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

Andreas Blume† Ernest K. Lai‡ Wooyoung Lim§

January 21, 2015

Abstract

This paper explores plausible deniability theoretically and experimentally in a communication game motivated by Warner's (1965) randomized response technique (RRT). It thus links game-theoretic approaches to noisy communication with survey practice in the field and a novel implementation in the lab. Consistent with equilibria of our game and in line with Warner, the frequency of truthful responses in the lab is significantly higher with randomization than without. Contrary to the intended use of RRT, however, there are equilibria that translate into lower and even invalid (negative) population estimates, and these are supported by both prior and our own experimental findings.

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

JEL classification: C72; C92; D82; D83

*We are grateful to Yeon-Koon Che, Navin Kartik, Kohei Kawamura, Sangmok Lee, Shih En Lu, John Morgan, and Joel Sobel for valuable comments and suggestions. For helpful comments and discussions, we thank seminar participants at Chinese University of HK, City University of HK, Lehigh University (Economics and Psychology Departments), HKUST, IUPUI, Korea University, Manhattan College, Nanyang Technological University, National Taiwan University, National University of Singapore, Özyeğin University, Rutgers University, Seoul National University, Shanghai University of Finance and Economics, Sogang University, Sungkyunkwan University, University of Arizona, and Yeungnam 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, the 1st Haverford Meeting on Behavioral and Experimental Economics, the 24th International Conference on Game Theory, 2013 North-American ESA Conference at Santa Cruz, 2014 China Meeting of Econometric Society, the USC Experimental Economics Conference 2014 and the European Summer Symposium in Economic Theory 2014 at Gerzensee. 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

We explore the role of plausible deniability both theoretically and experimentally in a communication game motivated by Warner's (1965) randomized response technique.

Plausible deniability may be used to deflect responsibility[1], distort information[2], and can reduce incentives for pro-social behavior.[3] On the positive side, it may provide protection for whistleblowers (Chassang and Padró i Miquel, 2013) and make it possible to communicate useful first-order information while withholding destructive higher-order information (Ayres and Nalebuff, 1996). From a design perspective, a natural question then is how to balance the benefits of plausible deniability, e.g., privacy protection, with its costs, e.g., from contamination of information channels.

The literature has studied a range of devices that can generate plausible deniability, such as restrictions on *ex post* information about behavior (Tadelis, 2011; Dana et al, 2006), indirect or ambiguous language (Ayres and Nalebuff, 1996; Pinker, Nowak and Lee, 2008; Mialon and Mialon, 2013) and commitment to random intervention (Chassang and Padró i Miquel, 2013). 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 stochastic continuations (Krishna and Morgan, 2004) in variants of the Crawford-Sobel (1982) model, can also be viewed in this vein: injecting randomness into communication environments 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

[1] An example is Admiral John Poindexter's assumption of responsibility for the diversion of some of the proceeds from arms sales to Iran to support the Contras in Nicaragua and his withholding of documents from President Reagan to provide him with deniability (Bogen and Lynch, 1989).

[2] Calomiris (2009) notes that in the case of novel financial instruments the lack of a track record is a source of deniability for ratings agencies with an interest in ratings inflation.

[3] Tadelis (2011) reports data from a trust-game experiment in which trustees returned less when their decisions could not be distinguished from random events. Dana et al. (2006) find that experimental subjects frequently are willing to avoid playing a \$10 dictator game in favor of a \$9 exit option, when using the exit option ensures that the (potential) receiving player does not learn that otherwise a dictator game would have been played.

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 whose interests are not perfectly aligned with listeners to provide some but not all of their information. In short, randomness has the potential to encourage information transmission by providing plausible deniability.

This potential of randomness providing plausible deniability 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.[4]

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

[4]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 (1972). 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. In a related technique called the item count technique (also known as the list response technique) proposed by Miller (1984), respondents are asked to report how many of $N + 1$ items are true when among them only one item is sensitive (see, e.g., Karlan and Zinman (2012) with an application to measuring the use of micro finance loan proceeds and Coffman, Coffman, and Ericson (2013) on LGBT populations).

is some faith in the theoretical prediction that introducing randomness may aid communication.

Is this faith justified? The use of RRT is predicated 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 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 primarily 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.

Recently, and independently, John, Loewenstein, Acquisti and Vosgerau (2013) have conducted laboratory experiments to evaluate RRT. They are motivated by empirical findings according to which there is frequent non-adherence to RRT instructions, direct questioning often yields more valid responses than RRT and sometimes RRT generates invalid prevalence estimates, with negative frequencies or frequencies above 100% of the population. They suggest that those "paradoxical findings" might be rooted in either unclear instructions or "protective behavior" of participants who worry that innocuous responses are viewed as admissions. Consistent with those rationales they find that non-adherence to RRT instructions can be alleviated and estimates improved by framing jeopardizing

responses as socially desirable and by clarifying that jeopardizing responses do not amount to admissions.

We also provide experimental evidence for systematic non-adherence to RRT instructions. Consistent with the prior evidence cited by John et al. (2013), in our data RRT systemically underestimates the population proportion of the stigmatizing trait, and when stigmatization aversion is high we get negative and thus invalid estimates. In line with the findings of John et al. (2013) we experimentally demonstrate that non-adherence can be explained with protective behavior and that protective behavior becomes more pronounced with increasing concerns about stigmatization relative to truth-telling incentives.

In addition we provide a game-theoretic framework that accounts for these behaviors: the key is that even when adhering to RRT instructions is incentive compatible, i.e., there is an equilibrium where responses are truthful, there is a host of other equilibria with only partially truthful responses. There are two competing focal principles, truth and privacy, which makes it problematic to use focalness of truth for equilibrium selection. As a result, even under ideal conditions the rationale for RRT, which implicitly appeals to the focalness of truth, competes with an equally coherent narrative, in which privacy is focal. This may encourage respondents to engage in protective behaviors that avoid jeopardizing answers. If so, evaluations of RRT responses that ignore that agents will strategize in this manner will be misleading.

Ljungqvist (1993) has treated RRT from the perspective of utility maximizing agents and investigated the conditions for truth-telling to be incentive compatible.[5] Following in his footsteps, in this paper, we use RRT as a vehicle for illustrating the potential benefits and limitation of randomness for 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 men-

[5]Kawamura (2013) studies information transmission in social surveys where a welfare maximizing decision maker communicates with a random sample of individuals who have heterogeneous preferences.

tioned 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 permits us to compare RRT equilibria with equilibria of the game in which questions are known, which we refer to as the direct response technique (DRT). It brings into focus the possibility of multiple equilibria, especially of equilibria that are informative without being truthful and might lead to misleading inferences from RRT.

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. The key insight from the formal model is that adhering to RRT instructions is only one of many equilibria. There are other equally coherent narratives, in which responses are informative but not truthful. Those narratives are supported by equilibria in which respondents protect themselves by failing to give jeopardizing answers. Importantly, the equilibrium analysis supports the frequent empirical finding that RRT can deliver inaccurately low and sometimes invalid (negative) estimates of the proportion of stigmatized traits in a population.

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 – there is *over-communication* with DRT; 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)—there is *under-communication* with RRT; and, non-stigmatized types are less truthful under RRT than under DRT. The key experimental result is that while RRT improves truth-telling relative to DRT, it may result in lower and possibly invalid estimates of the proportion of the population with the stigmatized trait.

These distorted and possibly invalid population inferences are due to departures from truth-telling in line with the informative but not truthful equilibria of our game theoretic model. In these equilibria as in our experimental data, respondents answer truthfully when the answer is innocuous and lie some of the time when the answer is jeopardizing.

The paper is organized as follows. Section 2 sets up and motivates our model. In Section 3 we fully characterize the equilibrium sets of both the RRT and the DRT versions of our model. Section 4 lays out our experimental design and formulates hypotheses based on our characterization of equilibrium sets in the theoretical model. In Section 5 we report our experimental findings. We conclude in section 6 with an assessment of findings and suggestions for future research.

2 A Model of Survey Response

Our theoretical analysis examines how equilibrium behavior restricts information transmission under different survey techniques. We consider a simple model with two players and two types, where there is an interviewer and a respondent. While in practice a survey typically involves many respondents, our modeling choice using a two-player game allows us to focus in a stark setting on the most important issue surrounding the use of RRT—the incentive to respond informatively.

The respondent privately observes his type $\theta \in \{s, t\}$, where s is designated as the *stigmatized* type and t the *regular* type. Players have a common prior that the two types are equally likely.[6] The assumption that the prior is uniform is only

[6]Given that the raison d'être for RRT is to improve the estimation of an unknown population parameter, the standard common prior assumption that we make in our model, that respondents draw their type from a commonly known distribution parametrized by the prior probability of an s type, may seem stronger than usual. It is tempting to model the uncertainty about the probability of s-types and the interviewer's inference problem explicitly. For our purposes this would be overkill: (1) It would detract from our prime objective, to understand the operation of the key rationale for RRT, that credibly introducing noise offers privacy protection; (2) the privacy protection feature is logically prior to the inference problem and therefore worth studying on its own (if privacy protection does not work in the interaction between a single respondent and the interviewer, the rationale for RRT evaporates); (3) the complexity of the analysis of our model suggests that an added layer of uncertainty would make the problem of a full characterization of the equilibrium set intractable; (4) an implementation in the lab would be less transparent, with less salient incentives for subjects; (5) whereas our model has a clear antecedent in Ljungqvist's (1993) analysis, there is presently no published attempt at a complete

made for convenience. None of the qualitative features of the analysis depend on it. The principal reason for adopting a particular choice for the prior is that we are interested in generating predictions for a lab experiment. Choosing a different prior would give us similar predictions. Fully characterizing all equilibria for all priors is intractable.

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, *direct 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 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 direct response regime corresponds to the degenerate case in which $p_s \in \{0, 1\}$; in the randomized response regime, $p_s \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$.[7]

We assume that the respondent's incentive is shaped by two considerations: *stigmatization aversion* and *lying aversion*. Stigmatization aversion may be the result of individuals directly caring about perceptions or be instrumental, when the trait is associated with an illegal activity. Evidence for lying aversion has been documented in several experimental studies of communication games (e.g., Sánchez-Pagés and Vorsatz, 2007; Gneezy, 2005).

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,

game-theoretic study of the full-blown inference problem in which both the interviewer and the respondents are uncertain about the proportion of the population who have the stigmatized trait (see, however, Flannery (2015), who provides a partial equilibrium characterization in a model in which the proportion of the stigmatized types is uncertain; he shows that the truthful responding condition in our model is sufficient for truthful responses in his model); (6) in order to yield predictions that can be taken to the lab, a model that did address the inference problem directly would have to make common knowledge assumption at some level, e.g, by assuming that the population parameter of interest is drawn from a common knowledge distribution, for otherwise Weinstein and Yildiz' (2007) negative conclusions about game theoretic predictions would apply; and, (7) as we will see, we will be able to think about population inference even in our setting by simply applying the usual procedure to our data and to the frequencies predicted by our game model.

[7]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.

that he is of the stigmatized type s. The designation of s and t as stigmatized type and regular type is thus set by referring 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.[8] A higher value of λ means that the respondent is more lying averse. Similarly, a more stigmatization averse respondent will have a higher ξ.

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 dominates lying aversion. The *lying-stigmatization aversion ratio*, $\frac{\lambda}{\xi}$, will serve an important role in our equilibrium characterizations.

We assume that the only "action" the interviewer performs in the model is to update her beliefs. A belief function of the interviewer is $\mu_s : \{y, n\} \to [0, 1]$, which specifies for each received response a probability that $\theta = s$.[9]

[8] 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).

[9] 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.

3 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*.

The following property of the interviewer's posterior beliefs in equilibrium plays a crucial role in the analysis of both 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," $\xi|\mu_s(y) - \mu_s(n)|$.

3.1 Direct Response

In the direct response regime, $p_s \in \{0, 1\}$. 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. Without loss of generality, for the direct response regime, we consider that "Are you a t?" is the question being asked:

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

1. *for every lying-stigmatization aversion ratio* $\frac{\lambda}{\xi} \in (0, \frac{1}{2}]$, *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;*[10]

[10] For the details of the D1 criterion, see Banks and Sobel (1987) and Cho and Krep (1987).

2. *for* $\frac{\lambda}{\xi} \in (\frac{1}{2}, 1)$*, 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*.*

All proofs are in the appendix. Proposition 1 gives 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 direct.[11]

3.2 Randomized Response

In the randomized response regime, where p_s is non-degenerate, the question q that the respondent responds to 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 characterizes the set of RRT equilibria:

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*

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

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

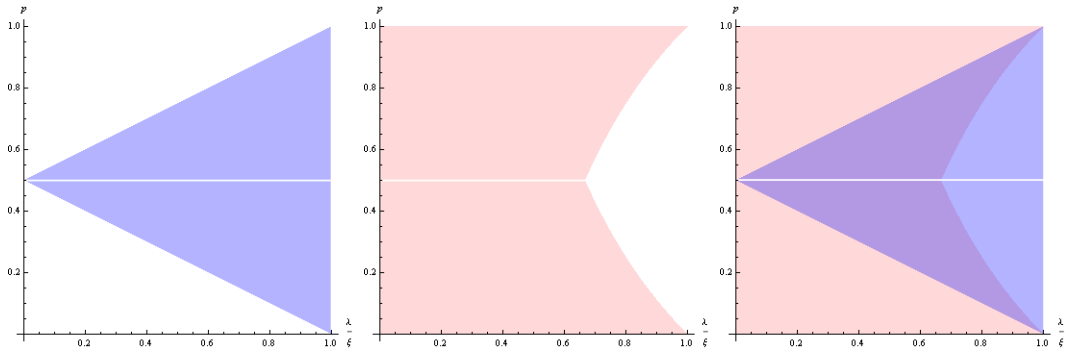
Since $\lambda < \xi$, it follows that the union of the sets in the first and second parts of

[11] 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.

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. In a typical non-truthful informative equilibrium s and t respond truthfully if the response is not jeopardizing and randomize otherwise.



(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 direct 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.

3.2.1 Population Estimates

The essence of what can go wrong with using RRT is captured by the fact that RRT admits informative equilibria that lead to distorted and sometimes invalid (i.e. negative) population estimates. To see this note that if the question "Are you

an s?" is asked with probability p_s, everyone in a population answers truthfully and the fraction who have the trait s is E, then the expected fraction of "yes" answers is

$$Y = p_s E + (1 - E)(1 - p_s),$$

solving for E and replacing Y with its sample equivalent $\hat{Y}$ we obtain the prevalence estimator proposed by Warner

$$\hat{E} = \frac{\hat{Y} + p_s - 1}{2p_s - 1}. \tag{1}$$

Consider an environment in which $\frac{\lambda}{\xi} = \frac{1}{8}$ and suppose that $p_s = 0.4$. In this case there is no truthful equilibrium and thus Warner's estimate for the prevalence of the stigmatized trait would be biased for any data generated from an equilibrium. There is an informative but non-truthful equilibrium in which the respondent's equilibrium strategy satisfies $\sigma(y|s, q_t) = 0.37$, $\sigma(y|t, q_s) = 0.33$, $\sigma(y|s, q_s) = 1$, $\sigma(y|t, q_t) = 1$.[12] For that strategy, the implied proportion of "yes" answer is $\hat{Y} = 0.6794$ and therefore Warner's prevalence estimate equals

$$\hat{E} = \frac{\hat{Y} + p_s - 1}{2p_s - 1} = \frac{0.6794 + 0.4 - 1}{0.8 - 1} = -0.397$$

which is negative and therefore invalid. If we change the environment by reducing stigmatization aversion so that $\frac{\lambda}{\xi} = \frac{1}{4}$ there is a truthful equilibrium, but also an informative but non-truthful equilibrium in which the respondent's equilibrium strategy satisfies $\sigma(y|s, q_t) = 0.11$, $\sigma(y|t, q_s) = 0.27$, $\sigma(y|s, q_s) = 1$, $\sigma(y|t, q_t) = 1$. For that strategy, the implied proportion of "yes" answer is $\hat{Y} = 0.5873$ and therefore Warner's prevalence estimate equals $\hat{E} = 0.0635$, far from the true population frequency of $E = 0.5$. This leads to the following observation:

Observation 1. *The randomized response regime admits informative but non-truthful equilibria that imply distorted and sometimes even invalid (negative) expected estimates of the fraction of the population with the stigmatized trait. Equilibria that result in extremely distorted estimates exists even under the most favorable conditions for RRT, i.e. when there are truthful equilibria.*

[12] The calculation of the mixed-strategy probabilities can be found in the proof of Proposition 2 in Appendix A.

3.2.2 Information-Eliciting Performance

Observation 1 reminds us that ignoring strategic behavior under RRT may result in misleading and nonsensical inference. One possible response is to try to investigate and then incorporate deviations from adherence to RRT instructions. This has been proposed, for example, by Clark and Desharnais (1998). Our model can be used to explore the potential of this approach.

We will focus on the most favorable condition, where the researcher knows the respondents' strategies both for DRT and RRT. In that case of course if sample size is not an issue it is possible to obtain an accurate estimate of the proportion of members of the population with the stigmatized trait. If sample size is an issue, it becomes of interest how much information can be extracted from each observation.

To this end, we evaluate the information-eliciting performance of the two different response regimes using a standard measure from information theory (Shannon, 1948). Specifically, we measure the maximal transmittable information, as expressed by *mutual 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 $0 \log 0 = 0$ 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 about θ due to the observation of r; it ranges from zero to one.[13] [14]

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 the maximal mutual information by $\bar{I}_D(\frac{\lambda}{\xi})$ for the direct response regime and $\bar{I}_R(\frac{\lambda}{\xi})$ for the randomized response regime.

Figure 2 summarizes our findings (the details can be found in Appendix B): The upper left-hand panel shows that the maximal mutual information achievable with DRT is zero for high relative stigmatization aversion, $\frac{\lambda}{\xi} < \frac{1}{2}$, becomes

[13] Mutual information is also referred to as relative entropy or Kullback-Leibler divergence, the divergence between the joint and product distributions of the random variable 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).

[14] 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 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).

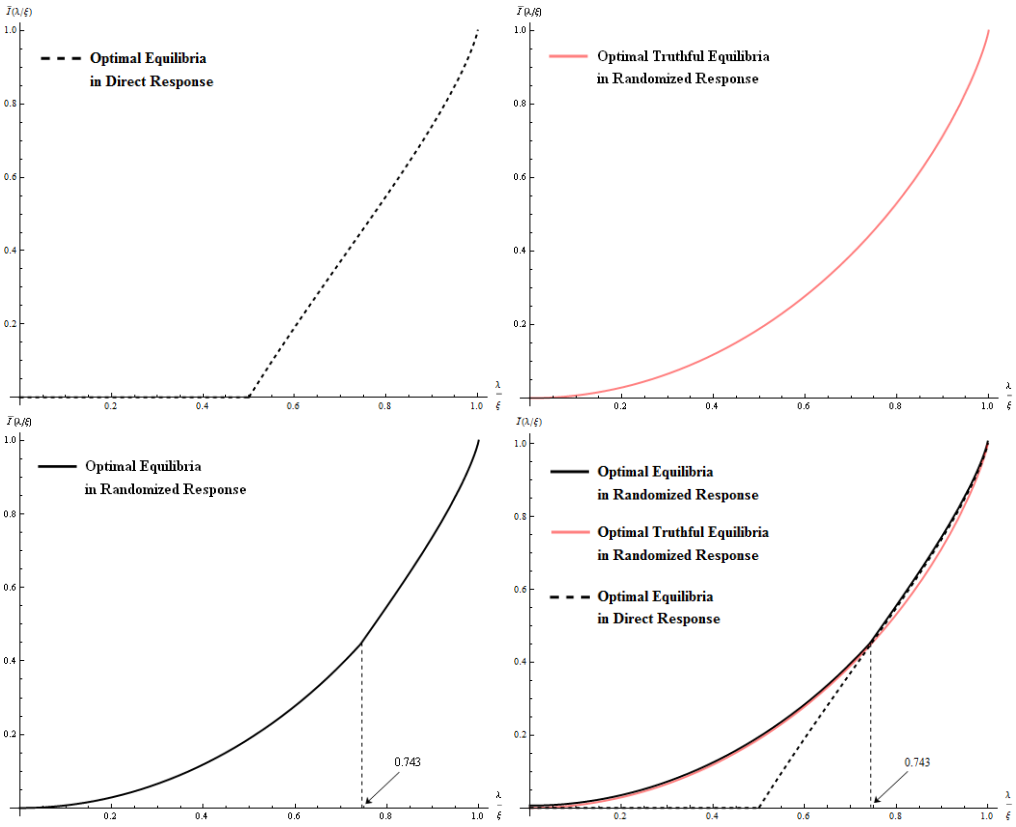


Figure 2: Maximal Mutual Information

positive for moderate relative stigmatization aversion, $\frac{\lambda}{\xi} > \frac{1}{2}$, and increases to one as $\frac{\lambda}{\xi} \to 1$. The upper right-hand panel shows that, in contrast, the maximal mutual information achievable with truthful RRT equilibria is strictly positive regardless of the relative stigmatization aversion and thus improves on DRT when stigmatization aversion is relatively high. While DRT is dominated by RRT for relatively high stigmatization aversion, the same is not the case for moderate stigmatization aversion, as shown in the two lower panels. Below $\frac{\lambda}{\xi} \approx 0.743$ RRT dominates DRT and the optimal mutual information is achieved with a truthful RRT equilibrium. Beyond that value RRT and DRT are tied and the optimal truth-telling RRT equilibrium is dominated by the optimal DRT equilibrium (and also an informative non-truthtelling RRT equilibrium).

4 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 4.1, discussing the rationales behind the adoption of the model parameters for the treatments. We then describe in Section 4.2 the laboratory environments in which these treatments were conducted.

4.1 Treatments and Hypotheses

Table 1 describes our treatments. It shows a 2×2 design, where the rows of the matrix correspond to the values of $\frac{\lambda}{\xi} \in \left\{\frac{1}{4}, \frac{1}{8}\right\}$ 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 direct response regime and $p_s = 0.4$ to an instance of the randomized response regime.

Our choice of parameters is guided by two considerations: (1) We are interested in a sharp distinction between RRT and DRT and (2) we like to evaluate RRT under conditions where it is most relevant in the field. Both of these considerations suggest to look at parameterizations for which stigmatization aversion

Table 1: Experimental Treatments

	$p_s = \text{Prob}(q_s) = 0$	$p_s = \text{Prob}(q_s) = 0.4$
$\frac{\lambda}{\xi} = \frac{1}{4}$	*DirectLow*: Direct Response / Low Relative Stigmatization Aversion (Equilibrium Prediction: No Informative Equilibrium)	*RandomLow*: Randomized Response / Low Relative Stigmatization Aversion (Equilibrium Prediction: Truthful Equilibrium Exists)
$\frac{\lambda}{\xi} = \frac{1}{8}$	*DirectHigh*: Direct 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, No Truthful Equilibria Exist)

is high (relative to lying aversion). For the parameters in our experiment, theory predicts that there is no information transmission under DRT whereas there is always an informative equilibrium under RRT and under one condition an equilibrium with truthful responses, thus inducing the sharp separation we are looking for. High stigmatization aversion corresponds to more sensitive traits in the field, the case for which RRT is of most interest.

With these parameter values, the equilibrium analysis of our model predicts that there will be no information transmission in the direct 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 stigmatization aversion is not too high.

In the direct response treatments, *DirectLow* and *DirectHigh*, "Are you a t?" is always asked ($p_s = 0$). 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 its maximum 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 direct response treatments, we round up

0.375 and use $p_s = 0.4$.[15] 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 direct 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 direct response treatments.*

Our second hypothesis focuses on the direct 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.[16] The D1 pooling equilibria predict the same respondent's behavior in *DirectLow* and *DirectHigh*, 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 DirectLow and DirectHigh, both stigmatized types and regular types respond with "yes" significantly more often than with "no," and there is no significant difference in the uses of "yes" between them. 2) The uses*

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

[16] 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*.

of responses by both stigmatized types and regular types do not differ significantly between DirectLow and DirectHigh.

We turn next to the interviewer's beliefs within each treatment, after "yes" and "no":

Hypothesis 3. *1) In the direct 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 direct 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.

4.2 Design and Procedures

Our experiment was conducted at the Pittsburgh Experimental Economics Lab. A total of 304 subjects with no prior experience in these 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 $8 - 10$ randomly matched groups.[17] The experiment was programmed and conducted using z-Tree (Fischbacher, 2007).

[17]We set a recruiting target of 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.

In each session, half the subjects were randomly assigned the role of Member A (respondent) and the other half the role of Member B (interviewer), with role assignments remaining fixed throughout the session. They participated in 40 rounds of decisions in groups of two.[18] After each and every round, subjects were randomly rematched, i.e., we used *random matching*. 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 direct 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 responded to the question being 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 units (ECU).[19] If Member A's response to the question truthfully reported which shape was selected, he/she would receive 300 ECU in *DirectLow*/*RandomLow* and 275 ECU in *DirectHigh*/*RandomHigh*; otherwise with untruthful responses, he/she would receive 250 ECU. Lying aversion was thus induced as earning either 50 ECU or

[18]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.

[19]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.

25 ECU for being 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 *DirectLow* and *RandomLow* $\frac{\lambda}{\xi} = \frac{1}{4}$ was implemented as $\frac{50}{200}$ and in *DirectHigh* and *RandomHigh* $\frac{\lambda}{\xi} = \frac{1}{8}$ was implemented as $\frac{25}{200}$.[20]

Member B's rewards revolved around the provision of incentives for truthful reporting of beliefs.[21] We used a belief-elicitation mechanism in which, irrespective of risk attitudes, truthfully reporting one's beliefs is a dominant strategy (Karni, 2009).[22] 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 B.

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/TRI-

[20]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 *DirectHigh*/*RandomHigh* and \$30 in *DirectLow*/*RandomLow*. Had we varied the absolute level of stigmatization aversion, we would have had to adjust the base earning upward for *DirectHigh* and *RandomHigh* resulting in a considerably higher upper bound of payments or, with no such upward adjustment, accept the possibility of negative earnings.

[21]Given the passive role of the interviewer in the model, a conceivable, alternative way to implement it in the lab is to replace Member B (and thus remove the need of using rather complex belief-elicitation mechanism and any concern over Member B's other-regarding preferences) with a program that mechanically updates beliefs following Bayes' rule. However, given that our games have multiple equilibria with no clear theoretical guidance for selection, even with the restriction of Bayes' rule it is unclear how the program should update beliefs upon inputs from Member A. Using such a program may therefore not be appropriate.

[22]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 dër Weele (2009) and Hossain and Okui (2013).

ANGLE, 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, thus providing the incentives for eliciting true beliefs.[23]

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.

5 Experimental Findings

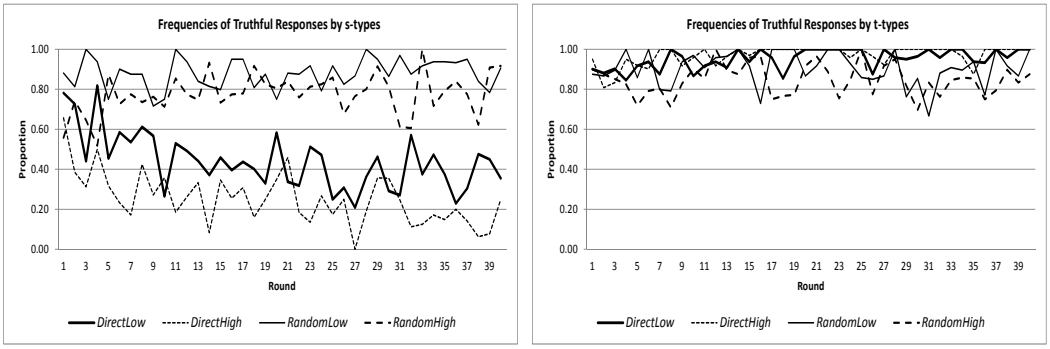
5.1 Respondents' Responses and Interviewers' Beliefs

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 direct response treatment:

Result 1. *1) Stigmatized types provided truthful responses decidedly more often in the randomized response treatments than in the direct response treatments. 2) Regular types provided truthful responses more often in the direct response treatments than in the randomized response treatments.*

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 *DirectLow*

[23]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.



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

Figure 3: Trends of Truthful Response Frequencies

and 19% in *DirectHigh*. 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 aversion ($p = 0.0143$ for all four possible comparisons, Mann-Whitney tests).[24] For t-types, the truthful response frequencies were significantly higher in the direct 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 *DirectLow* and *DirectHigh*, 90% in *RandomLow*, and 84% in *RandomHigh* ($p = 0.0143$ for all four possible comparisons, Mann-Whitney tests).

The fact that t-types become less truthful with RRT, while s types become more truthful, is consistent with the form informative non-truthful equilibria take in the game we analyzed. Unlike with DRT, with RRT being truthful requires regular types sometimes to give jeopardizing answers. The intuition that they may want to avoid doing so and engage in non-truthful protective behavior instead is confirmed by both our formal analysis and our experimental data.

We noted in Observation 1 of our theoretical analysis that the protective be-

[24] All aggregate data reported and used for statistical testings are from the last 20 rounds. The qualitative aspects of our findings remain unchanged if we use, for example, data from the last 30 or even all 40 rounds. However, the frequency trends, especially those for types s in the direct response treatments where convergence was most conspicuous, suggest that the 20th round provides a reasonable cutoff for behavior having settled down. Using data from the 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.

havior of regular types may result in distorted and even invalid estimates of the proportion of stigmatized types in the population. This concern is validated by our experimental data, as shown in Table 2. There we report for each session the estimated prevalence of the stigmatized type in the population, using the estimator proposed by Warner (see equation (1)). We find that regardless of whether DRT or RRT is used, the estimated population proportion of the stigmatized type is underestimated. Furthermore, on average RRT yields a less accurate estimate. With higher stigmatization aversion the estimated proportion drops and in three out of four sessions with RRT the estimate becomes invalid.

Table 2: Actual and Estimated Proportions of Stigmatized Types

	Actual	Estimated	Actual	Estimated
	DirectLow		*RandomLow*	
Session 1	0.51	0.21	0.55	0.28
Session 2	0.50	0.18	0.47	0.14
Session 3	0.45	0.09	0.50	0.08
Session 4	0.46	0.25	0.50	0.11
Mean	0.48	0.19	0.50	0.15
	DirectHigh		*RandomHigh*	
Session 1	0.49	0.11	0.46	-0.45
Session 2	0.44	0.04	0.59	-0.22
Session 3	0.45	0.09	0.56	-0.58
Session 4	0.50	0.17	0.48	0.08
Mean	0.47	0.10	0.52	-0.29

Note: Data are from last 20 rounds of each session. The means for treatments are calculated using each group in each round as an observation.

In principle it is conceivable that respondents did not answer truthfully because they did not understand the experimental instructions. There are, however, two reasons to believe that this is not the case. First, as we have seen, departures from truth are in line with incentives for privacy protection by regular types. Second, there is no apparent time trend in the estimated proportion, as shown in Figure 4, and therefore greater familiarity with the setup did not lead to improved prevalence estimates.

Given that in the direct 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:

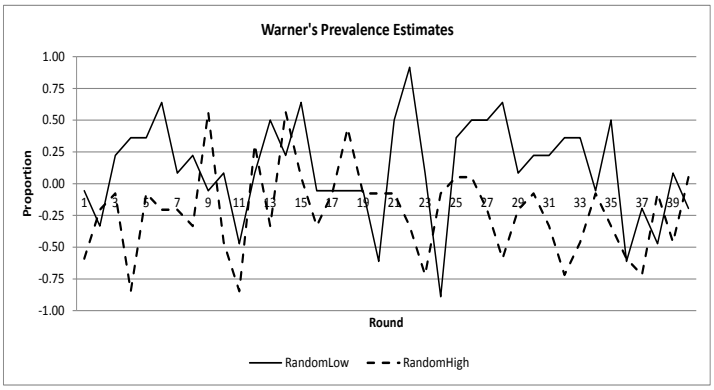


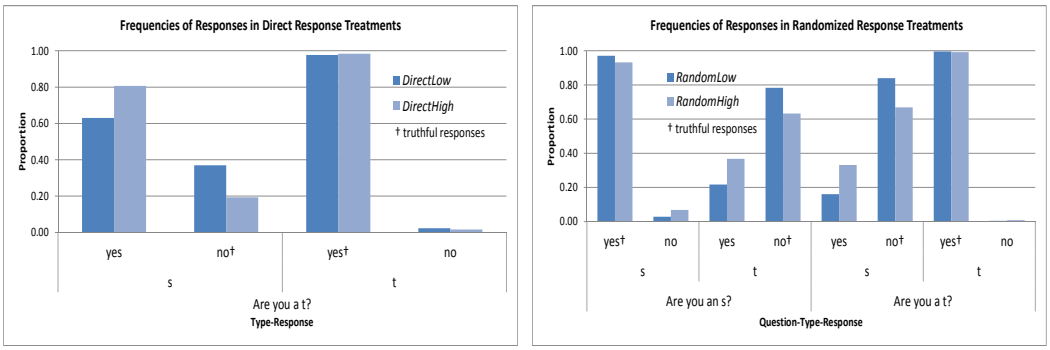
Figure 4: Trends of Prevalence Estimates

Result 1a. *1) In DirectLow, regular types and, to a lesser degree, stigmatized types responded with "yes" significantly more often than with "no." In DirectHigh, 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 5 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 *DirectLow* and *DirectHigh* 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, in contrast to 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.[25] The qualitative prediction that "yes" is used more often than "no" by s-types was, however, largely confirmed ($p = 0.0625$ for *DirectHigh* and $p = 0.125$ for *Direct-*

[25]We conducted an additional session for robustness check, where the parameters were the same as *DirectLow* except that $p_s = 1$ (i.e., the direct question became "Are you an s?"). Compared to *DirectLow* 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%.

Low, Wilcoxon signed-rank tests).



(a) Direct Response Treatments (b) Randomized Response Treatments

Figure 5: Frequencies of Responses

Despite equilibrium predicting no behavior difference between *DirectLow* and *DirectHigh*, in light of the over-communication observed, a natural question is how the difference 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:

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 DirectLow and DirectHigh.*

For the direct response treatments, the second part of Hypothesis 2 is confirmed for t-types but not for s-types: the stronger relative stigmatization aversion in *DirectHigh* 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 5(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*. 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*.[26]

The significantly lower frequencies of negative truthful responses for the question "Are you an s?" in both $RandomHigh$ and $RandomLow$ treatments represent 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.[27]

[26]Given the rather unusual randomization of questions from subjects' perspective, it is conceivable that an experimenter effect existed in which subjects expected that the experimenter expected them to deceive the interviewer when questions were randomized. One would, however, expect that the effect existed more or less uniformly across different questions. The observation that s-types' frequencies of truthful responses varied across questions (more affirmative truthful responses than negative truthful responses) suggested that the potential of experimenter effect is likely to be small.

[27]An additional contributing factor for observing protective behaviors in the field may be heterogeneity in individual weighting of truth-telling and stigmatization concerns. Those with stronger stigmatization concerns might be expected to engage in protective behaviors even if

To further explore what drove the respondents' observed behavior, we bring the interviewers' beliefs into the picture. The "Elicited" columns in Table 3 present the interviewers' elicited beliefs.[28] 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 3: Elicited and Empirical Beliefs Assigned to Type s

Response	"yes"		"no"		"yes"		"no"	
Beliefs	Elicited	Empirical	Elicited	Emp	Elicited	Emp	Elicited	Emp
	DirectLow				*DirectHigh*			
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 direct 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 direct 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 *DirectLow* and

others are content with being truthful.

[28]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.

0.36 in *DirectHigh*, 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 *DirectLow* and 10% of the time in *DirectHigh*. The corresponding elicited beliefs assigned to s were 0.79 in *DirectLow* and 0.70 in *DirectHigh*, significantly higher than when "yes" was received ($p = 0.0625$ for both treatments, Wilcoxon signed-rank tests). In fact, in *DirectLow* 44% of the time the elicited beliefs were equal or larger than 0.9, while it was 31% in *DirectHigh*. 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.[29] Prior experimental studies have documented that subjects have 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.[30] Indeed, the respondents' observed behavior resembled

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

[30]Note that since the experimental procedure involves no real stigmatization, with stigmatization aversion induced through preferences over revealing context-free SQUARES and TRI-

the informative equilibrium under a higher lying-stigmatization aversion ratio.[31]

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 direct 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 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 equi-

ANGLES, no homegrown stigmatization aversion is expected.

[31]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.

libria. 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 ±5% of the observed frequencies.[32]

It is instructive to reverse perspectives and try to determine which aversion ratios, $\frac{\lambda}{\xi}$, are implied by the observed response frequencies if one identifies those frequencies with the mixing probabilities in an informative non-truthful equilibrium. In the case of RandomLow s responds to q_t with "yes" with a frequency of 0.16 and t responds to q_s with "yes" with a frequency of 0.22. The implied aversion ratio is approximately 0.2, compared to the induced ratio of 0.25. In the case of RandomHigh s responds to q_t with "yes" with a frequency of 0.33 and t responds to q_s with "yes" with a frequency of 0.37. The implied aversion ratio is approximately 0.17, compared to the induced ratio of 0.125. While those aversion ratios are not exact matches for the ones we were trying to induce, they are in the right range and preserve the order of the intended ratios. We take this calibration exercise as further evidence that the informative non-truthful equilibria give a sensible account of behavior in our randomized response treatments.[33]

[32]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*.

[33]One can also perform the calibration exercise for the direct response treatments. Consistent with the over-communication we found there, the implied aversion ratios are markedly higher than the induced ratios: In DirectLow the implied aversion ratio is 0.61, compared to an induced ratio of 0.25. In DirectHigh the implied aversion ratio is 0.55, compared to an induced ratio of 0.125. An interesting open question is how to reconcile the difference between the implied aversion ratios for DRT and RRT. One possibility is that respondents develop homegrown perceptions about p_s. If they have an exaggerated sense of the difference between p_s and p_t, i.e. perceive p_s to be lower than it is, the implied aversion ratio increases. Another possibility is that psychologically the RRT procedure might not feel safe, as John et al. (2013) have suggested. Finally, it might be that truth-telling is more salient under DRT since there is a more definite sense of what constitutes truth. The latter might be especially interesting from an applied perspective, as it suggests a potentially adverse effect of RRT on lying aversion.

5.2 Information-Eliciting Performance

The mutual information measures implied by our data are reported in Table 4. There we also report a decomposition of the difference between the DRT and RRT mutual information measures into the contributions from randomizing questions, which we refer to as "Noise," the protective strategic response by t types to randomizing questions, referred to as "Protective behavior of t," and the primary intended strategic effect of RRT—to induce s types to truthfully answer with "no" in response to the question "Are you a t?", which we refer to as the "Positive Effect of RRT."

Table 4: Mutual Information

	DirectLow	Noise	Protective Behavior of t	Positive Effect of RRT	Remaining	*RandomLow*
Session 1	0.134					0.039
Session 2	0.213					0.046
Session 3	0.080					0.017
Session 4	0.267					0.003
Mean	0.173	−0.171	+0.012	+0.023	−0.011	0.026
	DirectHigh	Noise	Protective Behavior of t	Positive Effect of RRT	Remaining	*RandomHigh*
Session 1	0.055					0.006
Session 2	0.012					0.027
Session 3	0.056					0.030
Session 4	0.198					0.009
Mean	0.080	−0.076	+0.036	+0.007	−0.029	0.018

Note: The figures are obtained by applying an alternative formula of mutual information, $\sum P(r'|\theta')P(\theta')\log\frac{P(r'|\theta')}{P(r')}$, to the empirical counterparts (last 20 rounds aggregates) of $P(\theta')$, $P(q_{\theta'})$, and $P(r'|q_{\theta'},\theta')$. The effects of noise are obtained by, starting with the mutual information in DRT, replacing $P(q_s)$ and $P(q_t)$ in DRT with those in RRT. (In the same step, negation of the answer to q_t is also used for the frequency of responses to the unasked question q_s in DRT, i.e., we assume that for DRT $P(r'|q_s,\theta') = 1 - P(r'|q_t,\theta')$.) The effects of types-$t$'s protective behavior are obtained by replacing $P(r'|q_s,t)$ in DRT with those in RRT. The positive effects of RRT are obtained by replacing $P(r'|q_t,s)$ in DRT with those in RRT. The remaining effects are obtained by replacing the remaining $P(r'|\cdot,\cdot)$ in DRT with those in RRT.

Theory predicts that for the parameters in questions RRT performs strictly better than DRT. The direct response regime, however, dominates in the lab due to the *over-communication* with DRT and *under-communication* with RRT. At the same time, a comparative statics prediction from the theory was recovered in the laboratory: relative to the direct response regime, the randomized response regime performed better when there was stronger stigmatization aversion.

The decomposition of the difference of the mutual information measures for

direct and randomized response reveals that the direct effect of randomizing questions ("Noise") nearly completely eliminates the information gain from randomized response whereas the intended strategic effect of RRT ("Positive Effect of RRT") makes a positive, although small, contribution to informativeness.

6 Discussion and Conclusion

Randomness can be a source of plausible deniability. This has the potential to improve information transmission, as demonstrated in the literature on strategic communication through noisy channels. It remains an open question, however, whether and how this potential can be translated into gains in applications.

Randomized response is an ideal starting point to begin the investigation of this question. It relies on the same basic idea, that randomness shields the sender, and it has been used with some success in the field. The idea to use randomness to generate privacy protection while retaining the ability to obtain prevalence estimates at the population level is ingeniously simple. It can be straightforwardly modeled as a game and implemented in an experimental laboratory.

The theoretical analysis of the randomized-response game yields two principal novel insights: (1) there are informative equilibria that are not truthful, and (2) those equilibria can imply distorted and possibly invalid estimates of the prevalence of the stigmatizing trait in the population. Non-truthful informative equilibria are robust to standard belief-based refinements since the randomization ensures that there are no off-equilibrium responses. Furthermore, they plausibly express the focal principle of privacy protection. The model thus delivers that there is no need for auxiliary explanations of observed non-compliance with randomized response instructions. Non-truthful responses are entirely rational and vary in a predictable fashion with incentives.

Our experimental findings are best accounted for by the informative but non-truthful equilibria of the randomized-response game. Consistent with these equilibria, randomized response improves truth-telling but results in systematic departures from full truth-telling for jeopardizing answers and distorted prevalence estimates. While we find over-communication with direct questioning, consistent with the experimental literature on strategic information transmission, we

find *under-communication* with randomized response. Observed departures from truthful responding tend to be rational, occur for jeopardizing answers, and respond to changing incentives.

The research strategy we propose, to build fully specified game models of RRT and to take them to the experimental laboratory, can be straightforwardly applied to alternate versions of RRT. Any variant of RRT requires that there are answers that result in less favorable posterior than prior beliefs about the respondent. This creates incentives for non-compliance. Therefore, we expect that the key message of our paper, that one should expect systematic non-compliance with RRT instructions for jeopardizing answers, applies to all variants of RRT.

One might wish to mitigate this effect by designing RRT so that it minimally moves posterior beliefs. Since this requires increasing the noise level it will have to be accompanied by increasing sample size. Even this might not be enough, and it might be worthwhile using the laboratory to investigate whether it is indeed the case that responses become approximately truthful when truthful responding is made nearly uninformative; in our version of RRT this would correspond to letting the probability of asking each question converge to one half.

A key takeaway from our game analysis and lab implementation is that non-compliance with RRT instructions is to be expected. For implementations of randomized response in the field this suggests incorporating non-compliance into the data analysis. A simple way of recognizing the possibility of non-compliance would be to report an array of different prevalence estimates corresponding to different assumed compliance rates. In addition, it will be important to continue the current effort to estimate non-compliance rates (as, for example, in Cruyff, van den Hout, van der Heijden and Böckenholt (2007)). This will involve developing appropriate identifying assumptions. Clark and Desharnais (1998), for example, propose to estimate non-compliance rates using the assumption that these rates do not vary with changes in the probability of asking each question. Alternative identifying assumptions are possible and our model provides guidance by relating compliance rates to fundamentals (aversion ratio) and design parameters (the probabilities with which questions are asked) and by making qualitative predictions about the form of noncompliance.

References

[1] Allen, Franklin. 1987. "Discovering Personal Probabilities When Utility Functions are Unknown." *Management Science*, 33: 452-454.

[2] Ayres, Ian and Barry Nalebuff. 1996. "Common Knowledge as a Barrier to Negotiation," *UCLA Law Review*, 44: 1631-1659.

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

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

[5] 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.

[6] 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.

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

[8] 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.

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

[10] Bogen, David and Michael Lynch. 1989. "Taking Account of the Hostile Native: Plausible Deniability and the Production of Conventional History in the Iran-Contra Hearings," *Social Problems*, 36: 197-224

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

[12] 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.

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

[14] Calomiris, Charles W. 2009. "The Debasement of Ratings: What's Wrong and How We Can Fix It." Columbia University Working Paper.

[15] Chassang, Sylvain and Gerard Padró i Miquel. 2013. "Corruption, Intimidation and Whistleblowing: A Theory of Inference from Unverifable Reports," Princeton University Working Paper.

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

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

[18] Clark, Stephen J. and Robert A. Desharnais. 1998. "Honest answers to embarrassing questions: Detecting cheating in the randomized response model." *Psychological Methods*, 3: 160-168.

[19] Coffman, Katherine B, Lucas C Coffman, Keith M Marzilli Ericson. 2013. "The Size of the LGBT Population and the Magnitude of Anti-Gay Sentiment are Substantially Underestimated." Ohio State University Working Paper.

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

[21] Crawford, Vincent and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica*, 50: 1431-1451.

[22] Cruyff, Maarten JLF, Ardo van den Hout, Peter GM van der Heijden, and Ulf Böckenholt. 2007. "Log-linear randomized-response models taking self-protective response behavior into account." *Sociological Methods & Research* 36: 266-282.

[23] Dana, Jason, Daylian M. Cain and Robyn Dawes. 2006. "What You Don't Know Won't Hurt Me: Costly (but Quiet) Exit in Dictator Games," *Organizational Behavior and Human Decision Processes*, 100:193-201.

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

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

[26] 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.

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

[28] Flannery, Tim. 2015. "A Game Theoretic Analysis of the Randomized Response Technique." University of Arizona Working Paper.

[29] 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.

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

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

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

[33] 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.

[34] Hossain, Tanjim and Okui Ryo. 2013. "The Binarized Scoring Rule of Belief Elicitation." *Review of Economic Studies*, 80: 984-1001.

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

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

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

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

[39] Karlan, Dean S., and Jonathan Zinman. 2012. "List randomization for sensitive behavior: An application for measuring use of loan proceeds." *Journal of Development Economics*, 98: 71-75.

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

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

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

[43] Kawamura, Kohei, 2013. "Eliciting information from a large population." *Journal of Public Economics*, 103: 44-54.

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

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

[46] 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.

[47] 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.

[48] Mialon, Hugo M., and Sue H. Mialon. 2013. "Go Figure: The Strategy of Nonliteral Speech," *American Economic Journal: Microeconomics*, 5: 186-212.

[49] Miller, JD. 1984. "A new survey technique for studying deviant behavior." PhD Dissertation, George Washington University.

[50] 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.

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

[52] Pinker, Steven, Martin A. Nowak, and James J. Lee. 2008. "The Logic of Indirect Speech," *Proceedings of the National Academy of Sciences*, 105: 833-838.

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

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

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

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

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

[58] St John, Freya A. V. , Aidan M. Keane, Gareth Edwards-Jones, Lauren Jones, Richard W. Yarnell, and Julia P. G. Jones. 2012. "Identifying Indicators of Illegal Behaviour: Carnivore Killing in Human-Managed landscapes." *Proceedings of Royal Society Biological Science*.

[59] 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.

[60] Tadelis, Steven. 2011. "The Power of Shame and the Rationality of Trust." UC Berkeley Working Paper.

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

[62] Weinstein, Jonathan, and Muhamet Yildiz. 2007. "A Structure Theorem for Rationalizability with Application to Robust Predictions of Refinements." *Econometrica*, 75: 365-400.

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

Appendices

For Online Publication

Online 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 direct 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 $\sigma(y|s,q_t)\in[0,1)$ 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

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

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

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

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

$$\sigma(y|s,q_s) = \frac{-\xi+\sqrt{4\lambda^2-4\lambda\xi[1-2(1-p_s)\sigma(y|t,q_t)]+\xi^2}+2\lambda[1-(1-p_s)\sigma(y|t,q_t)]}{2p_s\lambda}. \tag{A.11}$$

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.11) 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.11) 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.11) 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.11) 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 randomized 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.11) 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 $p_s \in (0,\frac{1}{2}) \cup (\frac{1}{2},1)$, $p_s > 2 - \frac{\xi}{\lambda}$.

□

Online Appendix B – Mutual Information

B.1 Characterizations

Recall that in the direct 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 direct 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,*

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 direct 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 direct 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 down 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 mid-point of $[\frac{\xi-\lambda}{2\xi}, 1 - \frac{\xi-\lambda}{2\xi}]$ and 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 direct response regime, with strict dominance for $\frac{\lambda}{\xi} \in (0,c)$, where $c \approx 0.743$.*

B.2 Proofs

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

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

Note that as a function, (B.1) 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

$$\mu_s(y) = \frac{\Pr(y|s)}{\Pr(y|s) + \Pr(y|t)} \Leftrightarrow \Pr(y|s) + \Pr(y|t) = \frac{\Pr(y|s)}{\mu_s(y)}, \tag{B.2}$$

$$\mu_s(n) = \frac{\Pr(n|s)}{\Pr(n|s) + \Pr(n|t)} \Leftrightarrow \Pr(n|s) + \Pr(n|t) = \frac{\Pr(n|s)}{\mu_s(n)}. \tag{B.3}$$

Substituting (B.2) and (B.3) into (B.1), we obtain

$$\begin{aligned} I(\theta; r) = 1 + & \left[\frac{\Pr(y|s)}{2\mu_s(y)}\right] (\mu_s(y) \log \mu_s(y) + [1 - \mu_s(y)] \log[1 - \mu_s(y)]) \\ & + \left[\frac{\Pr(n|s)}{2\mu_s(n)}\right] (\mu_s(n) \log \mu_s(n) + [1 - \mu_s(n)] \log[1 - \mu_s(n)])]. \end{aligned} \tag{B.4}$$

We use the fact that $\Pr(n|\cdot) = 1 - \Pr(y|\cdot)$ to further eliminate $\Pr(n|s)$ and $\mu_s(n)$. Note that (B.3) can be rewritten as

$$\mu_s(n) = \frac{1 - \Pr(y|s)}{2 - [\Pr(y|s) + \Pr(y|t)]} = \frac{\mu_s(y)[1 - \Pr(y|s)]}{2\mu_s(y) - \Pr(y|s)}, \tag{B.5}$$

where in the second equality we use (B.2) for $\Pr(y|s) + \Pr(y|t)$. Using (B.5) and

the fact that $\frac{\Pr(n|s)}{2\mu_s(n)} = 1 - \frac{\Pr(y|s)}{2\mu_s(y)}$, (B.4) becomes

$$\begin{aligned} I(\theta;r) = 1+ & \left[\frac{\Pr(y|s)}{2\mu_s(y)}\right](\mu_s(y)\log\mu_s(y) + [1-\mu_s(y)]\log[1-\mu_s(y)]) \\ & + \left[1 - \frac{\Pr(y|s)}{2\mu_s(y)}\right]\left[\left(\frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)}\right)\log\left(\frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)}\right)\right. \\ & \left. + \left(1 - \frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)}\right)\log\left(1 - \frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)}\right)\right]. \end{aligned} \tag{B.6}$$

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 (B.5), the constraint becomes

$$\frac{\mu_s(y)[1-\Pr(y|s)]}{2\mu_s(y)-\Pr(y|s)} - \mu_s(y) = \frac{\lambda}{\xi} \Leftrightarrow \Pr(y|s) = \mu_s(y)\left(\frac{\xi}{\lambda}\right)\left(2\left[\frac{\lambda}{\xi} + \mu_s(y)\right] - 1\right). \tag{B.7}$$

Substituting (B.7) into (B.6), we obtain the following function in terms of $\mu_s(y)$ only:

$$\begin{aligned} I(\theta;r) =& \hat{I}(\mu_s(y)) \\ =& 1 + \left(1 - \frac{\xi}{2\lambda}[1-2\mu_s(y)]\right)(\mu_s(y)\log\mu_s(y) + [1-\mu_s(y)]\log[1-\mu_s(y)]) \\ & + \left(\frac{\xi}{2\lambda}[1-2\mu_s(y)]\right)\left[\left(\mu_s(y) + \frac{\lambda}{\xi}\right)\log\left(\mu_s(y) + \frac{\lambda}{\xi}\right)\right. \\ & \left. + \left(1 - \mu_s(y) - \frac{\lambda}{\xi}\right)\log\left(1 - \mu_s(y) - \frac{\lambda}{\xi}\right)\right]. \end{aligned} \tag{B.8}$$

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

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

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

The first-order condition for an extremum is

$$\left[1-\frac{2\lambda}{\xi}-4\mu_s(y)\right]\ln\left(\frac{\mu_s(y)+\frac{\lambda}{\xi}}{\mu_s(y)}\right)=\left[3-\frac{2\lambda}{\xi}-4\mu_s(y)\right]\ln\left(\frac{1-\mu_s(y)-\frac{\lambda}{\xi}}{1-\mu_s(y)}\right). \tag{B.9}$$

Equation (B.9) 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:

$$\begin{aligned}\hat{I}'''(\mu_s(y))=&\frac{1}{([\mu_s(y)][1-\mu_s(y)][\mu_s(y)+\frac{\lambda}{\xi}][1-\mu_s(y)-\frac{\lambda}{\xi}])^2\ln 4}\\&\times\left(1-2\mu_s(y)-\frac{\lambda}{\xi}\right)\left[2\left(\frac{\lambda}{\xi}\right)^3[1-2\mu_s(y)]\right.\\&-3\left(\frac{\lambda}{\xi}\right)^2(1-2\mu_s(y)[1-\mu_s(y)])\\&+\left(\frac{\lambda}{\xi}\right)(1-2\mu_s(y)(2-\mu_s(y)[3-2\mu_s(y)])\\&\left.+2\mu_s(y)[1-\mu_s(y)](1-\mu_s(y)[1-\mu_s(y)])\right].\end{aligned} \tag{B.10}$$

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:

$$\hat{\mu}_s(y) = \frac{1}{2}\left(1 - \frac{\lambda}{\xi} - \sqrt{2\sqrt{4\left(\frac{\lambda}{\xi}\right)^4 - \left(\frac{\lambda}{\xi}\right)^2 + 1} - 3\left(\frac{\lambda}{\xi}\right)^2 - 1}\right), \tag{B.11}$$

$$\bar{\mu}_s(y) = \frac{1}{2}\left(1 - \frac{\lambda}{\xi}\right), \tag{B.12}$$

$$\tilde{\mu}_s(y) = \frac{1}{2}\left(1 - \frac{\lambda}{\xi} + \sqrt{2\sqrt{4\left(\frac{\lambda}{\xi}\right)^4 - \left(\frac{\lambda}{\xi}\right)^2 + 1} - 3\left(\frac{\lambda}{\xi}\right)^2 - 1}\right). \tag{B.13}$$

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 (B.12) is the only point in $(0, 1 - \frac{\lambda}{\xi})$ at which the third derivative vanishes. Evaluating the expression in (B.10) 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 (B.12) 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 (B.11)-(B.13) coincide. Evaluating the expression in (B.10) for $\frac{\lambda}{\xi} \in [\sqrt{3/7}, 1]$ then gives that $\hat{I}'''(\mu_s(y)) \gtreqless 0$ 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 (B.11) satisfies that $\hat{\mu}_s(y) \in (0, \frac{1}{2}(1-\frac{\lambda}{\xi}))$ and the solution in (B.13) 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 (B.10) 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 (B.8), 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 (B.8) 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 direct 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.

□

Online Appendix C – Experimental Instructions

C.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.

C.2 z-Tree Screen Shots

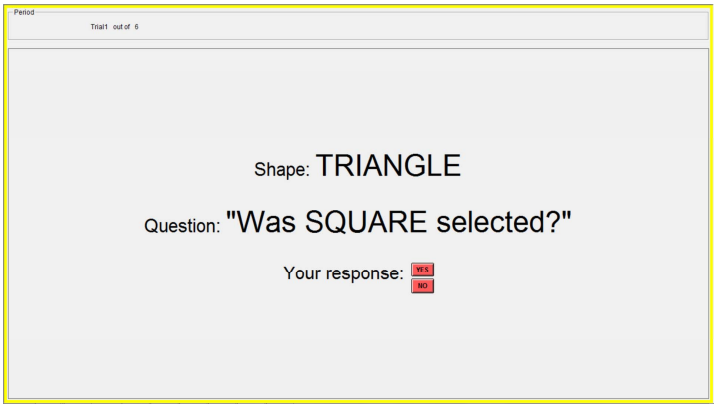


Figure 6: Member A's Response Screen

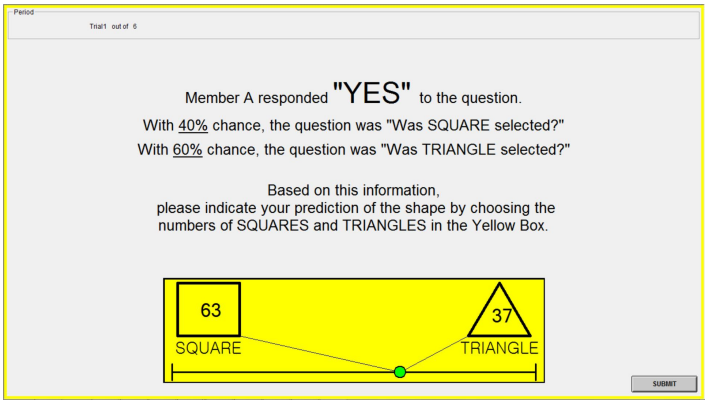


Figure 7: Member B's Prediction Screen

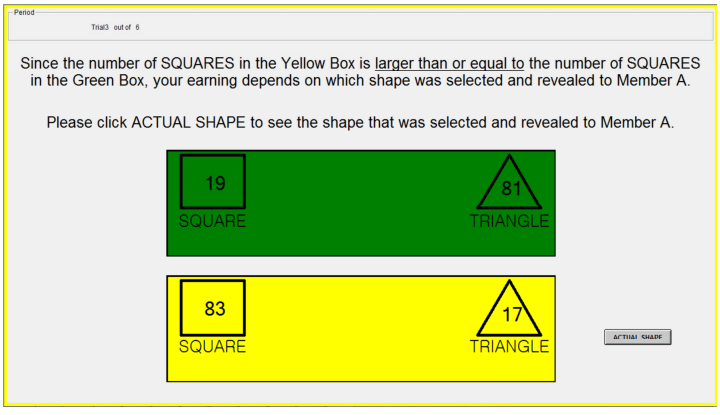


Figure 8: Member B's Reward Screen