all values of the sender bias there is a level of noise for which the corresponding equilibrium in the noisy game achieves higher *ex-ante* welfare than that of the most efficient equilibrium in the noise-free game.

Goltsman et al. (2009) consider general mediated communication and identify an efficiency bound for payoffs achievable through communication equilibria. It turns out that for all values of the sender bias this bound can be implemented in the noisy game of Blume et al. (2007), through biased mediator (Ivanov, 2010), or through strategy correlation (Blume, 2012). The bound can also be achieved, for a range of sender biases, through multiple rounds of simultaneous unmediated communication (Krishna and Morgan, 2004).

RRT was first proposed by Warner (1965) to improve inference when there is a need to ask sensitive questions that may be perceived as threatening to respondents. Since Warner (1965), improvements have been made to enhance the method's efficiency and reliability, and numerous variations of RRT have been developed and applied (see Lensvelt-Mulders, Hox, van der Heijden and Maas, 2005, for a meta-analysis of validations of RRT).[3] There have also been instances

[3]There are also extensions of Warner's method to deal with non-binary data (like the number of abortions) and to infer distributions of continuous random variables (like income). See for example Warner (1971) and Poole (1974) and the references therein.

in which use of RRT has resulted in counterintuitive results, which have raised questions about the efficacy of the method. John et al. (2013) survey the literature on failures of RRT, provide further evidence and suggest possible remedies.

The conceptual foundations of RRT have been formally investigated by Ljungqvist (1993), who treats respondents as utility maximizing agents trading off lying aversion against stigmatization aversion. From this perspective, finding the proper way of implementing RRT becomes a mechanism design problem: the investigator pursues her objective of minimizing the bias and variance of the estimator of the population proportion carrying the stigmatizing trait subject to respondents' incentive compatibility constraints that ensure that truth-telling preferences outweigh privacy concerns. Randomization or noise plays a similar role in Ljungqvist's model as in the above cited literature on noisy information transmission, the principal difference being that in the information transmission literature noise is a property of the communication channel, whereas in Ljungqvist's model and in implementations of RRT, the noise generating device is only observed by the respondent. A crucial component of making the noise matter in Ljungqvist's model is therefore that respondents comply with the device because of their postulated lying aversion.

3 A Model of Survey Response

The objective of our theoretical analysis is to examine, in a simple setting amenable to experimental implementations, how equilibrium behavior restricts the information transmission outcomes that may emerge under different survey techniques. We consider, building on the environment in Ljungqvist (1993), a simple model with two players and two types. There is an interviewer and a respondent. The respondent privately observes his type $\theta \in \{s, t\}$, where s is designated as the *stigmatized* type and t the *regular* type. The common prior is that the two types are equally likely.[4]

The interviewer elicits the respondent's private type with a question, q, which could either be "Are you an s?" (q_s) or "Are you a t?" (q_t). We compare two response regimes, *deterministic response* and *randomized response*. A general setup that encompasses both regimes has q_s and q_t drawn, respectively, with commonly known probabilities p_s and $1 - p_s$. The outcome of the

[4]We deliberately abstract away from some aspects of actual surveys. A survey, for example, involves more than one respondent and has the objective to estimate the proportion of the population belonging to groups with certain characteristics. In the future, it may be interesting to consider a model in which there is a non-trivial population of responders, the population proportion of the trait of interest is a random variable, both the interviewer and the responders receive private signals about the realization of that random variable, and the interviewer's objective is to correctly estimate the proportion of the population having the trait. For now, focusing on one respondent and assuming a common prior, as we do, helps us highlight the incentives behind individual response behavior and facilitates comparison with the literature on information transmission, noisy channels and mediated communication cited above. The common prior assumption is standard in applications of games with incomplete information; having the prior be uniform is a choice of convenience.

draw (i.e., which question the respondent responds to) is known to the respondent but not to the interviewer. The respondent responds to a question with $r \in \{y, n\}$, where the exogenous semantics of y is "yes" and that of n "no." The deterministic response regime corresponds to the degenerate case in which $p_s \in \{0, 1\}$; in the randomized response regime, $p_s \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$.[5]

We assume that the respondent's incentive is shaped by two considerations: *stigmatization aversion* and *lying aversion*. Consider a survey on tax evasion. The first piece of information that the interviewer wants to obtain may pertain to whether the respondent has or has not evaded tax. Given that tax evasion is typically not socially approved, a respondent who has evaded tax may be reluctant to reveal it, for fear that it will entail stigmatization. On the other hand, given that the respondent has self-selected into participating in the survey, it is reasonable to expect that he is otherwise willing to cooperate with the interviewer by honestly answering questions. More generally, people may be driven by an intrinsic aversion to lying to tell the truth, which is documented in several experimental studies of communication games (e.g., Sánchez-Pagés and Vorsatz, 2007; Gneezy, 2010).

Formally, we model stigmatization aversion as a belief-dependent preference. The respondent's payoff is a decreasing function of the interviewer's belief, μ_s, that he is of the stigmatized type s. The designation of s and t as stigmatized type and regular type is thus chosen with respect to the respondent's payoff. For lying aversion, we consider the triple $(\theta, q, r) \in \{s, t\} \times \{q_t, q_s\} \times \{y, n\}$ and define a "truthful set" $\mathscr{H} = \{(s, q_s, y), (t, q_t, y), (s, q_t, n), (t, q_s, n)\}$. The respondent obtains a payoff gain if, for example, the event (s, q_s, y) occurs in which his type is s and he responds to "Are you an s?" with "yes." We assume that the respondent's payoff function takes the following form:

$$U((\theta, q, r), \mu_s) = \lambda \mathbb{I}_{\mathscr{H}}(\theta, q, r) - \xi \mu_s,$$

where $\lambda, \xi \geq 0$ are parameters measuring, respectively, the degrees of lying aversion and stigmatization aversion, and $\mathbb{I}_{\mathscr{H}}(\theta, q, r)$ is an indicator function that takes the value of 1 if $(\theta, q, r) \in \mathscr{H}$ and 0 otherwise.[6] A higher value of λ means that the respondent is more lying averse, i.e., he has a stronger preference for truth-telling. Similarly, a more stigmatization averse respondent will have a higher ξ. Note that both s and t have the same degrees of aversions. For stigmatization, this means that t dislikes being identified as s as much as s does. Furthermore, the two payoff components do not interact with each other so that, for example, the respondent's payoff gain from honesty is independent of how certain the interviewer is about his type.

To make sure that the information transmission problem is not trivial, we further restrict the aversion parameters to satisfy $0 \leq \lambda < \xi$, or $\frac{\lambda}{\xi} \in [0, 1)$, so that stigmatization aversion strictly

[5] We rule out $p_s = \frac{1}{2}$ from consideration because it corresponds to the uninteresting case where the interviewer obtains no information no matter how the respondent responds.

[6] When $\lambda > 0$, "talk is not cheap" in our model. For work that introduces exogenous preference for honesty into cheap-talk models, see, e.g., Chen (2011), Kartik, Ottaviani and Squintani (2007), and Kartik (2009).

dominates lying aversion. The *lying-stigmatization aversion ratio*, $\frac{\lambda}{\xi}$, will serve an important role in our equilibrium characterizations. As will become apparent below, if $\frac{\lambda}{\xi} \geq 1$, full information transmission will be feasible even in the deterministic response regime, defeating the very purpose of using RRT.

We assume that the only "action" the interviewer performs in the model is to update her beliefs, which corresponds to the situations where survey conductors obtain estimates of the population parameters of interests. A belief function of the interviewer is $\mu_s : \{y, n\} \to [0, 1]$, which specifies for each received response a probability that $\theta = s$. The function will be determined as part of an equilibrium.[7]

4 Equilibrium Characterization

The solution concept is perfect Bayesian equilibrium (henceforth equilibrium), i.e., strategies are optimal given beliefs and beliefs are derived from Bayes' rule whenever possible. When a type responds truthfully according to the exogenous semantics of y and n (e.g., s responds to q_t with n), he is said to give a *truthful response*. When a truthful response involves y (n), it is called an *affirmative* (*negative*) truthful response. An equilibrium is said to be *informative* if both y and n are used with positive probability in equilibrium and $\mu_s(y) \neq \mu_s(n)$; if both types of the respondent give truthful responses with probability one, the equilibrium is *truthful*.

We begin by establishing a property of the interviewer's posterior beliefs in equilibrium that will play a crucial role in the analysis of the equilibria and the information-eliciting performance of the response regimes:

Lemma 1. *On the equilibrium path of any equilibrium of the survey response model, the interviewer's belief differential after the two different responses satisfies* $|\mu_s(y) - \mu_s(n)| \leq \frac{\lambda}{\xi}$.

This is a simple consequence of the fact that in equilibrium the "benefit", λ, from a truthful response must outweigh the "cost" $\xi|\mu_s(y) - \mu_s(n)|$.

[7] Alternatively, one could make the respondent's payoff a function of his belief about the belief of the interviewer, e.g. the expected belief of the interviewer. Our modeling choice is guided by simplicity, conformity with our experimental design, and the observation that in our setting in equilibrium the distinction between the interviewer's belief and the respondent's expectation of that belief disappears. For a discussion of possible modeling choices in psychological games (Geanakoplos, Pearce and Stacchetti, 1989), see Battigalli and Dufwenberg (2009). Ottaviani and Sørensen (2006) consider a cheap-talk game in which a sender cares about his reputation, modeled as the discrepancy between the receiver's belief about the state and the actual state. In our model, the respondent's (sender) payoff depends on how likely the interviewer believes the state to be s. See also Bernheim (1994) for a model of conformity in which agents' esteem, derived from the opinion of others as in our case, is modeled via belief-dependent preferences.

4.1 Deterministic Response

In the deterministic response regime, we have a degenerate $p_s \in \{0, 1\}$, and thus it is common knowledge which question is asked. A behavior strategy of the respondent, $\sigma : \{s, t\} \to \Delta\{y, n\}$, specifies for each θ the distribution of responses to the commonly known question, q_s or q_t.

For the sake of concreteness and without loss of generality, in characterizing the equilibria of the deterministic response regime, we consider that "Are you a t?" is the question asked:

Proposition 1. *In the deterministic response regime in which $q = q_t$ ($p_s = 0$),*

1. *for every lying-stigmatization aversion ratio $\frac{\lambda}{\xi} \in (\frac{1}{2}, 1)$, there exists a unique equilibrium; this equilibrium is informative, with t always giving an affirmative truthful response y and s randomizing between y and n;*

2. *for $\frac{\lambda}{\xi} \in (0, \frac{1}{2}]$, there are exactly two equilibrium outcomes; in one both types respond with y and in the other both respond with n; and, only the outcome where both types respond with y survives the D1 criterion.*

An obvious corollary of the proposition is that there is no truthful equilibrium in the deterministic response regime in the relevant range of the lying-stigmatization aversion ratio.[8]

Whenever $\mu_s(y) \neq \mu_s(n)$, the type whose truthful response results in the lower probability value strictly prefers to respond truthfully, for doing so brings the dual benefits of being less stigmatized as well as avoiding lying. For an informative equilibrium with $q = q_t$, we must have $\mu_s(n) > \mu_s(y)$ on the equilibrium path so that it is type t who strictly prefers to respond truthfully with y and thus always does so.[9] We cannot, however, also have type s always truthfully respond, in this case with n, because then $\mu_s(n) = 1$ and $\mu_s(y) = 0$, violating the restriction put on the belief differential (Lemma 1). Intuitively, whenever s is willing to respond with n under belief profile $\mu_s(n) > \mu_s(y)$, he is trading off the benefit from avoiding lying against the benefit of being less stigmatized. Given that $\lambda < \xi$, the former can never dominate the latter for a belief differential of one. Accordingly, the informative equilibrium can never be truthful, which leaves the only possibility that s randomizes between y and n.

If type s randomizes, since he now sometimes responds with y the value of $\mu_s(y)$ increases, narrowing the belief differential ($\mu_s(n)$ remains at one since it is still exclusively used by s). In equilibrium, the randomization requires that the benefit of avoiding lying be equal to the benefit of being less stigmatized, which happens precisely when the restriction on belief differential binds

[8] When $\frac{\lambda}{\xi} = 0$, the game becomes one of cheap talk, and there is also a babbling equilibrium in which s and t completely randomize between y and n with the same probabilities. The proof of Proposition 1 and that of the upcoming Proposition 2 (Appendix A) contain complete characterizations of the sets of equilibria in both response regimes.

[9] With $\mu_s(y) > \mu_s(n)$, s strictly prefers to respond with n, which creates a contradiction that $\mu_s(y) = 0$.

at $\mu_s(n) - \mu_s(y) = \frac{\lambda}{\xi}$, in which s responds with n with probability $2 - \frac{\xi}{\lambda}$. The fact that the probability has to be strictly positive in turn yields the restriction that $\frac{\lambda}{\xi} > \frac{1}{2}$.[10]

For the uninformative equilibria, in which both types respond identically, the belief after the unused response has to put sufficiently high probability on s to justify the response of the lying type. In particular, the probability has to be at least $\frac{\lambda}{\xi}$ higher than $\frac{1}{2}$, the interviewer's posterior after the uninformative response. Given the upper bound of one for probabilities, this yields the restriction that $\frac{\lambda}{\xi} \leq \frac{1}{2}$. A common response of n to question q_t means that type t but not type s is lying. If s is ever willing to deviate to respond with y despite forgoing the benefit of avoiding lying, t will have a stronger incentive to deviate due to the benefit of avoiding lying. Consequently, the equilibrium outcome with common response n does not survive the D1 Criterion (Banks and Sobel, 1987; Cho and Kreps, 1987), which prescribes an out-of-equilibrium belief $\mu_s(y) = 0 < \frac{1}{2}$. When the common response is y, the role of t and s in the above argument is reversed; the D1 Criterion prescribes $\mu_s(n) = 1$, high enough to support y as the equilibrium common response.

Proposition 1 gives a formal expression to the predicament that calls for the use of RRT: if the respondent is sufficiently stigmatization averse (relative to the degree of lying aversion), no information can be transmitted when the survey question is deterministic.

4.2 Randomized Response

In the randomized response regime, when p_s is non-degenerate, the question q that the respondent answers becomes part of his private information, which now consists of two components: (θ, q). Accordingly, a behavior strategy of the respondent is $\sigma : \{s, t\} \times \{q_s, q_t\} \to \Delta\{y, n\}$.

The following proposition confirms the intuition for the use of RRT: the possibility of information transmission is enhanced and for appropriate choices of the probability p_s of asking the sensitive question, q_s, truthful responding is an equilibrium.

Proposition 2. *In the randomized response regime,*

1. *there exists a truthful equilibrium if and only if*

$$p_s \in \left[\frac{\xi - \lambda}{2\xi}, 1 - \frac{\xi - \lambda}{2\xi}\right];$$

2. *there exists a non-truthful informative equilibrium if and only if*

$$p_s \in \left(\left(0, \frac{\xi - \lambda}{\lambda}\right] \cup \left[1 - \frac{\xi - \lambda}{\lambda}, 1\right)\right) \bigcap \left(\left(0, \frac{1}{2}\right) \cup \left(\frac{1}{2}, 1\right)\right); \textit{ and,}$$

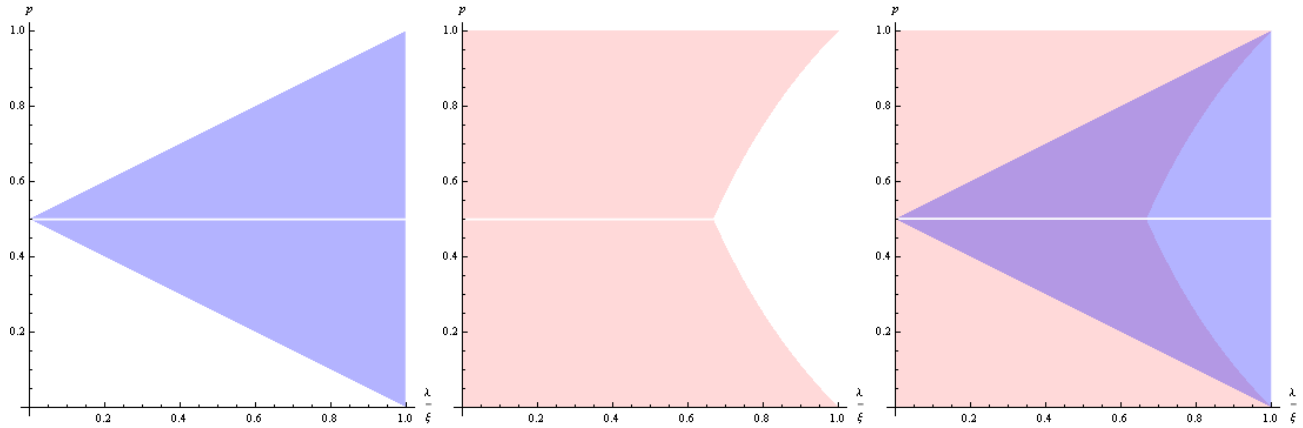
[10] The case for $q = q_s$ is symmetric, where in the informative equilibrium t always responds with n and s responds with y with probability $2 - \frac{\xi}{\lambda}$, again yielding the restriction that $\frac{\lambda}{\xi} > \frac{1}{2}$.

3. there exist uninformative equilibria for all $p_s \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$ *if and only if* $\frac{\lambda}{\xi} \in [0, \frac{1}{2}]$.

Since $\lambda < \xi$, it follows that the union of the sets in the first and second parts of Proposition 2 is all of $(0, 1)$, and therefore an immediate implication of this result is:

Corollary 1. *In the randomized response regime with* $\frac{\lambda}{\xi} \in (0, 1)$*, there exists an informative equilibrium for every* $p_s \in \left(0, \frac{1}{2}\right) \cup \left(\frac{1}{2}, 1\right)$.

Figure 1 depicts the regions of the parameter space in which informative equilibria exist in the randomized response regime. An example of non-truthful informative equilibria has s and t always respond truthfully to a question, different for each of them, and completely randomize between y and n to the respective other questions.



(a) Truthful Equilibria (b) Non-Truthful Informative Equilibria (c) All Informative Equilibria

Figure 1: Existence of Informative Equilibria for $(p_s, \frac{\lambda}{\xi}) \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1) \times (0, 1)$

Similar to the deterministic response regime, lying aversion ($\lambda > 0$) is a necessary condition for the existence of informative equilibria. In addition to the change that truthful responding can be sustained in equilibrium, another significant difference of randomized response is that information can now be transmitted for all $\frac{\lambda}{\xi} \in (0, 1)$. The interviewer's uncertainty about which question is asked alleviates the negative impact of stigmatization aversion, making information transmission possible even when ξ is arbitrarily large. However, the benefit does not come without cost, because the very same uncertainty decreases the amount of information transmitted.

Despite supporting the intuition that RRT sometimes enables truthful responding, the fully game-theoretic analysis that is summarized in Proposition 2 also hints at a possible problem with comparative validations of the randomized response method that has thus far eluded the literature: satisfaction of the more-is-better criterion could simply be an indicator of a non-truthful informative equilibrium if $p_s \notin \left[\frac{\xi-\lambda}{2\xi}, 1 - \frac{\xi-\lambda}{2\xi}\right]$. A researcher who does not know the

value of $\frac{\lambda}{\xi}$ and who relies on the more-is-better criterion to validate the procedure might therefore mistakenly assume that respondents tell the truth, when in fact there is some degree of randomization.

A further potential issue for using the procedure in practice is that even if $p_s \in \left[\frac{\xi-\lambda}{2\xi}, 1-\frac{\xi-\lambda}{2\xi}\right]$, truthful equilibria may coexist with other informative equilibria. According to Proposition 2, this will be the case whenever $p_s \in [\frac{\xi-\lambda}{2\xi}, \frac{\xi-\lambda}{\lambda}] \cup [1-\frac{\xi-\lambda}{\lambda}, 1-\frac{\xi-\lambda}{2\xi}]$. Thus, even for values of p_s for which truthful responding is an equilibrium, having the more-is-better criterion satisfied is not necessarily an indicator of respondents acting according to a truthful equilibrium.

5 Information-Eliciting Performance

In this section, we evaluate the information-eliciting performance of the two different response regimes. Using the criterion of mutual information from information theory (Shannon, 1948), we measure the maximal transmittable information from the respondent to the interviewer that is consistent with equilibrium behavior.

We begin with a brief discussion of the nature and properties of mutual information in the context of our environment. Suppose $\Pr(\theta') > 0$ is the prior of the respondent's type θ' and $\Pr(\theta'|r')$ is the posterior upon observation of response r'. When r' is observed at θ', there is an informational gain if $\Pr(\theta'|r') > \Pr(\theta')$ or $\frac{\Pr(\theta'|r')}{\Pr(\theta')} > 1$. Similarly, an informational loss occurs at θ' if $\frac{\Pr(\theta'|r')}{\Pr(\theta')} < 1$. One can assign numerical values $v\left(\frac{P(\theta'|r')}{P(\theta')}\right)$ to the informational gains and losses by introducing a function $v : \mathbb{R} \to \mathbb{R}$ that is strictly monotonic, continuous and satisfies $v(1) = 0$. One such function is the logarithm. Using $\log(\cdot)$ for $v(\cdot)$, the expected net informational gain about the random variable θ due to the observation of the random variable r is thus

$$I(\theta; r) = \sum_{(\theta', r') \in \{s,t\} \times \{y,n\}} P(\theta', r') \log \frac{P(\theta'|r')}{P(\theta')},$$

which is precisely the definition of mutual information, where by continuity the convention of $0 \log 0$ is adopted. Note that the above expression can be rewritten as $I(\theta; r) = H(\theta) - H(\theta|r)$, where $H(\theta) = -\sum_{\theta' \in \{s,t\}} \Pr(\theta') \log \Pr(\theta')$ is the entropy of the respondent's type and $H(\theta|r) = -\sum_{r' \in \{y,n\}} \Pr(r') \sum_{\theta' \in \{s,t\}} \Pr(\theta'|r') \log \Pr(\theta'|r')$ is the conditional entropy of the respondent's type given r. Entropy is a measure of the uncertainty of a random variable. Mutual information therefore measures, quite intuitively, the reduction in the uncertainty of θ due to the observation of r.[11] It serves as our criterion to evaluate the information-elicitation performance of the

[11] Mutual information is also referred to as relative entropy or Kullback-Leibler divergence, one between the joint and product distributions of the random variables in question. If the base of the logarithm is 2, which is commonly adopted in information theory, then the unit of the entropy is in bits; if the base is e, the unit is in nats. Given that our model has a binary type space, we use 2 as our base. For an excellent reference in information theory, see Cover and Thomas (1991).

two response regimes: the higher the mutual information, i.e., the higher the reduction in the uncertainty, the higher the performance, with a maximum value of 1 and a minimum of 0.[12]

Note that $\Pr(\theta'|r')$ is nothing but the interviewer's posterior beliefs, which, together with $\Pr(r')$, are determined by the respondent's strategy and, in the case of randomized response, the probabilities of the questions. We evaluate the mutual information implied by the respondent's equilibrium strategies. In light of the multiple equilibria, we focus on the question: for a given lying-stigmatization aversion ratio $\frac{\lambda}{\xi}$, what is the mutual information of the respective most informative equilibria in the two responses regimes, with "informativeness" evaluated with respect to mutual information? We denote such maximal mutual information by $\bar{I}_D(\frac{\lambda}{\xi})$ for the deterministic response regime and $\bar{I}_R(\frac{\lambda}{\xi})$ for the randomized response regime.

Recall that in the deterministic response regime, the uninformative and the informative equilibria exist under complementary ranges of $\frac{\lambda}{\xi} \in [0,1)$ divided by $\frac{1}{2}$. Accordingly, we have the following evaluation:

Proposition 3. *In the deterministic response regime, the maximal mutual information allowed by any equilibrium is*

$$\bar{I}_D(\tfrac{\lambda}{\xi}) = \begin{cases} 0, & \text{if } \frac{\lambda}{\xi} \in (0, \frac{1}{2}], \\ 1 + \frac{1}{2}[(\frac{\xi}{\lambda} - 1)\log(1 - \frac{\lambda}{\xi}) + \log\frac{\lambda}{\xi}], & \text{if } \frac{\lambda}{\xi} \in (\frac{1}{2}, 1). \end{cases}$$

Given our specification that s and t are equally likely, the entropy of θ is 1, which is the maximum entropy possible. The uncertainty that remains for the interviewer in the informative equilibrium is therefore $-\frac{1}{2}[(\frac{\xi}{\lambda} - 1)\log(1 - \frac{\lambda}{\xi}) + \log\frac{\lambda}{\xi}] \in (0,1)$.

With the continuum of informative equilibria, the determination of the maximal performance is less straightforward for the randomized response regime. To facilitate the exposition, we start with the following lemma:

Lemma 2. *In the randomized response regime,*

1. *for $\frac{\lambda}{\xi} \in (\frac{1}{2}, 1)$ and probability of q_s set at $p_s = \frac{\xi - \lambda}{\lambda}$ or $p_s = 1 - \frac{\xi - \lambda}{\lambda}$, there exist equilibria whose mutual information coincides with $\bar{I}_D(\frac{\lambda}{\xi})$; and,*

[12] Given that in our model no payoff function is specified for the interviewer, there is no obvious candidate for defining a value of information that would be less arbitrary than using mutual information. Also, pursuing the goal of maximizing the precision of the estimator of the population frequency of stigmatization subject to a truth-telling constraint, as in Ljungqvist (1993), is compromised by the presence of multiple equilibria. This, and the fact that mutual information is widely used in information theory, motivate us to adopt it as our measure of informational gain. Jose, Nau and Winkler (2008) investigate how entropy measures of information relate to utility. Kelly (1956) links information-theoretic measures with the value of information in the case of a gambler who receives information through a noisy channel. Donaldson-Matasci, Bergstrom and Lachmann (2010) identify uncertain environments in which the biological fitness value of information corresponds exactly to mutual information and show more generally that mutual information is an upper bound on the fitness value of information. Information-theoretic measures of information have recently been used in macroeconomics to study the consequences of information processing constraints (Sims, 2003), and in organization theory to capture the idea that organizations have limited communication capacity (Dessein, Galeotti and Santos, 2013).

2. *for* $\frac{\lambda}{\xi} \in (0,1)$*, the maximal mutual information among the truthful equilibria is* $\bar{I}_{R-T}(\frac{\lambda}{\xi}) = \frac{1}{2}[(1-\frac{\lambda}{\xi})\log(1-\frac{\lambda}{\xi}) + (1+\frac{\lambda}{\xi})\log(1+\frac{\lambda}{\xi})]$*, achieved at* $p_s = \frac{\xi-\lambda}{2\xi}$ *or* $p_s = 1 - \frac{\xi-\lambda}{2\xi}$.

Furthermore, there exists a $c \approx 0.743$ *such that* $\bar{I}_{R-T}(\frac{\lambda}{\xi}) > \bar{I}_D(\frac{\lambda}{\xi})$ *for* $\frac{\lambda}{\xi} \in (0,c)$ *and* $\bar{I}_{R-T}(\frac{\lambda}{\xi}) \leq \bar{I}_D(\frac{\lambda}{\xi})$ *for* $\frac{\lambda}{\xi} \in [c,1)$ *with strict inequality except at* $\frac{\lambda}{\xi} = c$.

In the deterministic response regime, the mutual information is determined by the respondent's strategy, which, in the informative equilibrium with $q = q_t$, consists of truthful response by type t, $\sigma(y|t) = 1$, and randomization by type s, $\sigma(n|s) = 2 - \frac{\xi}{\lambda}$. In the randomized response regime, the probabilities of the questions also contribute to determining the mutual information. This suggests the possibility that the non-degenerate question probabilities may serve as an exogenous randomization to mimic the equilibrium randomization in the deterministic response regime, resulting in the same set of response probabilities and posteriors that enter into the computation of mutual information. The first part of Lemma 2 says that this is indeed the case. The analysis boils to finding p_s, $\sigma(y|t,q_s)$, $\sigma(y|t,q_s)$, $\sigma(n|s,q_s)$, and $\sigma(n|s,q_t)$ in the randomized response regime so that $p_s\sigma(y|t,q_s) + (1-p)\sigma(y|t,q_t) = 1$ and $p_s\sigma(n|s,q_s) + (1-p)\sigma(n|s,q_t) = 2 - \frac{\xi}{\lambda}$. These conditions are satisfied by $p_s = \frac{\xi-\lambda}{\lambda}$ coupled with the strategy $\sigma(y|t,q_s) = \sigma(y|t,q_t) = \sigma(n|s,q_t) = 1$ and $\sigma(n|s,q_s) = 0$ which form an equilibrium in the randomized response regime if and only if p_s is at that exact value. The two equilibria in the two different response regimes result in the same posteriors; this is no coincidence because the incentive conditions behind one equilibrium carry over to the other.

The intuition behind the second part of Lemma 2 can be easily seen from the fact that when $p_s = \frac{1}{2}$, the uninteresting case which we ruled out by definition, no information is transmitted regardless of how the respondent responds; the interviewer's posteriors will remain at $\frac{1}{2}$. More information is transmitted, and thus the mutual information is higher, when p_s moves away from $\frac{1}{2}$. Given the constraint that the truthful equilibria can be supported only for $p_s \in [\frac{\xi-\lambda}{2\xi}, 1-\frac{\xi-\lambda}{2\xi}]$, the maximal mutual information of this class of equilibria is achieved when p_s is at the boundaries of the interval.

We proceed to characterize the maximal mutual information under the randomized response regime, covering all equilibria:

Proposition 4. *In the randomized response regime, the maximal mutual information allowed by any equilibrium is*

$$\bar{I}_R(\tfrac{\lambda}{\xi}) = \begin{cases} \frac{1}{2}[(1-\frac{\lambda}{\xi})\log(1-\frac{\lambda}{\xi}) + (1+\frac{\lambda}{\xi})\log(1+\frac{\lambda}{\xi})], & \text{if } \frac{\lambda}{\xi} \in (0,c), \\ 1 + \frac{1}{2}[(\frac{\xi}{\lambda}-1)\log(1-\frac{\lambda}{\xi}) + \log\frac{\lambda}{\xi}], & \text{if } \frac{\lambda}{\xi} \in [c,1), \end{cases}$$

where $c \approx 0.743$.

The essence behind Proposition 4 is that the two values of mutual information in Lemma 2 form an upper envelope of the mutual information of all equilibria in the randomized response

regime. The following corollary, which compares the maximal information-eliciting performance of the two response regimes, is immediate:

Corollary 2. *For given* $\frac{\lambda}{\xi} \in (0,1)$*, the maximal mutual information under the randomized response regime weakly dominates that under the deterministic response regime, with strict dominance for* $\frac{\lambda}{\xi} \in (0,c)$*, where* $c \approx 0.743$*.*

Figure 2 provides a visualization of the comparison.

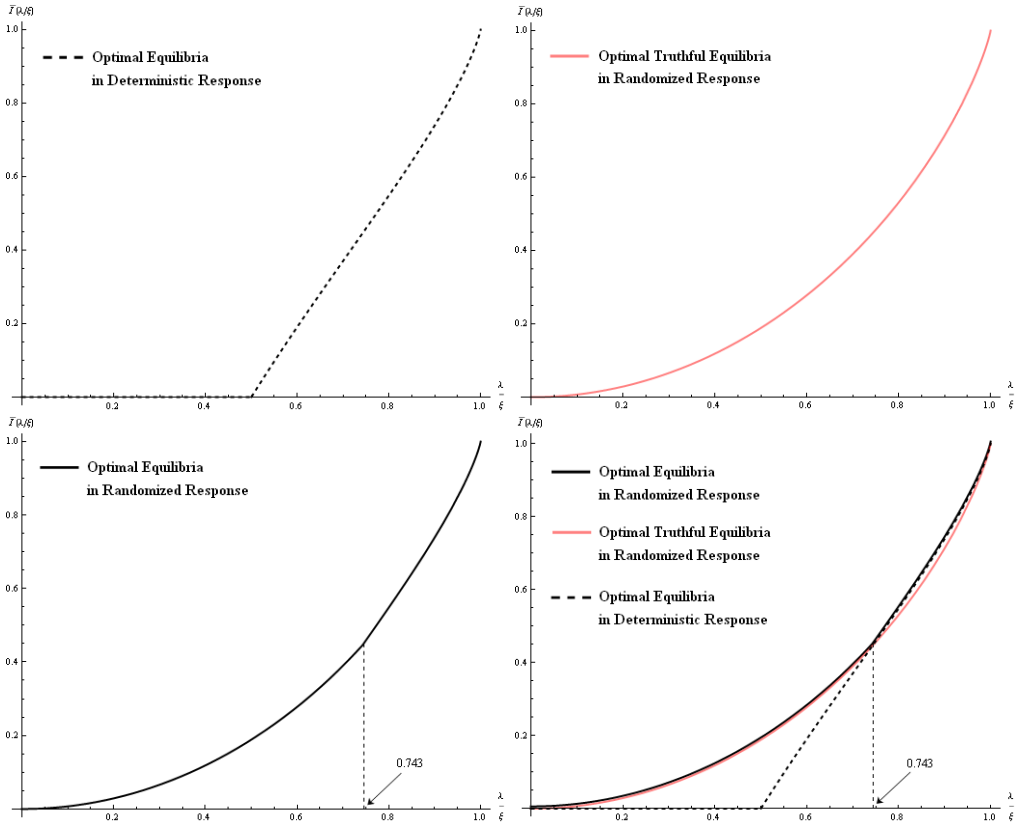


Figure 2: Maximal Mutual Information

6 Experimental Implementation

We experimentally implement the two response regimes, using monetary incentives to induce laboratory environments that are faithful to the theoretical model. We begin by describing our experimental treatments and hypotheses in Section 6.1, discussing the rationales behind the adoption of the model parameters for the treatments. We then describe in Section 6.2 the laboratory environments with which these treatments were conducted.

6.1 Treatments and Hypotheses

Propositions 1–4 serve as the guide for our treatment design, with the lying-stigmatization aversion ratio ($\frac{\lambda}{\xi}$) and the probability of "Are you an s?" (p_s) being the treatment variables. The objective of our experimental investigation is to explore if behavior in the laboratory is consistent with the different behavior predicted under different response regimes. To this end, we select parameter values that lead to pronounced behavior differences across the two response regimes according to the theory.

Table 1: Experimental Treatments

	$p_s = \text{Prob}(q_s) = 0$	$p_s = \text{Prob}(q_s) = 0.4$
$\frac{\lambda}{\xi} = \frac{1}{4}$	*DeterLow*: Deterministic Response/ Low Relative Stigmatization Aversion (Equilibrium Prediction: No Informative Equilibrium)	*RandomLow*: Randomized Response/ Low Relative Stigmatization Aversion (Equilibrium Prediction: Truthful Equilibrium Exists)
$\frac{\lambda}{\xi} = \frac{1}{8}$	*DeterHigh*: Deterministic Response/ High Relative Stigmatization Aversion (Equilibrium Prediction: No Informative Equilibrium)	*RandomHigh*: Randomized Response/ High Relative Stigmatization Aversion (Equilibrium Prediction: Informative But Not Truthful Equilibria Exist)

Table 1 presents our treatments; it describes a 2×2 design, where the rows of the matrix correspond to the values of $\frac{\lambda}{\xi} \in \{\frac{1}{4}, \frac{1}{8}\}$ and the columns to the values of $p_s \in \{0, 0.4\}$. The value $\frac{\lambda}{\xi} = \frac{1}{4}$ represents low relative stigmatization aversion and $\frac{\lambda}{\xi} = \frac{1}{8}$ high relative stigmatization aversion. The value $p_s = 0$ corresponds to an instance of the deterministic response regime and $p_s = 0.4$ to an instance of the randomized response regime.

With these parameter values, there will be no information transmission in the deterministic response treatments, and in the randomized response treatments truthfulness of the respondent is possible only with $\frac{\lambda}{\xi} = \frac{1}{4}$, i.e., when the relative stigmatization aversion is low.

In the deterministic response treatments, *DeterLow* and *DeterHigh*, "Are you a t?" is always asked ($p_s = 0$). The consideration of pronounced behavior differences leads to our choices of $\frac{1}{4}$ and $\frac{1}{8}$ for $\frac{\lambda}{\xi}$; for these values we expect treatment effects when comparing the two randomized response treatments to each other and when comparing randomized response to deterministic response. In the randomized response treatments, *RandomLow* and *RandomHigh*, "Are you an s?" is asked 40% of the time ($p_s = 0.4$). Being truthful arguably represents the most distinct behavior in the randomized response regime, and this becomes one of our criteria in choosing the value of p_s. According to Proposition 2, in *RandomLow* with $\frac{\lambda}{\xi} = \frac{1}{4}$ there exists a truthful equilibrium when $p_s \in [0.375, 0.625]$. Furthermore, the performance of the truthful equilibrium

is at the maximal when p_s is at the boundaries of the range. For convenience in implementations while maintaining as much difference (in terms of performance) as possible from the deterministic response treatments, we round up 0.375 and use $p_s = 0.4$.[13] Note that non-truthful informative equilibria also exist under the chosen parameters. In *RandomHigh* with $\frac{\lambda}{\xi} = \frac{1}{8}$, the existence of truthful equilibrium requires a different set of values of p_s that does not include 0.4. However, in order to facilitate clean comparison between the two randomized response treatments with change in only one treatment variable, we keep $p_s = 0.4$ for *RandomHigh*. Theoretically, with $\frac{\lambda}{\xi} = \frac{1}{8}$ and $p_s = 0.4$, there are non-truthful informative equilibria.

Our theoretical results also inspire our experimental hypotheses. Given the multiplicity of equilibria, we formulate hypotheses only when definite qualitative comparisons can be backed by the predictions of equilibrium and the D1 criterion. We begin with the behavior of the stigmatized type in different response regimes, comparing across the columns of the treatment matrix. The unique D1 pooling equilibria predict that in the deterministic response treatments type s always lies. On the other hand, in the randomized response treatments equilibrium predicts that type s either always tells the truth or does so with positive probability. This gives us our first hypothesis:

Hypothesis 1. *Stigmatized types provide truthful responses significantly more often in the randomized response treatments than in the deterministic response treatments.*

While the D1 equilibria also predict that in the deterministic response treatments type t always tell the truth, no definite qualitative comparison can be made because in the randomized response treatments the truthful/informative equilibria predict that type t either always tells the truth or does so with positive probability less than one. We thus consider our experiments exploratory in this regard.[14]

Our second hypothesis focuses on the deterministic response treatments, covering explicitly the prediction of the D1 criterion and the effect of different levels of relative stigmatization aversion. The latter pertains to comparison across the rows of the treatment matrix.[15] The D1 pooling equilibria predict the same respondent's behavior in *DeterLow* and *DeterHigh*, where both s and t always respond with "yes" to "Are you a t?" This suggests the following hypothesis:

Hypothesis 2. *1) In each of DeterLow and DeterHigh, both stigmatized types and regular types respond with "yes" significantly more often than with "no," and there is no significant difference*

[13]The decision to round up the lower boundary value instead of rounding down 0.625 is motivated by consistency with the deterministic response treatments in which, between $p_s = 0$ and $p_s = 1$, the lower value is used.

[14]Note that the randomization of the non-stigmatized type t in informative but not truthful equilibria of the randomized response regime is consistent with the observation in John et al. (2013) that reliance on RRT may prompt non-stigmatized types to engage in protective behaviors and thus possibly compromise the inferences made when employing RRT.

[15]The multiple equilibria under the randomized response treatments do not provide definite comparisons. We thus do not hypothesize on the comparisons between *RandomLow* and *RandomHigh*.

in the uses of "yes" between them. 2) The uses of responses by both stigmatized types and regular types do not differ significantly between DeterLow and DeterHigh.

The first part of the hypothesis revolves around the predicted uses of responses under the D1 criterion, which is implicitly used in Hypothesis 1. A hypothesis translated directly from the point predictions of the D1 pooling equilibria will almost surely be refuted, especially since the predicted points are the upper and lower bounds of all possible frequencies. We thus use a weaker hypothesis that involves qualitative comparisons consistent with the equilibrium predictions. Note that the information property of the pooling equilibria—that no information is transmitted—is captured in the first part of Hypothesis 2. Also informed by the D1 pooling equilibria, the second part posits that varying between the two levels of relative stigmatization aversion has no significant impact on the behavior of both types.

We turn next to the interviewer's beliefs. As will be discuss below, we elicit the interviewer-subjects' beliefs about their opponents' types. While theory predicts that there are differences between the interviewer's beliefs in different treatments, the differences after different responses are of more empirical relevance because they represent how much information is provided by the responses as perceived by the subjects. Furthermore, these belief differences allow us to evaluate the incentives faced by the respondent-subjects. We thus formulate our third hypothesis by comparing, within each treatment, beliefs after "yes" and "no":

Hypothesis 3. *1) In the deterministic response treatments, the interviewers' elicited beliefs assign significantly higher probability to s after "no" than after "yes," and the beliefs after "yes" are not significantly different from the prior* 0.5*. 2) In the randomized response treatments, the interviewers' elicited beliefs after "yes" and after "no" are significantly different.*

In the deterministic response treatments, the D1 pooling equilibria predict that the belief after "yes" is that s and t are equally likely, while the out-of-equilibrium belief after "no" has to be sufficiently higher than 0.5 to support the equilibrium. With the anticipation that both responses will be observed, this serves as the basis of the first part of the hypothesis. In the randomized response treatments, a higher belief assigned to s after "yes" or after "no" are both consistent with equilibrium. We thus do not hypothesize beyond the fact that the beliefs are different. And which belief profile will prevail in the laboratory—a question that will be relevant to equilibrium selection—is an empirical issue that we explore with the experiments.

Our last hypothesis pertains to the information-eliciting performance of the response regimes:

Hypothesis 4. *The mutual information implied by observed behavior is significantly higher in the randomized response treatments than in the deterministic response treatments.*

The mutual information implied by our experimental data will be used to evaluate the hypothesis. Note that even though theoretically the performances of the randomized response

treatments are positive but not maximal under our parameter choices, the hypothesis is still backed by our theoretical results because with the pooling equilibria the predicted conditional relative entropies are zero in the deterministic response treatments.

6.2 Design and Procedures

Our experiment was conducted at the Pittsburgh Experimental Economics Lab. A total of 304 subjects with no prior experience in the experiments were recruited from the undergraduate/graduate population of the University of Pittsburgh to participate in 16 experimental sessions, four per each treatment. A *between-subject* design was used, and each session involved $16-20$ distinct subjects making decisions in $8-10$ groups.[16] The experiment was programmed and conducted using z-Tree (Fischbacher, 2007).

In each session, half the subjects was randomly assigned the role of Member A (respondent) and the other half Member B (interviewer), with role assignments remaining fixed throughout the session; they participated in 40 rounds of decisions in groups of two formed by using *random matching*.[17] In each group and each round, the computer randomly drew either SQUARE (s) or TRIANGLE (t). Both members were informed about the fact that each shape would have an equal chance to be drawn, but the selected shape would be revealed only to Member A. In the deterministic response treatments, Member A was presented with the question "Was TRIANGLE selected?" (q_t), which was known to Member B. In the randomized response treatments, the computer would draw a question from either "Was SQUARE selected?" (q_s) or "Was TRIANGLE selected?" Both members were informed about the fact that the former question would have a 40% chance to be drawn, but the selected question would be revealed only to Member A. In both sets of treatments, Member A answered to the question asked, either with "yes" or "no." The response was revealed to Member B, who was then asked to predict the likelihood that SQUARE or TRIANGLE was drawn. Member B was asked to allocate 100 shapes between SQUARE and TRIANGLE, where the number of SQUARES would represent the predicted likelihood that SQUARE was selected.

We used monetary incentives to induce lying and stigmatization aversions. Subjects were rewarded in each round in experimental currency unit (ECU).[18] If Member A's response to

[16] We targeted at recruiting 20 subjects (10 groups) for a session and set a minimum of 16 in case of insufficient show-ups. We met our target for 10 sessions, with the remaining six sessions four conducted with 18 subjects and two conducted with 16 subjects.

[17] Before the 40 official rounds, subjects participated in 6 rounds of practice, in which they assumed the role of Member A for three rounds and Member B for another three rounds. The objective of subjects assuming both roles in the practice rounds was to familiarize them with the computer interface and the flow of the whole decision process.

[18] We randomly selected three rounds and used the average earning in the selected rounds for real payments at the exchange rate of 10 ECU for 1 USD. As will be discussed below, there was a rather large discrepancy of what a Member B could earn in a round. The use of three round average was intended to smooth out the variations. Payments to subjects ranged from, including a $5 show-up fee, $10 to $35, with an average of $29.7.

the question truthfully reported which shape was selected, he/she would receive 300 ECU in *DeterLow*/*RandomLow* and 275 ECU in *DeterHigh*/*RandomHigh*; otherwise with untruthful responses, he/she would receive 250 ECU. Lying aversion was thus induced as earning either 50 ECU or 25 ECU for truthful.

Stigmatization aversion was induced as follows: in all treatments Member A's ECU earned from giving the response would be reduced by, in ECU, twice the number of SQUARES allocated by Member B; thus, compared to the case when Member B predicted a zero probability of SQUARE, Member A's earning was 200 ECU lower than when Member B predicted a probability of one. Note that in implementing different levels of relative stigmatization aversion, our design varied the absolute level of lying aversion instead of the absolute level of stigmatization aversion: in *DeterLow* and *RandomLow* $\frac{\lambda}{\xi} = \frac{1}{4}$ was implemented as $\frac{50}{200}$ and in *DeterHigh* and *RandomHigh* $\frac{\lambda}{\xi} = \frac{1}{8}$ was implemented as $\frac{25}{200}$.[19]

Member B's rewards revolved around the provision of incentives for truthful reporting of beliefs. We used a belief-elicitation mechanism in which, irrespective of risk attitudes, truthfully reporting one's beliefs is a dominant strategy (Karni, 2009).[20] In the following, we describe the essence of our reward procedure that implements the mechanism; the details of the presentation to subjects can be found in the experimental instructions in Appendix D.

The procedure enlisted the use of two binary lotteries. We used the upper and lower bounds of Member A's earnings, 300 ECU and 50 ECU, as the lotteries' monetary outcomes. After Member B predicted the likelihood of SQUARE/TRIANGLE, he/she would be presented with a lottery that also involved SQUARE and TRIANGLE. The probability of drawing a SQUARE in this lottery has been randomly determined out of 100 uniform possibilities with $\frac{1}{100}$ increments and was revealed to Member B at this point. If the probability of drawing a SQUARE in this lottery turned out to be higher than Member B's predicted likelihood of SQUARE having been selected for Member A, he/she would draw from this lottery, receiving 300 ECU for drawing a SQUARE and 50 ECU for a TRIANGLE. Otherwise, Member B's earning would depend on Member A's shape, which constituted another binary lottery: he/she would earn 300 ECU if it was a SQUARE and 50 ECU if it was a TRIANGLE. Under this reward procedure, making predictions according to true beliefs always guaranteed Member B a draw from one of two lotteries where the (subjective) probability of earning the higher "prize," 300 ECU, was higher,

[19]This design approach was necessitated by maintaining reasonable bounds on earnings which did not differ by too much across treatments. The base earning of 250 ECU ensured, with the induced $\xi = 200$, that subjects received a minimum of 50 ECU in a round; subjects were thus guaranteed, excluding the show-up fee, a positive payment of \$5. On the other hand, the maximum ECU that a subject could earn in a round was $275 - 300$; subjects' pre-show-up-fee payments were thus capped by \$27.5 in *DeterHigh*/*RandomHigh* and \$30 in *DeterLow*/*RandomLow*. Had we varied the absolute level of stigmatization aversion, we would have had to adjust the base earning upward for *DeterHigh* and *RandomHigh* resulting in a considerably higher upper bound of payments or, with no such upward adjustment, accept the possibility of negative earnings.

[20]Other efforts to attenuate biases caused by risk attitudes in belief elicitation include Allen (1987), Offerman, Sonnemans, van de Kuilen and Wakker (2009), Schlag and van der Weele (2009) and Hossain and Okui (forthcoming).

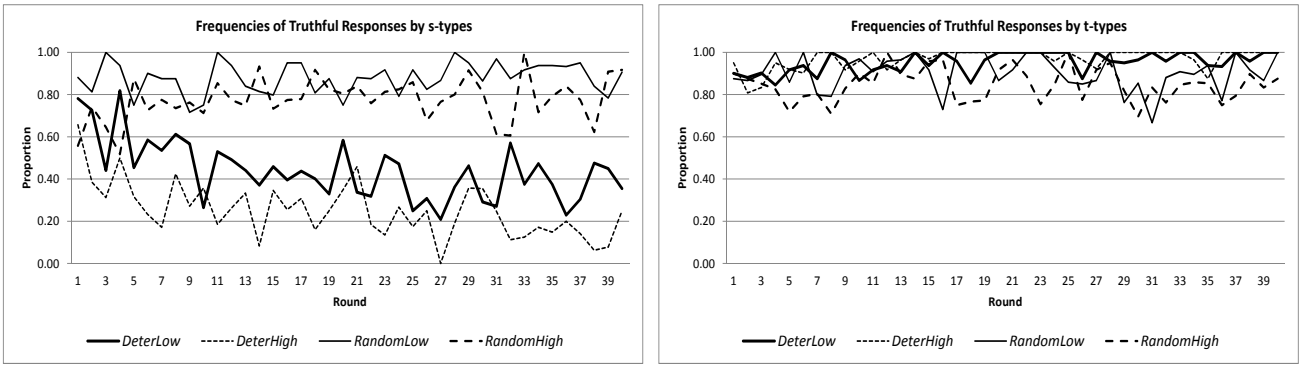
thus providing the incentives for eliciting true beliefs.[21]

At the end of each round, we provided information feedback on which shape and, for the randomized response treatments, which question were selected and revealed to Member A, Member A's response, Member B's prediction, and the subject's own earning.

7 Experimental Findings

7.1 Respondents' Responses and Interviewers' Beliefs

Figure 3 presents the trends of truthful response frequencies. Our first result compares, across the columns of the treatment matrix, the randomized response treatment with the deterministic response treatment:



(a) s-types (b) t-types

Figure 3: Trends of Truthful Response Frequencies

Result 1. *1) Stigmatized types provided truthful responses decidedly more often in the randomized response treatments than in the deterministic response treatments. 2) Regular types provided truthful responses more often in the deterministic response treatments than in the randomized response treatments; the differences, while statistically significant, was of much smaller magnitudes compared to the opposite differences of stigmatized types.*

Result 1 confirms Hypothesis 1. The frequencies of truthful responses by s-types, aggregated across the last 20 rounds of all sessions, were 37% in *DeterLow* and 19% in *DeterHigh*. The corresponding frequencies were 89% in *RandomLow* and 79% in *RandomHigh*. Using session-level data as independent observations, statistical tests confirm that the frequencies are significantly higher in the randomized response treatments irrespective of the levels of relative stigmatization

[21] Using induced beliefs, Hao and Houser (2012) experimentally evaluate the mechanism in Karni (2009). The way we presented the mechanism to the subjects was similar to theirs.

aversion ($p = 0.0143$ for all four possible comparisons, Mann-Whitney tests).[22] For t-types, the truthful response frequencies were significantly higher in the deterministic response treatments, but in aggregate the magnitudes of the differences were at most one third of those of s-types: the frequencies were 98% in both *DeterLow* and *DeterHigh*, 90% in *RandomLow*, and 84% in *RandomHigh* ($p = 0.0143$ for all four possible comparisons, Mann-Whitney tests).

Given that in the deterministic response treatments s's truthful response involves "no" and t's truthful response involve "yes," the frequencies reported above imply the following which addresses the first part of Hypothesis 2:

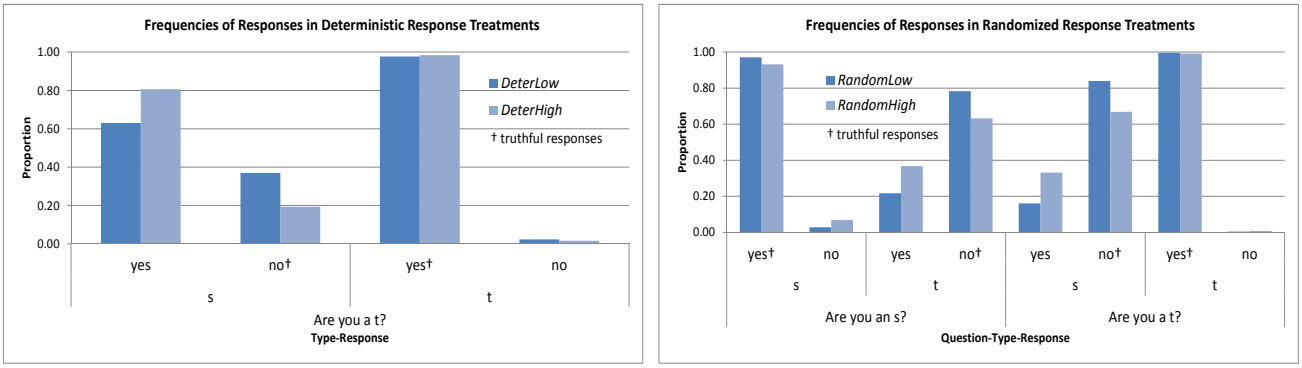
Result 1a. *1) In DeterLow, regular types and, to a lesser degree, stigmatized types responded with "yes" significantly more often than with "no." In DeterHigh, both stigmatized types and regular types responded with "yes" significantly more often than with "no." 2) In both treatments, regular types responded with "yes" significantly more often than did stigmatized types.*

Figure 4 presents the aggregate frequencies of responses. The behavior of t-types was very close to the point prediction of the D1 pooling equilibrium, where in both *DeterLow* and *DeterHigh* the frequencies of "yes" were 98%. On the other hand, s-types used "yes" less often than did t-types, rejecting the hypothesis that there is no significant difference between their behavior ($p = 0.0625$ for both treatments, Wilcoxon signed-rank tests). Given that t-types almost always responded with "yes," s-types' non-negligible uses of "no" transmitted information, contrasting the prediction of the pooling equilibrium. Over-communication, a common finding in the experimental literature of communication games (e.g., Forsythe, Lundholm and Rietz, 1999; Blume, Dejong, Kim and Sprinkle, 1998, 2001; Cai and Wang, 2006), was thus also observed in our experiments.[23] The qualitative prediction that "yes" is used more often than "no" by s-types was, however, largely confirmed ($p = 0.0625$ for *DeterHigh* and $p = 0.125$ for *DeterLow*, Wilcoxon signed-rank tests).

Despite equilibrium predicting no behavior difference between *DeterLow* and *DeterHigh*, in light of the over-communication observed, a natural question is how its extent responds to the different incentives under alternative levels of relative stigmatization aversion. Our next result addresses the question by comparing across the row of the treatment matrix, covering also the randomized response treatments:

[22] All aggregate data reported and used for statistical testings are from last 20 rounds. The qualitative aspects of our findings remain unchanged if we use, for example, data from last 30 or even all 40 rounds. However, the frequency trends, especially those for types s in the deterministic response treatments where convergence was most conspicuous, suggest that the 20th round provides a reasonable cutoff for the settlement of behavior. Using data from last 20 rounds thus allows us to give more weight to converged behavior. Unless otherwise indicated, the reported p-values are from one-sided tests.

[23] We conducted an additional session for robustness check, where the parameters were the same as *DeterLow* except that $p_s = 1$ (i.e., the deterministic question became "Are you an s?"). Compared to *DeterLow* with $p_s = 0$, a higher instance of over-communication by s-types was observed: the frequency of truthful "yes" response was 46%. There was almost no difference for t-types, where the frequency of truthful "no" response was 99%.



(a) Deterministic Response Treatments (b) Randomized Response Treatments

Figure 4: Frequencies of Responses

Result 2. *1) Stigmatized types provided truthful responses significantly more often in the low relative stigmatization treatments than in the high relative stigmatization treatments. 2) To a lesser degree, regular types provided truthful response significantly more often in RandomLow than in RandomHigh; there was no significant difference in regular types' truthful response frequencies between DeterLow and DeterHigh.*

For the deterministic response treatments, the second part of Hypothesis 2 is confirmed for t-types but not for s-types: the stronger relative stigmatization aversion in *DeterHigh* had no impact on t-types' behavior (two-sided $p = 1$, the Mann-Whitney test), whereas s-types over-communicated less when it was more costly to do so ($p = 0.0286$, Mann-Whitney test).

In the randomized response treatments, the different levels of relative stigmatization aversion affected the truthful behavior of both s-types and t-types, with a slightly stronger effect on the former ($p = 0.0143$ for s-types and $p = 0.0571$ for t-types, Mann-Whitney tests). Figure 4(b) shows that the frequencies of affirmative truthful responses were largely the same in *RandomLow* and *RandomHigh*, and the effects of stronger relative stigmatization aversion were exerted through negative truthful responses. For the question "Are you an s?" s-types responded affirmatively with "yes" with frequencies 97% in *RandomLow* and 93% in *RandomHigh*; t-types responded negatively with "no" with frequencies 78% in *RandomLow* and 63% in *RandomHigh*.[24]

[24]This represents the kind of protective behavior by non-stigmatized types that John et al. (2013) make responsible for occasional non-intuitive data obtained with RRT. Since in our experiment "Are you a t?" is the more frequently asked question, in a putative truthful equilibrium a "no" response is more jeopardizing: "no" is the response that moves posterior beliefs in the direction of giving more weight to the stigmatized s-type. Thus t-types (as well as s-types), all else equal, have an incentive to avoid giving "no" responses. In a truthful equilibrium this incentive is balanced by the incentive to be truthful. As our equilibrium analysis reveals, however, a complicating feature is that there are multiple equilibria and we therefore face an equilibrium selection problem. It is not implausible that the balance of stigmatization and truthfulness concerns also affects equilibrium selection; from this perspective the focal principle of privacy protection may undermine that of truthfulness and push equilibrium behavior away from the extreme of pure truth telling. An additional contributing factor for observing protective behaviors in the field may be heterogeneity in individual weightings of truth-telling and stigmatization concerns. Those with stronger stigmatization concerns might be expected to engage in protective

For the question "Are you a t?" t-types responded affirmatively with "yes" with frequencies higher than 99% in both *RandomLow* and *RandomHigh*; s-types responded negatively with "no" with frequencies 84% in *RandomLow* and 67% in *RandomHigh*.

To explore what drove the respondents' observed behavior, we bring the interviewers' beliefs into the picture. The "Elicited" columns in Table 2 present the interviewers' elicited beliefs.[25] Observations from individual subjects were fairly noisy as can be seen by the high standard deviations. We proceed to our next result, which addresses Hypothesis 3:

Table 2: Elicited and Empirical Beliefs Assigned to Type s

Response	"yes"		"no"		"yes"		"no"	
Beliefs	Elicited	Empirical	Elicited	Empirical	Elicited	Empirical	Elicited	Empirical
	DeterLow				*DeterHigh*			
Session 1	0.39 (0.21)	0.39	0.71 (0.25)	0.87	0.31 (0.16)	0.44	0.77 (0.15)	0.86
Session 2	0.33 (0.19)	0.39	0.87 (0.12)	1.00	0.40 (0.23)	0.42	0.69 (0.35)	0.75
Session 3	0.37 (0.31)	0.40	0.87 (0.16)	0.93	0.38 (0.14)	0.40	0.71 (0.28)	0.88
Session 4	0.43 (0.24)	0.30	0.71 (0.24)	0.96	0.37 (0.16)	0.39	0.65 (0.20)	1.00
Mean	0.38 (0.04)	0.37	0.79 (0.09)	0.94	0.36 (0.04)	0.42	0.70 (0.05)	0.87
	RandomLow				*RandomHigh*			
Session 1	0.42 (0.21)	0.45	0.65 (0.16)	0.68	0.43 (0.20)	0.43	0.51 (0.19)	0.50
Session 2	0.52 (0.24)	0.37	0.63 (0.21)	0.62	0.33 (0.22)	0.52	0.67 (0.20)	0.71
Session 3	0.45 (0.26)	0.44	0.64 (0.26)	0.59	0.40 (0.13)	0.50	0.54 (0.12)	0.71
Session 4	0.48 (0.17)	0.47	0.58 (0.14)	0.54	0.33 (0.21)	0.44	0.57 (0.23)	0.54
Mean	0.46 (0.04)	0.43	0.63 (0.03)	0.61	0.37 (0.05)	0.47	0.57 (0.07)	0.61

Note: Data are from last 20 rounds of each session. For the elicited beliefs, the parentheses contain standard deviations. The standard deviations for each session are calculated using each group in each round as an observation. Standard deviations for treatments are calculated using each session as an observation. For the empirical beliefs, the numbers are obtained by applying Bayes' rule to the observed frequencies of the respondents' types, the questions, and the respondents' responses conditional on types, aggregated across the last 20 rounds of each session.

Result 3. *1) In all treatments, the probabilities assigned to s according to the elicited beliefs were significantly higher after "no" than after "yes." 2) In the deterministic response treatments, the probabilities assigned to s according to the elicited beliefs after "yes" were significantly below* 0.5.

The interviewers' elicited beliefs were consistent with the over-communication observed in the deterministic response treatments. While the D1 pooling equilibrium predicts that the interviewer believes s and t to be equally likely after receiving "yes," the aggregate elicited beliefs assigned to s were 0.38 in *DeterLow* and 0.36 in *DeterHigh*, significantly lower than 0.5 ($p = 0.0625$ for both treatments, Wilcoxon signed-rank tests). It did, however, indicate that the interviewer-subjects believe – correctly – that their opponents were transmitting information.

The out-of-equilibrium "no" was received, on average, 19% of the time in *DeterLow* and 10% of the time in *DeterHigh*. The corresponding elicited beliefs assigned to s were 0.79 in *DeterLow*

behaviors even if others are content with being truthful.

[25]We use the 20th round as the cutoff for aggregations so as to maintain consistency with the aggregations of respondents' data. In most cases, the trends were stable over round.

and 0.70 in *DeterHigh*, significantly higher than when "yes" was received ($p = 0.0625$ for both treatments, Wilcoxon signed-rank tests). In fact, in *DeterLow* 44% of the time the elicited beliefs were equal or larger than 0.9, while it was 31% in *DeterHigh*. The low but positive frequencies observed for "no" provided a window to investigate how the interviewer-subjects assigned beliefs for events that in theory are off the equilibrium path. Although it did not require a sophisticated reasoning to assign a higher probability to s given that the interviewer-subjects were receiving "no" to "Are you a t?" the elicited beliefs reflected the forward-induction reasoning that s was more likely than t to respond with "no."

Note that with higher probabilities assigned to s after "no" than after "yes," responding with "yes" provided t-types with two monetary rewards, one from telling the truth and one from inducing a lower probability assigned to s. This accounted for why t-types almost always provided truthful responses. On the other hand, when s-types told the truth with "no," they were trading the truthful response reward for a higher probability assigned to s. Given the magnitudes of elicited beliefs, the latter on average was sufficient to outweigh the former, suggesting that considerations other than monetary rewards might be driving the over-communication on the respondents' part, depriving us of equilibrium behavior insofar as monetary incentives are concerned.[26] Prior experimental studies have documented that subjects had intrinsic preference for honesty (e.g., Gneezy, 2005; Sánchez-Pagés and Vorsatz, 2007). In our case, it is conceivable that home-grown lying aversion was brought into the laboratory which added on to the one we induce with monetary rewards, resulting in a lower effective level of relative stigmatization aversion. Indeed, the respondents' observed behavior resembled the informative equilibrium under a higher lying-stigmatization aversion ratio.[27]

The elicited beliefs assigned to s after "no" were 0.63 in *RandomLow* and 0.57 in *RandomHigh*, significantly higher than the beliefs after "yes," which were 0.46 in *RandomLow* and 0.37 in *RandomHigh* ($p = 0.0625$ for both treatments, Wilcoxon signed-rank tests). In all the equilibria under the adopted parameters, the interviewer's beliefs assigned to s were in the neighborhoods of 0.6 after one response and 0.4 after another. Whether the higher probability occurred after "yes" or after "no" was consistent with equilibrium. With the aid of more sophisticated reasoning than was required in the deterministic response treatments but nothing close to a full-blown use of Bayes' rule, the probabilities of the questions might have provided a focal point for subjects to

[26] To support the uninformative equilibria, the out-of-equilibrium beliefs assigned to s were required to be ≥ 0.75 in *DeterLow* and ≥ 0.625 in *DeterHigh*.

[27] Risk aversion might also have played a role. To avoid additional layer of complexity to our already involved experimental instructions, we did not use binary lotteries (Roth and Malouf, 1979; Berg, Dickhaut and O'Brien, 1986) to induce risk neutrality. The respondents were trading a certain sum from truthful responses for a risky prospect of lower probability assigned to s. Given the high variations in elicited beliefs, risk aversion might have favored truthful responses. Note, however, that this does not undermine the conclusion that the use of random questions led to more truthful responses, as highly varied elicited beliefs were also observed in the randomized response treatments; risk aversion was largely controlled for in the comparisons between the two sets of response treatments.

form beliefs that are close to the predicted values and with the higher values assigned after "no." Upon receiving "yes," if an interviewer-subject considers that the respondent is very likely telling the truth, it is fairly straightforward to reason that with probability around 0.4 the respondent's type is s, because such is the probability that "Are you an s?" is asked. Similar reasoning would lead one to conclude that the probability of s should be around 0.6 after "no." While there were considerable variations in individual observations so that the aggregate numbers close to the predicted values might not be representative, the above reasoning may have at least led subjects to believe correctly that "no" was more likely to come from s than is "yes."

Equilibria that are consistent with the elicited belief profiles are: a truthful equilibrium in *RandomLow*; informative equilibria in both *RandomLow* and *RandomHigh* in which the respondent always gives affirmative truthful responses but randomizes between "yes" and "no" when truthful responses are negative. The observed aggregate behavior resembled the non-truthful informative equilibria. The frequencies of "yes" by s-types to "Are you an s?" and by t-types to "Are you a t?" were highly consistent with the predicted affirmative truthful responses, where in both *RandomLow* and *RandomHigh* the frequencies were close to 100%. In the cases where truthful responses were negative, randomizations between "yes" and "no" consistent with the informative equilibria generated predictions that are within $\pm 5\%$ of the observed frequencies.[28]

While the multiple equilibria provide large degrees of freedom for interpreting observed behavior, the distinctive profile of response frequencies provided evidence that subjects' behavior was shaped by the specific incentives behind the equilibrium constructions. The key observation was again related to the presence and absence of tradeoff in giving truthful responses. In the equilibria in which the interviewer's beliefs assigned to s are higher after "no," s and t strictly prefer to give affirmative truthful responses: they obtain a higher payoff from being truthful, and "yes" results in a lower probability assigned to s. When s is responding to "Are you a t?" or t is responding to "Are you an s?" they are, however, facing a tradeoff: responding truthfully with "no" brings a higher payoff at the expense of a higher value of the interviewer's beliefs. In equilibrium, either the payoff from truthful responses dominates and we have a truthful equilibrium, or the two balances out so that s and t are willing to randomize between "yes" and "no" for a non-truthful informative equilibrium.

In the experiments, the respondent-subjects behaved in a manner that precisely matched the non-truthful informative equilibrium: given the elicited beliefs, they overwhelming chose to be truthful with affirmative responses, but when a truthful response involved "no" they were truthful less often, which in aggregate resembled randomization between "yes" and "no." The elicited beliefs in turn reflected the observed behavior of the respondents, closing the loop for

[28]For $\mu_s(n) > \mu_s(y)$, we have that $\sigma(y|s,q_s) = \sigma(y|t,q_t) = 1$, and the remaining equilibrium strategies satisfy $\sigma(n|t,q_s) \in (0,1)$ and $\sigma(n|s,q_t) = [\sqrt{25+80\sigma(n|t,q_s)} - 2\sigma(n|t,q_s) - 5]/3 \in (0,1]$ in *RandomLow* and $\sigma(n|t,q_s) \in (0,1]$ and $\sigma(n|s,q_t) = [\sqrt{225+160\sigma(n|t,q_s)} - 2\sigma(n|t,q_s) - 15]/3 \in (0,1)$ in *RandomHigh*. The formulae for equilibrium strategies can generate $\sigma(n|s,q_t) \approx 0.89$ (84% observed) and $\sigma(n|t,q_s) \approx 0.73$ (78% observed) in *RandomLow* and $\sigma(n|s,q_t) \approx 0.63$ (67% observed) and $\sigma(n|t,q_s) \approx 0.67$ (63% observed) in *RandomHigh*.

equilibrium.[29] Note also that when the elicited beliefs were higher after "no" than after "yes," respondent-subjects' optimal choice was to give affirmative truthful responses regardless of the level of relative stigmatization aversion. However, for negative truthful responses, a higher level of relative stigmatization tilted the tradeoff in favor of the non-truthful "yes." This accounted for the observation that the lower frequencies of truthful responses in *RandomHigh* mainly consisted of lower frequencies of negative truthful responses.[30]

7.2 Empirical Beliefs and Information-Eliciting Performance

In this subsection, we analyze the observed performance of the response regimes by computing the mutual information implied by our data. An important component of the mutual information are the probabilities of the respondent's types conditional on the responses, i.e., the interviewer's updated beliefs. The "Empirical" columns in Table 2 presents the *empirical beliefs*, where we apply Bayes' rule to the observed frequencies of the respondents' types, the questions, and the respondents' responses to obtain the conditional frequencies of types. Other than being instrumental to the calculation of the mutual information, the empirical beliefs are of interests on their own. They provide a summary of the respondents' observed behavior in terms that can be directly compared with the elicited beliefs, allowing us to systematically explore how well the elicited beliefs reflected the respondents' actual play.

The empirical beliefs shared the same qualitative properties with the elicited beliefs in that the beliefs assigned to s were significantly higher after "no" than after "yes" ($p = 0.0625$ for all four treatments, Wilcoxon signed-rank tests), suggesting that the respondents' qualitative behavior was reflected by the elicited beliefs. Quantitatively, the absence of significant differences between the two kinds of beliefs in most cases in the randomized response treatments (two-sided $p \geq 0.375$ except for "yes" in *RandomHigh* in which one-sided $p = 0.0625$, Wilcoxon signed-rank tests) suggests that the interviewers behaved as if they were best responding. There was also no significant difference between the two kinds of beliefs after "yes" in *DeterLow* (two-sided $p = 0.875$, Wilcoxon signed-rank test), indicating that in aggregate the elicited beliefs accurately reflected the actual information contained in the response. However, in the other cases in the deterministic response treatments, the empirical beliefs were significantly higher

[29] We explore this side of the best responses in more details in the next subsection.

[30] To further explore if subjects were responding to the incentives resulting from the elicited belief differentials or they simply had a stronger intrinsic preference for truthfully responding with "yes" than with "no," we conducted another robustness session with the same parameters as *RandomHigh* except that now $p_s = 0.6$ (i.e., "Are you an s?" instead of "Are you a t?" is asked with 60% chance). Supporting our argument that the probabilities of the questions provided a focal point for belief formation, the elicited beliefs assigned to s were higher after "yes." Accordingly, the incentives for the two kinds of truthful responses were reversed: strict preference exists for negative truthful responses while there is a tradeoff for affirmative truthful responses. The respondent-subjects responded to these different incentives. For the question "Are you an s?" s-types responded with "yes" with frequency 44%; t-types responded with "no" with frequency 94%. For the question "Are you a t?" t-types responded with "yes" with frequency 70%; s-types responded with "no" with frequency 94%.

than the elicited beliefs ($p = 0.0625$ for beliefs after "no" in *DeterLow* and after both responses in *DeterHigh*, Wilcoxon signed-rank tests). The profiles of the differences suggest that in these cases the interviewer-subjects tended to overestimate the information contained in "yes" and underestimated the information contained in "no": the elicited beliefs after "yes" were farther away from the prior than were the empirical beliefs, while the opposite was true for beliefs after "no."

The mutual information implied by our data are reported in Table 3. Our next and final result compares the mutual information under different response regimes and different levels of relative stigmatization aversion:

Table 3: Mutual Information

	DeterLow	*DeterHigh*
Session 1	0.134	0.055
Session 2	0.213	0.012
Session 3	0.080	0.056
Session 4	0.267	0.198
Mean	0.173	0.080
	RandomLow	*RandomHigh*
Session 1	0.038	0.002
Session 2	0.045	0.027
Session 3	0.017	0.029
Session 4	0.003	0.007
Mean	0.026	0.016

Note: The numbers are obtained by applying the formula of mutual information to the empirical beliefs in Table 2 and the observed frequencies (aggregated across the last 20 rounds) of the respondents' types and responses.

Result 4. *1) The observed mutual information was significantly higher in DeterLow than in RandomLow. To a lesser degree, it was also significantly higher in DeterHigh than in RandomHigh. 2) The observed mutual information was, also to a lesser degree, significantly higher in DeterLow than in DeterHigh, while there was no significant difference between RandomLow and RandomHigh.*

Theory predicts that for the parameters in questions the randomized response regime performs strictly better than the deterministic response regime. However, rejecting Hypothesis 4, the deterministic response regime dominated in the laboratory due to the over-communication. Yet, as the extent of over-communication decreased in response to higher degree of relative stigmatization aversion, the dominance also became weaker and less significant ($p = 0.0143$ for the comparison between *DeterLow* and *RandomLow* and $p = 0.0571$ between *DeterHigh* and *RandomHigh*, the Mann-Whitney tests). The conclusion can also be reached by the fact that

stronger stigmatization had a negative impact on the performance of the deterministic response regime but had no effect on the randomized response regime (one-sided $p = 0.0571$ for the comparison between *DeterLow* and *DeterHigh* and two-sided $p = 0.4857$ between *RandomLow* and *RandomHigh*, Mann-Whitney tests). Despite the discrepancies between the theory and the data regarding their relative performance, a comparative statics prediction from the theory was recovered in the laboratory: relative to the deterministic response regime, the randomized response regime performed better when there was a stronger stigmatization involved.

8 Discussion and Conclusion

The literature on communication games suggests that injecting noise into the communication protocol may improve information transmission.

Myerson's (1991) messenger pigeon example nicely illustrates the underlying rationale. In a sender-receiver game with two sender types, type 2 prefers mimicking type 1 to being identified. As a result, there is no equilibrium in which both types truthfully send messages "I am type 1" and "I am type 2." If, however, communication passes through a noisy channel that sometimes replaces the message "I am type 1" with the message "I am type 2" (in the example the message "I am type 1" is the act of sending a messenger pigeon, and message replacement is failure of the pigeon to arrive) then receipt of the message "I am type 2" no longer unequivocally attaches the undesirable type 2 label to the sender, and in the example separation at the message sending stage, or being truthful in terms of the message labels "I am type 1" and "I am type 2", becomes possible.

When conducting surveys about sensitive issues with a binary state space, analogous to Myerson's messenger pigeon example, there are two types of respondents, one of which is stigmatized and therefore prefers not to be identified. If the stigmatized type prefers that there remain uncertainty about its identity to being misidentified as the non-stigmatized type, i.e., the stigmatized type's preference order is $\textit{uncertainty} \succ \textit{misidentification} \succ \textit{identification}$ – then the payoff structure is exactly as in Myerson's example and the same logic applies. Communication through a noisy channel helps mask the stigma and improves the incentive for being truthful.

In the survey setting it is perhaps more plausible and therefore assumed that the stigmatized type's preference is monotonic in the probability that the interviewer's belief assigns to the stigmatized trait, and therefore $\textit{misidentification} \succ \textit{uncertainty} \succ \textit{identification}$. In that case, some degree of lying aversion can restore the former preference order, provided the respondent's preference order satisfies $\textit{uncertainty and telling the truth} \succ \textit{misidentification and lying} \succ \textit{identification and telling the truth}$. Then a request to be truthful combined with a random mechanism for generating and privately posing questions, as in Warner's (1965) formulation of RRT, generates the same incentive structure as in Myerson's example, with the accompanying

promise that there may be a truth-telling equilibrium.

We have put forward a simple game model that lets us unpack the different components of this narrative. In this model stigmatization aversion is counterbalanced by lying aversion, albeit not to the degree that truth-telling becomes a dominant strategy. Truth-telling is only achieved if questions are randomized, and even then only for a subset of the parameter space.

The theoretical analysis of the game yields two principal novel insights: there are informative equilibria that are not truthful, and truthful equilibria need not be the maximally informative equilibria. The first of these observations undermines the rationale for trying to validate RRT with studies that show that using it yields higher estimates of the prevalence of stigmatized traits than other approaches. The second observation is a reminder that sometimes the direct informational loss from randomizing questions will be too steep a price to pay for truthfulness.

The laboratory implementation of our game on one hand confirms the theoretical prediction that RRT has the capacity to increase the incidence of truth-telling relative to direct questioning. At the same, we find that there are systematic deviations from truth-telling when according to the induced parameters truth-telling is an equilibrium. This is despite the fact that there is a body of experimental evidence suggesting that experimental subjects bring prior preferences for truth-telling to the lab (e.g., Sánchez-Pagés and Vorsatz, 2007; Gneezy, 2010), which would only reinforce incentives for truth-telling in our experiment. Contrary to the prediction of the theory, we find that in the lab the informativeness of RRT falls short of that of direct questioning.

Overall, our theoretical and experimental results suggest that while the potential for injecting noise into the communication protocol to enhance truth-telling is real, it may be difficult to utilize this effect to achieve (approximately) complete truth-telling and to improve informativeness.

Appendix A – Proofs

We list the following payoff profiles that will be used in the proofs.

$$U((s,q_s,y),\mu_s(y)) = U((t,q_t,y),\mu_s(y)) = \lambda - \xi\mu_s(y), \tag{A.1}$$

$$U((s,q_s,n),\mu_s(n)) = U((t,q_t,n),\mu_s(n)) = -\xi\mu_s(n), \tag{A.2}$$

$$U((s,q_t,n),\mu_s(n)) = U((t,q_s,n),\mu_s(n)) = \lambda - \xi\mu_s(n), \tag{A.3}$$

$$U((s,q_t,y),\mu_s(y)) = U((t,q_s,y),\mu_s(y)) = -\xi\mu_s(y). \tag{A.4}$$

Proof of Lemma 1. Suppose there is an equilibrium in which $|\mu_s(y) - \mu_s(n)| > \frac{\lambda}{\xi}$ on the equilibrium path. If $\mu_s(y) - \mu_s(n) > \frac{\lambda}{\xi}$, then $-\xi\mu_s(n) > \lambda - \xi\mu_s(y)$ and $\lambda - \xi\mu_s(n) > -\xi\mu_s(y)$. Regardless of whether it is q_s or q_t, (A.1)–(A.4) indicate that both s and t strictly prefer to respond with n. This implies that $\mu_s(y)$ is not on the equilibrium path, a contradiction. If $\mu_s(y) - \mu_s(n) < -\frac{\lambda}{\xi}$, then $\lambda - \xi\mu_s(y) > -\xi\mu_s(n)$ and $-\xi\mu_s(y) > \lambda - \xi\mu_s(n)$. (A.1)–(A.4) then indicate that both s and t strictly prefer to respond with y, which again leads to the contradiction that $\mu_s(n)$ is not on the equilibrium path. □

Proof of Proposition 1. We characterize all equilibria in the deterministic response regime with $q = q_t$. We first show that there exists no truthful equilibrium, which follows immediately from Lemma 1. If in an equilibrium both s and t give truthful responses with probability one, then $|\mu_s(y) - \mu_s(n)| = 1$. Given that $\frac{\lambda}{\xi} \in [0,1)$, this contradicts that in any equilibrium $|\mu_s(y) - \mu_s(n)| \leq \frac{\lambda}{\xi}$ on the equilibrium path.

Note that in any informative equilibrium with $q = q_t$, we must have that $\mu_s(n) > \mu_s(y)$; if $\mu_s(y) > \mu_s(n)$, it follows from (A.3) and (A.4) that s strictly prefers to respond with n, which implies that $\mu_s(n) \geq \mu_s(y)$, a contradiction. With $\mu_s(n) > \mu_s(y)$, it follows from (A.1) and (A.2) that t strictly prefers to respond with y. Thus, in any informative equilibrium, t must give truthful response with probability one and s must randomize between y and n. The indifference of s between y and n implies, from (A.3) and (A.4), that $\lambda = \xi[\mu_s(n) - \mu_s(y)]$. Given that n is used exclusively by s, we have $\mu_s(y) = 1 - \frac{\lambda}{\xi}$, which holds if and only if $\sigma(n|s) = 2 - \frac{\xi}{\lambda}$. Hence, if an informative equilibrium exists, it is unique. Since $\xi > \lambda \geq 0$, the requirement that $\sigma(n|s) \in (0,1)$ imposes the restriction that $\frac{\lambda}{\xi} > \frac{1}{2}$. Thus, if $\frac{\lambda}{\xi} \in (\frac{1}{2},1)$, one can construct an informative equilibrium; if an informative equilibrium exists, we must have $\frac{\lambda}{\xi} \in (\frac{1}{2},1)$.

In any uninformative equilibrium, either (a) only one response is used in equilibrium or (b) both responses are used and $\mu_s(y) = \mu_s(n) = \frac{1}{2}$. In case (b), it requires that $\sigma(y|s) = \sigma(y|t) \in (0,1)$, i.e., both s and t are indifferent between y and n. And given that $\mu_s(y) = \mu_s(n)$, they are indifferent if and only if $\lambda = 0$. Thus, an uninformative equilibrium with $\sigma(y|s) = \sigma(y|t) \in (0,1)$ exists if and only if $\frac{\lambda}{\xi} = 0$.

Consider next case (a). Suppose both s and t respond with y with probability one so that

$\mu_s(y) = \frac{1}{2}$ on the equilibrium path. For this to constitute an equilibrium, we require, from (A.1) and (A.2), that $\lambda \geq \xi[\frac{1}{2} - \mu_s(n)]$ for t and, from (A.3) and (A.4), that $\xi[\mu_s(n) - \frac{1}{2}] \geq \lambda$ for s, where $\mu_s(n)$ is an out-of-equilibrium belief. Only the second inequality binds, and thus the out-of-equilibrium belief required to support the equilibrium is that $\mu_s(n) \geq \frac{\lambda}{\xi} + \frac{1}{2}$. That $\mu_s(n) \in [0, 1]$ imposes the restriction that $\frac{\lambda}{\xi} \leq \frac{1}{2}$. Suppose next that both s and t respond with n with probability one so that $\mu_s(n) = \frac{1}{2}$ on the equilibrium path. By a similar argument, for this to constitute an equilibrium, we require that the out-of-equilibrium belief $\mu_s(y) \geq \frac{\lambda}{\xi} + \frac{1}{2}$, which again imposes the restriction that $\frac{\lambda}{\xi} \leq \frac{1}{2}$. Thus, if $\frac{\lambda}{\xi} \in [0, \frac{1}{2}]$, one can construct uninformative equilibria with outcomes where either both types respond with y or both respond with n; for $\frac{\lambda}{\xi} \in (0, \frac{1}{2}]$, these are the only uninformative equilibrium outcomes. Conversely, if an uninformative equilibrium exists, we must have $\frac{\lambda}{\xi} \in [0, \frac{1}{2}]$. This also implies that for any $\frac{\lambda}{\xi} \in (\frac{1}{2}, 1)$ there is no uninformative equilibrium and hence in that range the unique informative equilibrium is the only equilibrium.

We next apply the D1 criterion to the two equilibrium outcomes in which only one response is used. Let $U^*(\theta)$ be the equilibrium payoff of type-θ respondent. For the equilibrium outcome in which both types respond with n, we have that $U^*(s) = \lambda - \frac{\xi}{2}$ and $U^*(t) = -\frac{\xi}{2}$. If types s and t deviate to y, their payoffs will be, respectively, $\tilde{U}(s) = -\xi\mu_s(y)$ and $\tilde{U}(t) = \lambda - \xi\mu_s(y)$. Note that $\tilde{U}(s) - U^*(s) \geq 0$, i.e., type s weakly prefers deviating to y, if and only if $\mu_s(y) \in [0, \frac{1}{2} - \frac{\lambda}{\xi}]$. On the other hand, $\tilde{U}(t) - U^*(t) > 0$, i.e., type t strictly prefers deviating to y, if and only if $\mu_s(y) \in [0, \frac{1}{2} + \frac{\lambda}{\xi})$. Note that if $\frac{\lambda}{\xi} > 0$, $[0, \frac{1}{2} - \frac{\lambda}{\xi}] \subset [0, \frac{1}{2} + \frac{\lambda}{\xi})$; s is deleted for y under the D1 criterion, and thus the equilibrium outcome does not survive the selection criterion if $\frac{\lambda}{\xi} > 0$. Turning to the equilibrium outcome in which both types respond with y, note that type t weakly prefers to deviating to n if and only if $\mu_s(n) \in [0, \frac{1}{2} - \frac{\lambda}{\xi}]$ and type s strictly prefers to deviating to n if and only if $\mu_s(n) \in [0, \frac{1}{2} + \frac{\lambda}{\xi})$. By a similar argument, if $\frac{\lambda}{\xi} > 0$, the D1 criterion deletes t for n. The equilibrium outcome with both types responding with y can be supported by the resulting belief that $\mu_s(n) = 1$; the outcome thus survives the criterion for $\frac{\lambda}{\xi} > 0$. Finally, if $\frac{\lambda}{\xi} = 0$, the D1 criterion puts no restriction on the interviewer's out-of-equilibrium beliefs, and thus both outcomes survive the D1 criterion.

□

Proof of Proposition 2. We establish the result by verifying the following claim, which characterizes all equilibria in the randomized response regime:

In the randomized response regime in which $p_s \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$,

1. *there exist uninformative equilibria if and only if $\frac{\lambda}{\xi} \in [0, \frac{1}{2}]$; the class of uninformative equilibria in which all types (s, q_s), (t, q_t), (s, q_t) and (t, q_s) completely randomize between y and n in the same manner exists if and only if $\frac{\lambda}{\xi} = 0$;*

2. *there exists a truthful equilibrium if and only if $p_s \in [\frac{1}{2} - \frac{\lambda}{2\xi}, \frac{1}{2} + \frac{\lambda}{2\xi}]$; and,*

3. *the set of non-truthful informative equilibria is completely described by the following statements:*

   (a) *there exists an informative equilibrium in which* (s, q_s) *and* (t, q_t) *always give a truthful response and*

      i. (s, q_t) *always gives a truthful response and* (t, q_s) *randomizes between* y *and* n *if and only if* $\frac{1}{2} - \frac{\lambda}{2\xi} < p_s < \frac{\xi}{\lambda} - 1$;

      ii. (s, q_t) *randomizes between* y *and* n *and* (t, q_s) *always gives a truthful response if and only if* $p_s < \frac{1}{2} - \frac{\lambda}{2\xi}$;

      iii. (s, q_t) *randomizes between* y *and* n *and* (t, q_s) *always gives a non-truthful response if and only if* $p_s < \frac{\xi}{\lambda} - 1 < 1$;

      iv. (s, q_t) *always give a truthful response and* (t, q_s) *always gives a non-truthful response if and only if* $p_s = \frac{\xi}{\lambda} - 1$;

      v. (s, q_t) *and* (t, q_s) *randomize between* y *and* n *if and only if* $p_s < \frac{\xi}{\lambda} - 1$;

   (b) *there exists an informative equilibrium in which* (s, q_t) *and* (t, q_s) *always give a truthful response and*

      i. (s, q_s) *always gives a truthful response and* (t, q_t) *randomizes between* y *and* n *if and only if* $2 - \frac{\xi}{\lambda} < p_s < \frac{1}{2} + \frac{\lambda}{2\xi}$;

      ii. (s, q_s) *randomizes between* y *and* n *and* (t, q_t) *always gives a truthful response if and only if* $p_s > \frac{1}{2} + \frac{\lambda}{2\xi}$;

      iii. (s, q_s) *randomizes between* y *and* n *and* (t, q_t) *always gives a non-truthful response if and only if* $p_s > 2 - \frac{\xi}{\lambda} > 0$;

      iv. (s, q_s) *always gives a truthful response and* (t, q_t) *always gives a non-truthful response if and only if* $p_s = 2 - \frac{\xi}{\lambda}$;

      v. (s, q_s) *and* (t, q_t) *randomize between* y *and* n *if and only if* $p_s > 2 - \frac{\xi}{\lambda}$.

Given that the respondent has four types, (s, q_s), (t, q_t), (s, q_t) and (t, q_s) and each type can either respond with y with probability one, respond with n with probability one, or completely randomize between the two, there are in total 81 classes of strategy profiles as candidates for equilibrium. We proceed by either characterizing the condition under which a class of strategy profiles constitutes equilibria or eliminating one as equilibrium candidate, until we exhaust all 81 possibilities.

We begin with the uninformative equilibria in part 1 of the claim. In any such equilibrium, either (a) only one response is used in equilibrium or (b) both responses are used and $\mu_s(y) = \mu_s(n) = \frac{1}{2}$. In case (b), it requires that $\sigma(y|s, q_s) = \sigma(y|t, q_t) = \sigma(y|s, q_t) = \sigma(y|t, q_s) \in (0, 1)$, i.e., all types are indifferent between y and n. And given that $\mu_s(y) = \mu_s(n)$, they are indifferent if

and only if $\lambda = 0$. Thus, an uninformative equilibrium with $\sigma(y|s,q_s) = \sigma(y|t,q_t) = \sigma(y|s,q_t) = \sigma(y|t,q_s) \in (0,1)$ exists if and only if $\frac{\lambda}{\xi} = 0$. For case (a), consider first that all types respond with y with probability one so that $\mu_s(y) = \frac{1}{2}$ on the equilibrium path. For this to constitute an equilibrium, we require, from (A.1) and (A.2), that $\lambda \geq \xi[\frac{1}{2} - \mu_s(n)]$ for (s,q_s) and (t,q_t) and, from (A.3) and (A.4), that $\xi[\mu_s(n) - \frac{1}{2}] \geq \lambda$ for (s,q_t) and (t,q_s), where $\mu_s(n)$ is an out-of-equilibrium belief. Only the second inequality binds, and thus the out-of-equilibrium belief required to support the equilibrium is that $\mu_s(n) \geq \frac{\lambda}{\xi} + \frac{1}{2}$. That $\mu_s(n) \in [0,1]$ imposes the restriction that $\frac{\lambda}{\xi} \leq \frac{1}{2}$. Consider next that all types respond with n with probability one so that $\mu_s(n) = \frac{1}{2}$ on the equilibrium path. By a similar argument, for this to constitute an equilibrium, we require that the out-of-equilibrium belief $\mu_s(y) \geq \frac{\lambda}{\xi} + \frac{1}{2}$, which again imposes the restriction that $\frac{\lambda}{\xi} \leq \frac{1}{2}$. Thus, if $\frac{\lambda}{\xi} \in [0, \frac{1}{2}]$, one can construct uninformative equilibria where either all types respond with y or all respond with n; for $\frac{\lambda}{\xi} \in (0, \frac{1}{2}]$, these are the only uninformative equilibria. Conversely, if an uninformative equilibrium exists, we must have $\frac{\lambda}{\xi} \in [0, \frac{1}{2}]$.

We are left with 78 possibilities. We proceed to eliminate candidates for informative equilibria. Recall that in an informative equilibrium, y and n are used with positive probability and $\mu_s(y) \neq \mu_s(n)$. Note that whenever $\mu_s(y) \neq \mu_s(n)$, at least two types strictly prefer their truthful response that also results in lower μ_s. If $\mu_s(n) > \mu_s(y)$, it follows from (A.1) and (A.2) that $\sigma(y|s,q_s) = \sigma(y|t,q_t) = 1$. If $\mu_s(y) > \mu_s(n)$, it follows from (A.3) and (A.4) that $\sigma(y|s,q_t) = \sigma(y|t,q_s) = 0$. The condition that either $\sigma(y|s,q_s) = \sigma(y|t,q_t) = 1$ or $\sigma(y|s,q_t) = \sigma(y|t,q_s) = 0$ eliminates 63 classes of strategy profiles, leaving 15 distinct possibilities. Consider that $\mu_s(n) > \mu_s(y)$ so that $\sigma(y|s,q_s) = \sigma(y|t,q_t) = 1$. The interviewer's beliefs are

$$\mu_s(n) = \frac{(1-p_s)(1-\sigma(y|s,q_t))}{p_s(1-\sigma(y|t,q_s)) + (1-p_s)(1-\sigma(y|s,q_t))}, \tag{A.5}$$

$$\mu_s(y) = \frac{p_s + (1-p_s)\sigma(y|s,q_t)}{1 + p_s\sigma(y|t,q_s) + (1-p_s)\sigma(y|s,q_t)}. \tag{A.6}$$

If $\sigma(y|s,q_t) = 1$, both (s,q_s) and (s,q_t) respond with y with probability one, leading to the contradiction that $\mu_s(n) = 0$. Thus, two additional classes of strategy profiles, which prescribe $\sigma(y|s,q_s) = \sigma(y|t,q_t) = \sigma(y|s,q_t) = 1$ coupled with either $\sigma(y|t,q_s) = 0$ or $\sigma(y|t,q_s) \in (0,1)$, are ruled out. Consider next that $\mu_s(y) > \mu_s(n)$ so that $\sigma(y|s,q_t) = \sigma(y|t,q_s) = 0$. The interviewer's beliefs are

$$\mu_s(y) = \frac{p_s\sigma(y|s,q_s)}{p_s\sigma(y|s,q_s) + (1-p_s)\sigma(y|t,q_t)}, \tag{A.7}$$

$$\mu_s(n) = \frac{1 - p_s + p_s(1-\sigma(y|s,q_s))}{1 + p_s(1-\sigma(y|s,q_s)) + (1-p_s)(1-\sigma(y|t,q_t))}. \tag{A.8}$$

If $\sigma(y|s,q_s) = 0$, both (s,q_s) and (s,q_t) respond with n with probability one, leading to the

contradiction that $\mu_s(y) = 0$. Thus, two more classes of strategy profiles, which prescribe $\sigma(y|s,q_t) = \sigma(y|t,q_s) = \sigma(y|s,q_s) = 0$ coupled with either $\sigma(y|t,q_t) = 1$ or $\sigma(y|t,q_t) \in (0,1)$, are further eliminated.

The rest of the proof verifies and characterizes the remaining 11 classes of strategy profiles as informative equilibria. Consider first the truthful equilibrium in part 2 of the claim, in which $\sigma(y|s,q_s) = \sigma(y|t,q_t) = 1$ and $\sigma(y|s,q_t) = \sigma(y|t,q_s) = 0$. The resulting interviewer's beliefs are $\mu_s(y) = p_s$ and $\mu_s(n) = 1 - p_s$. Suppose that $p_s < \frac{1}{2}$. It follows from (A.1) and (A.2) that (s,q_s) and (t,q_t) strictly prefer y to n. For (s,q_t) and (t,q_s) to weakly prefer n to y, it follows from (A.3) and (A.4) that we require $p_s \geq \frac{1}{2} - \frac{\lambda}{2\xi}$. Suppose next that $p_s > \frac{1}{2}$. It follows from (A.3) and (A.4) that (s,q_t) and (t,q_s) strictly prefer n to y. For (s,q_s) and (t,q_t) to weakly prefer y to n, it follows from (A.1) and (A.2) that we require $p_s \leq \frac{1}{2} + \frac{\lambda}{2\xi}$. Truthful equilibria thus exist if and only if $p_s \in [\frac{1}{2} - \frac{\lambda}{2\xi}, \frac{1}{2} + \frac{\lambda}{2\xi}]$.

We proceed to non-truthful informative equilibria. We divide the remaining 10 cases according to the magnitudes of the interviewer's beliefs. Consider first that $\mu_s(n) > \mu_s(y)$. The strategies $\sigma(y|s,q_s) = \sigma(y|t,q_t) = 1$ are to be coupled with $\sigma(y|s,q_t) = [0,1)$ and $\sigma(y|t,q_s) \in [0,1]$, accounting for five remaining classes of strategy profiles. All of them require, from (A.3) and (A.4), that $\lambda = \xi[\mu_s(n) - \mu_s(y)] > 0$. Substituting (A.5) and (A.6) into $\lambda = \xi[\mu_s(n) - \mu_s(y)]$ and solving for $\sigma(y|s,q_t)$, we obtain

$$\sigma(y|s,q_t) = \frac{\xi \pm \sqrt{4\lambda^2 - 4\lambda\xi[1 - 2p_s(1 - \sigma(y|t,q_s))] + \xi^2} - 2\lambda p_s \sigma(y|t,q_s)}{2\lambda(1-p_s)}.$$

Note that for $0 \leq \lambda < \xi$, $\sqrt{4\lambda^2 - 4\lambda\xi[1 - 2p_s(1 - \sigma(y|t,q_s))] + \xi^2} - 2\lambda p_s \sigma(y|t,q_s) \geq \sqrt{(2\lambda - \xi)^2} - 2\lambda p_s$. Thus, we have that

$$\frac{\xi + \sqrt{4\lambda^2 - 4\lambda\xi[1 - 2p_s(1 - \sigma(y|t,q_s))] + \xi^2} - 2\lambda p_s \sigma(y|t,q_s)}{2\lambda(1-p_s)} \geq \frac{\xi + \sqrt{(2\lambda - \xi)^2} - 2\lambda p_s}{2\lambda(1-p_s)}.$$

Given that $\frac{\xi + \sqrt{(2\lambda-\xi)^2} - 2\lambda p_s}{2\lambda(1-p_s)} = 1$ for $2\lambda - \xi \geq 0$ and $\frac{\xi + \sqrt{(2\lambda-\xi)^2} - 2\lambda p_s}{2\lambda(1-p_s)} > 1$ for $2\lambda - \xi < 0$, the above solution for $\sigma(y|s,q_t)$ is not relevant ($\sigma(y|s,q_t) = 1$ is ruled out above). The relevant solution is thus

$$\sigma(y|s,q_t) = \frac{\xi - \sqrt{4\lambda^2 - 4\lambda\xi[1 - 2p_s(1 - \sigma(y|t,q_s))] + \xi^2} - 2\lambda p_s \sigma(y|t,q_s)}{2\lambda(1-p_s)}. \tag{A.9}$$

Consider the following five cases, which correspond to part 3(a) of the claim:

1. Suppose $\sigma(y|s,q_t) = 0$. For $\sigma(y|t,q_s) \geq 0$, (A.9) reduces to $\sigma(y|t,q_s) = \frac{\sqrt{4\lambda^2 - 4\lambda\xi(1-2p_s) + \xi^2} - \xi}{2p_s\lambda}$. Thus, there exists an equilibrium with $\sigma(y|s,q_s) = \sigma(y|t,q_t) = 1$, $\sigma(y|s,q_t) = 0$ and $\sigma(y|t,q_s) \in (0,1)$ if and only if, for $p_s \in (0,\frac{1}{2}) \cup (\frac{1}{2},1)$, $0 < \frac{\sqrt{4\lambda^2 - 4\lambda\xi(1-2p_s) + \xi^2} - \xi}{2p_s\lambda} < 1$ or equivalently $\frac{1}{2} - \frac{\lambda}{2\xi} < p_s < \frac{\xi}{\lambda} - 1$.

2. Suppose $\sigma(y|t,q_s)=0$. Solution (A.9) reduces to $\sigma(y|s,q_t)=\frac{\xi-\sqrt{4\lambda^2-4\lambda\xi(1-2p_s)+\xi^2}}{2\lambda(1-p_s)}$. Thus, there exists an equilibrium with $\sigma(y|s,q_s)=\sigma(y|t,q_t)=1$, $\sigma(y|s,q_t)\in(0,1)$ and $\sigma(y|t,q_s)=0$ if and only if, for $p_s\in(0,\frac{1}{2})\cup(\frac{1}{2},1)$, $0<\frac{\xi-\sqrt{4\lambda^2-4\lambda\xi(1-2p_s)+\xi^2}}{2\lambda(1-p_s)}<1$. Note that for $0\leq\lambda<\xi$ and $p_s\in(0,1)$, $\frac{\xi-\sqrt{4\lambda^2-4\lambda\xi(1-2p_s)+\xi^2}}{2\lambda(1-p_s)}$ is strictly decreasing in p_s. Thus, we have that $\frac{\xi-\sqrt{4\lambda^2-4\lambda\xi(1-2p_s)+\xi^2}}{2\lambda(1-p_s)}<\frac{\xi-\sqrt{(2\lambda-\xi)^2}}{2\lambda}$. Note that $\frac{\xi-\sqrt{(2\lambda-\xi)^2}}{2\lambda}<1$ for $2\lambda-\xi>0$ and $\frac{\xi-\sqrt{(2\lambda-\xi)^2}}{2\lambda}=1$ for $2\lambda-\xi\leq 0$. Thus, $\frac{\xi-\sqrt{4\lambda^2-4\lambda\xi(1-2p_s)+\xi^2}}{2\lambda(1-p_s)}<1$ is satisfied for all parameter values. The remaining inequality $\frac{\xi-\sqrt{4\lambda^2-4\lambda\xi(1-2p_s)+\xi^2}}{2\lambda(1-p_s)}>0$ reduces to $p_s<\frac{1}{2}-\frac{\lambda}{2\xi}$.

3. Suppose $\sigma(y|t,q_s)=1$. Solution (A.9) reduces to $\sigma(y|s,q_t)=\frac{\xi-\sqrt{(2\lambda-\xi)^2-2\lambda p}}{2\lambda(1-p_s)}$. Note that if $2\lambda-\xi\leq 0$, $\sigma(y|s,q_t)=1$, which is ruled out above. This implies that for $\sigma(y|s,q_t)<1$, we must have $2\lambda-\xi>0$, in which case $\sigma(y|s,q_t)=\frac{1}{1-p_s}\left(\frac{\xi}{\lambda}-1-p_s\right)$. Thus, there exists an equilibrium with $\sigma(y|s,q_s)=\sigma(y|t,q_t)=\sigma(y|t,q_s)=1$ and $\sigma(y|s,q_t)\in(0,1)$ if and only if, for $p_s\in(0,\frac{1}{2})\cup(\frac{1}{2},1)$, $0<\frac{1}{1-p_s}\left(\frac{\xi}{\lambda}-1-p_s\right)<1$ or equivalently $p_s<\frac{\xi}{\lambda}-1<1$.

4. Suppose $\sigma(y|s,q_t)=0$ and $\sigma(y|t,q_s)=1$. Solution (A.9) reduces to $\xi-\sqrt{(2\lambda-\xi)^2}-2p_s\lambda=0$. Note that if $2\lambda-\xi\leq 0$, $p_s=1$, which violates $p_s<1$ for randomized response. This implies that for the stated strategy profile to constitute an equilibrium in the randomized response regime, we must have $2\lambda-\xi>0$, in which case $p_s=\frac{\xi}{\lambda}-1$. Thus, there exists an equilibrium with $\sigma(y|s,q_s)=\sigma(y|t,q_t)=\sigma(y|t,q_s)=1$ and $\sigma(y|s,q_t)=0$ if and only if, for $p_s\in(0,\frac{1}{2})\cup(\frac{1}{2},1)$, $p_s=\frac{\xi}{\lambda}-1$.

5. It can be verified from (A.9) that $\sigma(y|s,q_t)\geq 1$ if and only if $2\lambda-\xi\leq 0$ and $\sigma(y|t,q_s)=1$. Thus, if $\sigma(y|t,q_s)\in(0,1)$, we must have $\sigma(y|s,q_t)<1$. On the other hand, $\sigma(y|s,q_t)>0$ if and only if $\xi-\sqrt{4\lambda^2-4\lambda\xi[1-2p_s(1-\sigma(y|t,q_s))]+\xi^2}-2p_s\lambda\sigma(y|t,q_s)>0$, which can be verified to hold for $\sigma(y|t,q_s)\in(0,1)$ if and only if $p_s<\frac{\xi}{\lambda}-1$. Thus, there exists an equilibrium with $\sigma(y|s,q_s)=\sigma(y|t,q_t)=1$, $\sigma(y|s,q_t)\in(0,1)$ and $\sigma(y|t,q_s)\in(0,1)$ if and only if, for $p_s\in(0,\frac{1}{2})\cup(\frac{1}{2},1)$, $p_s<\frac{\xi}{\lambda}-1$.

Consider next that $\mu_s(y)>\mu_s(n)$. The strategies $\sigma(y|s,q_t)=\sigma(y|t,q_s)=0$ are to be coupled with $\sigma(y|s,q_s)=(0,1]$ and $\sigma(y|t,q_t)\in[0,1]$, accounting for the last five cases. All of them require, from (A.1) and (A.2), that $\lambda=\xi[\mu_s(y)-\mu_s(n)]>0$. Substituting (A.7) and (A.8) into $\lambda=\xi[\mu_s(y)-\mu_s(n)]$ and solving for $\sigma(y|s,q_s)$, we obtain

$$\sigma(y|s,q_s)=\frac{-\xi\pm\sqrt{4\lambda^2-4\lambda\xi[1-2(1-p_s)\sigma(y|t,q_t)]+\xi^2}+2\lambda[1-(1-p_s)\sigma(y|t,q_t)]}{2p_s\lambda}.$$

Note that for $0\leq\lambda<\xi$, $-\sqrt{4\lambda^2-4\lambda\xi[1-2(1-p_s)\sigma(y|t,q_t)]+\xi^2}+2\lambda[1-(1-p_s)\sigma(y|t,q_t)]\leq$

$-\sqrt{(2\lambda-\xi)^2}+2\lambda$. Thus, we have that

$$\frac{-\xi-\sqrt{4\lambda^2-4\lambda\xi[1-2(1-p_s)\sigma(y|t,q_t)]+\xi^2}+2\lambda[1-(1-p_s)\sigma(y|t,q_t)]}{2p_s\lambda}$$
$$\leq \frac{-\xi-\sqrt{(2\lambda-\xi)^2}+2\lambda}{2p_s\lambda}.$$

Given that $\frac{-\xi-\sqrt{(2\lambda-\xi)^2}+2\lambda}{2p_s\lambda}=0$ for $2\lambda-\xi\geq 0$ and $\frac{-\xi-\sqrt{(2\lambda-\xi)^2}+2\lambda}{2p_s\lambda}<0$ for $2\lambda-\xi<0$, the above solution for $\sigma(y|s,q_s)$ is not relevant ($\sigma(y|s,q_s)=0$ is ruled out above). The relevant solution is thus

$$\sigma(y|s,q_s)=\frac{-\xi+\sqrt{4\lambda^2-4\lambda\xi[1-2(1-p_s)\sigma(y|t,q_t)]+\xi^2}+2\lambda[1-(1-p_s)\sigma(y|t,q_t)]}{2p_s\lambda}. \quad \text{(A.10)}$$

Consider the following five cases, which correspond to part 3(b) of the claim:

1. Suppose $\sigma(y|s,q_s)=1$. For $\sigma(y|t,q_t)\leq 1$, we have (A.10) reducing to $\sigma(y|t,q_t)=1+\frac{\xi-\sqrt{4\lambda^2+4\lambda\xi(1-2p_s)+\xi^2}}{2\lambda(1-p_s)}$. Thus, there exists an equilibrium with $\sigma(y|s,q_t)=\sigma(y|t,q_s)=0$, $\sigma(y|s,q_s)=1$ and $\sigma(y|t,q_t)\in(0,1)$ if and only if, for $p_s\in(0,\frac{1}{2})\cup(\frac{1}{2},1)$, $0<1+\frac{\xi-\sqrt{4\lambda^2+4\lambda\xi(1-2p_s)+\xi^2}}{2\lambda(1-p_s)}<1$ or equivalently $2-\frac{\xi}{\lambda}<p<\frac{1}{2}+\frac{\lambda}{2\xi}$.

2. Suppose $\sigma(y|t,q_t)=1$. Solution (A.10) reduces to $\sigma(y|s,q_s)=1-\frac{\xi-\sqrt{4\lambda^2+4\lambda\xi(1-2p_s)+\xi^2}}{2p_s\lambda}$. Thus, there exists an equilibrium with $\sigma(y|s,q_t)=\sigma(y|t,q_s)=0$, $\sigma(y|t,q_t)=1$ and $\sigma(y|s,q_s)\in(0,1)$ if and only if, for $p_s\in(0,\frac{1}{2})\cup(\frac{1}{2},1)$, $0<1-\frac{\xi-\sqrt{4\lambda^2+4\lambda\xi(1-2p_s)+\xi^2}}{2p_s\lambda}<1$. Note that for $0\leq\lambda<\xi$ and $p_s\in(0,1)$, $1-\frac{\xi-\sqrt{4\lambda^2+4\lambda\xi(1-2p_s)+\xi^2}}{2p_s\lambda}$ is strictly decreasing in p_s. Thus, we have that $1-\frac{\xi-\sqrt{4\lambda^2+4\lambda\xi(1-2p_s)+\xi^2}}{2p_s\lambda}>1-\frac{\xi-\sqrt{(2\lambda-\xi)^2}}{2\lambda}$. Note that $1-\frac{\xi-\sqrt{(2\lambda-\xi)^2}}{2\lambda}>0$ for $2\lambda-\xi>0$ and $1-\frac{\xi-\sqrt{(2\lambda-\xi)^2}}{2\lambda}=0$ for $2\lambda-\xi\leq 0$. Thus, $1-\frac{\xi-\sqrt{4\lambda^2+4\lambda\xi(1-2p_s)+\xi^2}}{2p_s\lambda}>0$ is satisfied for all parameter values. The remaining inequality $1-\frac{\xi-\sqrt{4\lambda^2+4\lambda\xi(1-2p_s)+\xi^2}}{2p_s\lambda}<1$ reduces to $p_s>\frac{1}{2}+\frac{\lambda}{2\xi}$.

3. Suppose $\sigma(y|t,q_t)=0$. Solution (A.10) reduces to $\sigma(y|s,q_s)=\frac{2\lambda-\xi+\sqrt{(2\lambda-\xi)^2}}{2p_s\lambda}$. Note that if $2\lambda-\xi\leq 0$, $\sigma(y|s,q_s)=0$, which is ruled out above. This implies that for $\sigma(y|s,q_s)>0$, we must have $2\lambda-\xi>0$, in which case $\sigma(y|s,q_s)=\frac{2\lambda-\xi}{p_s\lambda}$. Thus, there exists an equilibrium with $\sigma(y|s,q_t)=\sigma(y|t,q_s)=0$, $\sigma(y|t,q_t)=0$ and $\sigma(y|s,q_s)\in(0,1)$ if and only if, for $p_s\in(0,\frac{1}{2})\cup(\frac{1}{2},1)$, $0<\frac{2\lambda-\xi}{p\lambda}<1$ or equivalently $p_s>2-\frac{\xi}{\lambda}>0$.

4. Suppose $\sigma(y|s,q_s)=1$ and $\sigma(y|t,q_t)=0$. Solution (A.10) reduces to $2\lambda-\xi+\sqrt{(2\lambda-\xi)^2}-2p\lambda=0$. Note that if $2\lambda-\xi\leq 0$, $p_s=0$, which violates $p_s>0$ for randomized response. This implies that for the stated strategy profile to constitute an equilibrium in the ran-

domized response regime, we must have $2\lambda - \xi > 0$, in which case $p_s = 2 - \frac{\xi}{\lambda}$. Thus, there exists an equilibrium with $\sigma(y|s, q_t) = \sigma(y|t, q_s) = \sigma(y|t, q_t) = 0$ and $\sigma(y|s, q_s) = 1$ if and only if, for $p_s \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$, $p_s = 2 - \frac{\xi}{\lambda}$.

5. It can be verified from (A.10) that $\sigma(y|s, q_s) \leq 0$ if and only if $2\lambda - \xi \leq 0$ and $\sigma(y|t, q_t) = 0$. Thus, if $\sigma(y|t, q_t) \in (0, 1)$, we must have $\sigma(y|s, q_s) > 0$. On the other hand, $\sigma(y|s, q_s) < 1$ if and only if $2\lambda(1-p_s)(1-\sigma(y|t, q_t)) - \xi + \sqrt{4\lambda^2 - 4\lambda\xi[1 - 2(1 - p_s)\sigma(y|t, q_t)] + \xi^2} < 0$, which can be verified to hold for $\sigma(y|t, q_t) \in (0, 1)$ if and only if $p_s > 2 - \frac{\xi}{\lambda}$. Thus, there exists an equilibrium with $\sigma(y|s, q_t) = \sigma(y|t, q_s) = 0$, $\sigma(y|s, q_s) \in (0, 1)$ and $\sigma(y|t, q_t) \in (0, 1)$ if and only if, for $p_s \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$, $p_s > 2 - \frac{\xi}{\lambda}$.

□

Proof of Proposition 3. From items 3 and 4 of Proposition 1, if $\frac{\lambda}{\xi} \in [0, \frac{1}{2}]$, any equilibrium must be uninformative. The interviewer's posterior beliefs are the same as the prior, which implies that $H(\theta|r) = H(\theta) = 1$, and thus $I(\theta; r) = 0$. Note that for the equilibria with common response, the out-of-equilibrium beliefs do not enter into the calculation because for the unused response r', $\Pr(r') = 0$.

From item 2 of Proposition 1, if $\frac{\lambda}{\xi} \in (\frac{1}{2}, 1)$, in the unique equilibrium $\sigma(y|t) = 1$ and $\sigma(y|s) = \frac{\xi}{\lambda} - 1$, and thus $\Pr(y) = \frac{\xi}{2\lambda}$. Bayes' rule implies that $\mu_s(y) = 1 - \frac{\lambda}{\xi}$ and $\mu_s(n) = 1$. Accordingly, $H(\theta|r) = -(\frac{\xi}{2\lambda})[(1 - \frac{\lambda}{\xi}) \log(1 - \frac{\lambda}{\xi}) + \frac{\lambda}{\xi} \log \frac{\lambda}{\xi}]$, where $0 \log 0 = 0$ is used. Thus, for the unique equilibrium under $\frac{\lambda}{\xi} \in (\frac{1}{2}, 1)$, $I(\theta|r) = 1 + \frac{1}{2}[(\frac{\xi}{\lambda} - 1) \log(1 - \frac{\lambda}{\xi}) + \log \frac{\lambda}{\xi}]$. Note finally that while the above argument is made assuming $q = q_t$, the case for $q = q_s$ is symmetric.

□

Proof of Lemma 2. For item 1, note that from Proposition 3, we have that for $\frac{\lambda}{\xi} \in (\frac{1}{2}, 1)$, $\bar{I}(\frac{\lambda}{\xi}) = 1 + \frac{1}{2}[(\frac{\xi}{\lambda} - 1) \log(1 - \frac{\lambda}{\xi}) + \log \frac{\lambda}{\xi}]$, which, with $q = q_t$, is derived from the equilibrium in which $\sigma(y|t) = 1$ and $\sigma(y|s) = \frac{\xi}{\lambda} - 1$. The strategy profile implies the following components for mutual information, $\mu_s(y) = 1 - \frac{\lambda}{\xi}$, $\mu_s(n) = 1$ and $\Pr(y) = \frac{\xi}{2\lambda}$. We first show that there is an equilibrium in the randomized response regime that has the same components. Consider the equilibrium in which $\sigma(y|s, q_s) = \sigma(y|t, q_t) = \sigma(y|t, q_s) = 1$ and $\sigma(y|s, q_t) = 0$, which exists if and only if $p_s = \frac{\xi}{\lambda} - 1$ and $\frac{\lambda}{\xi} > \frac{1}{2}$. It is immediate from (A.5) that $\mu_s(n) = 1$, from (A.6) that $\mu_s(y) = \frac{p_s}{1+p_s} = 1 - \frac{\lambda}{\xi}$, and that $\Pr(y) = \frac{1}{2}(1 + p) = \frac{\xi}{2\lambda}$. We show next that there is another equilibrium in the randomized response regime that has the same components up to rotation of the responses, and thus has the same mutual information. Consider the equilibrium in which $\sigma(y|s, q_t) = \sigma(y|t, q_s) = \sigma(y|t, q_t) = 0$ and $\sigma(y|s, q_s) = 1$, which exists if and only if $p_s = 2 - \frac{\xi}{\lambda}$ and $\frac{\lambda}{\xi} > \frac{1}{2}$. It is immediate from (A.7) that $\mu_s(y) = 1$, from (A.8) that $\mu_s(n) = \frac{1-p_s}{2-p_s} = 1 - \frac{\lambda}{\xi}$, and that $\Pr(n) = \frac{1}{2}(2 - p_s) = \frac{\xi}{2\lambda}$. Thus, for $\frac{\lambda}{\xi} \in (\frac{1}{2}, 1)$ and $p_s \in \{\frac{\xi}{\lambda} - 1, 2 - \frac{\xi}{\lambda}\}$, there exist equilibria in the randomized response regime whose mutual information is $1 + \frac{1}{2}[(\frac{\xi}{\lambda} - 1) \log(1 - \frac{\lambda}{\xi}) + \log \frac{\lambda}{\xi}]$.

For item 2, consider the truthful equilibria in which $\sigma(y|s,q_s)=\sigma(y|t,q_t)=1$ and $\sigma(y|s,q_t)=\sigma(y|t,q_s)=0$, which exist if and only if $p_s\in[\frac{1}{2}-\frac{\lambda}{2\xi},\frac{1}{2}+\frac{\lambda}{2\xi}]$. The strategy profiles imply that $\mu_s(y)=p_s$, $\mu_s(n)=1-p_s$, and $\Pr(y)=\Pr(n)=\frac{1}{2}$. The resulting mutual information is thus $1+p_s\log p_s+(1-p_s)\log(1-p_s)$, which attains its minimum at $p_s=\frac{1}{2}$ and is strictly convex in p_s. This implies that for $p_s\in[\frac{1}{2}-\frac{\lambda}{2\xi},\frac{1}{2}+\frac{\lambda}{2\xi}]$, the mutual information attains maxima when $p_s\in\{\frac{1}{2}-\frac{\lambda}{2\xi},\frac{1}{2}+\frac{\lambda}{2\xi}\}$. Substituting $p_s\in\{\frac{1}{2}-\frac{\lambda}{2\xi},\frac{1}{2}+\frac{\lambda}{2\xi}\}$ into $1+p_s\log p_s+(1-p_s)\log(1-p_s)$, we obtain $\bar{I}_{R-T}(\frac{\lambda}{\xi})=\frac{1}{2}[(1-\frac{\lambda}{\xi})\log(1-\frac{\lambda}{\xi})+(1+\frac{\lambda}{\xi})\log(1+\frac{\lambda}{\xi})]$.

Finally, we compare the two values of mutual information, $\frac{1}{2}[(1-\frac{\lambda}{\xi})\log(1-\frac{\lambda}{\xi})+(1+\frac{\lambda}{\xi})\log(1+\frac{\lambda}{\xi})]$ and $1+\frac{1}{2}[(\frac{\xi}{\lambda}-1)\log(1-\frac{\lambda}{\xi})+\log\frac{\lambda}{\xi}]$. We define, subtracting the latter from the former, $\Delta\bar{I}(\frac{\lambda}{\xi})=\frac{1}{2}[(2-\frac{\lambda}{\xi}-\frac{\xi}{\lambda})\log(1-\frac{\lambda}{\xi})+(1+\frac{\lambda}{\xi})\log(1+\frac{\lambda}{\xi})-\log\frac{\lambda}{\xi}]-1$ for $\frac{\lambda}{\xi}=[\frac{1}{2},1]$, using the fact that the expression is well-defined at the endpoints of the interval $[\frac{1}{2},1]$. Note that $\Delta\bar{I}(\frac{1}{2})=\frac{3}{4}\log 3-1>0$, $\Delta\bar{I}(1)=0$, and $\frac{d\Delta\bar{I}(\frac{\lambda}{\xi})}{d(\frac{\lambda}{\xi})}=\frac{[1-(\frac{\lambda}{\xi})^2]\ln(1-\frac{\lambda}{\xi})+(\frac{\lambda}{\xi})^2\ln(1+\frac{\lambda}{\xi})}{(\frac{\lambda}{\xi})^2\ln 4}>0$ at $\frac{\lambda}{\xi}=1$. Hence, there exists $x\in(0,\frac{1}{2})$ for which $\Delta\bar{I}(x)<0$, and, by the intermediate value theorem, there exists a $c\in(\frac{1}{2},1)$ with $\Delta\bar{I}(c)=0$. Since $\frac{d^2\Delta\bar{I}(\frac{\lambda}{\xi})}{d(\frac{\lambda}{\xi})^2}=-\frac{(\frac{\lambda}{\xi})(1+\frac{2\lambda}{\xi})+2(1+\frac{\lambda}{\xi})\ln(1-\frac{\lambda}{\xi})}{(\frac{\lambda}{\xi})^3(1+\frac{\lambda}{\xi})\ln 4}>0$ for $\frac{\lambda}{\xi}\in[\frac{1}{2},1]$, this c is unique. It can be verified numerically that $c\approx 0.743$.

□

Proof of Proposition 4. We solve a constrained maximization problem, where the objective function is the mutual information and the constraint comes from the restriction of equilibria, i.e., the maximal belief differential that $|\mu_s(y)-\mu_s(n)|\leq\frac{\lambda}{\xi}$ (Lemma 1). Since our objective is to find the maximal mutual information allowed by any equilibria in the randomized response regime, it follows from Lemma 2 that for truthful equilibria we can focus on the cases where $p_s\in\{\frac{1}{2}-\frac{\lambda}{2\xi},\frac{1}{2}+\frac{\lambda}{2\xi}\}$, which imply that $|\mu_s(y)-\mu_s(n)|=\frac{\lambda}{\xi}$; for the other equilibria in which at least two types randomize between responses, the indifference also requires that $|\mu_s(y)-\mu_s(n)|=\frac{\lambda}{\xi}$. Accordingly, for our purpose it is without loss of generality to consider that the constraint binds.

The objective function is

$$\begin{aligned}I(\theta;r)=1&+\left[\frac{\Pr(y|s)+\Pr(y|t)}{2}\right](\mu_s(y)\log\mu_s(y)+[1-\mu_s(y)]\log[1-\mu_s(y)])\\&+\left[\frac{\Pr(n|s)+\Pr(n|t)}{2}\right](\mu_s(n)\log\mu_s(n)+[1-\mu_s(n)]\log[1-\mu_s(n)]).\end{aligned}\tag{A.11}$$

Note that as a function, (A.11) has six variables. We use the fact that these are probabilities to reduce the number of variables. First of all, by Bayes' rule, we have that

$$\mu_s(y)=\frac{\Pr(y|s)}{\Pr(y|s)+\Pr(y|t)}\Leftrightarrow\Pr(y|s)+\Pr(y|t)=\frac{\Pr(y|s)}{\mu_s(y)},\tag{A.12}$$

$$\mu_s(n)=\frac{\Pr(n|s)}{\Pr(n|s)+\Pr(n|t)}\Leftrightarrow\Pr(n|s)+\Pr(n|t)=\frac{\Pr(n|s)}{\mu_s(n)}.\tag{A.13}$$

Substituting (A.12) and (A.13) into (A.11), we obtain

$$\begin{aligned} I(\theta;r) = 1+ &\left[\frac{\Pr(y|s)}{2\mu_s(y)}\right](\mu_s(y)\log\mu_s(y) + [1-\mu_s(y)]\log[1-\mu_s(y)]) \\ &+ \left[\frac{\Pr(n|s)}{2\mu_s(n)}\right](\mu_s(n)\log\mu_s(n) + [1-\mu_s(n)]\log[1-\mu_s(n)])]. \end{aligned} \tag{A.14}$$

We use the fact that $\Pr(n|\cdot) = 1 - \Pr(y|\cdot)$ to further eliminate $\Pr(n|s)$ and $\mu_s(n)$. Note that (A.13) can be rewritten as

$$\mu_s(n) = \frac{1-\Pr(y|s)}{2-[\Pr(y|s)+\Pr(y|t)]} = \frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)}, \tag{A.15}$$

where in the second equality we use (A.12) for $\Pr(y|s)+\Pr(y|t)$. Using (A.15) and the fact that $\frac{\Pr(n|s)}{2\mu_s(n)} = 1 - \frac{\Pr(y|s)}{2\mu_s(y)}$, (A.14) becomes

$$\begin{aligned} I(\theta;r) = 1+ &\left[\frac{\Pr(y|s)}{2\mu_s(y)}\right](\mu_s(y)\log\mu_s(y) + [1-\mu_s(y)]\log[1-\mu_s(y)]) \\ &+ \left[1-\frac{\Pr(y|s)}{2\mu_s(y)}\right]\left[\left(\frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)}\right)\log\left(\frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)}\right)\right. \\ &\left.+ \left(1-\frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)}\right)\log\left(1-\frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)}\right)\right]. \end{aligned} \tag{A.16}$$

Finally, we eliminate $\Pr(y|s)$ by using the belief constraint. Without loss of generality, we consider the case where $\mu_s(n) > \mu_s(y)$ so that the constraint is $\mu_s(n) - \mu_s(y) = \frac{\lambda}{\xi}$. Using (A.15), the constraint becomes

$$\frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)} - \mu_s(y) = \frac{\lambda}{\xi} \Leftrightarrow \Pr(y|s) = \mu_s(y)\left(\frac{\xi}{\lambda}\right)\left(2\left[\frac{\lambda}{\xi}+\mu_s(y)\right]-1\right). \tag{A.17}$$

Substituting (A.17) into (A.16), we obtain the following function in terms of $\mu_s(y)$ only:

$$\begin{aligned} I(\theta;r) =& \hat{I}(\mu_s(y)) \\ =& 1 + \left(1-\frac{\xi}{2\lambda}[1-2\mu_s(y)]\right)(\mu_s(y)\log\mu_s(y) + [1-\mu_s(y)]\log[1-\mu_s(y)]) \\ &+ \left(\frac{\xi}{2\lambda}[1-2\mu_s(y)]\right)\left[\left(\mu_s(y)+\frac{\lambda}{\xi}\right)\log\left(\mu_s(y)+\frac{\lambda}{\xi}\right)\right. \\ &\left.+ \left(1-\mu_s(y)-\frac{\lambda}{\xi}\right)\log\left(1-\mu_s(y)-\frac{\lambda}{\xi}\right)\right]. \end{aligned} \tag{A.18}$$

Note that there are also the box constraints that $\mu_s(y) \in [0,1]$ and $\mu_s(n) \in [0,1]$. And given the belief constraint, these box constraints are satisfied if and only if $\mu_s(y) \in [0, 1-\frac{\lambda}{\xi}]$. Thus,

our maximization problem is

$$\underset{\mu_s(y)\in[0,1-\frac{\lambda}{\xi}]}{\text{Max}}\ \hat{I}(\mu_s(y)).$$

Note that $\hat{I}(\cdot)$ is symmetric at $\frac{1}{2}\big(1-\frac{\lambda}{\xi}\big)$, i.e., $\hat{I}\big(\frac{1}{2}\big(1-\frac{\lambda}{\xi}\big)+x\big)=\hat{I}\big(\frac{1}{2}\big(1-\frac{\lambda}{\xi}\big)-x\big)$.

The first-order condition for an extremum is

$$\left[1-\frac{2\lambda}{\xi}-4\mu_s(y)\right]\ln\left(\frac{\mu_s(y)+\frac{\lambda}{\xi}}{\mu_s(y)}\right)=\left[3-\frac{2\lambda}{\xi}-4\mu_s(y)\right]\ln\left(\frac{1-\mu_s(y)-\frac{\lambda}{\xi}}{1-\mu_s(y)}\right). \tag{A.19}$$

Equation (A.19) is satisfied at the point of symmetry, $\mu_s(y)=\frac{1}{2}\big(1-\frac{\lambda}{\xi}\big)$, which is the mid-point of the range $[0,1-\frac{\lambda}{\xi}]$. The second derivative of $\hat{I}(\mu_s(y))$ is

$$\hat{I}''(\mu_s(y))=\frac{\frac{1-2\mu_s(y)}{2(\mu_s(y)+\frac{\lambda}{\xi})(1-\mu_s(y)-\frac{\lambda}{\xi})}-\frac{1-2[\mu_s(y)+\frac{\lambda}{\xi}]}{2\mu_s(y)[1-\mu_s(y)]}-2\ln\Big(\Big[\frac{1-\mu_s(y)}{\mu_s(y)}\Big]\Big[\frac{\mu_s(y)+\frac{\lambda}{\xi}}{1-\mu_s(y)-\frac{\lambda}{\xi}}\Big]\Big)}{\frac{\lambda}{\xi}\ln 2}.$$

It can be verified that

$$\hat{I}''(\tfrac{1}{2}(1-\tfrac{\lambda}{\xi}))=\frac{4\left[\frac{\frac{\lambda}{\xi}}{(1-\frac{\lambda}{\xi})(1+\frac{\lambda}{\xi})}+\ln\left(\frac{1-\frac{\lambda}{\xi}}{1+\frac{\lambda}{\xi}}\right)\right]}{\frac{\lambda}{\xi}\ln 2}\gtreqless 0\quad\text{for}\quad\frac{\lambda}{\xi}\gtreqless d,$$

where $d\approx 0.796$. Thus, $\mu_s(y)=\frac{1}{2}(1-\frac{\lambda}{\xi})$ corresponds to a local maximum for $\frac{\lambda}{\xi}<d$ and a local minimum for $\frac{\lambda}{\xi}>d$.

We further derive the third derivative:

$$\begin{aligned}\hat{I}'''(\mu_s(y))=&\frac{1}{([\mu_s(y)][1-\mu_s(y)][\mu_s(y)+\frac{\lambda}{\xi}][1-\mu_s(y)-\frac{\lambda}{\xi}])^2\ln 4}\\&\times\left(1-2\mu_s(y)-\frac{\lambda}{\xi}\right)\left[2\left(\frac{\lambda}{\xi}\right)^3[1-2\mu_s(y)]\right.\\&-3\left(\frac{\lambda}{\xi}\right)^2(1-2\mu_s(y)[1-\mu_s(y)])\\&+\left(\frac{\lambda}{\xi}\right)(1-2\mu_s(y)(2-\mu_s(y)[3-2\mu_s(y)])\\&\left.+2\mu_s(y)[1-\mu_s(y)](1-\mu_s(y)[1-\mu_s(y)])\right].\end{aligned} \tag{A.20}$$

We evaluate the values of the third derivative for $\frac{\lambda}{\xi}\in[0,1)$, which in turns allows us to infer the properties of the second derivative and to establish the global maxima of the objective function.

Solving $\hat{I}'''(\mu_s(y)) = 0$ gives three real solutions:

$$\hat{\mu}_s(y) = \frac{1}{2}\left(1 - \frac{\lambda}{\xi} - \sqrt{2\sqrt{4\left(\frac{\lambda}{\xi}\right)^4 - \left(\frac{\lambda}{\xi}\right)^2 + 1} - 3\left(\frac{\lambda}{\xi}\right)^2 - 1}\right), \tag{A.21}$$

$$\bar{\mu}_s(y) = \frac{1}{2}\left(1 - \frac{\lambda}{\xi}\right), \tag{A.22}$$

$$\tilde{\mu}_s(y) = \frac{1}{2}\left(1 - \frac{\lambda}{\xi} + \sqrt{2\sqrt{4\left(\frac{\lambda}{\xi}\right)^4 - \left(\frac{\lambda}{\xi}\right)^2 + 1} - 3\left(\frac{\lambda}{\xi}\right)^2 - 1}\right). \tag{A.23}$$

We first consider $\frac{\lambda}{\xi} \in [0, \frac{1}{2}]$. Note that for $\frac{\lambda}{\xi} \in [0, \frac{1}{2}]$, $\bar{\mu}_s(y) = \frac{1}{2}(1 - \frac{\lambda}{\xi})$ in (A.22) is the only point in $(0, 1 - \frac{\lambda}{\xi})$ at which the third derivative vanishes. Evaluating the expression in (A.20) for $\frac{\lambda}{\xi} \in [0, \frac{1}{2}]$ then gives that $\hat{I}'''(\mu_s(y)) \gtreqless 0$ for $\mu_s(y) \lesseqgtr \frac{1}{2}(1 - \frac{\lambda}{\xi})$. And for $\frac{\lambda}{\xi} \in [0, \frac{1}{2}]$, $\lim_{\mu_s(y)\to 0} \hat{I}''(\mu_s(y)) = \lim_{\mu_s(y)\to(1-\frac{\lambda}{\xi})} \hat{I}''(\mu_s(y)) = -\infty$. Accordingly, with $\frac{1}{2} < d$, for $\frac{\lambda}{\xi} \in [0, \frac{1}{2}]$, $\hat{I}''(\mu_s(y)) \leq \hat{I}''(\frac{1}{2}(1 - \frac{\lambda}{\xi})) < 0$ for all $\mu_s(y) \in [0, 1 - \frac{\lambda}{\xi}]$. Thus, $\hat{I}(\mu_s(y))$ is strictly concave on $[0, 1 - \frac{\lambda}{\xi}]$ for $\frac{\lambda}{\xi} \in [0, \frac{1}{2}]$, and $\mu_s(y) = \frac{1}{2}(1 - \frac{\lambda}{\xi})$ corresponds to a global maximum for $\frac{\lambda}{\xi} \in [0, \frac{1}{2}]$.

We consider next $\frac{\lambda}{\xi} \in [\sqrt{3/7}, 1]$. Note that for $\frac{\lambda}{\xi} \in (\sqrt{3/7}, 1)$, $\bar{\mu}_s(y) = \frac{1}{2}(1 - \frac{\lambda}{\xi})$ in (A.22) is the only point in $[0, 1 - \frac{\lambda}{\xi}]$ at which the third derivative vanishes. And for $\frac{\lambda}{\xi} \in \{\sqrt{3/7}, 1\}$, the three solutions in (A.21)-(A.23) coincide. Evaluating the expression in (A.20) for $\frac{\lambda}{\xi} \in [\sqrt{3/7}, 1]$ then gives that $\hat{I}'''(\mu_s(y)) \lesseqgtr 0$ for $\mu_s(y) \gtreqless \frac{1}{2}(1 - \frac{\lambda}{\xi})$. Accordingly, with $\sqrt{3/7} < d$, for $\frac{\lambda}{\xi} \in (d, 1]$, $\hat{I}''(\mu_s(y)) \geq \hat{I}''(\frac{1}{2}(1 - \frac{\lambda}{\xi})) > 0$ for all $\mu_s(y) \in [0, 1 - \frac{\lambda}{\xi}]$. Thus, $\hat{I}(\mu_s(y))$ is strictly convex on $[0, 1 - \frac{\lambda}{\xi}]$ for $\frac{\lambda}{\xi} \in (d, 1]$, and the global maxima lie at, given the symmetry at $\frac{1}{2}(1 - \frac{\lambda}{\xi})$, the two boundaries, $\mu_s(y) = 0$ or $\mu_s(y) = 1 - \frac{\lambda}{\xi}$.

We further divide the remaining case $\frac{\lambda}{\xi} \in (\frac{1}{2}, d]$ into two sub-cases, when $\frac{\lambda}{\xi} \in (\frac{1}{2}, \sqrt{3/7})$ and when $\frac{\lambda}{\xi} \in [\sqrt{3/7}, d]$. We consider the latter case first. It follows from the above that for $\frac{\lambda}{\xi} \in [\sqrt{3/7}, d]$, we have that $\hat{I}''(\mu_s(y)) \geq \hat{I}''(\frac{1}{2}(1 - \frac{\lambda}{\xi}))$ for all $\mu_s(y) \in [0, 1 - \frac{\lambda}{\xi}]$. Given the symmetry of $\hat{I}(\mu_s(y))$, we without loss of generality focus on its behavior for $\mu_s(y) \in [0, \frac{1}{2}(1 - \frac{\lambda}{\xi})]$. Note that for $\frac{\lambda}{\xi} \in [\sqrt{3/7}, d]$, $\lim_{\mu_s(y)\to 0} \hat{I}''(\mu_s(y)) = \infty$ and recall that for $\frac{\lambda}{\xi} \leq d$, $\hat{I}''(\frac{1}{2}(1 - \frac{\lambda}{\xi})) \leq 0$. Given that for $\frac{\lambda}{\xi} \in [\sqrt{3/7}, 1]$, $\hat{I}'''(\mu_s(y)) < 0$ for $\mu_s(y) < \frac{1}{2}(1 - \frac{\lambda}{\xi})$, there exists a unique $k \in (0, \frac{1}{2}(1 - \frac{\lambda}{\xi})]$ such that $\hat{I}''(k) = 0$. This further implies that there is at most one point in $(0, \frac{1}{2}(1 - \frac{\lambda}{\xi}))$ such that the first-order condition is satisfied, in which case it corresponds to a local minimum; $\mu_s(y) = \frac{1}{2}(1 - \frac{\lambda}{\xi})$ thus corresponds to a unique local maximum. Given that, for $\frac{\lambda}{\xi} \in [\sqrt{3/7}, d]$, $\hat{I}(\mu_s(y))$ is strictly convex for $\mu_s(y)$ sufficiently close to zero and concave (strictly concave for $\frac{\lambda}{\xi} < d$) in the neighborhood of $\frac{1}{2}(1 - \frac{\lambda}{\xi})$, the global maximum is achieved either at the unique local maximum at $\mu_s(y) = \frac{1}{2}(1 - \frac{\lambda}{\xi})$ or at the boundary $\mu_s(y) = 0$ or, by symmetry, $\mu_s(y) = 1 - \frac{\lambda}{\xi}$.

Finally, we consider $\frac{\lambda}{\xi} \in (\frac{1}{2}, \sqrt{3/7})$. Note that for $\frac{\lambda}{\xi} \in (\frac{1}{2}, \sqrt{3/7})$, the solution in (A.21)

satisfies that $\hat{\mu}_s(y) \in (0, \frac{1}{2}(1-\frac{\lambda}{\xi}))$ and the solution in (A.23) satisfies that $\tilde{\mu}_s(y) \in (\frac{1}{2}(1-\frac{\lambda}{\xi}), 1-\frac{\lambda}{\xi})$. Similar to the above paragraph, the following argument focuses on $\mu_s(y) \in [0, \frac{1}{2}(1-\frac{\lambda}{\xi})]$ under the symmetry. Evaluating the expression in (A.20) gives that $\hat{I}'''(\mu_s(y)) \leq 0$ for $\mu_s(y) \leq \hat{\mu}_s(y)$ and $\hat{I}'''(\mu_s(y)) > 0$ for $\mu_s(y) \in (\hat{\mu}_s(y), \frac{1}{2}(1-\frac{\lambda}{\xi}))$. Recall that for $\frac{\lambda}{\xi}$ in this range, we have that $\hat{I}''(\frac{1}{2}(1-\frac{\lambda}{\xi})) < 0$. Then, the fact that $\hat{I}'''(\mu_s(y)) > 0$ for $\mu_s(y) \in (\hat{\mu}_s(y), \frac{1}{2}(1-\frac{\lambda}{\xi}))$ implies that $\hat{I}''(\mu_s(y)) < 0$ for $\mu_s(y) \in (\hat{\mu}_s(y), \frac{1}{2}(1-\frac{\lambda}{\xi}))$. Note that for $\frac{\lambda}{\xi} \in (\frac{1}{2}, \sqrt{3/7})$, $\lim_{\mu_s(y)\to 0} \hat{I}''(\mu_s(y)) = \infty$. Thus, given that $\hat{I}'''(\mu_s(y)) \leq 0$ for $\mu_s(y) \leq \hat{\mu}_s(y)$, there exists a unique $v \in (0, \hat{\mu}_s(y)]$ such that $\hat{I}''(v) = 0$. The argument from the above paragraph then applies to establish that the global maximum is again achieved either at the unique local maximum at $\mu_s(y) = \frac{1}{2}(1-\frac{\lambda}{\xi})$ or at the boundary $\mu_s(y) = 0$ or, by symmetry, $\mu_s(y) = 1-\frac{\lambda}{\xi}$.

Substituting $\mu_s(y) = \frac{1}{2}(1-\frac{\lambda}{\xi})$ into (A.18), we obtain $\frac{1}{2}[(1-\frac{\lambda}{\xi})\log(1-\frac{\lambda}{\xi}) + (1+\frac{\lambda}{\xi})\log(1+\frac{\lambda}{\xi})]$, which is precisely the mutual information of the truthful equilibrium; substituting $\mu_s(y) = 0$ or $\mu_s(y) = 1-\frac{\lambda}{\xi}$ into (A.18) and using $0\log 0 = 0$, we obtain $1 + \frac{1}{2}[(\frac{\xi}{\lambda}-1)\log(1-\frac{\lambda}{\xi}) + \log\frac{\lambda}{\xi}]$, which is precisely the mutual information of the informative equilibrium in the deterministic response, which can be replicated in the randomized response. The result follows from the fact that $c < d$, where $c \approx 0.743$ is the critical value in Lemma 2.

□

Appendix B – Experimental Instructions

B.1 Instructions (*RandomLow*)

INSTRUCTION

Welcome to the experiment. This experiment studies decision making between two individuals. In the following two hours or less, you will participate in 40 rounds of decision making. Please read the instructions below carefully; the cash payment you will receive at the end of the experiment depends on how well you make your decisions according to these instructions.

Your Role and Decision Group

Half of the participants will be randomly assigned the role of Member A and the other half the role of Member B. Your role will remain fixed throughout the experiment. In each round, one Member A will be paired with one Member B to form a group of two. The two members in a group make decisions that will affect their rewards in the round. Participants will be randomly rematched after each round to form new groups.

Your Decision in Each Round

In each round and for each group, the computer will randomly select, with equal chance, either SQUARE or TRIANGLE. The selected shape will be revealed to Member A. Independently, the computer will also randomly select one of the following two questions for Member A: "Was SQUARE selected?" or "Was TRIANGLE selected?" The chance that "WAS SQUARE selected?" will be drawn is 40%, and the chance that "Was TRIANGLE selected?" will be drawn is 60%. Note that the two pieces of information—which shape and which question are selected—is only known to Member A; Member B is not provided with such information.

Member A's Decision

At the beginning of each round, the selected shape and question will be shown on your screen. You respond to the selected question by clicking either "Yes" or "No", and your decision in the round is completed. You are free to choose your response; it is not part of the instructions that you have to respond to indicate the actual shape selected.

Once you click the button, your response will be shown on the screen of the Member B that you are paired with in the round. Be reminded again that he/she will only see your "Yes"/"No" response and will not know which question you are responding to nor which shape was selected.

Member B's Decision

Based on the "Yes"/"No" response of Member A, you will be asked to predict the shape that was selected by the computer. You state your prediction in percentage terms, similar to how rain forecasts are typically reported, i.e., there is an $X\%$ chance of rain (so with $(100 - X)\%$ chance there will be no rain). You will be rewarded according to the accuracy of your prediction.

In each round, you will be presented with a Yellow Box that contains 100 shapes. You will be asked to decide how many shapes are SQUARES and how many are TRIANGLES. The numbers of SQUARES and TRIANGLES in the Yellow Box represent your prediction. For example, if the number of SQUARES is 70 (so the number of TRIANGLES is 30), it means that you predict that there is a 70% (30%) chance that the computer has selected SQUARE (TRIANGLE). You input your prediction by clicking on a line with a green ball on it that lies inside the Yellow Box. The left end of the line represents 0 SQUARES and 100 TRIANGLES; the right end represents 100 SQUARES and 0 TRIANGLES. You can choose any integer point in between. When you click on the line, the green ball will move to the point you click on, and the corresponding numbers of SQUARES and TRIANGLES will be shown inside $\square$ and $\triangle$ in the Yellow Box.

You adjust your click until you arrive at your desired numbers, after which you click the submit button. Your decision in the round is then completed. (You still have to perform some manual task to have your reward in the round determined. More information will provided below.)

Your Reward in Each Round

Your reward in the experiment will be expressed in terms of experimental currency unit (ECU). The following describes how your reward in each round is determined.

Member A's Reward

The amount of ECU you earn in a round depends on two factors. The first is whether your "Yes"/"No" response to the selected question indicates which shape was actually selected by the computer. If it does, you will receive 300 ECU; if it does not, you will receive 250 ECU.

The second factor is Member B's prediction of the chance that SQUARE was selected. The amount of ECU you earn from responding to the question (either 300 or 250) will be reduced by twice the number of SQUARES in Member B's Yellow Box.

Here is an example of two different scenarios in which your earnings will both be 160 ECU:

1. The computer selected SQUARE and "Was TRIANGLE selected?" You responded "No". Since your response indicates which shape was actually selected, you receive 300 ECU for the first part. If Member B predicts a 70% chance of SQUARE by having 70 SQUARES in the Yellow Box, your earning in the round will be $300 - (2 \times 70) = 160$ ECU.

2. The computer selected TRIANGLE and "Was SQUARE selected?" You responded "Yes". Since your response does not indicate which shape was actually selected, you receive 250 ECU for the first part. If Member B predicts a 45% chance of SQUARE by having 45 SQUARES in the Yellow Box, your earning in the round will be $250-(2\times 45)=160$ ECU.

Member B's Reward

The amount of ECU you earn in a round, either 300 ECU or 50 ECU, is determined by the procedure described below. The reward procedure provides incentives to you to state your prediction according to what you truly believe is the chance that SQUARE/TRIANGLE was selected: your earning in expected terms will be highest if you state your true belief.

You will be presented with another box, a Green Box, that helps determine your earning. The Green Box also contains 100 shapes. At the beginning of each round, a number is randomly drawn with equal chance from 1 to 100 to determine the number of SQUARES in the Green Box (100 minus the number drawn is the number of TRIANGLES). Since this happens at the beginning of the round, it is not influenced by any decision made during the round. It is also independent of the shape and question that are selected for Member A. The numbers of SQUARES and TRIANGLES in the Green Box will be revealed to you only after you submit the numbers for the Yellow Box. Your earning in the round will be determined as follows:

1. If the number of SQUARES in the Yellow Box is larger than or equal to the numbers of SQUARE in the Green Box, your earning will depend on which shape was selected and revealed to Member A at the beginning of the round:
   (a) If it was SQUARE, you will receive 300 ECU.
   (b) If it was TRIANGLE, you will receive 50 ECU.
2. If the number of SQUARES in the Yellow Box is smaller than the numbers of SQUARE in the Green Box, you will randomly draw a shape from the Green Box:
   (a) If the randomly drawn shape is a SQUARE, you will receive 300 ECU.
   (b) If the randomly drawn shape is a TRIANGLE, you will receive 50 ECU.

Information Feedback

At the end of each round, the computer will provide a summary for the round: which shape and question were selected and revealed to Member A, Member A's response, the number of SQUARES in Member B's Yellow Box, and your earning in ECU.

Your Cash Payment

The experimenter randomly selects 3 rounds out of 40 to calculate your cash payment. (So it is in your best interest to take each round seriously.) Your total cash payment at the end of the experiment will be the average amount of ECU you earned in the 3 selected rounds divided by 10 (i.e., 10 ECU = 1 USD) plus a $5 show-up fee.

Quiz and Practice

To ensure your understanding of the instructions, we will provide you with a quiz and practice rounds. We will go through the quiz after you answer it on your own. You will then participate in 6 practice rounds, where you will have a chance to play both Member A (3 rounds) and Member B (3 rounds). The practice rounds are part of the instructions which are not relevant to your cash payment; its objective is to get you familiar with the computer interface and the flow of the decisions in each round.

Once the practice rounds are over, the computer will tell you "The official rounds begin now!" You will be randomly assigned the role of either Member A or Member B, which will not change during the 40 official rounds.

Adminstration

Your decisions as well as your monetary payment will be kept confidential. Remember that you have to make your decisions entirely on your own; please do not discuss your decisions with any other participants.

Upon finishing the experiment, you will receive your cash payment. You will be asked to sign your name to acknowledge your receipt of the payment (which will not be used for tax purposes). You are then free to leave.

If you have any question, please raise your hand now. We will answer your question individually. If there is no question, we will proceed to the quiz.

B.2 z-Tree Screen Shots

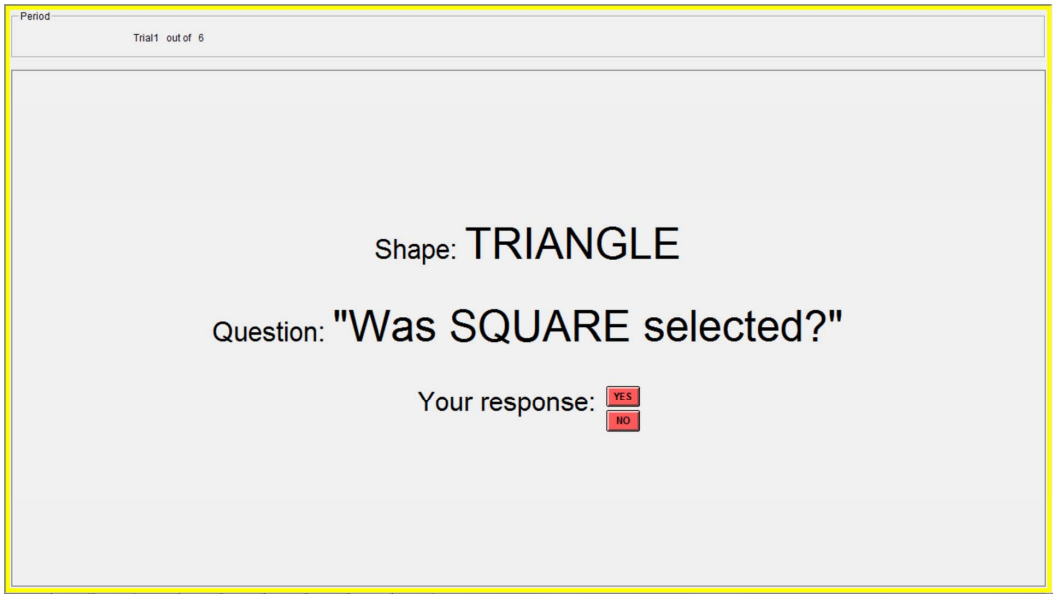


Figure 5: Member A's Response Screen

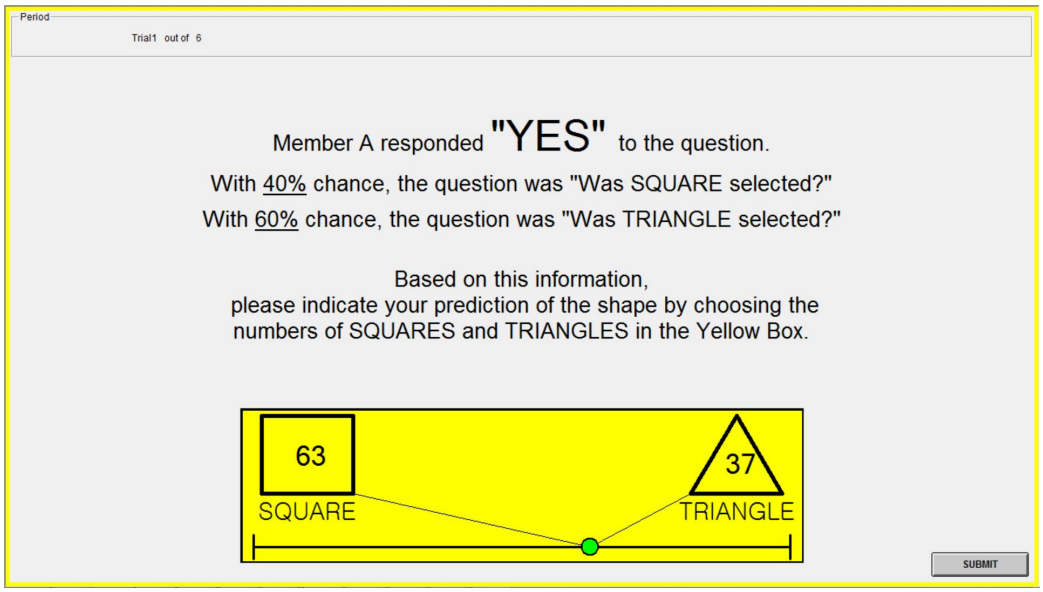


Figure 6: Member B's Prediction Screen

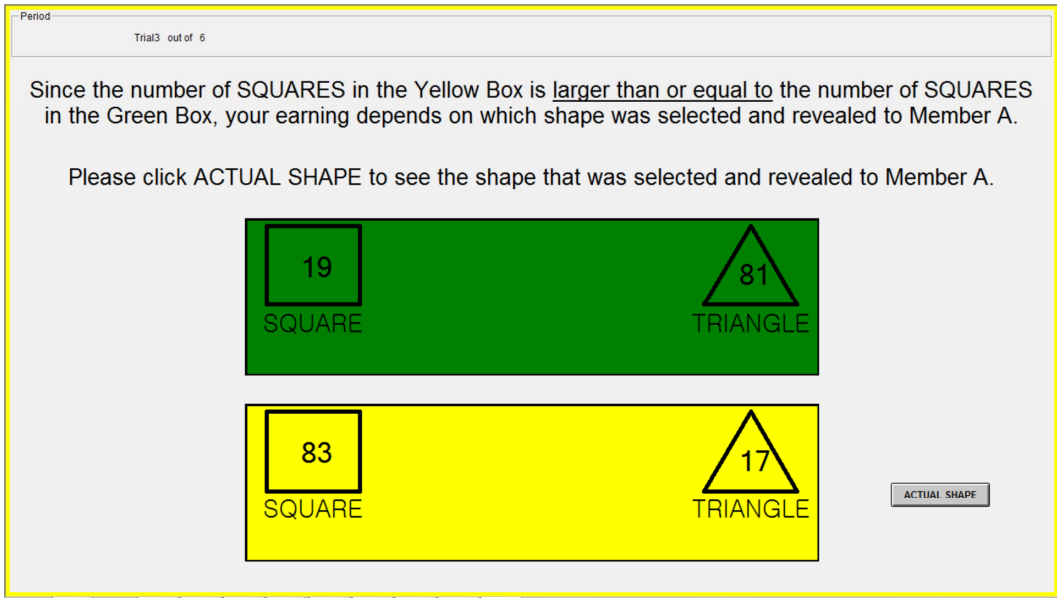


Figure 7: Member B's Reward Screen

References

[1] Banks, Jeffrey S., and Joel Sobel. 1987. "Equilibrium Selection in Signaling Games." *Econometrica*, 55: 647-661.

[2] Battigalli, Pierpaolo, and Martin Dufwenberg. 2009. "Dynamic Psychological Games." *Journal of Economic Theory*, 144: 1-35.

[3] Beldt, Sandra F., Wayne W. Daniel, and Bikramjit S. Garcha. 1982. "The Takahasi-Sakasegawa Randomized Response Technique: A Field Test." *Sociological Methods & Research*, 11: 101-111.

[4] Berg, J. E., L. Daley, J. Dickhaut, and J. O'Brien. 1986. "Controlling Preferences for Lotteries on Units of Experimental Exchange." *Quarterly Journal of Economics*, 101: 281-306.

[5] Bernheim, B. Douglas. 1994. "A Theory of Conformity." *Journal of Political Economy*, 102: 841-877.

[6] Blume, Andreas. 2012. "A Class of Strategy-Correlated Equilibria in Sender-Receiver Games." *Games and Economic Behavior*, 75: 510-517

[7] Blume, Andreas, Douglas V. Dejong, Yong-Gwan Kim, and Geoffrey B. Sprinkle. 1998. "Experimental Evidence on the Evolution of the Meaning of Messages in Sender-Recevier Games." *American Economic Review*, 88: 1323-1340.

[8] Blume, Andreas, Douglas V. Dejong, Yong-Gwan Kim, and Geoffrey B. Sprinkle. 2001. "Evolution of Communication with Partial Common Interest." *Games and Economic Behavior*, 37: 79-120.

[9] Blume, Andreas, Oliver J. Board, and Kohei Kawamura. 2007. "Noisy Talk." *Theoretical Economics*, 2: 395-440.

[10] Boruch, Robert F. 1972. "Strategies for Eliciting and Merging Confidential Social Research Data." *Policy Sciences*, 3: 275-297.

[11] Buchman, Thomas A., and John A. Tracy. 1982. "Obtaining Responses to Sensitive Questions: Conventional Questionnaire versus Randomized Response Technique." *Journal of Accounting Research*, 20: 263-271.

[12] Cai, Hongbin, and Joseph Tao-Yi Wang. 2006. "Overcommunication in Strategic Information Transmission Games." *Games and Economic Behavior*, 56: 7-36.

[13] Chen, Ying. 2011. "Perturbed Communication Games with Honest Senders and Naive Receivers." *Journal of Economic Theory*, 146: 401-424.

[14] Cho, In-Koo, and David Kreps. 1987. "Signaling Games and Stable Equilibria." *Quarterly Journal of Economics*, 102: 179-221.

[15] Cover, Thomas M., and Joy A. Thomas. 1991. *Elements of Information Theory*. John Wiley and Sons: New York, NY.

[16] Dessein, Wouter, Andrea Galeotti, and Tano Santos. 2013. "Rational Inattention and Organizational Focus." Columbia University Working Paper.

[17] Donaldson-Matasci, Matina C., Carl T. Bergstrom, and Michael Lachmann. 2010. "The Fitness Value of Information." *Oikos*. 119: 219-230.

[18] Elffers, Henk, Peter van der Heijden, and Merlijn Hezemans. 2003. "Explaining Regulatory Non-Compliance: A Survey Study of Rule Transgression for Two Dutch Instrumental Laws, Applying the Randomized Response Method." *Journal of Quantitative Criminology*, 19: 409-439.

[19] Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics*, 10, 171-178.

[20] Forges, Francoise. 1986. "An Approach to Communication Equilibria." *Econometrica*, 54: 1375-1385.

[21] Forsythe, Robert, Russell Lundholm, and Thomas Rietz. 1999. "Cheap Talk, Fraud and Adverse Selection in Financial Markets: Some Experimental Evidence." *Review of Financial Studies*, 12: 481-518.

[22] Goltsman, Maria, Johannes Hörner, Gregory Pavlov, and Francesco Squintani. 2009. "Mediation, Arbitration and Negotiation." *Journal of Economic Theory*, 144: 1397-1420.

[23] Geanakoplos, John, David Pearce, and Ennio Stacchetti. 1989. "Psychological Games and Sequential Rationality." *Games and Economic Behavior*, 1: 60-79.

[24] Gneezy, Uri. 2005. "Deception: The Role of Consequences." *American Economic Review*, 95: 384-394.

[25] Hao, Li, and Daniel Houser. 2012. "Belief Elicitation in the Presence of Naive Participants: An Experimental Study." *Journal of Risk and Uncertainty*, 44: 161-180.

[26] Hossain, Tanjim and Okui Ryo. Forthcoming. "The Binarized Scoring Rule of Belief Elicitation." *Review of Economic Studies*.

[27] Houston, Jodie, and Alfred Tran. 2001. "A Survey of Tax Evasion using the Randomized Response Technique." *Advances in taxation*, 13: 69-94.

[28] Ivanov, Maxim. 2010. "Communication via a Strategic Mediator." *Journal of Economic Theory*, 145: 869-884.

[29] John Leslie K., George Loewenstein, Alessandro Acquisti and Joachim Vosgerau. 2013. "Paradoxical Effects of Randomized Response Techniques." Carnegie Mellon University Working Paper.

[30] Jose, Victor Richmond R., Robert F. Nau, and Robert L. Winkler. 2008. "Scoring Rules, Generalized Entropy, and Utility Maximization." *Operations Research*, 56: 1146-1157.

[31] Karni, Edi. 2009. "A Mechanism for Eliciting Probabilities." *Econometrica*, 77: 603-606.

[32] Kartik, Navin. 2009. "Strategic Communication with Lying Costs." *Review of Economic Studies*, 76: 1359-1395.

[33] Kartik, Navin, Marco Ottaviani, and Francesco Squintani. 2007. "Credulity, Lies, and Costly Talk." *Journal of Economic Theory*, 134: 93-116.

[34] Kelly, J. L. 1956. "A New Interpretation of Information Rate." *Bell System Tech. J.* 35: 917-926.

[35] Krishna, Vijay and John Morgan, 2004. "The Art of Conversation: Eliciting Information from Experts through Multi-Stage Mommunication." *Journal of Economic Theory*, 117: 147-179.

[36] Lensvelt-Mulders, Gerty J. L. M., Joop J. Hox, Peter G. M. van der Heijden, Cora J. M. Maas 2005. "Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation." *Sociological Methods Research*, 33: 319-348.

[37] Ljungqvist, Lars. 1993. "A Unified Approach to Measures of Privacy in Randomized Response Models: A Utilitarian Perspective." *Journal of the American Statistical Association*, 88: 97-103.

[38] Myerson, Roger B. 1986. "Multistage Games with Communication." *Econometrica*, 54: 323-358.

[39] Myerson, Roger B. 1991. Game Theory: Analysis of Conflict. Harvard University Press, Cambridge, Massachusetts.

[40] Offerman, Theo, Joep Sonnemans, Gijs van de Kuilen and Peter P. Wakker. 2009. "A Truth-Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes." *Review of Economic Studies*, 76: 1461-1489.

[41] Ottaviani, Marco, and Peter Norman Sørensen. 2006. "Reputational Cheap Talk." *Rand Journal of Economics*, 37: 155-175.

[42] Poole, W. Kenneth. 1974. "Estimation of the Distribution of a Continuous Type Random Variable Through Randomized Response." *Journal of the American Statistical Association*, 69: 1002-1005.

[43] Roth, Alvin, and M. Malouf. 1979. "Game-Theoretic Models and the Role of Bargaining." *Psychological Review*, 86: 574-594.

[44] Savage, Leonard J. 1971. "Elicitation of Personal Probabilities and Expectations." *Journal of American Statistical Association*, 66: 783-801.

[45] Sánchez-Pagés, Santiago, and Marc Vorsatz. 2007. "An Experimental Study of truthful-responding in Sender-Receiver Game." *Games and Economic Behavior*, 61: 86-112.

[46] Schlag, Karl H. and Joël van der Weele. 2009. "Efficient Interval Scoring Rules." Working Paper, Universitat Pompeu Fabra.

[47] Shannon, Claude. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal*, 27: 379-423, 623-656.

[48] Sims, Christopher A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics*, 50: 665-690.

[49] St John, Freya A. V. , Aidan M. Keane, Gareth Edwards-Jones, Lauren Jones, Richard W. Yarnell, and Julia P. G. Jones. 2011. "Identifying Indicators of Illegal Behaviour: Carnivore Killing in Human-Managed landscapes." *Proceedings of Royal Society Biological Science.*

[50] Striegel, Heiko, Rolf Ulrich, and Perikles Simon. 2010. "Randomized Response Estimates for Doping and Illicit Drug Use in Elite Athletes." *Drug and Alcohol Dependence*, 106: 230-232.

[51] Warner, Stanley L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association*, 60: 63-69.

[52] Warner, Stanley L. 1971. "The Linear Randomized Response Model." *Journal of the American Statistical Association*, 66: 884-888.

[53] Wimbush, Dan C. and Donald R. Dalton. 1997. "Base Rate for Employee Theft: Convergence of Multiple Methods." *Journal of Applied Psychology*, 82: 756-63.