

NBER WORKING PAPER SERIES

SECURE SURVEY DESIGN IN ORGANIZATIONS:
THEORY AND EXPERIMENTS

Sylvain Chassang
Christian Zehnder

Working Paper 25918
<http://www.nber.org/papers/w25918>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2019

Chassang gratefully acknowledges funding from Princeton University's Griswold center. Zehnder gratefully acknowledges funding from the Swiss Science Foundation under Grant No. 100018 152903. The paper greatly benefited from discussions with Andrew Caplin, Ernst Fehr, Guillaume Fréchette, Daniel Gottlieb, Tatiana Homonoff, Michel Maréchal, Ariel Rubinstein, Andy Schotter, Larry Samuelson and Heather Sarsons, as well as seminar audiences at Duke, Middlesex University, Lausanne, Munich, Zurich, Hamburg, Essex, NYU, Sciences Po, UC Santa Barbara, UC Berkeley, USC, and WashU. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2019 by Sylvain Chassang and Christian Zehnder. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Secure Survey Design in Organizations: Theory and Experiments
Sylvain Chassang and Christian Zehnder
NBER Working Paper No. 25918
June 2019
JEL No. C72,C81,C92,D23,D73,D74,D82,D86

ABSTRACT

We study the impact of secure survey designs ensuring plausible deniability on information transmission in organizations. We are interested in settings in which fear of retaliation makes potential informants reluctant to reveal the truth. Theory predicts that: (i) popular randomized-response designs fail to induce informative reports, because they are strategically equivalent to non-secure direct-elicitation designs; (ii) hard-garbling designs that exogenously distort survey responses improve information transmission; and (iii) unbiased estimates of the impact of survey design on information transmission can be obtained in equilibrium. Laboratory experiments qualify these predictions. While hard-garbling does improve information transmission over direct-elicitation, other predictions fail: randomized response performs much better than expected; and false accusations lead to a small but persistent bias in treatment effect estimates. We show that these deviations from equilibrium can be accounted for in an off-the-shelf model of boundedly rational play, and that this model of play makes specific predictions over the bias of treatment effect estimators. Additional experiments reveal that play converges to equilibrium if players can (socially) learn from cross-sectional data. These results suggest that randomized response cannot be used systematically in organizational settings, whereas hard garbling improves survey quality even under long-run equilibrium conditions.

Sylvain Chassang
Department of Economics
New York University
19 West 4th Street
New York, NY 10012
and NBER
sylvain.chassang@gmail.com

Christian Zehnder
Faculty of Business and Economics
University of Lausanne
Quartier UNIL-Chamberonne
Internef 612
CH-1015 Lausanne
Switzerland
christian.zehnder@unil.ch

1 Introduction

This paper studies the effectiveness of different survey designs in eliciting sensitive information in an organizational context. Getting early information about potential misbehavior or fraud is a key challenge for organizations. As a result, they often expend considerable efforts to improve information transmission: from 360° evaluations, to compliance monitoring, to external anonymous hotlines. However, transmitting sensitive information (e.g. fraudulent or inappropriate behavior by one’s boss) remains challenging.¹ Our goal in this paper is to better understand how survey designs meant to guarantee plausible deniability can improve the transmission of information. Among others, we are interested in the effectiveness of randomized-response surveys and related designs (Warner, 1965, Greenberg et al., 1969) used in social sciences to ask sensitive questions. We place special emphasis on the problem of identifying treatment effects from noisy and unverifiable reports.

Our model broadly follows Chassang and Padró i Miquel (2018). We consider a Principal-Agent-Monitor framework in which the monitor can report whether the Agent’s type is Good (G) or Bad (B). Reporting a Bad agent improves the welfare of both the monitor and the principal, but is costly to the agent. The difficulty is that the agent can commit to retaliate against the monitor in the event an incriminating report is recorded. This model captures environments in which: (i) the number of potential informants is small, so that they can be subject to retaliation in spite of nominal anonymity; (ii) corrective actions can be taken against Bad agents even on the basis of noisy information²; and (iii) potential informants do not have strong incentives to make false accusations.

We model survey design as a (possibly random) mapping from monitor reports r to survey records \tilde{r} that may be equal or different from the monitor’s report. We explore the impact

¹Kaplan and Schultz (2007) and Kaplan et al. (2009) argue that anonymous external reporting channels mandated by the Sarbanes-Oxley act failed to increase intention-to-report rates. As Chassang and Padró i Miquel (2018) observe: “this can be explained by the fact that in many cases the set of people informed about misbehavior is small, so that formal anonymity offers little actual protection.”

²The noisiness of information (specifically the fact that signals of misbehavior may include false positives) clearly constrains the severity of organizational responses, but effective action is still plausible. A manager may be required to undergo sensitivity training. An internal investigation may be triggered, and so on.

of three different survey designs on information transmission:

1. A direct-elicitation design **DE** in which the survey record is equal to the response of the agent to the question

Q0: “True or False: The agent’s type is Bad.”

2. A randomized-response design **RR** following Warner (1965) and Greenberg et al. (1969): With probability $1 - \pi$, the survey record is the response of the agent to the question Q0. With probability π , the agent is asked the unrelated question

Q1: “What is the color of the sky: Blue or Red?”

The survey record is *True* if the response is Blue, and *False* if the response is Red.

3. A hard-garbling design **HG** in which the survey record is equal to the monitor’s response to question Q0 with probability $1 - \pi$, and to *True* with probability π .

The rationale behind designs **RR** and **HG** is the following: because a recorded report $\tilde{r} = \textit{True}$ (incriminating the agent) can occur even though the monitor sends the preliminary report $r = \textit{False}$, it becomes costly to threaten the monitor with punishment in the event of record $\tilde{r} = \textit{True}$. Costly punishment (at least seemingly) occurs on the equilibrium path. The crucial difference between the two mechanisms is that **RR** relies on the monitors’ compliance with the injunction to answer the unrelated question correctly, whereas **HG** imposes an exogenous information garbling. Our main questions of interest are: Does garbling the information content of messages improve reporting in practice? Can one build data-driven estimates of treatment effects permitting cost-benefit analysis? Which mechanism is likely to be most effective depending on the circumstances?

The paper proceeds in three steps. First, we clarify theoretical predictions from a strategic analysis of the games induced by different survey designs. We establish that in equilibrium, direct elicitation **DE** fails to sustain information flows when retaliation costs are high. In contrast, hard garbling **HG** can sustain informative reports, even when the magnitude of retaliation is large. The case of randomized response **RR** is ambiguous. If it is common knowledge that monitors are *obedient* and give the objectively correct answer to unrelated

question Q1, then RR is strategically equivalent to HG. If it is common knowledge that monitors answer the unrelated question Q1 to maximize their payoffs, then RR is strategically equivalent to DE. A plausible guess is that play under RR will be a mixture of play under DE and HG, perhaps starting close to HG during initial play, and approaching DE as the players gain experience. Finally, we show that it is possible to construct an estimator of the impact of survey design on the reporting of Bad agents. This estimator is unbiased whenever monitors are rational, but may be biased if monitors are irrational and falsely accuse Good agents with different propensities across survey designs. As a result, understanding deviations from equilibrium is an important input for inference.

Second, we evaluate our theoretical predictions in the lab. Experimental findings confirm that indeed, hard garbling HG can improve reporting against Bad agents compared to direct elicitation DE, but—as expected—this comes at the cost of inefficient intervention against Good agents. Contrary to expectations, the frequency of reports against Good agents is not zero and does not vanish over time. As a result our estimate of treatment effects exhibits a small amount of bias. Also contrary to expectations, behavior under randomized response RR is not a simple mixture of behavior under DE and HG. As in the case of HG threats by agents against monitors less frequent than under DE. Unlike the case of HG, monitors rarely report Good agents when they are asked the unrelated question Q1. As a result, RR gets the best of both HG and DE: frequent reporting of Bad agents, and infrequent reporting of Good agents. This pattern does not vanish over time.

Third, we investigate our experimental data through the lens of an off-the-shelf model of boundedly strategic play, adapting the Quantal Response Level- k model of Camerer et al. (2016) to our dynamic setting.³ Since DE and RR are strategically equivalent up to framing effects, differences in behavior between these two games are interpreted as differences in the effective rationality of players. Specifically, agents have a harder time predicting the

³See Stahl (1993) and McKelvey and Palfrey (1995) for seminal work on Level- k and quantal response models, as well as Camerer et al. (2004) and Crawford and Iriberri (2007) for an investigation of their performance as predictors of empirical play in games and mechanisms. See Ho and Su (2013) for an extension of cognitive hierarchy models to dynamic settings.

response of monitors to the threat of retaliation under RR than DE. We show that this non-equilibrium behavior has unambiguous implications about bias in treatment effect estimates. In addition, it suggests that the good performance of RR in our data may not be robust to learning. While players do not seem to learn from their own experience in our first set of experiments, a second set of experiments reveals that providing players with cross-sectional outcome data (as would arise from social learning) induces convergence to equilibrium. This convergence has two important consequences. First, false accusations against Good agents are diminished, so that the bias in the estimated treatment effect of HG on the reporting rate against Bad agents disappears almost completely. Second, equilibrium threats and reporting behavior under RR and DE converge, so that RR no longer outperforms DE. This cautions against the use of randomized-response designs RR in stable organizational settings. Hard-garbling designs, in contrast, improves information transmission even when players obtain feedback from cross-sectional data.

This paper contribute to multiple strands of literature. First, it contributes to a large body of work on garbled survey design, and inference from garbled surveys. Randomized response techniques (and its many variants) were pioneered by Warner (1965) and Greenberg et al. (1969) as a means to elicit higher quality information on sensitive topics.⁴ Recently, Blair et al. (2015) has provided an analysis of identification under such designs, but assumes that key behavioral aspects of the response are known. Rosenfeld et al. (2016) offers a field validation of randomized response techniques in the context of a vote on abortion law taking place in Mississippi. On the negative side, John et al. (2018) report evidence from a series of experiments showing that randomized-response techniques can backfire if respondents are concerned over response misinterpretation. Chuang et al. (2019) also provide evidence that respondents often disregard the injunction to privately garble their response, and simply provide the least sensitive answer. The current paper tackles theoretically and empirically the problem of identification when, response choices are endogenous, and take place in an

⁴The related computer science literature on differential privacy (Dinur and Nissim, 2003, Dwork et al., 2006) studies ways to reveal population statistics without compromising individual anonymity.

organizational setting that microfound the why some answers are considered sensitive.

The paper also fits in a literature that seeks to better understand how information transmission protocols affect strategic information transmission in an equilibrium setting. Chassang and Padró i Miquel (2018) is closest to our approach, and considers the problem of whistleblowing when misbehaving agents can commit to effective threats.⁵ Blume et al. (2013) proposes a signaling model with lying aversion in which randomized response techniques can improve information transmission in equilibrium. Lab implementation, including direct costs for lying, supports their theoretical predictions. Ayres and Unkovic (2012) consider the problem of whistleblowing when many parties are informed but face an incomplete information coordinated action problem in view of possible retaliation. They advocate the use of information escrow accounts as a way to solve the coordination problem. Mechtenberg et al. (2017) is concerned with the problem of false reporting in the context of whistleblowing. In a lab experiment, they show that whistleblower protection based on facts rather than subjective evidence may increase informativeness. This work echoes our concerns regarding the false reporting of Good agents.⁶

Finally, the paper contributes to the growing literature on mechanism design with boundedly rational or behavioral players. Crawford and Iriberri (2007) explore the extent to which Level- k models can explain overbidding in auctions. Fehr et al. (2017) show that social preferences have a significant impact on the design of extensive form mechanisms. The work of Glazer and Rubinstein (2014) is also closely connected. It suggests that complex questionnaires can induce boundedly rational players to reveal their type. The surprising effectiveness of randomized response RR suggests the idea may be even more effective in extensive form strategic environments: a player's naïve beliefs over the behavior of others may be used to attain outcomes that are not implementable under common knowledge of rationality. This observation echoes the work of Fudenberg and Levine (2006) on effective superstitions.

⁵See also Ortner and Chassang (2018) for an analysis of how endogenous asymmetric information can improve information transmission when the agent can bribe potential informants.

⁶Note that Mechtenberg et al. (2017) focus on environments where informants strictly benefit from false accusations, while we concentrate on settings without material incentives to misreport. The fact that false reporting occurs even in our setting suggests that the issue cannot be ignored.

The paper is organized as follows. Section 2 describes our framework. Section 3 establishes benchmark results on equilibrium play and the identification of treatment effects. Sections 4, 5, and 6 take these theoretical predictions to experimental data and explore failures of equilibrium predictions. Section 7 interprets experimental findings and probes the data using an off-the-shelf model of bounded rationality. Section 8 concludes with a discussion of alternative interpretations, and implications for survey design in organizations.

2 The Reporting Game

Players, actions and payoffs. We consider a Principal-Agent-Monitor setting closely related to the framework of Chassang and Padró i Miquel (2018). An agent A affects the welfare of stakeholders S as a function of her type $\tau \in \{G, B\}$, where types G and B are respectively referred to as Good and Bad. Let $q \in (0, 1)$ denote the share of Bad agents in the population.

An informed monitor M observes the type τ of agent A and chooses a report r from some set R . Intended reports $r \in R$ lead to realized reports $\tilde{r} \in \{0, 1\}$ according to transmission channels $\phi : R \rightarrow \Delta(\{0, 1\})$ that will be described shortly. Note that the set of possible reports R may be different from the set of received reports $\{0, 1\}$. A received report $\tilde{r} = 1$ automatically leads to a publicly observed intervention.

The agent can affect the behavior of the monitor by committing to punish her in the event a positive realized report $\tilde{r} = 1$ (and the corresponding intervention) occurs. The agent's commitment decision is denoted by $c \in \{0, 1\}$.⁷ Successfully reporting a Bad agent allows the organization to reduce her negative impact on both the monitor and stakeholders. Payoffs U_A , U_M and U_S to the agent, monitor and stakeholders take the form

$$U_A = -\tilde{r}D - c\tilde{r}K_A; \quad U_M = Y - c\tilde{r}K_M; \quad U_S = Y,$$

⁷Commitment may arise from external reputational concerns vis à vis other monitors and other agents, or through an internal reputation, for instance, the self image of delivering on one's word.

with output $Y = Y_0 - (1 - \gamma\tilde{r})\mathbf{1}_{\tau=B}L_B - \tilde{r}\mathbf{1}_{\tau=G}L_G$, where Y_0 is the maximum output, L_B represents the efficiency loss from facing a Bad agent, γL_B is the amount by which that loss can be reduced if the Bad agent is reported, and L_G is the efficiency loss from reporting a Good agent. Parameters D , K_A , and K_M respectively denote the agent's losses given report $\tilde{r} = 1$, her cost for punishing the monitor, and the damage incurred by the monitor in case of punishment.

Finally, with experimental investigation in mind, we allow for social preferences on the side of the monitor, so that her actual preferences are described by $U_M^\alpha \equiv U_M + \alpha U_S$, where α is an altruism parameter.

Reporting mechanisms. The focus of our analysis is the mapping ϕ from intended reports r to realized reports \tilde{r} . We are interested in the following three possibilities.

- (i) *Direct Elicitation DE*. Possible reports take values in $R = \{0, 1\}$. The realized report \tilde{r} is equal to the intended report r . The monitor is asked to report the type of the agent, and her report is recorded without alteration.
- (ii) *Hard Garbling HG*. Under hard garbling, possible reports take values in $R = \{0, 1\}$. If $r = 1$, then $\tilde{r} = 1$. If instead $r = 0$, then $\tilde{r} = 0$ with probability $1 - \pi$, and $\tilde{r} = 1$ with probability π . The monitor is asked to report the type of the agent. Whenever she sends report $r = 0$, her answer is *exogenously* switched to $\tilde{r} = 1$ with probability π .
- (iii) *Randomized Response RR*. Randomized response techniques, pioneered by Warner (1965), take multiple forms. We focus on the frequently used unrelated-question implementation. More precisely, with probability $1 - \pi$, the report space is $R = \{0, 1\}$ and realized report is $\tilde{r} = r$. The monitor is asked to report the type of the agent, and that response is recorded without alteration. With probability π , $R = \{Blue, Red\}$, $\tilde{r} = \mathbf{1}_{r=Blue}$. Additionally, this reporting problem is framed with the unrelated question: “What is the color of the sky?”

For expositional reasons, it is useful to define an *obedient* version of the randomized-response design game, denoted by **oRR**, in which *the monitor is compelled to answer the unrelated question truthfully*.

Solution Concept. The reference solution concept for our analysis is Subgame Perfect Equilibrium (SPE). We consider alternative solution concepts, including level- k models (Stahl, 1993, Nagel, 1995, Camerer et al., 2004, Crawford and Iriberri, 2007, Ho and Su, 2013), and Quantal Response models (McKelvey and Palfrey, 1995) when we analyze experimental data. We bring up self-confirming equilibrium (Fudenberg and Levine, 1993) when we investigate learning.

3 Theoretical predictions

We first characterize equilibrium behavior, and provide benchmark identification results, under the assumption that players are rational.

3.1 Equilibrium behavior

Proposition 1 (garbling and retaliation). *Take parameters of the direct elicitation, and hard garbling games as given, except for punishment costs to the agent and the monitor K_A and K_M . There exists $K > 0$ sufficiently large such that whenever $K_A, K_M > K$,*

- (i) *under direct elicitation: all SPEs are such that Bad agents commit to punish; the monitor sends report $r = 0$;*
- (ii) *under hard garbling: there exists a unique SPE; the agent does not commit to punish; the monitor reports a Bad agent but not a Good one.*

In both games, regardless of costs of punishment K_A and K_M , it is never rational for the monitor to report Good agents.

The key force behind this result is that following a threat, sufficiently large punishments are off of the equilibrium path under direct elicitation DE, but remain on the equilibrium path under hard garbling HG.

Definition 1 (outcome equivalence). *We say that two survey games G_0 and G_1 (corresponding to survey technologies ϕ_0, ϕ_1) are outcome equivalent in SPE if for every SPE σ_0 of G_0 (respectively G_1), there exists a SPE σ_1 of G_1 (respectively G_0) such that σ_0 and σ_1 induce the same joint distribution over triplets (τ, c, \tilde{r}) corresponding to agent type, commitment to punish, and realized report.*

Proposition 2. *The randomized-response game RR is outcome equivalent in SPE to the direct elicitation game DE .*

The obedient randomized-response game oRR is outcome equivalent in SPE to the hard garbling game HG .

This implies that if players are obedient and answer the unrelated question correctly, randomized response and hard garbling should lead to the same distribution over type τ , commitment to punish c , and realized report \tilde{r} . If instead players are not obedient, randomized response should have led to the same distribution over outcomes as direct elicitation. If players are in-between, it is reasonable to expect that outcomes under randomized response should be a mixture of outcomes under hard garbling, and direct elicitation. Together Propositions 1 and 2 suggest that the reporting of Good and Bad agents across games will follow the patterns summarized in Table 1. We confront these predictions to experimental data in Section 4.

Table 1: anticipated realized reporting across different games.

	realized reporting of Bad types	realized reporting of Good types
DE	low	low
HG	high	high
RR	low	low
oRR	high	high

3.2 Measurement and identification

Lab experiments allow to measure the impact of survey design on reporting because the experimenter gets to observe the true type of each agent. This is not possible in the field since types τ are not observed, and neither is the proportion q of Bad types. Still, as Chassang and Padró i Miquel (2018) emphasize, under assumptions about play, it is possible to recover underlying parameters of interest from reporting data alone. Lab experiments can then serve to test the validity of our estimators under realistic play.

Let μ denote a distribution over (τ, c, r) : the agent's type $\tau \in \{Good, Bad\}$, her commitment to retaliate $c \in \{0, 1\}$, and intended reports by the monitor $r \in R$. Distribution μ may or may not be an equilibrium. Consider an i.i.d. sample of play of size N drawn from μ . Let $R_\tau = \mathbb{E}_\mu \left[\frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\tau_i=\tau} r_i \right]$ denote the mean of *intended* reports against agents of type τ . Let $\widehat{R} = \frac{1}{N} \sum_{i=1}^N \widetilde{r}_i$ denote the aggregate sample-mean of *realized* reports. Let $\widetilde{R}_\tau = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\tau_i=\tau} \widetilde{r}_i$ denote the sample mean of realized reports against agents of type τ . Note that only the sample mean \widehat{R} is observable to an external observer. For any $\tau \in \{G, B\}$, R_τ and \widetilde{R}_τ are not directly accessible to the observer. By convention, we use $\widehat{\cdot}$ to indicate that a variable is computed using data available to the econometrician.

We are interested in the following underlying outcomes:

- The amount of intended reports against Bad agents R_B .
- The share of realized reports against Good agents, $\mathbb{E}_\mu \widetilde{R}_G$.
- The share of unintended incriminating reports $\widetilde{r} = 1$, $R_\dagger \equiv \mathbb{E}_\mu \widehat{R} - R_B - R_G$.

Proposition 3 (identification). *Under the hard garbling game HG , for any distribution μ over play, μ -almost-surely, the following hold:*

$$R_B = \lim_{N \rightarrow \infty} \frac{\widehat{R} - \pi}{1 - \pi} - R_G \quad (1)$$

$$R_\dagger = \lim_{N \rightarrow \infty} (1 - \widehat{R}) \frac{\pi}{1 - \pi}. \quad (2)$$

If play by the monitor is rational, then $R_G = 0$, and $\mathbb{E}_\mu \tilde{R}_G \leq R_\dagger$. If all Bad agents are reported, i.e. $R_B = q$, then the last inequality is tight: $\mathbb{E}_\mu \tilde{R}_G = R_\dagger$.

The case of direct elicitation DE corresponds to setting $\pi = 0$ under hard garbling HG. Proposition 3 implies that R_B and R_\dagger are identified provided the monitor is rational. The rate of realized complaints against Good agents $\mathbb{E}_\mu \tilde{R}_G$ need not be identified even if the monitor is rational. The reason for this is that the share q of Good agents is not known. However, R_\dagger , which can be estimated from data, is a tight upper bound to $\mathbb{E}_\mu \tilde{R}_G$.

This lets us use $\hat{R}_B \equiv \frac{\hat{R} - \pi}{1 - \pi}$ as an estimator of the mass of intended reports R_B against Bad agents. Under equilibrium, this is a consistent estimator of R_B . If we deviate from equilibrium, the estimator may be upward biased since by construction $R_G \geq 0$. Whether \hat{R}_B is consistent depends on whether the monitor is rational and chooses not to report Good agents. This is ultimately an empirical question that experiments are well suited to inform.

The statistic $\hat{R}_\dagger \equiv \frac{\pi}{1 - \pi}(1 - \tilde{R})$ is a mechanically consistent estimator of R_\dagger , regardless of whether play is in equilibrium or not. For this reason our investigation of experimental results focuses on the consistency of \hat{R}_B .

Treatment effects. In practice, to perform cost-benefit analyses, we are primarily interested in computing the treatment effect of different reporting mechanisms on the reporting of Bad agents:

$$\hat{\Delta}R_B = \hat{R}_B^{\text{treatment}} - \hat{R}_B^{\text{control}}.^8 \quad (3)$$

The potential bias of $\hat{\Delta}R_B$ is equal to $R_G^{\text{treatment}} - R_G^{\text{control}}$. Even if the monitor is not rational, so that $R_G > 0$, the treatment effect estimators $\hat{\Delta}R_B$ need not be biased if the non-rational false reporting of Good agents is independent of the reporting mechanism. Inversely, the estimator may be inconsistent if deviations from equilibrium play differ systematically across games. In Section 7 we evaluate potential bias out-of-equilibrium using an off-the-shelf

⁸The impact of treatment on undesired reports $\hat{\Delta}R_\dagger = \hat{R}_\dagger^{\text{treatment}} - \hat{R}_\dagger^{\text{control}}$ can be recovered mechanically.

model of boundedly rational play.

4 Experimental Design

The analysis of Section 3 clarifies the impact of different survey designs under equilibrium, and shows it is possible to estimate the impact of survey design on reporting using realized reporting data alone. This begs the question: what happens if players do not behave according to equilibrium? More specifically:

- Does garbling increase the reporting of Bad types?
- If there are deviations from equilibrium play, how do they affect the bias and consistency of treatment effect estimator $\hat{\Delta}R_B$?
- Are deviations from equilibrium play amenable to modeling?
- How long-lived are deviations from equilibrium play?

We use experimental data on play in survey games to address these questions.

The baseline game. We recruited 20 participants for each session of our experiment. Half of the participants were randomly allocated to the role of agents, the other half were assigned the role of monitors. Participants remained in their roles for all 25 periods of the experiment.⁹ At the beginning of each of the 25 identical periods of the game, participants are randomly re-matched into $n = 10$ agent-monitor pairs and agents are randomly allocated one of two equally likely types $\tau \in \{G, B\}$. To avoid fully passive participants in the laboratory, we did not have separate subjects representing stakeholders. Instead other monitors play the role of stakeholders, so that the agent's type directly affects the welfare of all monitors. After having observed their types, agents decide whether or not to commit $c \in \{0, 1\}$ to punish the monitor if the monitor's realized report turns out to be positive ($\tilde{r} = 1$). Monitors

⁹This choice reflects our belief that role reversals in power relationships are infrequent. We expect that role reversals would improve convergence to equilibrium, but the question is ultimately an empirical one.

observe the commitment decision of the agent they have been paired with and then pick their intended report $r \in R$.

Agent i 's contribution to total output $Y \equiv \sum_{i=1}^n y_i$ is

$$y_i = y_G - (1 - \gamma\tilde{r})\mathbf{1}_{\tau=B}L_B - \tilde{r}\mathbf{1}_{\tau=G}L_G,$$

where y_G is the output of an unreported Good agent (the maximum output), $L_B > L_G$ represents the efficiency loss created by an unreported Bad agent, $\gamma \in [0, 1]$ measures the extent to which reporting reduces the efficiency loss caused by a Bad agent and L_G represents the efficiency loss from reporting a Good agent.

Payoffs. Payoffs to participants are an affine transformation of the payoffs described in Section 2, and take the following form (payoffs in the experiment were calculated in points; at the end of the experiment points earned by participants are converted into cash using the exchange rate: 60 points = 1 Swiss Franc):

$$U_A = E_A - \tilde{r}D - c\tilde{r}K_A; \quad U_M = E_M + \frac{Y}{n} - c\tilde{r}K_M,$$

To avoid negative payoffs agents and monitors respectively received endowments $E_A = 80$ points and $E_M = 30$ points at the beginning of every period. An unreported Good agent's contribution to total output was $y_G = 400$ points. An unreported Bad agent generated a damage of $L_B = 800$ points (i.e., the net contribution to total output amounts to $y_G - L_B = -400$ points). Reporting a Bad agent reduced the implied damage by 37.5% ($\gamma = .375$). Reporting a Good agent reduced her contribution to total output by $L_G = 100$ points. The agent's payoff loss in case of a positive realized report was $D = 100$ points regardless of her type. The agent's cost for punishing the monitor was set to $K_A = 100$ points, and the damage imposed on the monitor by the agent's punishment was $K_M = 200$.

Treatments. We implemented three treatment conditions, corresponding to each of the reporting mechanisms HG, DE and RR. The probability with which negative intended reports were transformed into positive realized reports in the hard garbling treatment HG and with which monitors in the randomized response treatment RR were confronted with the unrelated question was set to $\pi = .25$.

Protocol. All subjects were students of the University of Lausanne (UNIL), the Swiss Federal Institute of Technology Lausanne (EPFL) or the Swiss Hotel Management School (EHL). We used the recruitment system ORSEE (Greiner, 2015). Each subject participated in one session only. All interactions of participants were completely anonymous. The experiment was programmed and conducted with z-Tree (Fischbacher, 2007).

To make sure that subjects fully understood the payoff consequences of available actions, each subject had to read a detailed set of instructions before the session started. Participants then had to answer several questions about feasible actions and their payoff consequences. A session started only after all subjects had correctly answered all questions.¹⁰

The data was collected in two waves. From March to May 2015, we conducted 6 sessions for each treatment without providing players further information about play. To investigate learning dynamics, from October to November 2015, we conducted another 6 sessions per treatment in an environment where players received cross sectional data about the payoff consequences of different actions. Overall we conducted 36 sessions with a total of 720 participants. Each session lasted approximately 90 minutes and subjects earned on average about 37 Swiss Francs (including a show-up fee of 10 Swiss Francs).

5 Findings for Direct Response and Hard Garbling

This section focuses on play in the direct elicitation and hard garbling games DE and HG, while Section 6 studies play under randomized response RR. We first describe the aggregate

¹⁰See Appendix D for an English translation of a sample of our French instructions (including control questions).

impact of survey design on outcomes before investigating the specific mechanics of threats and reporting.

5.1 Output and punishment

Proposition 1 establishes the main mechanism through which hard garbling can improve information flows relative to the direct elicitation method. If the punishment technology is effective (K_M sufficiently large), the equilibrium under DE is such that Bad agents always commit to punish and monitors therefore find it optimal to submit a negative report $r = 0$ irrespective of the agent’s type. As a consequence, punishment never occurs and no information is elicited in equilibrium. Under HG, in contrast, agents who commit to punish need to pay the cost of punishment K_A with positive probability. As a consequence, the expected cost of punishment increases and agents have weaker incentives to commit to punish. Reduced punishment threats, in turn, increase monitors’ propensity to submit positive reports $r = 1$ if they encounter Bad agents.

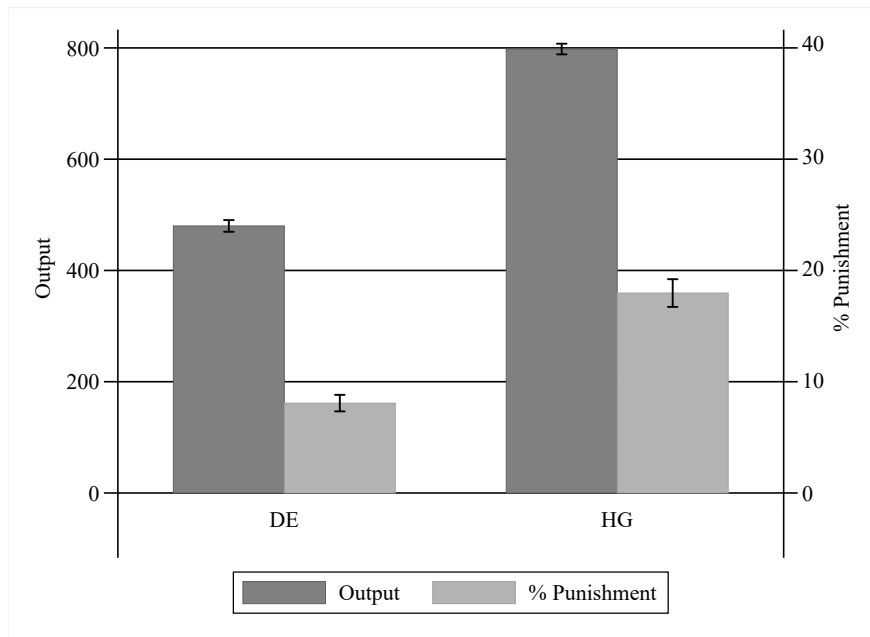
Hard garbling is expected to impose two types of costs on the elicitation system. First, garbling implies that some negative reports submitted for Good agents $r = 0$ will get turned into positive realized reports $\tilde{r} = 1$. Second, if punishment is not reduced to zero, garbling will trigger actual punishment on the equilibrium path. In practice, a principal would have to balance the value of greater information flows regarding Bad agents against these costs.

Figure 1 displays average output and the overall frequency of punishment in the direct response and the hard garbling treatments. On average, HG leads to higher output (798 points) than DE (480 points). This increase of about 66% is statistically significant (OLS: $p = 0.001$, RS: $p = 0.013$).¹¹ Punishment frequencies are positive in both treatments, which contradicts the prediction that there would be zero punishment under direct response. However,

¹¹For all treatment comparisons we report two tests. We perform OLS regressions in which we regress the dependent variable on treatment dummies. We perform these regressions using the data of all six treatments (i.e. including the three treatments reported later in the paper). Standard errors are adjusted for clustering at the session level (36 clusters). In addition, we also perform ranksum tests (RS) in which we use session averages as independent variables.

punishment occurs more often under hard garbling (18%) than under direct elicitation (8%). The effect of garbling on observed punishment is statistically significant (OLS: $p < 0.001$, RS: $p = 0.004$). We explore the mechanics underlying these findings below.

Figure 1: impact of reporting mechanism on output and punishment



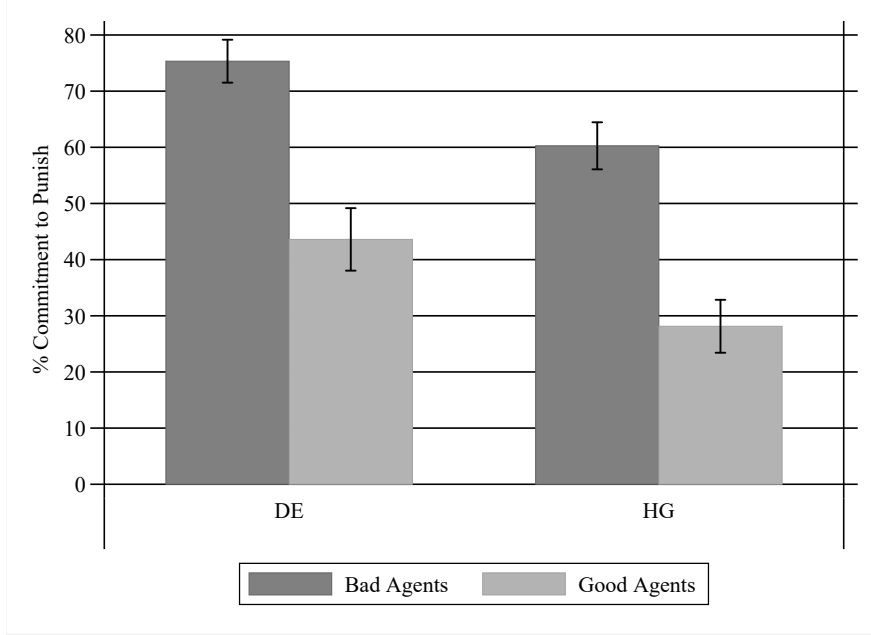
Note: The figure displays average output and the overall frequency of punishment under DE and HG. The variable *Output* corresponds to average per-period output at the session level. The maximally possible per-period output amounts to 1500 points. The variable *% Punishment* represents the percentage of agent-monitor pairs in which punishment occurred. Error bars mark ± 1 standard error from the mean (clustering at the individual level).

5.2 Commitment to punish

Figure 2 displays the frequency with which Good and Bad agents commit to punish under direct response and hard garbling. As expected, agents commit to punish less often under HG. Bad agents commit to punish with probability 75% under DE and probability 60% under HG. This reduction in the frequency of punishment threats is statistically significant (OLS: $p = 0.004$, RS: $p = 0.016$). Figure 3 shows that the frequency of threats by Bad agents

increases moderately over time under both HG and DE. However, the time trend (OLS) is only significant in DE (DE: $\beta = 0.007$, $p = 0.001$; HG: $\beta = 0.004$, $p = 0.234$).

Figure 2: agent's commitment to punish, by reporting mechanism, and agent type

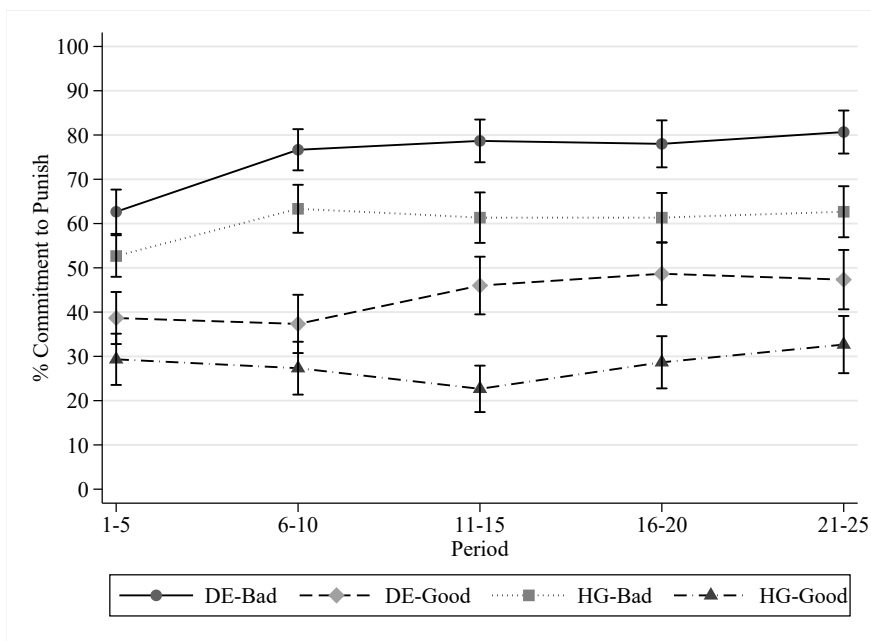


Note: The figure displays the frequency with which Good and Bad agents commit to punish under DE and HG. The variable *% Commitment to Punish* measures the within-type percentage of agents who commit to punish across all periods of a treatment. Error bars mark ± 1 standard error from the mean (clustering at the individual level).

Good agents also commit to punish rather frequently in both treatments. Under DE, Good agents make threats in 44% of cases. Under HG Good agents make threats in 28% of cases. The fact that Good agents commit to punish under DE is not surprising (it can occur in SPE), but committing to punish under HG is not consistent with equilibrium. Recall that it is strictly optimal for the monitor to submit a negative report $r = 0$ when paired with a Good agent. Given this, in equilibrium, a Good agent should strictly prefer not to commit to punish since she would have to pay cost K_A with positive probability even if though the monitor submits report $r = 0$. Thus, although hard garbling reduces Good agents propensity to make threats (OLS: $p = 0.040$, RS: $p = 0.128$), the remaining presence of punishment

commitments under HG is a notable failure of equilibrium. This behavior does not appear to diminish over time (see Figure 3). The point estimate of time trends under direct elicitation is positive ($\beta = 0.007$, $p = 0.036$), and there is essentially no trend under hard garbling ($\beta = 0.002$, $p = 0.697$). A possible reason for why Good agents commit to punish may be the fear of false accusations by (irrational) monitors. We return to this point when discussing the reporting behavior of monitors in the next subsection.

Figure 3: agent's commitment to punish over time



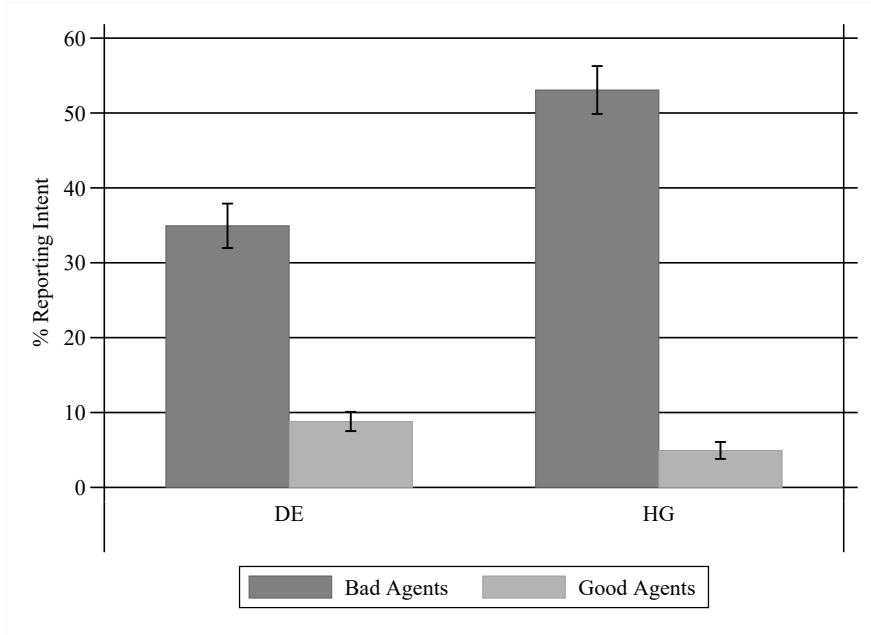
Note: The figure displays time trends in the frequency with which Good and Bad agents commit to punish under DE and HG. The variable *% Commitment to Punish* measures the within-type percentage of agents who commit to punish. Error bars mark ± 1 standard error from the mean (clustering at the individual level).

5.3 Reporting

Intended reporting. Figure 4 summarizes monitors' intended reporting irrespective of the agent's punishment commitment. Hard garbling improves intended reporting relative to direct elicitation. In particular, the frequency with which monitors intend to submit positive

reports $r = 1$ on Bad agents is significantly higher under hard garbling (53%) than under direct response (35%, OLS: $p = 0.001$, RS: $p = 0.013$). The difference in the reporting rates for Bad agents further increases over time. Whereas the reporting rate decreases only insignificantly under hard garbling ($\beta = -0.003$, $p = 0.390$), there is a significant negative time trend under direct elicitation ($\beta = -0.009$, $p < 0.001$).¹²

Figure 4: monitors' average intended reports conditional on the agent's type



Note: The figure summarizes monitors' intended reporting irrespective of the agent's punishment commitment. The variable *%Reporting Intent* measures the frequency with which monitors intend to submit a positive report $r = 1$ against an agent as a function of the agent's quality and the treatment (DE vs. HG). Error bars mark ± 1 standard error from the mean (clustering at the individual level).

Despite the fact that monitors have a strict incentive to submit a negative report $r = 0$ if matched with a Good agent, positive reports $r = 1$ against Good agents are observed under both HG (5%) and DE (9%). The presence of these false accusations in both treatments may at least partially explain why Good agents commit to punish (see Figure 2). Positive

¹²In the final five periods monitors intend to submit positive reports on Bad agents in 53% of the cases under HG and in 27% of the cases under DE (OLS: $p < 0.001$, RS: $p = 0.006$).

reporting rates against Good agents also imply that our estimator of the mass of intended reports against Bad agents \hat{R}_B (defined Section 3.2) is upward biased. At the same time, however, the difference between these reporting rates is not statistically significant (OLS: $p = 0.237$, RS: $p = 0.423$). We analyze the implications of these reporting rates for our treatment effect estimator $\hat{\Delta}R_B$ in detail in the next section.

A more detailed analysis of the reporting data reveals that monitors paired with Bad agents often submit positive reports $r = 1$ even if the agent committed to punish. When viewed through the lens of our theory, such behavior suggests that monitors exhibit significant altruism α towards other monitors. We observe the phenomenon in both treatments, but the inclination to report threatening agents turns out to be significantly larger under hard garbling (29%) than under direct elicitation (19%, OLS: $p = 0.001$, RS: $p = 0.004$).¹³ Positive reports against Good agents who committed to punish exist, but are rare (HG: 3%, DE: 4%). The greater reporting of Bad agents who have committed to punish in HG relative to DE also adds to the effect of information garbling on observed punishment. In fact, although fewer agents commit to punish under HG than under DE (see Figure 2), the fraction of monitors punished after intentionally reporting a threatening agent increases from 8% under DE to 9% under HG (the remainder of the aggregate punishment rate of 18% displayed Figure 1 is caused by the conversion of intended negative reports into positive realized reports under HG).

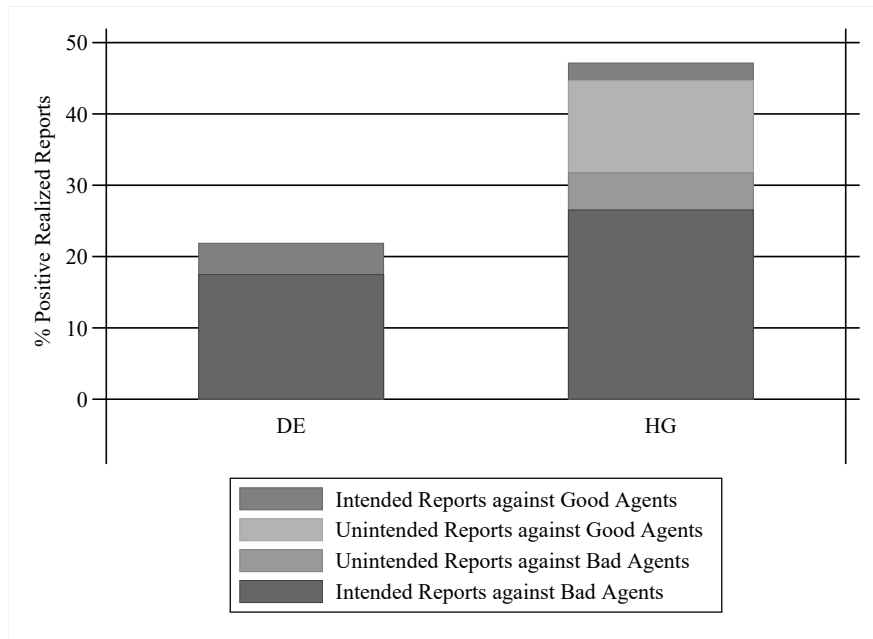
Realized reporting. The realized fraction of positively reported agents $\tilde{r} = 1$ is significantly higher under hard garbling (47%) than under direct elicitation (22%, OLS: $p < 0.001$, RS: $p = 0.004$). Garbling increases realized positive reports for two reasons. First, it increases the intended reporting of Bad agents (Figure 4), which increases the overall reporting rate by approximately 9 percentage points.¹⁴ Second, garbling increases the share

¹³Positive reporting of threatening Bad agents becomes less frequent over time in both treatments, but the trend is significant only under DE (DE: $\beta = -0.009$, $p < 0.001$, HG: $\beta = -0.004$, $p = 0.253$).

¹⁴The effect size reported here is half the size of the one reported in Figure 4 because we consider the overall reporting rate rather than the reporting rate within the subsample of Bad agents (which is 50% of the sample). Statistical significance is not affected by this scaling.

of realized reports $\tilde{r} = 1$ by turning 25% of intended negative reports into positive realized reports, which further increases the share of positive realized reports by 18 percentage points. Among those unintended positive reports roughly 13 percentage points are associated with Good agents, whereas the remaining 5 percentage points are associated with Bad agents. Figure 5 provides a graphical summary of the impact of garbling on reporting.

Figure 5: average realized reports



Note: The figure summarizes how garbling affects the composition of positive realized reports. The variable *%Positive Realized Reports* measures the frequency with which positive realized reports $\tilde{r} = 1$ against an agent are observed as a function of the treatment (DE vs. HG). The four categories distinguish whether a recorded report was intended or unintended (intended negative report that was turned into a positive realized report) and whether it was against a Good or a Bad agent.

5.4 Identification

The data lets us evaluate our identification assumptions, i.e. that there are no reports against Good agents in the case of estimator \hat{R}_B , or that reports against Good agents are independent of the survey design in the case of treatment-effect estimator $\hat{\Delta}R_B$.

The fact that monitors report Good types with positive probability (see section 5.3) means that our estimator \hat{R}_B of intended reports against Bad agents R_B is biased upwards. The difference in observed population fractions of reported Good agents across treatments (4.4% under DE and 2.5% under HG) imply that the treatment effect estimator $\hat{\Delta}R_B$ will be slightly biased as well.¹⁵

Table 2: estimated and true reporting rates

	estimator \hat{R}_B	true R_B	bias R_G
Direct Elicitation DE	21.9%	17.5%	4.4%
Hard Garbling HG	29.5%	26.5%	2.5%

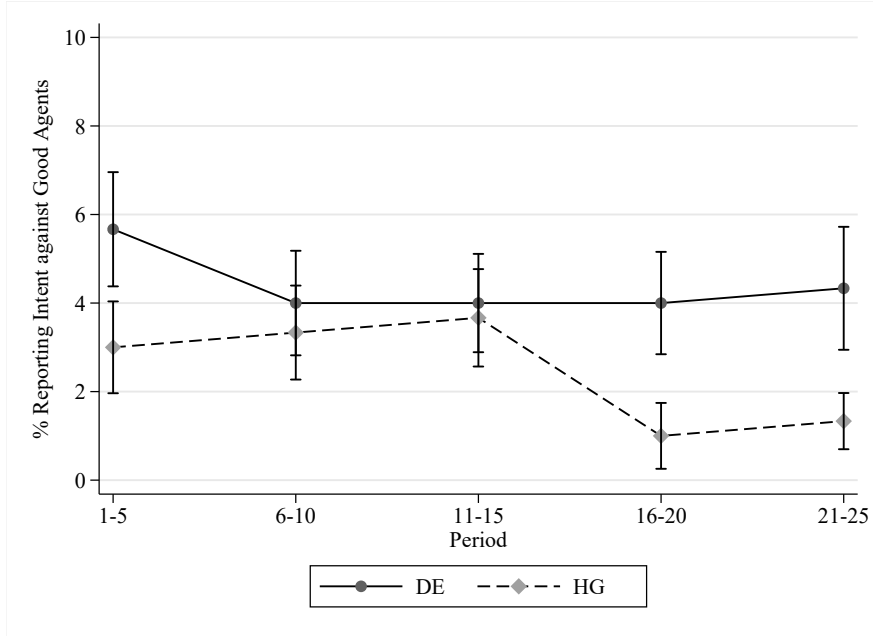
Table 2 summarizes estimated and true reporting rates. The treatment effect estimator yields $\hat{\Delta}R_B = \hat{R}_B^{\text{HG}} - \hat{R}_B^{\text{DE}} = 7.6$ percentage points which somewhat underestimates the true treatment effect $\Delta R_B = R_B^{\text{HG}} - R_B^{\text{DE}} = 9.1$ percentage points. The error in the treatment effect estimate ($\hat{\Delta}R_B - \Delta R_B = -1.5$ percentage points) is not equal to the difference between intended reporting rates of Good agents across treatments ($R_G^{\text{HG}} - R_G^{\text{DR}} = 2.5\% - 4.4\% = -1.9$ percentage points). The reason for this is that in small samples the empirical frequency with which intended negative reports $r = 0$ are turned into positive realized reports $\tilde{r} = 1$ can deviate slightly from π . In our particular case, the empirical frequency is 0.255 (instead of $\pi = 0.25$). This difference would obviously disappear in larger samples.

The difference between the rates at which Good agents are reported across treatments is not statistically significant (OLS: $p = 0.237$, RS: $p = 0.423$). Unfortunately, however, because of limited sample size the corresponding coefficient ($\beta = -0.019$) is not a precisely estimated zero. The 95% confidence interval is $[-0.052, 0.013]$. We can therefore not exclude the possibility that the estimated treatment effect is sizeably biased. In fact, as Figure 6 shows, differences in the false reporting of Good agents do not disappear over time. Under direct elicitation the reporting of Good agents remains roughly constant over

¹⁵Notice the frequencies with which Good agents are reported here are half of those described in Section 5.3. This is because reporting rates provided in Section 5.3 are conditioned on the subsample of Good agents.

time ($\beta = -0.0005$, $p = 0.423$), whereas the reporting rates decreases under hard garbling ($\beta = -0.0012$, $p = 0.010$). An estimation of the difference based on data of the final five periods alone therefore yields a larger (and marginally significant) difference: $\Delta R_G = 3$ percentage points (OLS: $p = 0.056$, RS: $p = 0.156$).

Figure 6: reporting of good agents over time



Note: The figure shows the development of intended reporting against Good agents under DE and HG over time. The variable *%Reporting Intent against Good Agents* measures the frequency with which monitors intend to submit a positive report $r = 1$ against a Good agent. Error bars mark ± 1 standard error from the mean (clustering at the individual level).

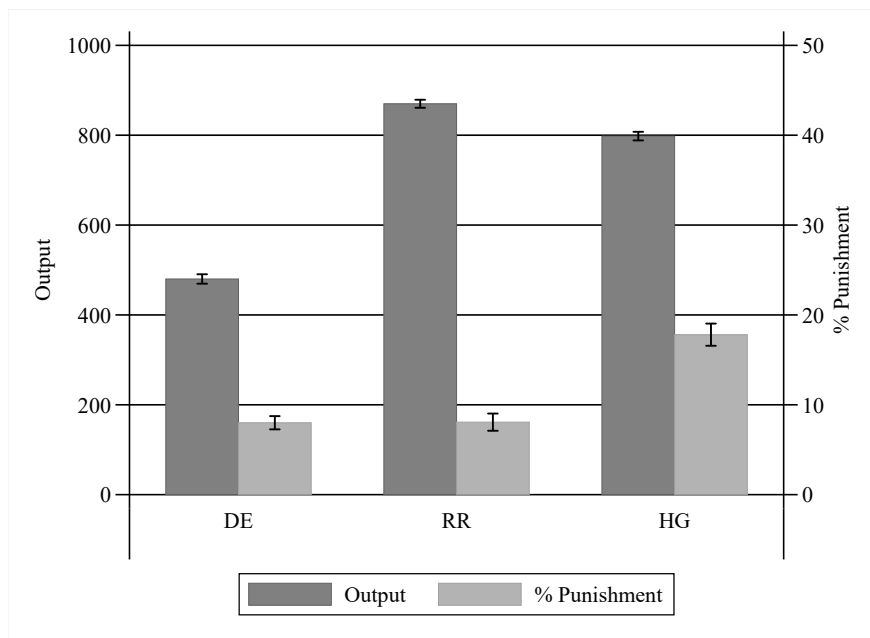
6 Findings for Randomized Response

We now turn to the case of randomized response RR. Note that since in this case the monitor perfectly controls the realized report $\tilde{r} \in \{0, 1\}$, we do not distinguish intended and realized report as in the case of hard garbling HG. We treat all realized reports as intended reports.

6.1 The best of both worlds

As Proposition 2 notes, the randomized response game RR is outcome equivalent to the direct elicitation game DE, while the obedient randomized response game oRR is outcome equivalent to the hard garbling game HG. A natural expectation is that experimental outcomes from the randomized response game would be a mixture of experimental outcomes from direct elicitation and hard garbling treatments. It turns out not to be the case. Randomized response seems to get the best of both survey procedures.

Figure 7: realized outcomes, randomized response

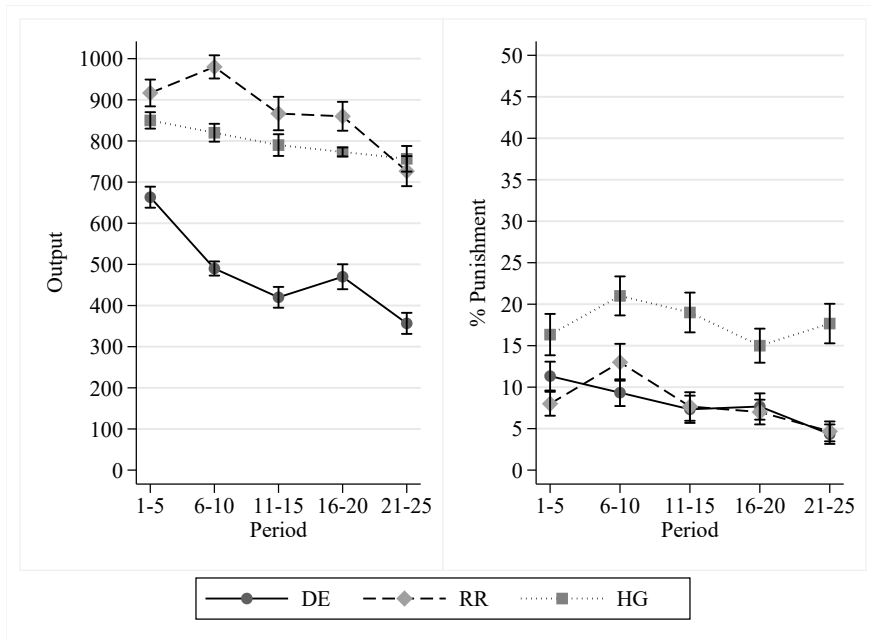


Note: The figure displays average output and the overall frequency of punishment under DE, RR and HG. The variable *Output* corresponds to average per-period output at the session level. The variable *% Punishment* represents the percentage of agent-monitor pairs in which punishment occurred. Error bars mark ± 1 standard error from the mean (clustering at the individual level).

Figure 7 compares outcomes and punishment rates across treatments. Randomized response manages to get the best of both games. Average output under RR (870 points) is higher than under both DE (480 points, OLS: $p = 0.003$, RS: $p = 0.025$), and HG (798,

although this difference is not significant: OLS: $p = 0.531$, RS: $p = 0.631$). Punishment in RR occurs with the same rate (8.0%) as in DE (8.0%, OLS: $p = 0.954$, RS: $p = 0.573$) and significantly less frequently than in HG (17.8%, OLS: $p < 0.001$, RS: $p = 0.004$). Figure 8 shows that these treatment differences remain very stable over the duration of the experiment: players are not learning to play according to SPE.¹⁶

Figure 8: realized outcomes over time, randomized response



Note: The figure displays the development of average output and the overall frequency of punishment over time under DE, RR and HG. The variable *Output* corresponds to average per-period output at the session level. The variable *% Punishment* represents the percentage of agent-monitor pairs in which punishment occurred. Error bars mark ± 1 standard error from the mean (clustering at the individual level).

¹⁶There are negative time trends in output in all three treatments. The time trends are significant under direct response ($\beta = -13.500$, $p < 0.001$) and randomized response ($\beta = -10.474$, $p = 0.021$), but not under hard garbling ($\beta = -6.192$, $p = 0.189$). The patterns are similar for punishment frequencies. Again there are significantly negative time trends under direct response ($\beta = -0.003$, $p < 0.001$) and randomized response ($\beta = -0.003$, $p = 0.021$), but not under hard garbling ($\beta = -0.001$, $p = 0.715$).

6.2 An intuitive explanation

The findings displayed in Figures 7 and 8 are consistent with bounded extensive-form rationality (see Ke (2018) for an axiomatization). Because monitors move last, they do not have to anticipate the responses of other players. As a result, their play is nearly subgame perfect. In contrast, agents need to predict the behavior of monitors to make decisions. This additional difficulty could make them more prone to take the unrelated question framing seriously. Our data confirm that this is the case.

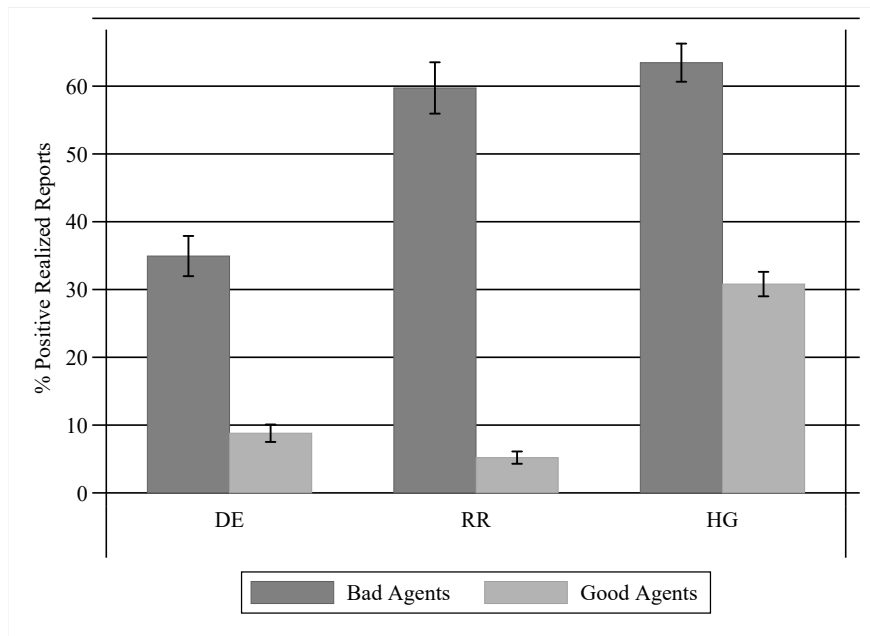
Monitors tend to treat the game as direct elicitation. Monitors do not answer the unrelated question correctly when the agent’s type is Good: positive reports occur in only 12% of the cases (13% if there is no punishment threat, 8% if there is a threat). The consequences of this behavior for overall reporting behavior are shown in Figure 9. The figure compares positive realized reports as a function of the agent’s type across treatments. Monitors’ reluctance to report Good agents under RR implies that realized reporting rates against Good agents are similar under RR and DE (5% vs. 9%, OLS: $p = 0.193$, RS: $p = 0.336$), but very different under RR and HG (5% vs. 31%, OLS: $p < 0.001$, RS: $p = 0.004$). This pattern remains very stable over time. Realized reports against Good agents do not exhibit a significant time trend under any treatment (DE: $\beta = -0.001$, $p = 0.424$, RR: $\beta = -0.002$, $p = 0.205$, HG: $\beta = 0.001$, $p < 0.264$), so that the reporting rates remain almost identical even if only the final five periods of the experiment are considered (DE: 9% RR: 5%, HG: 33%).

At the same time, when the agent’s type is Bad, the RR procedure does shift the monitors’ reporting behavior relative to what they do under DE. In fact, the overall rate of positive realized reports against Bad agents in RR (60%) is very similar to what we observe under HG (63%, OLS: $p = 0.614$, RS: $p = 0.748$) and substantially different from what we observe under DE (35%, OLS: $p < 0.004$, RS: $p = 0.019$).¹⁷ The reason for this effect is that monitors

¹⁷The reporting pattern against Bad agents across treatments does not change fundamentally over time. There are negative time trends under DE ($\beta = -0.009$, $p < 0.001$) and RR ($\beta = -0.008$, $p = 0.007$), whereas

tend to answer the unrelated question correctly when the agent is of the Bad type. It is not astonishing that answers to the unrelated question are mostly correct (94%) when there is no punishment threat, but many monitors also answer the unrelated question correctly when a Bad agent threatens to punish them (40%). As a consequence, Bad agents that commit to punish face incriminating realized reports more frequently under RR (29%) than under DE (19%, OLS: $p = 0.045$, RS: $p = 0.150$), but still less often than under HG (44%, OLS: $p = 0.002$, RS: $p = 0.055$).

Figure 9: monitors say the sky is red when the agent is good.



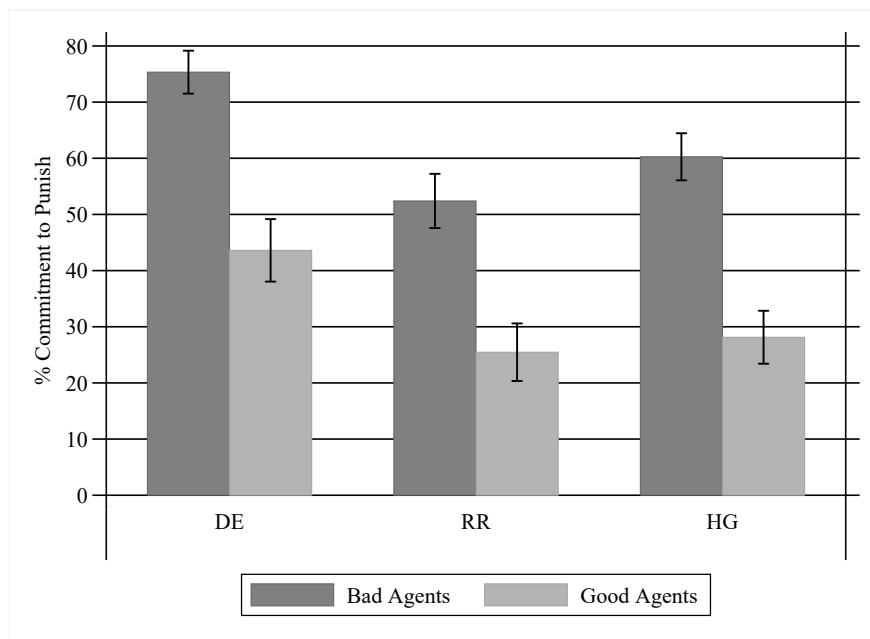
Note: The figure shows positive realized reports as a function of the agent's type under DE, RR and HG. The variable *Positive Realized Reports* measures the frequency with which a realized report against an agent occurs. Error bars mark ± 1 standard error from the mean (clustering at the individual level).

Overall, the observed reporting pattern favors performance under RR: monitors do not follow instructions when doing so would harm efficiency (they do not report Good agents), but they are more likely to obey the rules when doing so benefits efficiency (they report Bad agents). The reporting rate under HG is more stable ($\beta = -0.004$, $p = 0.206$). In the final five periods the reporting rates against Bad agents amount to: DE: 27% RR: 50%, HG: 61%.

agents even when under threat).

Agents treat the game as hard garbling. Threats by agents under randomized response RR are not significantly different from threats made under hard garbling HG (but they differ from those made under DE). Figure 10 shows the frequency of agents' commitment to punish contingent on the agent's type across treatments DE, RR, and HG.

Figure 10: agents commit to punish in the same way under RR and HG



Note: The figure displays the frequency with which Good and Bad agents commit to punish under DE, RR and HG. The variable *% Commitment to Punish* measures the within-type percentage of agents who commit to punish across all periods of a treatment. Error bars mark ± 1 standard error from the mean (clustering at the individual level).

The frequency with which Bad agents commit to punish is 52% under RR, versus 60% under HG (OLS: $p = 0.351$, RS: $p = 0.297$), and 75% under DE (OLS: $p = 0.010$, RS: $p = 0.030$). A similar result holds for Good agents. The frequency of commitments under RR is 25%, versus 28% under HG (OLS: $p = 0.625$, RS: $p = 0.748$), and 44% under DE (OLS: $p = 0.044$, RS: $p = 0.128$). Commitment to punish tends to increase slightly over time in

all three treatments (Bad agents: DE: $\beta = 0.007$, $p = 0.001$, RR: $\beta = 0.004$, $p = 0.086$, HG: $\beta = 0.004$, $p = 0.234$ / Good agents: DE: $\beta = 0.007$, $p = 0.036$, RR: $\beta = -0.002$, $p = 0.292$, HG: $\beta = 0.002$, $p = 0.697$), but the commitment frequencies in the final five periods confirm that the general pattern remains very stable (Bad agents: DE: 81% RR: 56%, HG: 63% / Good agents: DE: 47% RR: 23%, HG: 33%). These findings are consistent with the idea that agents believe they are playing oRR: the obedient version of RR.¹⁸

6.3 Learning

As Figure 8 highlights, output under RR remains close to output under HG over time. In other words, agents do not seem to learn how to play subgame perfect equilibrium. This finding is consistent with the fact that it is self-confirming (Fudenberg and Levine, 1993) for Bad agents to refrain from making threats if they believe that agents will answer the unrelated question truthfully.

Table 3: agents' future behavior is related to context-relevant experience only.

# late threats good	Coef.	Std.Err.	z	$P > z $	[0.025	0.975]
Intercept	-0.137	0.422	-0.330	0.758	-1.221	0.947
# early threats good	1.199	0.079	15.270	0.000	0.997	1.401
# early threats bad	0.082	0.133	0.610	0.567	-0.261	0.425
# late threats bad	Coef.	Std.Err.	z	$P > z $	[0.025	0.975]
Intercept	1.699	0.685	2.480	0.056	-0.062	3.461
# early threats good	0.123	0.086	1.440	0.209	-0.097	0.343
# early threats bad	0.858	0.137	6.280	0.002	0.507	1.210

Note: OLS estimation, standard errors clustered at the session level.

One objection to this interpretation is that agents can get evidence that monitors do not always take the unrelated question seriously (as Figure 9 shows): for instance, Good agents

¹⁸This data is also consistent with the finding that uncertainty makes contingent reasoning more difficult (Martínez-Marquina et al., 2018).

who do not commit to punish face a reporting rate of 6% rather than 25% under RR.¹⁹ This means that for the self-confirming interpretation to be correct, agents must not make successful inferences about continuation play when their type is Bad using data collected when their type is Good. The data show that this is indeed the case.

Table 3 reports (purely correlational) findings from regressing the number of threats given type in periods 11 to 20 on the number of threats given type in periods 1 to 10. Experience emitting threats when Good is not correlated to future threats conditional on being a Bad type, but is highly correlated to future threats conditional on being Good.

The impact of conditional payoff information. In order to confirm that insufficient information over conditional payoffs is the key limit to learning, we implement a second version of our treatments (HG, DE, RR), referred to as the *social learning* variant, in which we provide participants with such information. Specifically, agents are informed about the sample average of other agents' payoffs in previous sessions conditional on their type and commitment choice. Monitors learn sample averages of monitor profits in previous sessions conditional on agent quality, agent's commitment to punish, and the reporting decision.²⁰ While this is a rich informational setting, we believe it is not implausible within a stable organization: such information could plausibly arise from social learning.

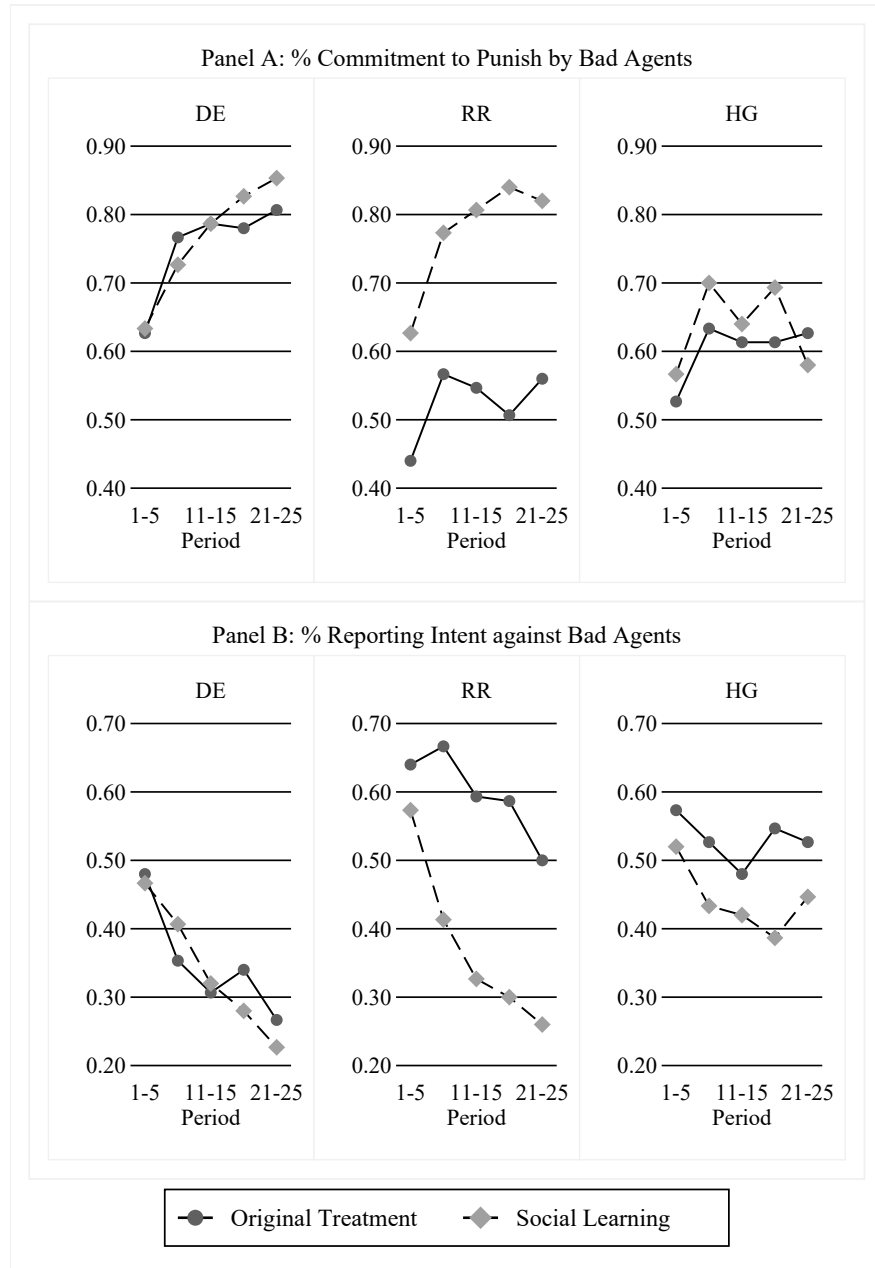
As Figure 11 illustrates, the agents' behavior under RR becomes indistinguishable from behavior under DE. This shows that failure to converge to SPE under RR was indeed due to limited information, rather than, say, design-specific social preferences. These results suggest that randomized response RR, while tempting on the face of evidence displayed Figure 7, should be used with great caution in stable organizations. If social learning allows decision makers to understand the procedure properly, the randomized response method no longer outperforms direct data elicitation (see Appendix A for a more detailed analysis of the impact

¹⁹It is important to keep in mind that agents do not see this aggregated information, but need to learn it over time. Such learning is difficult and slow, because agents only have very few observations at their disposal.

²⁰Under RR the information is provided separately for the direct and the unrelated question.

of the availability of conditional payoff information on performance under different elicitation mechanisms). In contrast, hard garbling continues to improve on information flows.

Figure 11: comparison of original and social learning treatments



Note: The figure compares time trends in commitment to punish by Bad agents and in reporting intent against Bad agents between the original treatments and the social learning treatments.

7 Modeling Boundedly Rational Behavior

In this Section we seek to formalize the intuitive discussion of non-equilibrium play under randomized response (Section 6.2) using off-the-shelf modeling tools. Specifically we use a natural variant of the Dynamic Level- k model of Ho and Su (2013) that allows for random utility shocks along the lines of Quantal Response Equilibrium (McKelvey and Palfrey, 1995).²¹ We emphasize that the model is wrong in significant ways, and bring up relevant issues explicitly. However, we believe that the opportunity for learning is maximized by sticking to an existing and well understood class of models rather than fixing the model in an ad hoc way. Our objectives for this section are threefold:

- (i) to formulate a theory of how bounded rationality can bias treatment-effect estimators by inducing patterns of false-reporting that differ across treatments;
- (ii) to allow for not-implausible counterfactual, or prospective analysis of how different survey mechanisms may work out in practice;
- (iii) to query the data through the lens of implied model parameters.

7.1 The Quantal Response Level- k model

We first provide a definition of the Quantum Response Level- k model (QRL- k) tailored for our survey games. Appendix B provides a more general definition encompassing all single-move extensive-form games (of which survey games are a special case), and shows that it's without loss of generality to set the highest level of rationality equal to the number of players.

Players are described by their level $k \in \{0, 1, 2\}$. Play in survey game $\mathcal{G} \in \{\text{DE}, \text{RR}, \text{HG}\}$ is described by the mass $\rho_{\mathcal{G}} \in [0, 1]$ of level 2 players, and the mass $1 - \rho_{\mathcal{G}}$ of level 1 players. Level 0 players exist only in the mind of level 1 players. For any decision node h_i , denote by A_{h_i} actions available to player i at h_i . Let $S_i \equiv \prod_{h_i} A_{h_i}$ denote the set of strategies available

²¹A possible alternative would be to use a variant of cursed equilibrium (Eyster and Rabin, 2005) in which players don't neglect the correlation between the play of others and an underlying state of the world, but instead neglect the correlation between the play of others and their own actions.

to player i , and by $u_i(s_i, s_{-i})$ the payoff to player i .²² The play of each level is described as follows:

- a level 0 player i follows instructions on how to play if any are given, and picks a strategy s_i with uniform probability over S_i if no instructions are provided;
- a level k player i best-responds to the distribution of strategies $\mu_{-i}^{k-1} \in \Delta(S_{-i})$ of an opponent of level $k-1$: for any decision node h_i , action a_i^k is a random variable solving

$$\max_{a_i \in A_{h_i}} \mathbb{E}_{\mu_{-i}^{k-1}} [u_i(a_i, s_{-i}) | h_i] + \varepsilon_i(a_i) \quad (4)$$

where error terms $(\varepsilon_i(a_i))_{i \in I, a_i \in A_{h_i}}$ are independent across players $i \in I$, and distributed as follows conditional on i :

- with probability $1 - \nu$, for all $a_i \in A_{h_i}$, $\varepsilon_i(a_i) = \sigma \varepsilon'_i(a_i)$, where $(\varepsilon'_i(a_i))_{a_i \in A_{h_i}}$ are i.i.d. and follow an extreme value type I distribution with mean 0, location 0, and scale parameter 1;
- with probability ν , $\varepsilon_i(a_i) = +\infty \mathbf{1}_{a_i = a_i^*}$, where a_i^* is uniformly distributed over A_{h_i} , and using the convention that $\infty \times 0 = 0$.²³

With probability ν , the shock is not-responsive to payoffs, i.e. error rates do not shrink as the payoffs and incentives get scaled up. With probability $1 - \nu$ the error is responsive to payoffs, and the degree of responsiveness is captured by parameter σ . This flexibility allows us to capture patterns of play in which a non-zero share of players fail to optimize $u_i(a_i, s_{-i})$ when stakes are large, and yet play is not uniformly random when stakes are small.²⁴

A QRL- k model of play described by $(\alpha, \sigma, \nu, \rho)$ induces a distribution $\mu^{\text{QRL}} \in \Delta(S)$ over strategy profiles $s \in S \equiv \prod_i S_i$.

²²This includes monitors' altruistic preferences, parameterized by α .

²³As Haile et al. (2008) highlight, the class of error terms used is a key determinant of the empirical content of Quantum Response Equilibrium.

²⁴If we impose $\nu = 0$, the scale parameter σ needed to explain failures to optimize when stakes are large predicts nearly uniformly random play when stake are small.

7.2 Theoretical Analysis

The QRL- k model is a useful benchmark for several reasons. The first is that it is identified from play. Let $\mu_{|A}^{\text{QRL}} \in \Delta(A)$ denote the projection of μ^{QRL} from strategies $s \in S$ to realized action profiles $a = (a_i, a_{-i})$. Importantly, $\mu_{|A}^{\text{QRL}}$ can be estimated using the sample distribution of play.

Proposition 4 (model identification). *For all survey games, HG, DE, and RR, parameters α , σ , ν and ρ are identified from the following moments:*

- *play by the monitor at all histories, $(\mu_{|M}^{\text{QRL}}(r = 1|c, \tau))_{\substack{c \in \{0,1\}, \\ \tau \in \{G,B\}}}$, and,*
- *play by the agent conditional on her type, $(\mu_{|A}^{\text{QRL}}(c = 1|\tau))_{\tau \in \{G,B\}}$.*

The next result shows that keeping social preferences α , and parameters (σ, ν) governing the distribution of error terms, the same across games, QRL- k can only capture the framing effects due to the unrelated question (i.e. the difference between RR and DE) through a change in the share ρ of level 2 players across games RR and DE.

Proposition 5 (rationality and framing). *Assume that parameters α , σ , ν are equal across RR and DE. QRL- k behavior μ_{RR}^{QRL} and μ_{DE}^{QRL} is such that:*

- (i) *Regardless of ρ_{RR} and ρ_{DE} , at every (type, commitment) node (τ, c) , the distribution of realized reports from the monitor is the same across RR and DE:*

$$\mu_{RR}^{\text{QRL}}(\tilde{r}|\tau, c) = \mu_{DE}^{\text{QRL}}(\tilde{r}|\tau, c).$$
- (ii) *Whenever $\rho_{RR} = \rho_{DE}$, the distribution of punishment-commitment choices by Bad agents is the same across RR and DE: $\mu_{RR}^{\text{QRL}}(c|B) = \mu_{DE}^{\text{QRL}}(c|B)$.*

Since in our data, Bad agents are less likely to commit to punishment under RR than DE, our experiment indicates that framing effects have an impact on the level of rationality of players. Specifically, it suggests that framing lowers the share of level 2 players under RR. This has implications about false reporting and the bias of treatment effect estimator $\hat{\Delta}R_B$ under boundedly rational play.

Proposition 6 (rationality and bias). *Assume that $\rho_{RR} \leq \rho_{DE}$, and keep parameters α , σ , ν equal across RR and DE. The following hold:*

(i) *Good agents commit to retaliate more often under DE than RR:*

$$\mu_{RR}^{QRL}(c = 1|G) \leq \mu_{DE}^{QRL}(c = 1|G).$$

(ii) *Monitors report Good agents more often under DE than under RR:*

$$\mu_{RR}^{QRL}(\tilde{r} = 1|G) \geq \mu_{DE}^{QRL}(\tilde{r} = 1|G).$$

This last result is intuitive. Because Good agents poorly anticipate the behavior of monitors under randomized response RR, they commit to punish less frequently under RR than DE. This means that monitors tend to have stronger incentives *not* to report Good agents under DE than RR. As a result, monitors report good agents more frequently under RR than DE. This leads us to overestimate the impact of treatment RR versus control DE on the reporting of Bad agents.

7.3 Structural investigation

We take the QRL- k model to the data (focusing on treatments without social learning) with three objectives: first, we assess in-sample fit; second, we explore the model's value in evaluating counterfactual scenarios; third, we examine the extent of bias due to false reporting.

In-sample fit. For each treatment HG, DE, and RR (original treatments without conditional payoff information), we estimate model parameters ρ (share of level 2 players), α (monitor altruism), σ (the scale of payoff-responsive shocks), and ν (the mass of payoff non-responsive shocks). Given identification (Proposition 4), we estimate parameters using the simulated method of moments (McFadden, 1989). Table 4 shows estimated parameters.

A first observation is that parameter estimates match the intuitive explanation for why RR performs so well: the share of level 2 players ρ is lower under RR than DE. This does not

Table 4: estimated parameters (original treatments).

	HG	DE	RR
ρ	1.	0.78	0.52
α	0.48	0.34	0.54
σ	16.7	21.6	8.8
ν	0.096	0.20	0.10

impact the behavior of monitors significantly but makes agents more careful about issuing threats since they believe that monitors may take the unrelated question seriously.

A second observation is that the rate of incompressible errors ν is large, especially under DE. As Table 4 clarifies, this is driven by the behavior of monitors under DE: they report Bad agents with probability 19.1% conditional on threats, which suggests that they may have fairly high altruism; however, they also report Good agents with probability 12.8% in the absence of threats, which suggests that they have low altruism. Ultimately these facts end up being rationalized through a high rate of payoff non-responsive errors.

A third observation is that the fit between simulated and empirical moments summarized by Table 5 is quite good. This implies that it would be difficult to differentiate this particular model of bounded rationality from a different one: the existing free parameters already explain the data fairly well. Therefore even though the high reporting rate $\tilde{r} = 1$ against Good agents under DE could potentially be rationalized in an expanded model, we believe that it is more productive for us to focus on the implications of the off-the-shelf QRL- k model, rather than come up with yet another framework.

Counterfactual scenarios. One use for models of bounded rationality is to help formulate quantitative counterfactual scenarios. We simulate behavior under RR, using parameters estimated under DE, but taking $\rho = .6$ reflecting the plausibly anticipated belief that people would be confused by the unrelated question.

Table 5: in-sample fit of empirical (E) and simulated (S) moments; “report” refers to a realized report $\tilde{r} = 1$

moment		Emp./Sim.	HG	DE	RR
threat given bad	E		0.603	0.753	0.524
		S	0.594	0.694	0.507
threat given good	E		0.281	0.436	0.255
		S	0.222	0.494	0.244
report bad given no threat	E		0.893	0.832	0.933
		S	0.952	0.898	0.950
report bad given threat	E		0.292	0.191	0.293
		S	0.262	0.133	0.309
report good given no threat	E		0.056	0.128	0.059
		S	0.110	0.192	0.051
report good given threat	E		0.033	0.037	0.031
		S	0.048	0.100	0.050

Table 6: extrapolated moments from DE to RR; “report” refers to a realized report $\tilde{r} = 1$

	empirical	simulated
threat given bad	0.524	0.541
threat given good	0.255	0.411
report bad given no threat	0.933	0.898
report bad given threat	0.293	0.133
report good given no threat	0.059	0.192
report good given threat	0.031	0.100

As Table 6 shows, the counterfactual simulation does a fairly good job of predicting actual behavior. In particular, it predicts that Bad agents will indeed refrain from making threats, but that monitors will rarely report either Good agents, or Bad agents conditional on threats. The main departures are driven by the fact that under DE, monitors report Good agents at a relatively high rate (13%) conditional on no threat. This results in a fairly high value of payoff-responsive shock-parameter σ , which makes emitting threats attractive under our simulated RR model. The corresponding low estimate of α under DE leads us to underpredict the reporting of Bad types, and overpredict the reporting of Good types.

False reporting and bias. As Section 3.2 highlights, the bias of treatment effect estimator $\hat{\Delta}R_B$ is equal to $R_G^{\text{treatment}} - R_G^{\text{control}}$: different false reporting of Good agents across different treatments biases the treatment effect estimator.²⁵ Proposition 6 shows that QRL- k makes specific predictions about bias: (i) Good agents should be less likely to make threats than under RR than DE; (ii) as a result, there should be more reporting of Good agents under RR than DE.

As Figure 12 shows, point (i) above is borne out in the data, but point (ii) is not (see also Table 5). The prediction that Good agents make threats more frequently under DE (44%) than RR (25%, OLS: $p = 0.044$, RS: $p = 0.128$) or HG (28%, OLS: $p = 0.040$, RS: $p = 0.128$) does hold (see Panel A). However, the share of intended reports against Good agents is higher under DE (9%) than either RR (5%, OLS: $p = 0.193$, RS: $p = 0.336$) or HG (5%, OLS: $p = 0.237$, RS: $p = 0.423$, see Panel B). The reason for this (as we have noted before) is that conditional on no-threat, monitors under DE report Good agents at a surprisingly high rate (13%) relative to monitors under HG (6%, OLS: $p = 0.160$, RS: $p = 0.337$) and RR (6%, OLS: $p = 0.183$, RS: $p = 0.197$, see Panel C).²⁶

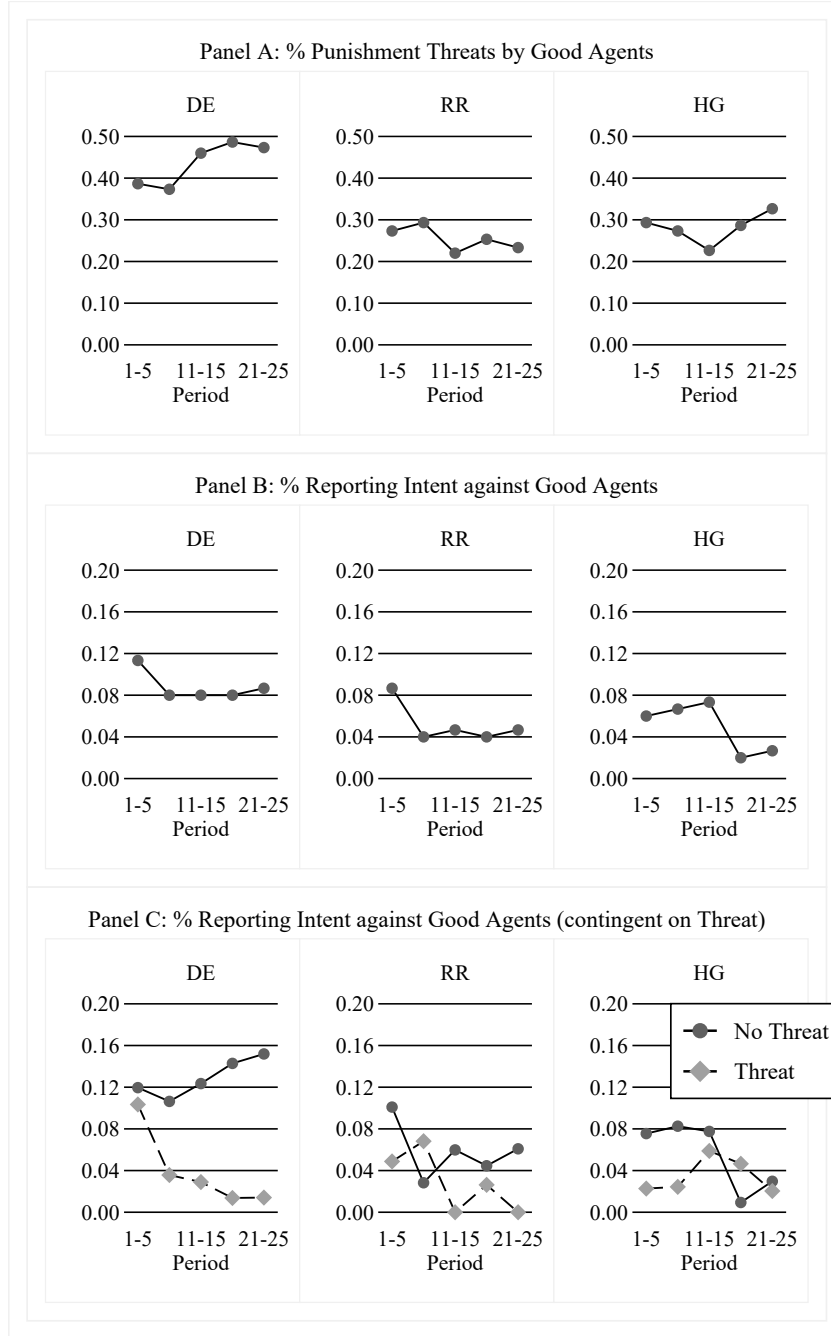
Impact of social learning on bias. Figure 13 shows that the bias of estimator $\hat{\Delta}R_B$ for the treatment effect of HG relative to DE disappears almost completely when players get feedback about conditional payoffs. The difference between reporting rates of Good agents across treatments amounts to $R_G^{HG} - R_G^{DE} = 1.4\% - 1.3\% = 0.1$ (or more precisely 0.07) percentage points (OLS: $p = 0.910$, RS: $p = 0.871$). The 95% confidence interval for the difference is $[-0.01, 0.01]$.²⁷ As a consequence, the estimated treatment effect is essentially

²⁵We abstract here from the fact that the empirical frequency with which information is garbled in RR and HG may slightly deviate from the theoretical probability in small samples. See section 5.4 for a discussion.

²⁶A potential explanation for this result is that play under direct-elicitation tends to be particularly favorable to agents, since threats are very effective. The false reporting of good agents who do not make threats may thus be an expression of spite from monitors. To account for this, the model would need to consider social preferences over the entire sequence of interactions.

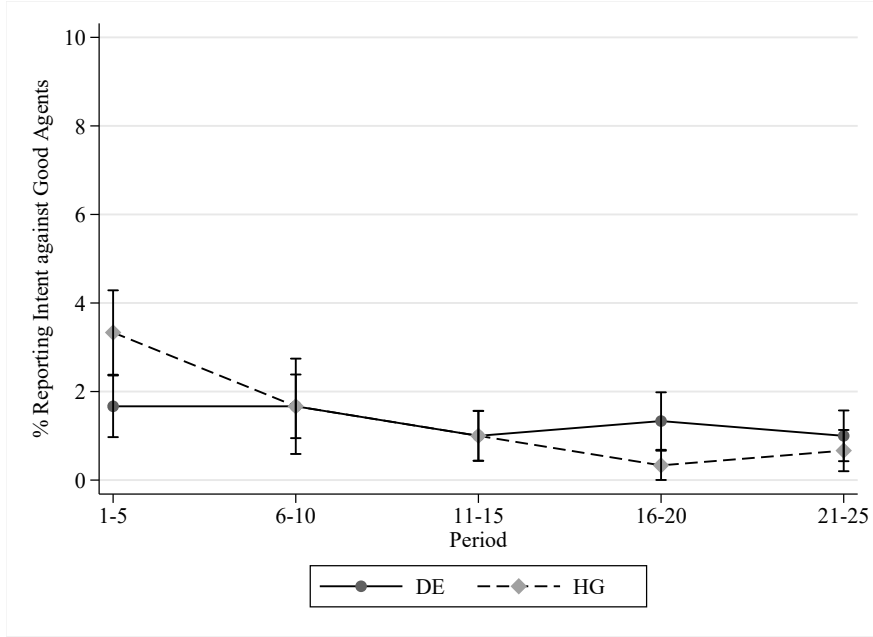
²⁷The difference becomes even smaller over time. An estimation of the difference based on data of the final 5 periods alone therefore yields: $\Delta R_G = 0.03$ percentage points (OLS: $p = 0.562$, RS: $p = 0.575$).

Figure 12: mechanics of false reporting of good agents



Note: The figure shows time trends in punishment threats by Good agents (Panel A) and intended reporting against Good agents (Panels B and C).

Figure 13: intended reporting of good agents over time in the social learning treatments



Note: The figure shows time trends in intended reporting against Good agents under DE and HG in the social learning experiment. The variable *%Reporting Intent against Good Agents* measures the frequency with which monitors intend to submit a positive report $r = 1$ against a Good agent. Error bars mark ± 1 standard error from the mean (clustering at the individual level).

unbiased when using the data obtained from the experiments in which social learning is possible. We note that the improved consistency of our treatment effect estimator under social learning is not caused by a higher frequency of threats from Good agents under DE (48% with information versus 44% without information, OLS: $p = 0.675$, RS: $p = 0.631$), but rather by a lower reporting rate of monitors against agents who do not threaten to punish (4% with information, versus 13% without information, OLS: $p = 0.083$, RS: $p = 0.200$).

8 Discussion

8.1 Summary

We study the value of survey methods achieving different forms of plausible deniability in equilibrium. We compare three survey designs of interest: direct elicitation DE, hard garbling HG and randomized response RR. Our theoretical analysis emphasizes that:

- randomized response RR and direct elicitation DE are outcome equivalent under SPE; obedient randomized response oRR (in which it is common knowledge that monitors take the unrelated question seriously) is outcome equivalent to HG in SPE.
- an estimator of treatment effects can be computed using unverifiable reports; it is consistent only when the false reporting of Good agents is independent of the survey design, which holds when monitors are rational.

Experimental investigation shows that subgame perfect equilibrium is not a successful model of empirical play in this setting. First, Good agents are being reported at positive and different rates across designs. Second, randomized response RR turns out to be surprisingly effective due to the fact that monitors do not take the unrelated question seriously, but agents seem to. This differential confusion is easily explained in a dynamic model of boundedly rational behavior, monitors act last and do not have to form beliefs about the play of others, while agents act first and have to form beliefs about the behavior of monitors. This simple insight has broader implications for extensive form mechanism design: for instance, the belief that complex questionnaires (Glazer and Rubinstein, 2014) work on others is sufficient to get relevant players to comply. The mechanics here are reminiscent of Fudenberg and Levine (2006) which clarifies the use of sufficiently off-the-equilibrium-path superstitious beliefs in the code of Hammurabi.

An off-the-shelf model, QRL- k , provides adequate quantitative fit and leads to new predictions about bias. Under bounded rationality, RR should lead to greater reporting of Good agents in equilibrium: because Good agents take the unrelated question seriously they refrain from making threats which, under QRL- k , leads to higher reporting of Good agents

under RR than DE. This prediction is not fully borne out in the data. The main source of prediction error is the fact that monitors report non-threatening Good agents at much higher rates under DE than under RR or HG.

From a modeling perspective, we note in passing that in two player extensive form games, the difference between QRL- k and cursed equilibrium is rather small. In a sense, under QRL- k level 1 agents underestimate the correlation between their play and the behavior of the monitor. Framing effects due to the unrelated question would be interpreted as an increase in correlation neglect.

We emphasize that the failure of SPE under RR is long lived, which is consistent the fact that the pattern of play we observe is broadly consistent with self-confirming equilibrium. This learning interpretation is confirmed by the fact that play does converge to SPE when players get information about the conditional payoffs obtained by other players like them. This causes the bias of treatment effect estimator $\hat{\Delta}R_B$ to vanish. We note that while self-confirming equilibrium explains why players do not learn SPE in treatments without social learning, self-confirming equilibrium does not help explain why play under DE and RR is different: the two games have the same set of self-confirming equilibria. Our view is that a model along the lines of QRL- k is needed to account for initial play across different games, while self-confirming equilibrium helps explain why initial play sticks.

Altogether, while the good performance of RR is striking in settings without social learning, we find that our evidence broadly confirms that one should be cautious when using randomized response RR in stable organizational settings where learning may occur. In addition, false positives cannot be entirely ignored, but may be partially anticipated using QRL- k .

We conclude with a prospective discussion of potential applications for secure survey design.

8.2 Applications

What environments? We evaluate the impact of survey design on the transmission of sensitive information in organizations. We study an environment in which an interested agent can commit to retaliate in order to suppress information. By forcing punishment on the equilibrium path, adding noise to surveys makes it more costly to suppress information. We note that a similar argument may hold when threats are not the issue, but instead, monitors are worried about the reputational effects. Indeed, inference about intended report r from realized report $\tilde{r} = 1$ satisfies

$$\log \left(\frac{\text{prob}(r = 1 | \tilde{r} = 1)}{\text{prob}(r = 0 | \tilde{r} = 1)} \right) = \log \left(\frac{\text{prob}(\tilde{r} = 1 | r = 1)}{\text{prob}(\tilde{r} = 1 | r = 0)} \right) + \log \left(\frac{\text{prob}(r = 1)}{\text{prob}(r = 0)} \right).$$

Garbling reduces term $\log \left(\frac{\text{prob}(\tilde{r}=1|r=1)}{\text{prob}(\tilde{r}=1|r=0)} \right)$ thereby diminishing the reputational impact of information transmission.

This logic applies in environments where there are no threats, but the agent exhibits spite and may retaliate in a manner commensurate to her belief that the monitor caused her harm (see Chassang and Zehnder, 2016, for a model along these lines). Garbled surveys reduce the impact of intervention on the agent’s posterior that the monitor transmitted information. As a result, it reduces the agent’s propensity to retaliate.

In many settings, instead of being concerned with potential retaliation, the monitor may be concerned with the impact information may have on the agent’s reputation. Consider the problem of detecting mental health or substance abuse issues for teams operating in high stakes environments, such as military and law enforcement units. High degrees of loyalty are essential for such teams. As a result, team members may be unwilling to signal that a teammate is experiencing issues: this may have a negative long-term impact on their teammate’s career. In such situations, suitably garbled information channels may help concerned team members get help for their teammates without endangering their teammates future careers. Because there is no embedded antagonism, this class of applications may also

exhibit reduced rates of false reporting.

Finally, note that the same argument may apply to information one submits about oneself: an agent may be unwilling to report that she is experiencing burn-out, or mental health issues, if she believes this will affect her career. The ability to send soft information improves opportunities for communication.

What interventions? By construction, the information provided through garbled surveys is soft information that cannot serve as the basis for heavy-handed corrective or punitive intervention. Plausible interventions could take the form of a thorough investigation potentially leading to actionable evidence. Alternatively, intervention could take the form of appropriate training that is especially helpful for Bad agents, and is only a minor inconvenience for Good agents. A manager may receive leadership or sensitivity training. A military or police officer potentially experiencing burnout may be placed on furlough for a few weeks, and so on.

References

- AYRES, I. AND C. UNKOVIC (2012): “Information Escrows,” *Mich. L. Rev.*, 111, 145.
- BLAIR, G., K. IMAI, AND Y.-Y. ZHOU (2015): “Design and analysis of the randomized response technique,” *Journal of the American Statistical Association*, 110, 1304–1319.
- BLUME, A., E. K. LAI, AND W. LIM (2013): “Eliciting private information with noise: the case of randomized response,” .
- CAMERER, C., S. NUNNARI, AND T. R. PALFREY (2016): “Quantal response and nonequilibrium beliefs explain overbidding in maximum-value auctions,” *Games and Economic Behavior*, 98, 243–263.
- CAMERER, C. F., T.-H. HO, AND J.-K. CHONG (2004): “A cognitive hierarchy model of games,” *The Quarterly Journal of Economics*, 119, 861–898.

- CHASSANG, S. AND G. PADRÓ I MIQUEL (2018): “Crime, Intimidation, and Whistleblowing: A Theory of Inference from Unverifiable Reports,” *Review of Economic Studies*.
- CHASSANG, S. AND C. ZEHNDER (2016): “Rewards and punishments: Informal contracting through social preferences,” *Theoretical Economics*, 11, 1145–1179.
- CHUANG, E., P. DUPAS, E. HUILLERY, AND J. SEBAN (2019): “Sex, Lies, and Measurement: Do Indirect Response survey methods work?” .
- CRAWFORD, V. P. AND N. IRIBERRI (2007): “Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner’s Curse and Overbidding in Private-Value Auctions?” *Econometrica*, 75, 1721–1770.
- DINUR, I. AND K. NISSIM (2003): “Revealing information while preserving privacy,” in *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, ACM, 202–210.
- DWORK, C., F. MCSHERRY, K. NISSIM, AND A. SMITH (2006): “Calibrating noise to sensitivity in private data analysis,” in *Theory of cryptography conference*, Springer, 265–284.
- EYSTER, E. AND M. RABIN (2005): “Cursed equilibrium,” *Econometrica*, 73, 1623–1672.
- FEHR, E., M. POWELL, AND T. WILKENING (2017): “Behavioral Constraints on the Design of Subgame-Perfect Implementation Mechanisms,” Tech. rep., Mimeo.
- FISCHBACHER, U. (2007): “z-Tree: Zurich toolbox for ready-made economic experiments,” *Experimental economics*, 10, 171–178.
- FUDENBERG, D. AND D. K. LEVINE (1993): “Self-confirming equilibrium,” *Econometrica: Journal of the Econometric Society*, 523–545.
- (2006): “Superstition and rational learning,” *American Economic Review*, 96, 630–651.

- GLAZER, J. AND A. RUBINSTEIN (2014): “Complex questionnaires,” *Econometrica*, 82, 1529–1541.
- GREENBERG, B. G., A.-L. A. ABUL-ELA, W. R. SIMMONS, AND D. G. HORVITZ (1969): “The unrelated question randomized response model: Theoretical framework,” *Journal of the American Statistical Association*, 64, 520–539.
- GREINER, B. (2015): “Subject pool recruitment procedures: organizing experiments with ORSEE,” *Journal of the Economic Science Association*, 1, 114–125.
- HAILE, P., A. HORTAÇSU, AND G. KOSENOK (2008): “On the empirical content of quantal response equilibrium,” *The American Economic Review*, 98, 180–200.
- HO, T.-H. AND X. SU (2013): “A dynamic level-k model in sequential games,” *Management Science*, 59, 452–469.
- JOHN, L. K., G. LOEWENSTEIN, A. ACQUISTI, AND J. VOSGERAU (2018): “When and why randomized response techniques (fail to) elicit the truth,” *Organizational Behavior and Human Decision Processes*, 148, 101 – 123.
- KAPLAN, S. E., K. PANY, J. A. SAMUELS, AND J. ZHANG (2009): “An Examination of the Effects of Procedural Safeguards on Intentions to Anonymously Report Fraud,” *Auditing: A Journal of Practice & Theory*, 28, 273–288.
- KAPLAN, S. E. AND J. J. SCHULTZ (2007): “Intentions to Report Questionable Acts: An Examination of the Influence of Anonymous Reporting Channel, Internal Audit Quality, and Setting,” *Journal of Business Ethics*, 71, 109–124.
- KE, S. (2018): “Boundedly rational backward induction,” Tech. rep., Theoretical Economics.
- MARTÍNEZ-MARQUINA, A., M. NIEDERLE, AND E. VESPA (2018): “Failures in Contingent Reasoning: The Role of Uncertainty,” *American Economic Review*.

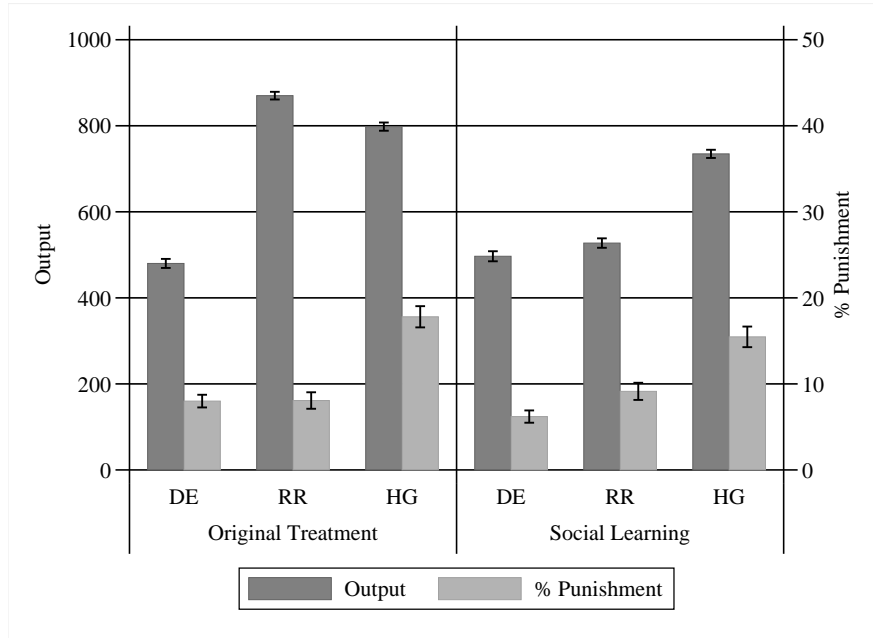
- McFADDEN, D. (1989): “A method of simulated moments for estimation of discrete response models without numerical integration,” *Econometrica: Journal of the Econometric Society*, 995–1026.
- McKELVEY, R. D. AND T. R. PALFREY (1995): “Quantal response equilibria for normal form games,” *Games and economic behavior*, 10, 6–38.
- MECHTENBERG, L., G. MUEHLHEUSSER, AND A. ROIDER (2017): “Whistle-Blower Protection: Theory and Experimental Evidence,” .
- NAGEL, R. (1995): “Unraveling in guessing games: An experimental study,” *The American Economic Review*, 85, 1313–1326.
- ORTNER, J. AND S. CHASSANG (2018): “Making Corruption Harder: Asymmetric Information, Collusion, and Crime,” *Journal of Political Economy*.
- ROSENFELD, B., K. IMAI, AND J. N. SHAPIRO (2016): “An empirical validation study of popular survey methodologies for sensitive questions,” *American Journal of Political Science*, 60, 783–802.
- STAHL, D. O. (1993): “Evolution of smartn players,” *Games and Economic Behavior*, 5, 604–617.
- WARNER, S. L. (1965): “Randomized response: A survey technique for eliminating evasive answer bias,” *Journal of the American Statistical Association*, 60, 63–69.

Appendix (for online publication)

A Further Empirical Analysis

In this section we describe in more detail the results of our social learning treatments. In those treatments we provide participants with conditional payoff information elicited in previous sessions of the same experiment.²⁸ We describe how conditional payoff information affects outcomes under our three elicitation mechanisms (DE, RR, HG).

Figure A.1: impact of social learning on output and punishment



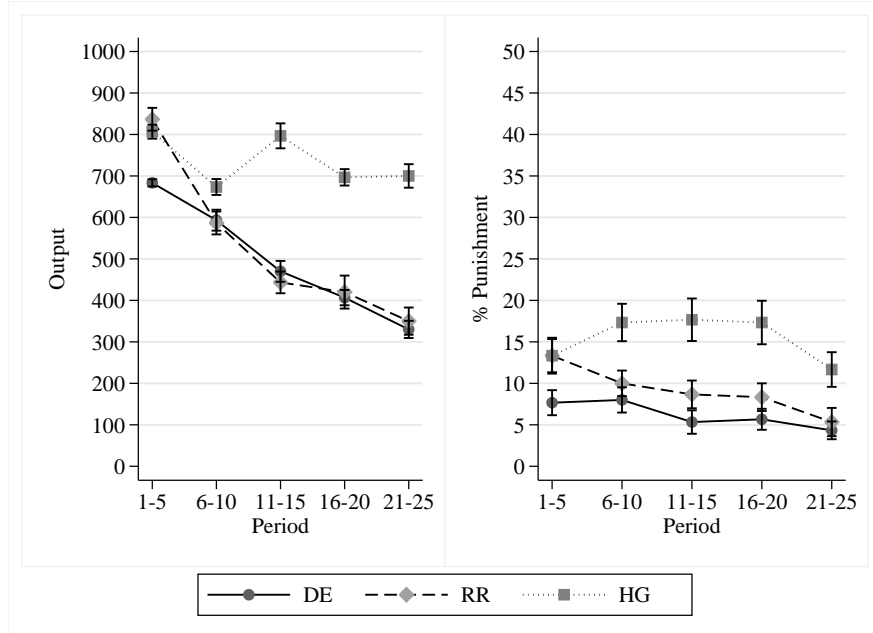
Note: The figure shows average output and the overall frequency of punishment. The variable *Output* corresponds to average per-period output at the session level. The variable *% Punishment* represents the percentage of agent-monitor pairs in which punishment occurred. Error bars mark ± 1 standard error from the mean (clustering at the individual level).

Output and punishment Figure A.1 displays average output and the overall frequency of punishment for all elicitation mechanisms in the absence (original treatments) and the presence of conditional payoff information (social learning treatments). The figure reveals

²⁸Agents are informed about sample averages of agent profits conditional on agent type and commitment to punish. Monitors learn sample averages of monitor profits conditional on agent quality, agent's commitment to punish, and the reporting decision.

that our finding that randomized response gets the best of both survey procedures (see section 6.1) no longer holds once social learning is possible. While there is no significant impact on average output under DE and HG²⁹, the availability of conditional payoff information reduces average output under RR from 870 points to 527 points (OLS: $p = 0.018$, RS: $p = 0.045$). As a consequence, there is no longer a significant difference in average output between DE (497) and RR (527) in the social learning treatments (OLS: $p = 0.769$, RS: $p = 0.873$).³⁰ Finally, while the availability of conditional payoff information leads to slightly lower punishment frequencies under DE (8.0% vs. 6.2%, OLS: $p = 0.067$, RS: $p = 0.127$) and HG (17.8% vs. 15.5%, OLS: $p = 0.048$, RS: $p = 0.037$), the punishment frequency under RR increases insignificantly (8.1% vs. 9.1%, OLS: $p = 0.546$, RS: $p = 0.521$).

Figure A.2: output and punishment under social learning



Note: The figure displays the development of average output and the overall frequency of punishment in all social learning treatments (DE, RR and HG). The variable *Output* corresponds to average per-period output at the session level. The variable *% Punishment* represents the percentage of agent-monitor pairs in which punishment occurred. Error bars mark ± 1 standard error from the mean (clustering at the individual level).

²⁹Under DE average output slightly increases from 480 points to about 497 points (OLS: $p = 0.846$, RS: $p = 1.000$) and under HG average output slightly decreases from 798 points to 735 points (OLS: $p = 0.410$, RS: $p = 0.631$).

³⁰In the presence of social learning average output under HG is significantly higher than under DE (OLS: $p = 0.003$, RS: $p = 0.025$) and under RR (OLS: $p = 0.064$, RS: $p = 0.092$).

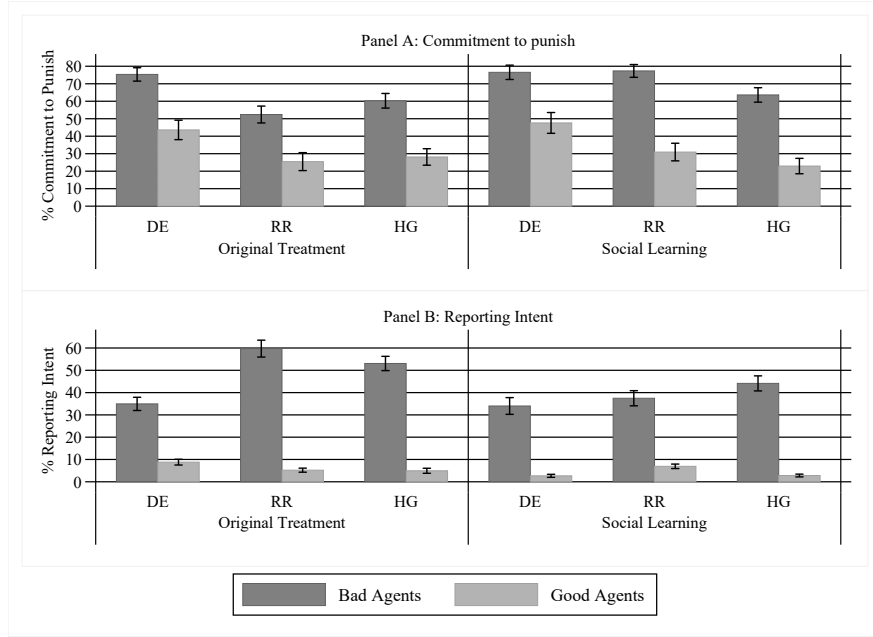
The fact that randomized response no longer outperforms direct elicitation in the presence of conditional payoff information is further confirmed by a dynamic analysis. Figure A.2 displays the development of average output and realized punishment in the social learning treatments over time. Average output under RR and DE converges across treatments after the first five periods, and then exhibit the same negative time-trend until the end of the experiment (DE: $\beta = -18.205$, $p < 0.001$, RR: $\beta = -22.705$, $p < 0.001$). Average output under HG, in contrast, experiences only a weak and non-significant negative time trend and stabilizes at a much higher level than in the other two treatments ($\beta = -3.744$, $p = 0.155$). In the final five periods average output under HG is roughly 700 points, compared to 330 points under DE (OLS: $p = 0.002$, RS: $p = 0.025$), and 350 points under RR (OLS: $p = 0.017$, RS: $p = 0.065$). The punishment frequencies show mildly negative time trends under DE ($\beta = -0.002$, $p < 0.001$) and RR ($\beta = -0.004$, $p = 0.002$), while the punishment frequency under HG remains by and large constant over time ($\beta = -0.001$, $p = 0.335$).

Commitment to punish, and reporting. Figure A.3 summarizes the impact of social learning on agents' commitment to punish and monitors' reporting intents. Regarding punishment commitments Panel A reveals that social learning almost exclusively affects agents' behavior under RR. In particular, the frequency with which Bad agents commit to punish under RR increases from 52% in the original treatment to 77% in the social learning treatment (OLS: $p = 0.009$, RS: $p = 0.030$). The rates at which Bad agents commit to punish under DE and HG and those of Good agents under all treatments do not significantly change in response to the presence of conditional payoff information.³¹ The increase in the frequency with which Bad agents commit to punish under RR in the social learning environment implies that there is no longer a difference in the commitment rate of Bad agents between RR (76.5%) and DE (77.3%, OLS: $p = 0.880$, RS: $p = 0.748$). Moreover, the commitment rate of Bad agents under HG (64%) is significantly lower than under both other elicitation mechanisms (HG vs. DE: OLS: $p = 0.006$, RS: $p = 0.020$, HG vs. RR: OLS: $p = 0.029$, RS: $p = 0.054$).³²

³¹In the following the first (resp. second) percentage corresponds to the rate at which agents commit to punish under the original (resp. social learning) treatment. DE: Bad agents (75% vs. 77%, OLS: $p = 0.778$, RS: $p = 0.574$), Good agents (44% vs. 48%, OLS: $p = 0.675$, RS: $p = 0.631$). RR: Good agents (25% vs. 31%, OLS: $p = 0.412$, RS: $p = 0.630$). HG: Bad agents (60% vs. 64%, OLS: $p = 0.512$, RS: $p = 0.575$), Good agents (28% vs. 23%, OLS: $p = 0.305$, RS: $p = 0.575$).

³²Note that even with conditional information, the commitment rate of Good agents is significantly higher under DE (48%) than under RR (31%, OLS: $p = 0.035$, RS: $p = 0.055$) and HG (23%, OLS: $p = 0.004$, RS: $p = 0.016$). The commitment rates of Good agents between RR and HG are not significantly different (OLS: $p = 0.209$, RS: $p = 0.575$).

Figure A.3: impact of social learning on commitment to punish and reporting intent

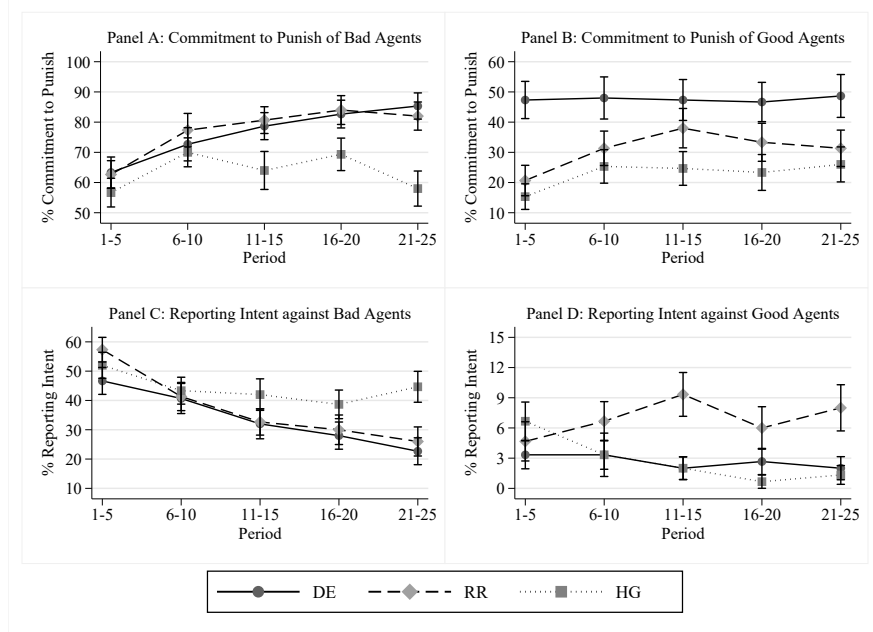


Note: The figure shows the observed frequencies of reporting intent and commitment to punish in all treatments. The variable *% Commitment to Punish* measures the within-type percentage of agents who commit to punish. The variable *% Reporting Intent* measures the frequency with which monitors intend to submit a positive report $r = 1$ against an agent as a function of the agent's quality and the treatment. Error bars mark ± 1 standard error from the mean (clustering at the individual level).

Panel B shows that the impact of social learning on monitors' reporting intents is also most pronounced under RR. While the introduction of conditional payoff information only leads to a moderate reduction in the frequency with which monitors intend to report Bad agents under DE (35% vs. 34%, OLS: $p = 0.861$, RS: $p = 0.520$) and HG (53% vs. 44%, OLS: $p = 0.086$, RS: $p = 0.261$), the frequency of intended reports under RR drops from 60% in the original treatment to 37% in the social learning treatment (OLS: $p = 0.018$, RS: $p = 0.037$). As a consequence, the rate of reporting intents against Bad agents is no longer different between RR and DE (OLS: $p = 0.610$, RS: $p = 0.810$). With respect to reporting intents against Good agents the presence of social learning opportunities implies that the rate of reporting intents drops to low levels under DE (9% vs. 3%, OLS: $p = 0.028$, RS: $p = 0.258$) and HG (5% vs. 3%, OLS: $p = 0.324$, RS: $p = 0.418$), but slightly increases under RR (5% vs. 7%, OLS: $p = 0.256$, RS: $p = 0.169$). These opposite developments imply that in the social learning treatments the rate of false reporting against Good agents is significantly

higher under RR than under DE (OLS: $p = 0.005$, RS: $p = 0.053$).³³

Figure A.4: commitment to punish and reporting intent under social learning



Note: The figure shows the development of reporting intent and commitment to punish in all social learning treatments (DE, RR and HG). The variable *% Commitment to Punish* measures the within-type percentage of agents who commit to punish. The variable *% Reporting Intent* measures the frequency with which monitors intend to submit a positive report $r = 1$ against an agent as a function of the agent's quality and the treatment. Error bars mark ± 1 standard error from the mean (clustering at the individual level).

Figure A.4 illustrates the dynamics of commitment to punish and reporting intents in the social learning treatment. The figure illustrates that both DE and RR suffer from increasingly undesirable behavior over time. In particular, Bad agents commit to punish more frequently over time (DE: $\beta = 0.007$, $p < 0.001$, RR: $\beta = 0.009$, $p < 0.001$, see Panel A) and are reported less frequently (DE: $\beta = -0.009$, $p < 0.001$, RR: $\beta = -0.015$, $p < 0.001$, see Panel C). Under HG these time trends are either absent (commitment rate of Bad agents:

³³Under RR the rate of reporting intents differs depending on whether monitors answer the relevant or the unrelated question. If monitors answer the relevant question, the introduction of conditional payoff information reduces reporting intents against Bad agents from 58% to 35% (OLS: $p = 0.020$, RS: $p = 0.055$) and reporting intents against Bad agents from 3% to 2% (OLS: $p = 0.594$, RS: $p = 0.337$). In case of the unrelated question, conditional payoff information reduces reporting intents against Bad agents from 64% to 45% (OLS: $p = 0.042$, RS: $p = 0.078$), but increases reporting intents against Bad agents from 12% to 21% (OLS: $p = 0.035$, RS: $p = 0.045$).

$\beta = 0.000$, $p = 0.960$) or weak (reporting of Bad agents: $\beta = -0.004$, $p = 0.086$). These findings reinforce our conclusion that the high performance of randomized response that we observed in our original treatments cannot be sustained in organizational settings in which social learning is feasible. The improved survey quality obtained under hard garbling, in contrast, remains stable.

B A More General QRL- k Model

Model. Consider the class of finite extensive-form games, with players $i \in I$, in which players move at most once, and past actions are public. The set of strategies of player i takes the form $S_i = \prod_{h_i \in H_i} A_{h_i}$ where A_{h_i} is the set of actions available to player i at history h_i . Let $S = \prod_{i \in I} S_i$ denote the set of strategies. For any sequence of marginal distributions over strategies $(\mu_i)_{i \in I} \in \prod_{i \in I} \Delta(S_i)$, we denote by μ_{-i} the product of independent distributions $\prod_{j \neq i} \mu_j$. Expected payoffs to player i are denoted by $u_i(s_i, s_{-i})$. Payoffs conditional on a decision node h_i and action $a_i \in A_{h_i}$ are simply denoted by $u_i(a_i, s_{-i})$.

Definition 2 (QRL- k model). *A quantal response level- k model of play consists of*

- (i) *A sequence $(\mu_{i,k})_{i \in I, k \in \mathbb{N}}$ of distributions of play $\mu_{i,k} \in \Delta(S_i)$, and independent noise terms $\varepsilon_i \in \mathbb{R}^{S_i}$ from a known parametric family, such that for all $s_i \in \text{supp} \mu_{i,k}$, $h_i \in H_i$, $a_i \in A_{h_i}$, and $k \geq 1$,*

$$\text{prob}_{\mu_{i,k}}(a_i = s_i(h_i)) = \text{prob}_{\varepsilon_i} \left(a_i = \arg \max_{a_i \in A_{h_i}} U_{h_i}(a_i, \mu_{-i,k-1}, \varepsilon_i) \right) \quad (5)$$

where $U_{h_i}(a_i, \varepsilon_i, \mu_{-i,k-1}) \equiv \mathbb{E}_{s_{-i} \sim \mu_{-i,k-1}} [u_i(a_i, s_{-i})] + \varepsilon_i(a_i)$.

- (ii) *A profile $(\lambda_i)_{i \in I}$ of distribution of levels $\lambda_i \in \Delta(\mathbb{N})$ describing the distribution of cognitive levels for each player.*

A QRL- k model of play induces a distribution $\mu^{\text{QRL}} \in \Delta(S)$ over strategy profiles $s = (s_i)_{i \in I}$ described by

$$\mu(s) = \sum_{(k_i)_{i \in I} \in \mathbb{N}^I} \prod_{i \in I} \lambda_i(k_i) \mu_{i,k}(-s_i). \quad (6)$$

Definition 3 (Common downward belief in rationality.). *We say that a player i of level- k exhibits common downward belief in rationality at history h_i if and only if*

- *h_i is a final decision node, and $k \geq 1$, or*

- *i* believes that any player *j* with a decision node h_j after h_i exhibits common downward belief in rationality at h_j .

Lemma 1 (limited impact of higher levels). *Consider an extensive-form single move game with N players and a QRL- k model of play. If $k \geq N$, then at any history h_i , player i exhibits common belief in downward rationality and $\mu_{i,k} = \mu_{i,k+1}$.*

Proof. Denote by $\#\text{succ}(h_i)$ the number of players that can playing after history h_i . Since the game is a single move game, whenever h' follows h , $\#\text{succ}(h') \leq \#\text{succ}(h) - 1$. Hence at any final decision node f (i.e. decision node that leads to final payoff realizations), there may have been at most $N - 1$ decisions taken. This implies common belief in downward rationality from the initial node.

The statement that $\mu_{i,k} = \mu_{i,k+1}$ follows from backward induction. ■

C Proofs

Proof of Proposition 1. Consider the case of direct elicitation. If the agent commits to punish, for K_M large enough, the monitor's best response is to report $r = 0$, so that the agent gets a payoff of 0 in equilibrium. Costs of punishment are not paid in equilibrium. If the agent does not commit to punish, the monitor finds it optimal send report $r = 1$ about the agent if and only if the agent is Bad, inducing a payoff of $-D$ for a Bad agent, and a payoff of 0 for a Good agent. As a result Bad agents find it optimal to commit to punish in SPE. Good agents are indifferent between committing to punish or not.

Under hard garbling, if the agent commits to punish, her payoff is bounded above by $-\pi K_A$: costs of punishment have to be paid on the equilibrium path. If the agent does not commit to punish, the monitor finds it optimal to send report $r = 1$ if and only if the agent is Bad. As a result, the agent gets a payoff greater than $-D$. It follows that no agent commits to punish.

Monitors always submit report $r = 0$ conditional on the agent type being Good, since this maximizes social surplus Y while minimizing expected potential costs $\mathbb{E}[d\tilde{r}K_M|r]$. ■

Proof of Proposition 2. The fact that DE and RR are outcome equivalent is immediate. The unrelated question is simply a relabeling of actions under direct elicitation.

Games **HG** and **oRR** differ only in the subgames after the agent commits to punish or not ($c \in \{0, 1\}$). Under **HG** the monitor sends message $r = 1$ if and only if

$$\begin{aligned} U_M(\tau, c, \tilde{r} = 1) &\geq \pi U_M(\tau, c, \tilde{r} = 1) + (1 - \pi) U_M(\tau, c, \tilde{r} = 0) \\ \iff U_M(\tau, c, \tilde{r} = 1) &\geq U_M(\tau, c, \tilde{r} = 0). \end{aligned}$$

Under **oRR**, when being asked to report the agent's type, the monitor sends message $r = 1$ if and only if $U_M(\tau, c, \tilde{r} = 1) \geq U_M(\tau, c, \tilde{r} = 0)$. By assumption, the monitor induces realized report $\tilde{r} = 1$ when asked the unrelated question. As a result, the equilibrium distribution of realized reports \tilde{r} conditional on any configuration (τ, c) coincide under **oRR**, and **HG**. As a result, any joint distribution of outcomes (τ, c, \tilde{r}) supported by equilibrium play in one game is supported by equilibrium play in the other game. ■

Proof of Proposition 3. We have that $\mathbb{E}_\mu[\tilde{R}] = R_B + R_G + (1 - R_B - R_G)\pi$, hence $R_B = \frac{\mathbb{E}_\mu[\tilde{R}] - \pi}{1 - \pi} - R_G$. By the law of large numbers, μ -a.s., $\lim_{N \rightarrow \infty} \tilde{R} = \mathbb{E}_\mu[\tilde{R}]$.

Since $R_\dagger = \mathbb{E}_\mu \tilde{R} - R_B - R_G$ substituting the expression for R_B above yields $R_\dagger = (1 - \mathbb{E}_\mu \tilde{R}) \frac{\pi}{1 - \pi}$. Equation (2) follows from the Law of Large Numbers.

It is immediate that $R_G = 0$ if the monitor is rational: regardless of whether the agent commits to punish or not, the monitor's payoff is maximized by sending report $r = 0$.

When $R_G = 0$, the expected mass of realized reports against Good agents $\mathbb{E}_\mu \tilde{R}_G$ satisfies

$$\begin{aligned} \mathbb{E}_\mu \tilde{R}_G &= R_G + (1 - q - R_G)\pi \\ &\leq R_G + (1 - R_B - R_G)\pi \\ &\leq \left(1 - \frac{\mathbb{E}_\mu \tilde{R} - \pi}{1 - \pi}\right) \pi = R_\dagger. \end{aligned}$$

This bound is tight whenever $q = R_B$, which occurs when all Bad types are reported. ■

Proof of Proposition 4. We first consider the hard-garbling game **HG**. Consider the behavior of a monitor after history (τ, c) . Since this is a final decision node, monitors behave rationally. The monitor chooses to send report $r = 1$ if and only if

$$U_M^\alpha(r = 1 | \tau, c) + \varepsilon_{r=1} \geq U_M^\alpha(r = 0 | \tau, c) + \varepsilon_{r=0}.$$

Let $\Delta U_M(\tau, c) \equiv U_M^\alpha(r = 1 | \tau, c) - U_M^\alpha(r = 0 | \tau, c)$ and $\bar{r}(\tau, c) \equiv \text{prob}(r = 1 | \tau, c)$. We have

that

$$\bar{r}(\tau, c) = .5\nu + (1 - \nu) \frac{\exp \frac{\Delta U_M(\tau, c)}{\sigma}}{1 + \exp \frac{\Delta U_M(\tau, c)}{\sigma}}.$$

Defining $\bar{\bar{r}}(\tau, c) \equiv \frac{\bar{r}(\tau, c) - .5\nu}{1 - \nu}$, we have that

$$\frac{\Delta U_M(\tau, c)}{\sigma} = \log \frac{\bar{\bar{r}}(\tau, c)}{1 - \bar{\bar{r}}(\tau, c)} = \log \frac{\bar{r}(\tau, c) - \nu/2}{1 - \bar{r}(\tau, c) - \nu/2}.$$

This implies that

$$\frac{\log \frac{\bar{r}(G, 0) - \nu/2}{1 - \bar{r}(G, 0) - \nu/2}}{\log \frac{\bar{r}(B, 0) - \nu/2}{1 - \bar{r}(B, 0) - \nu/2}} = -\frac{L_G}{\gamma L_B}. \quad (7)$$

It follows from the assumption that $\bar{r}(G, 0) < .5 < \bar{r}(B, 0)$ that the left-hand side of (7) is strictly decreasing in ν . Hence (7) has at most one solution. Given ν , values $\tilde{r}(G, c = 0)$ and $\tilde{r}(G, c = 1)$ pin-down σ :

$$\sigma = \frac{K_M(1 - \pi)}{\log \frac{\bar{\bar{r}}(G, 0)}{1 - \bar{\bar{r}}(G, 0)} - \log \frac{\bar{\bar{r}}(G, 1)}{1 - \bar{\bar{r}}(G, 1)}}.$$

Given σ , parameter α is given by

$$-\frac{\sigma \log \frac{\bar{\bar{r}}(G, 0)}{1 - \bar{\bar{r}}(G, 0)}}{L_G(1 - \pi)} - 1.$$

To pin down parameter ρ_{HG} , we need to consider play at non-terminal decision nodes. Consider a Bad agent's decision to commit to punish $c = 1$. The agent is level 1 with probability ρ_{HG} and level 2 with probability $1 - \rho_{\text{HG}}$. When the agent is level 1, she believes the monitor will send a report $r = 1$ whether or not she commits to punish. Hence an agent of level 1 commits to punish if and only if

$$-D - K_A + \varepsilon_{c=1} \geq -D + \varepsilon_{c=0}.$$

The probability of this event is

$$P_{c=1|lev1, B} = .5\nu + (1 - \nu) \frac{1}{1 + \exp(K_A/\sigma)}.$$

An agent of level 2 realizes that the monitor will be influenced by her threat, and anticipates

that depending on her commitment c , the monitor will send report $r = 1$ if and only if

$$\alpha Y_0 - cK_M + \varepsilon_{r=1} > \alpha(Y_0 - (1 - \pi)L_B) - \pi cK_M + \varepsilon_{r=0}$$

which occurs with probability

$$\mu_{r=1|c,B} = .5\nu + (1 - \nu) \frac{1}{1 + \exp\left(\frac{1-\pi}{\sigma}(cK_M - \alpha L_B)\right)}.$$

Hence a Bad agent of level 2 chooses to commit to punish whenever

$$-D\mu_{r=1|c=1,B} - K_A\mu_{r=1|c=1,B} + \varepsilon_{c=1} \geq -D\mu_{r=1|c=0,B} + \varepsilon_{c=0}.$$

This event occurs with probability

$$P_{c=1|lev2,B} = .5\nu + (1 - \nu) \frac{1}{1 + \exp\left(\frac{1}{\sigma}(D(\mu_{r=1|c=1,B} - \mu_{r=1|c=0,B}) + K_A\mu_{r=1|c=1,B})\right)}$$

Altogether, on average, a Bad agent chooses to commit to punish with probability

$$\mu_{c=1|B} = \rho_{HG}P_{c=1|lev1,B} + (1 - \rho_{HG})P_{c=1|lev2,B}$$

which pins down ρ_{HG} .

Game DE can be treated as a special case with $\pi = 1$. A similar proof holds for RR. ■

Proof of Proposition 5. Monitors of level 1 and 2 are both rational and act at final decision nodes. Up to a relabeling, the reports of the monitor have the same implied realized reports, and the same payoff consequences. As a result behavior by the monitor under the two games conditional on any final decision node must be identical. This yields point (i)

Consider now a Bad agent deciding whether or not to commit to punish:

- A Bad agent of level 1 believes the monitor is level 0 and complies with the framing of each game. As a result, under both RR and DE, the agent believes that with probability 1, the realized message will be $\tilde{r} = 1$. Hence under both RR and DE Bad agents of level 1 will behave identically.
- A Bad agent of level 2 realizes that the monitor is rational. We established under point (i) that a rational monitor would behave in a way that yields identical realized reports

under RR and DE. As a result, Bad agents of level 2 behave in identical ways across RR and DE.

If the share of level 1 and level 2 agents are the same across the two games, then the behavior of Bad agents should coincide across RR and DE. ■

Proof of Proposition 6. Consider the problem of a Good agent:

- A Good agent of level 1 believes that the monitor has level 0 and the the monitor's behavior is not influenced by threats. In addition, under DE, the agent believes that the realized report will be $\tilde{r} = 0$ with probability 1. Under RR, the agent believes that the realized report will be $\tilde{r} = 1$ with probability π . As a results, Good agents of level 1 will choose to commit to punish less frequently under RR than under DE.
- A Good agent of level 2 realizes that the monitor is rational. We established under point (i) that a rational monitor would behave in a way that yields identical realized reports under RR and DE. As a result, Good agents of level 2 behave in identical ways across RR and DE.

Whenever $\rho_{RR} \leq \rho_{DE}$ there are more level 1 agents under RR than DE. As a result, a smaller share of Good agents commits to retaliate under RR than DE. ■

D Instructions for Participants

We present an English translation of the original French instructions for participants in the randomized response treatment of our experiment. Instructions for participants in other treatments were very similar and are available from the authors on request.

These instructions were distributed on paper at the beginning of the experiments. The instructions were available to participants throughout the experiment.

At the end of this appendix we also show the section that we added to all instructions in the versions of the treatments with conditional payoff information. In addition, we also show how the information was displayed on participants' screens.

Instructions

Introduction

You are about to participate in an experiment of the University of Lausanne. During this experiment you have the opportunity to earn a sum of money that will be paid to you at the end of the experiment. The amount of money you earn may be more significant if

- you read the instructions carefully.
- you think carefully about the decisions you make.

If you have any questions while reading the instructions or while the experiment is in progress, feel free to call us by raising your hand. By contrast, any communication between participants—except through the channels offered as part of the experiment—is prohibited. In the event of non-compliance with these instructions, we will be obliged to exclude you from the experience without any payment.

In today's experiment, you will interact with other participants via your computer. The decisions you make will have an impact on your profit. Your decisions will also influence the profit of other participants, just as the decisions of other participants may influence your profit.

Your profit is calculated in points. At the end of the experiment your points will be converted into Swiss Francs according to the following exchange rate:

$$60 \text{ points} = 1 \text{ Swiss Franc}$$

Regardless of your decisions in the experiment, you will also receive a fixed amount of CHF 10 for your participation.

The experiment consists of several identical rounds. At the end of the session, your remuneration will be calculated as the sum of your income obtained in all these rounds.

I. Summary of the Experiment

There are 20 participants in this experiment. Each participant is randomly assigned to one of two roles: sender or reporter. There are 10 participants of each type.

You see your role displayed on your screen. Please write down your role here:

The experiment will last for 25 rounds. At the beginning of every round, each sender is randomly assigned to a new reporter with whom the sender will interact in this round. This interaction follows the same rules in each round. However, since the sender will be assigned to a new reporter in each round, he/she will interact with different reporters throughout the experiment.

The purpose of this first part of the instructions is to give you an overview of what will happen in the experiment. In the second part of these instructions, we will provide you with a much more detailed description of each step, including illustrations of how you will enter your decisions on the computer.

Interaction between the sender and the reporter

At the beginning of each round each sender receives a project. The sender's project can be of good quality, or of bad quality. The quality of the project is randomly determined and the sender cannot influence the quality.

After having been informed of the quality of his/her project, the sender must submit his/her project for inspection to the reporter. When the sender submits the project, he/she must send a message to the reporter. In this message the sender indicates whether or not he/she will reduce the reporter's profit if the project is blocked. This message is final and the sender cannot change his/her opinion later.

Subsequently, there are two possibilities: 1) Sometimes the reporter is asked to answer the question whether or not he/she wants to block the sender's project. 2) In other cases, the reporter is asked to answer "yes" to a question unrelated to the project. The sender never knows to which type of question the reporter has answered. The sender's project is always blocked if the reporter's answer is "yes" regardless of the question to which the reporter has answered. That is, if a sender's project has been blocked, the sender never knows for sure

whether the project has been blocked because the reporter wanted to block the project or because the reporter answered a question unrelated to the project.

If a project is implemented, the sender receives a bonus. This bonus does not depend on the quality of the project. The implementation of a project also has an impact on the total return, which is distributed among all reporters. If the quality of the project being implemented is good, the total return increases, if the quality is poor, the total return decreases. If a project is blocked, the sender must pay a penalty. In addition, blocking a project reduces the impact of a project on the total return (which is shared among the reporters). More specifically, blocking a good project reduces its positive impact, and blocking a bad project reduces its negative impact.

After the sender has been informed whether his/her project has been blocked or implemented, the sender's decision regarding the reduction of the reporter's profit is executed. The reporter's profit is reduced only if the project has been blocked and the sender has decided to reduce the reporter's profit in the event of a blocked project. If the reporter's profit is reduced, this also imposes a cost on the sender.

Finally, the profits are calculated. The sender's profit depends on the status of his/her project. If the project has been implemented the sender receives a bonus, but if the project has been blocked the sender must pay a penalty. In addition, the sender's profit also depends on whether or not he/she decides to reduce the reporter's profit (because a reduction of the reporter's profit is also costly for the sender). The reporter's profit depends on the total return that was created in the round. The greater the number of good projects that have been implemented and the greater the number of bad projects that have been blocked, the greater the profit of the reporter. In addition, the reporter's profit is reduced if the project has been blocked and the sender has decided to reduce the reporter's profit in the event of a blocked project. After the calculation of the profits the next round begins.

Remember: at the beginning of each new round, each sender is randomly assigned to a new reporter.

II. Detailed description of the experiment

The experiment is computerized. All decisions you make during the experiment must be entered via the computer in front of you.

In the second part of the instructions, we explain in detail what decisions you and other participants can make, how you can enter these decisions on the computer, and how these decisions affect your own profit and the profit of other participants. If you have any questions while reading the instructions, please raise your hand. An experimenter will come to you and answer your question.

1) **Assignment of the sender to a new reporter and initial endowment**

At the beginning of each round, each sender is randomly matched with a new reporter. The sender and the reporter each receive an initial endowment of 30 points. This initial endowment forms the basis for each participant's profit in each round. Depending on your own decisions and the decisions of other participants, your final profit in a round may be higher or lower than the initial allocation. It is possible that your profit is negative in some rounds. You have to cover such negative profits with the positive profits you earn in other rounds or, if necessary, with the fixed amount of CHF 10 that you receive for participation.

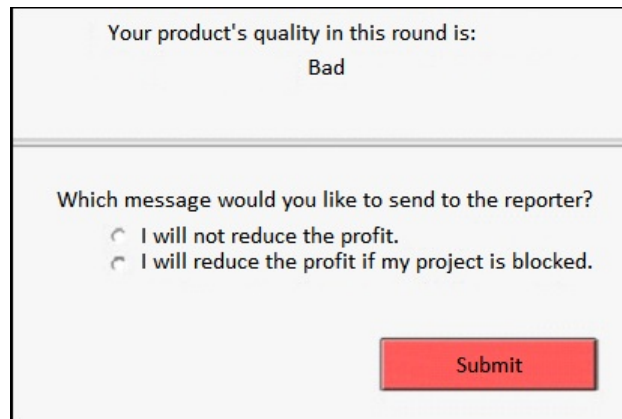
2) **Submission of the project by the sender and message to the reporter**

Each sender is assigned a new project in every round. This project can be of good or bad quality. Each quality is realized with a probability of 50%. The sender cannot influence the quality of the project. The quality of the project determines the impact of the project on the total return that is distributed among the reporters at the end of the round:

- A project of good quality increases the total return.
- A project of bad quality reduces the total return.

When the sender submits the project for inspection, he/she must attach a message in which he/she announces whether he/she will reduce the profit of the reporter in case of a blocked project, or not. This message is final and the sender cannot change this decision later. After choosing the message, the sender has to submit the project for inspection by clicking on the “submit” button.

The computer screen that provides project information to the sender and allows him/her to submit the project looks as follows:



A screenshot of a computer screen with a light gray background. The screen is divided into three horizontal sections. The top section contains the text "Your product's quality in this round is:" followed by "Bad" on the next line. The middle section contains the text "Which message would you like to send to the reporter?" followed by two radio button options: "I will not reduce the profit." and "I will reduce the profit if my project is blocked." The bottom section contains a red rectangular button with the word "Submit" in white text.

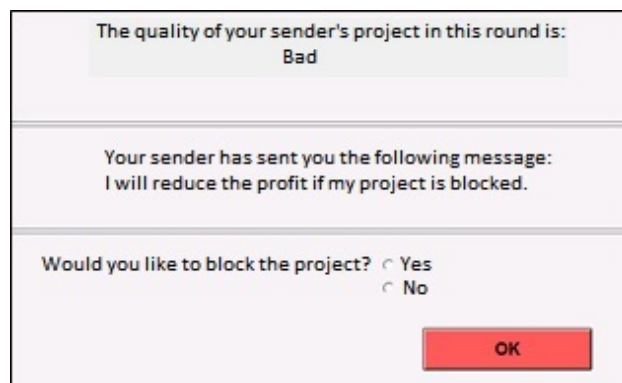
3) Evaluation of the project by the reporter

After the sender has submitted the project, the reporter is informed of the quality of the project and the sender's decision regarding the profit reduction.

Subsequently, there are two possibilities:

- i) **Evaluation:** The reporter is asked to answer the question whether he/she wants to block the sender's project or not. This possibility is realized with a probability of 75 percent.

The computer screen that asks the reporter whether or not he/she wants to block the project looks as follows:



A screenshot of a computer screen with a light gray background. The screen is divided into three horizontal sections. The top section contains the text "The quality of your sender's project in this round is:" followed by "Bad" on the next line. The middle section contains the text "Your sender has sent you the following message:" followed by "I will reduce the profit if my project is blocked." The bottom section contains the text "Would you like to block the project?" followed by two radio button options: "Yes" and "No". The bottom right corner contains a red rectangular button with the word "OK" in white text.

- ii) **Unrelated Question:** The reporter is asked to answer a question that has nothing to do with the sender's project (do you see the word "red" on your screen? Yes or no.) This possibility is realized with a probability of 25 percent. The correct answer to this question is always "yes", but the reporter can freely choose his/her answer.

The computer screen that shows the unrelated question looks as follows:



The screenshot shows a computer window with a light gray background. At the top, a dark gray header bar contains the text "The quality of your sender's project in this round is:" in white, followed by "Bad" in black. Below this, the main area contains the question "Do you see the word 'red' below?" in black. Underneath the question, the word "red" is displayed in a smaller font. At the bottom, there are two radio buttons: "Yes" and "No". The "Yes" radio button is selected. To the right of the radio buttons is a red rectangular button with the text "OK" in white.

Important:

The sender never knows whether the reporter has answered the evaluation question or the unrelated question. The sender's project is always blocked if the reporter's answer is "yes" regardless of the question to which the reporter has answered. If a sender's project is blocked, the sender cannot determine with certainty whether the project has been blocked because the reporter wanted to block the project or because the reporter answered a question unrelated to the project.

If the sender's project is not blocked, the project is implemented. In this case all its impact on the total return is realized:

- If a good project is implemented, it increases the total return by 400 points.
- If a bad project is implemented, it reduces the total return by 400 points.

If the sender's project is blocked, its impact on the total return is reduced:

- If a good project is blocked, the project increases the total return by only 300 points.
- If a bad project is blocked, the project reduces the total return by only 100 points.

4) Reduction of the reporter's profit by the sender

At the beginning of this phase the sender is informed if the project has been implemented or blocked.

If the project has been implemented the sender receives a 50 point bonus which is added to the initial 30 point endowment. The sender receives this bonus if and only if the project has been implemented, regardless of the quality of the project.

If the project is blocked, the sender not only loses 50 points bonus, but also has to pay a 50 points penalty which is deducted from the initial 30 points endowment. The payment of the penalty is also independent of the quality of the project and the sender must pay it in any case if the project has been blocked.

After observing whether the project has been implemented or blocked, the sender's decision regarding the reduction of the reporter's profit is executed. If the project has been blocked and the sender has decided to reduce the profit in case of a blocked project, the reporter's profit is reduced by 200 points. However, reducing the reporter's profit is also costly for the sender: he/she must pay 100 points from his/her own profit.

Important:

The sender's decision to reduce the reporter's profit in the event of a blocked project only has consequences if the project is blocked. If the project is implemented, nothing happens: the reporter's profit is not reduced by 200 points and the sender does not have to pay the 100 points for the reduction.

The computer screen that informs the sender whether or not the project has been blocked is as follows:

Your project has been implemented.
You have sent the following message to your reporter: I will reduce the profit if my project is blocked.
Your reporter's profit has not been reduced.
<input type="button" value="OK"/>

Subsequently, information about the projects that have been implemented and blocked in this round as well as the sender's profit and the reporter's profit are displayed on the screens.

III. Calculation of profits at the end of the round

In this third part of the instructions, we explain in detail how your decisions and the decisions of other participants in the experiment influence your profit and the profits of other participants.

The sender's profit

The sender's profit is calculated as follows:

Case 1: The sender's project has been implemented (in this case the sender's decision to reduce the reporter's profit in the event of a blocked project is not relevant):

$$\text{Sender Profit} = \text{Initial Endowment} + \text{Bonus}$$

Case 2: The sender decided not to reduce the reporter's profit and the sender's project was blocked:

$$\text{Sender Profit} = \text{Initial Endowment} - \text{Malus}$$

Case 3: The sender decided to reduce the reporter's profit in case of a blockage and the sender's project was blocked:

$$\text{Sender Profit} = \text{Init. Endowment} - \text{Malus} - \text{Cost of reducing reporter's profit}$$

Some examples:

- 1) Suppose that the sender has submitted a good quality project and has decided not to reduce the reporter's profit in the event of a blocked project. The project has been implemented.

The sender's profit is calculated as follows:

$$\text{Sender Profit} = 30 \text{ (Initial endowment)} + 50 \text{ (Bonus)}$$

$$\text{Sender Profit} = 80 \text{ points}$$

- 2) Suppose that the sender has submitted a poor quality project and has decided to reduce the reporter's profit in the event of a blocked project. The project has been implemented.

The sender's profit is calculated as follows:

$$\text{Sender Profit} = 30 \text{ (Initial endowment)} + 50 \text{ (Bonus)}$$

$$\text{Sender Profit} = 80 \text{ points}$$

- 3) Suppose that the sender has submitted a poor quality project and has decided not to reduce the reporter's profit in the event of a blocked project. The project has been blocked.

The sender's profit is calculated as follows:

$$\text{Sender Profit} = 30 \text{ (Initial endowment)} - 50 \text{ (Malus)}$$

$$\text{Sender Profit} = -20 \text{ points}$$

- 4) Suppose that the sender has submitted a good quality project and has decided to reduce the reporter's profit in the event of a blocked project. The project has been blocked.

The sender's profit is calculated as follows:

$$\text{Sender Profit} = 30 \text{ (Initial allocation)} - 50 \text{ (Malus)} - 100 \text{ (Cost of reduction)}$$

$$\text{Sender Profit} = -120 \text{ points}$$

The reporter's profit

The reporter's profit depends on the total return that has been generated by projects that have been implemented or blocked. The return increases with each good quality project and decreases with each bad quality project. Blocking a project reduces the impact of the project (positive or negative). The total return is calculated as follows:

$$\begin{aligned}\text{Total return} = & \text{Number of good projects implemented} \times 400 \text{ points} \\ & + \text{Number of good projects blocked} \times 300 \text{ points} \\ & - \text{Number of bad projects implemented} \times 400 \text{ points} \\ & - \text{Number of bad projects blocked} \times 100 \text{ points}\end{aligned}$$

For example:

- 1) Suppose that a total of three good projects have been implemented, two good projects have been blocked, two bad projects have been implemented and three bad projects have been blocked.

The total return is calculated as follows:

$$\text{Total yield} = 3 \times 400 + 2 \times 300 - 2 \times 400 - 3 \times 100 = 700 \text{ points}$$

- 2) Suppose that a total of five good projects have been implemented and five bad projects have been blocked.

The total return is calculated as follows:

$$\text{Total yield} = 5 \times 400 - 5 \times 100 = 1500 \text{ points}$$

The total return is distributed among all reporters, i.e. each reporter receives one tenth of the total return.

In addition, the reporter's profit also depends on whether or not the sender decides to reduce the reporter's profit.

The reporter's profit is calculated as follows:

Case 1: The sender's project has been implemented or the project has been blocked, but the sender has decided not to reduce the reporter's profit:

$$\text{Reporter Profit} = \text{Initial Endowment} + \text{Total Return} / 10$$

Case 2: The project was blocked and the sender decided to reduce the reporter's profit in the event of a blocked project:

$$\text{Reporter Profit} = \text{Initial Endowment} + \text{Total Return} / 10 - \text{Profit Reduction}$$

Some examples:

- 1) Suppose that the total return is 1000 points. The sender decided not to reduce the reporter's profit. The sender's project has been implemented. The reporter's profit is calculated as follows:

$$\text{Profit Reporter} = 30 \text{ (Initial allocation)} + 100 \text{ (Return} / 10)$$

$$\text{Profit Reporter} = 130 \text{ points}$$

- 2) Suppose the total return is 300 points. The sender decided to reduce the reporter's profit in the event of a blocked project. The sender's project has been implemented. The reporter's profit is calculated as follows:

$$\text{Profit Reporter} = 30 \text{ (Initial allocation)} + 30 \text{ (Return} / 10)$$

$$\text{Profit Reporter} = 60 \text{ points}$$

- 3) Suppose that the total return is 700 points. The sender decided to reduce the reporter's profit in the event of a blocked project. The sender's project has been blocked. The reporter's profit is calculated as follows:

$$\text{Profit Reporter} = 30 \text{ (Initial allocation)} + 70 \text{ (Return} / 10) - 200 \text{ (Reduction)}$$

$$\text{Profit Reporter} = -100 \text{ points}$$

At the end of each round, information about the types of projects that have been implemented and blocked, the sender's profit and the reporter's profit is displayed on the screen:

The screenshot shows a software interface with a yellow border. At the top left, it says "Period 1 of 25". Below this is a section titled "Summary of activities:". Inside this section is a table titled "Projects in this period:" with three columns: "Number" and "Return". The table lists: "Implemented good projects" (xxx, xxx), "Blocked good projects" (xxx, xxx), "Implemented bad projects" (xxx, xxx), "Blocked bad projects" (xxx, xxx), and "Total Return" (xxx). Below the table are two panels. The left panel, titled "Your profit", shows "Total Return xxx", "Profit Reduction xxx", and the formula "Your profit = Endowment + Return / 10 - Reduction = xxx". The right panel, titled "Sender's profit", shows "Sender's project Implemented Bonus: xxx", "Profit Reduction xxx", "Cost of reduction xxx", and the formula "Sender's profit = Endowment + Bonus - Cost of reduction = xxx". An "OK" button is at the bottom right.

Once the profit screen has disappeared, a new round begins in which the sender is randomly assigned to a new reporter.

Scenario:

To clarify the implications of the participants' decisions, we present a scenario. We will focus on a pair of players (a sender and a reporter) in a round of the experiment. We assume that the sender has a bad project in this round. In addition, we assume that the decisions of other participant pairs imply that five good projects and three bad projects have been implemented and one bad project has been blocked.

We now discuss all constellations of profits that can be realized:

Case 1: The sender decides not to reduce the reporter's profit.

a) The project is implemented.

$$\text{Total Return} = 5 \times 400 - 4 \times 400 - 1 \times 100 = 300 \text{ points}$$

$$\text{Sender Profit} = 30 \text{ (Endowment)} + 50 \text{ (Bonus)} = 80 \text{ points}$$

$$\text{Reporter Profit} = 30 \text{ (Endowment)} + 30 \text{ (Return / 10)} = 60 \text{ points}$$

b) The project is blocked.

$$\text{Total Return} = 5 \times 400 - 3 \times 400 - 2 \times 100 = 600 \text{ points}$$

$$\text{Sender Profit} = 30 \text{ (Endowment)} - 50 \text{ (Malus)} = -20 \text{ points}$$

$$\text{Profit Reporter} = 30 \text{ (Endowment)} + 60 \text{ (Return / 10)} = 90 \text{ points}$$

Case 2: The sender decides to reduce the reporter's profit in the event of a blocked project.

a) The project is implemented.

$$\text{Total Return} = 5 \times 400 - 4 \times 400 - 1 \times 100 = 300 \text{ points}$$

$$\text{Sender Profit} = 30 \text{ (Endowment)} + 50 \text{ (Bonus)} = 80 \text{ points}$$

$$\text{Reporter Profit} = 30 \text{ (Endowment)} + 30 \text{ (Return / 10)} = 60 \text{ points}$$

b) The project is blocked.

$$\text{Total Return} = 5 \times 400 - 3 \times 400 - 2 \times 100 = 600 \text{ points}$$

$$\text{Sender Profit} = 30 \text{ (Endowment)} - 50 \text{ (Malus)} - 100 \text{ (Cost of reduction)} = -120 \text{ points}$$

$$\text{Reporter Profit} = 30 \text{ (Endowm.)} + 60 \text{ (Return / 10)} - 200 \text{ (Reduction)} = -110 \text{ points}$$

Important :

Remember that the sender's project is always blocked if the reporter's answer is "yes" regardless of the question to which the reporter has answered. If a sender's project is blocked, the sender cannot determine with certainty whether the project has been blocked because the reporter wanted to block the project or because the reporter answered a question unrelated to the project.

IV. Control Questions

To ensure that you have understood the consequences of your decisions in this experience, we ask you to complete the following exercises. First, please write down all answers to the exercises on paper. Once you have completed the exercises, please enter your answers on the computer to verify that they are correct.

The experiment can only begin when everyone has answered these questions correctly.

If your screen is not yet on, simply move the mouse on your computer.

Exercise 1: Implementing or blocking projects

- a) With what probability will the reporter answer the question whether he/she wants to block the sender's project, or not?

Probability:

- b) With what probability will the reporter answer a question that is unrelated to the sender's project?

Probability:

Exercise 2: Calculation of total return

Suppose the sender has a good quality project.

- a) Suppose that in a round of the experiment five good projects were blocked and five bad projects were implemented. Please calculate the total return in this situation.

Total Return =

- b) Suppose that in a round of the experiment five good projects were implemented and five bad projects were blocked. Please calculate the total return in this situation.

Total Return =

- c) Suppose that in a round of the experiment four good projects and two bad projects were implemented and one good project and three bad projects were blocked. Please calculate the total return in this situation.

Total Return =

Exercise 3: Calculation of the reporter's profit

- a) Suppose the total return is 1000 points. The sender decided not to reduce the reporter's profit. The sender's project has been implemented. Please calculate the profit of the reporter.

Profit Reporter =

- b) Suppose the total return is 300 points. The sender decided to reduce the reporter's profit in the event of a blocked project. The sender's project has been blocked. Please calculate the profit of the reporter.

Profit Reporter =

- c) Suppose the total return is 1500 points. The sender decided to reduce the reporter's profit in the event of a blocked project. The sender's project has been implemented. Please calculate the profit of the reporter.

Profit Reporter =

Exercise 4: Calculating the sender's profit

- a) Suppose that the sender has received a good quality project. The sender decided to reduce the reporter's profit in the event of a blocked project. The project has been blocked. Please calculate the sender's profit.

Sender Profit =

- b) Suppose that the sender has received a good quality project. The sender decided not to reduce the reporter's profit. The project has been implemented. Please calculate the sender's profit.

Sender Profit =

- c) Suppose that the sender has been assigned a project of bad quality. The sender decided not to reduce the reporter's profit. The project has been blocked. Please calculate the sender's profit.

Sender Profit =

- d) Suppose that the sender has been assigned a project of bad quality. The sender decided to reduce the reporter's profit in the event of a blocked project. The project has been implemented. Please calculate the sender's profit.

Sender Profit =

Social Learning: Additional Section on Conditional Payoff Information

In all versions of our treatments with conditional payoff information the following section was added to the instructions right before the control questions (i.e., before section IV of the instructions):

Additional information on profits in the experiment:

This experiment has already been conducted with a substantial number of participants. In this session you have the possibility to benefit from the experience of previous participants. Before your first decision a table will appear on your screen.

The table will show you the average profits that other participants in the same role as you have realized with different decisions in this experiment. The displayed profits are based on decisions of 80 participants who have already taken part in the same experiment.

During the experiment you will always have the possibility to look at this table if you click on the “Information” button on your screen.

Screenshots: Conditional Payoff Information Displayed on Participant’s Screens

Sender’s screen:

Period 1 of 25		Remaining Time [sec]: 47
<p>The table below shows you average profits that previous participants in the same role as have realized with different decisions in this experiment.</p> <p>You are a sender. In each period you can decide whether or not you would like to reduce the profit of the reporter in case of a blocked project.</p> <p>The table shows the average profit for each possible decision as a function of the project's quality.</p>		
Quality of the project	Bad	Good
Do not reduce the profit in case of a blocked project.	xxx	xxx
Reduce the profit in case of a blocked project.	xxx	xxx
OK		

Reporter's screen:

Period		Remaining Time [sec]: 47
1 of 25		
<p>The table below shows you average profits that previous participants in the same role as have realized with different decisions in this experiment.</p> <p>You are a reporter. In each period you can decide whether or not you would like to block the sender's project.</p> <p>Before your decision you are informed about the quality of the project and the sender's decision regarding the reduction of your profit.</p> <p>The table shows the average profit for each possible decision.</p>		
Case 1: The sender has decided to not reduce the profit of the reporter in case of a blocked project.		
Quality of the project	Bad	Good
Do not block the project of the sender.	xxx	xxx
Block the project of the sender.	xxx	xxx
Case 2: The sender has decided to reduce the profit of the reporter in case of a blocked project.		
Quality of the project	Bad	Good
Do not block the project of the sender.	xxx	xxx
Block the project of the sender.	xxx	xxx
<div>OK</div>		

Remark: In the randomized response treatment the conditional payoff information of the reporter is displayed separately for the case in which the unrelated question is answered.