

# ONLINE APPENDIX

## Secure Survey Design in Organizations: Theory and Experiments

Sylvain Chassang and Christian Zehnder

### A Extensions

#### A.1 Contractual versus Reputational Incentives

In our model, the monitor is incentivized not to submit a report  $r = 1$  by ex ante threats from the agent: this is a contracting environment. In contrast, the survey methods inspired by Warner (1965) tend to be concerned with reputational incentives, i.e. what inferences people will draw about their behavior from realized reports. However, there is a close relationship between ex ante contractual threats, and ex post retaliation. In both cases, the effectiveness of threats depends on the informativeness of signal  $\tilde{r} = 1$ :

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

For this reason, garbling reports plays an essentially identical role under contractual and reputational incentives.

**Contractual incentives.** Consider a Bad agent, and assume for simplicity that altruism  $\alpha$  is less than  $\alpha^*$ , so that monitors send report  $r = 0$  whenever the agent commits to retaliate. Then it is optimal for the agent to commit to retaliate if and only if

$$D - K_A \times \text{prob}(\tilde{r} = 1|r = 0) > 0 \iff \frac{1}{\text{prob}(\tilde{r} = 1|r = 0)} > \frac{K_A}{D}.$$

Hence the monitor faces retaliation following  $\tilde{r} = 1$  if and only if  $\frac{1}{\text{prob}(\tilde{r}=1|r=0)}$  is large enough. Hence, sufficiently high garbling shuts down retaliation in equilibrium.

**Reputational incentives.** We now consider a setting in which the monitor is motivated by reputational incentives. There are no ex ante threats. Instead, 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).

Assume that the ex post expected punishment  $\bar{K}_M$  experienced by the monitor in the event  $\tilde{r} = 1$  is a function of the likelihood ratio of true reports conditional on the realization  $\tilde{r} = 1$ :

$$\bar{K}_M \left( \frac{\text{prob}(r = 1 | \tilde{r} = 1)}{\text{prob}(r = 0 | \tilde{r} = 1)} \right).$$

It follows from Bayes rule that

$$\begin{aligned} \log \left( \frac{\text{prob}(r = 1 | \tilde{r} = 1)}{\text{prob}(r = 0 | \tilde{r} = 1)} \right) &= \log \left( \frac{\text{prob}(\tilde{r} = 1 | r = 1)}{\text{prob}(\tilde{r} = 1 | r = 0)} \right) + \log \left( \frac{\text{prob}(r = 1)}{\text{prob}(r = 0)} \right) \\ &= \log \left( \frac{1}{\text{prob}(\tilde{r} = 1 | r = 0)} \right) + \log \left( \frac{\text{prob}(r = 1)}{\text{prob}(r = 0)} \right) \end{aligned}$$

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

**Reputation concerns without retaliation.** In many settings, instead of being concerned with potential retaliation or her own reputation, 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.

## A.2 Properties of QRL- $k$

**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  $h_i$ . For any profile of marginal distributions over strategies  $(\mu_i)_{i \in I} \in \prod_{i \in I} \Delta(S_i)$ , we denote by  $\mu_{-i}$  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 A.1** (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  $\varepsilon_i \in \mathbb{R}^{S_i}$  such that for all  $s_i \in \text{supp } \mu_{i,k}$ ,  $h_i \in H_i$ ,  $a_i \in A_{h_i}$ , and  $k \geq 1$ ,*

$$(7) \quad \text{prob}_{\mu_{i,k}}(a_i = s_i(h_i)) = \text{prob}_{\varepsilon_i} \left( a_i = \arg \max_{a_i \in A_{h_i}} U_{h_i}(a_i, \mu_{-i,k-1}, \varepsilon_i) \right)$$

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  $\mu^{\text{QRL}}$  over strategy profiles  $s = (s_i)_{i \in I}$  described by

$$(8) \quad \mu(s) = \sum_{(k_i)_{i \in I} \in \mathbb{N}^I} \prod_{i \in I} \lambda_i(k_i) \mu_{i,k}(s_i).$$

**Definition A.2** (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 A.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 play 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. ■

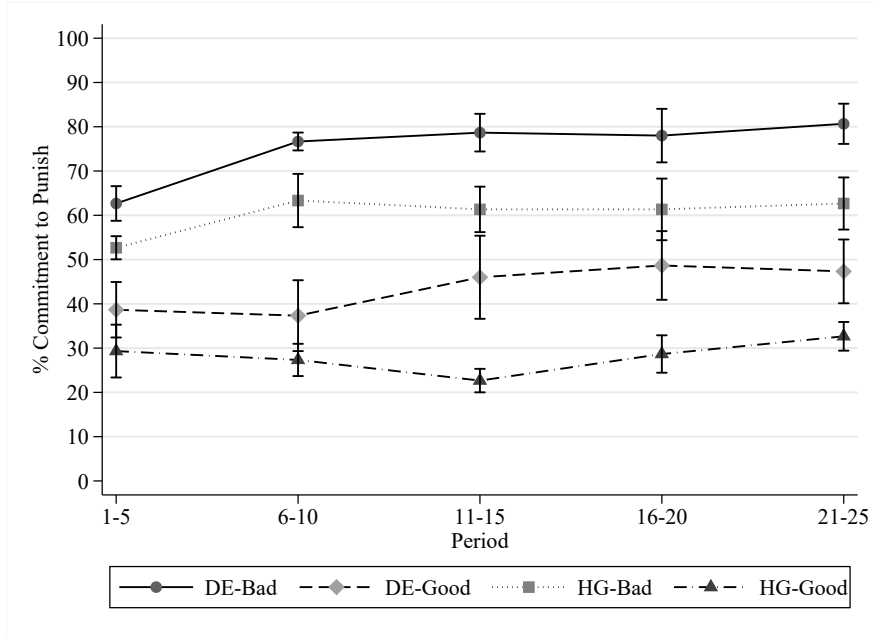
This implies that considering only players whose level of rationality is less than 2 is without loss of generality in the two-player survey games studied in Section VII.

## B Further Empirical Analysis

### B.1 Trends

This section reports trends in our first wave of experiments.

Figure B.1: agents' 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  $\pm 1$  standard error from the mean (clustered at the session level).

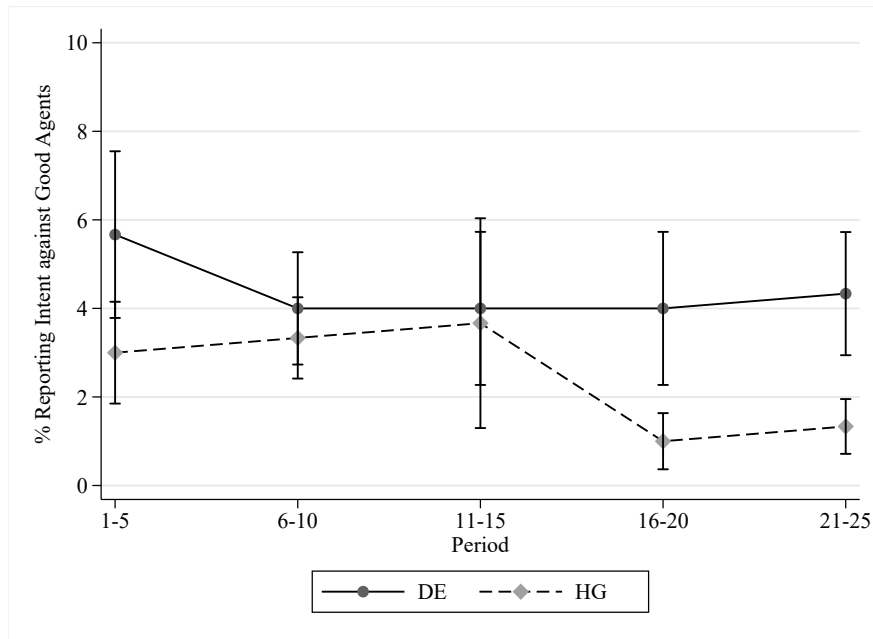
**Commitment to Punish.** Figure B.1 displays the time trends of the frequency with which Good and Bad agents commit to punish under direct response and hard garbling. The figure confirms that Bad agents consistently reduce the frequency of punishment threats in *HG* relative to *DE*. The frequency of 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:  $\beta = 0.004$ ,  $p = 0.234$ ).

In addition, Figure B.1 also shows that Good agents' commitment to punish under *HG* does not diminish over time in *HG* ( $\beta = 0.002$ ,  $p = 0.697$ ). This observation is important, because commitment to punish under *HG* is not consistent with equilibrium.

**Treatment effects.** The fact that monitors report Good agents more frequently under DE than HG biases treatment effect estimators. As Figure B.2 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 time ( $\beta = -0.0005$ ,  $p = 0.423$ ), whereas reporting 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:  $p = 0.210$ ).

Figure B.2: 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  $\pm 1$  standard error from the mean (clustered at the session level).

**Learning dynamics.** We note in Section VI that experimental behavior under RR is roughly self-confirming. Bad agents do not experiment with punishment threats, and fail to learn that monitors respond to incentives. One objection to this interpretation is that agents can get evidence that monitors do not always take the unrelated question seriously: Good agents who do not commit to punish face a reporting rate of 6% rather than 25% under RR.<sup>46</sup> 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 suggest that this is indeed the case.

Table B.1 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.

Table B.1: 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.

## B.2 Reporting of Good Agents

We noted in Section VII that monitors report Good agents at a greater rate under DE than either HG or RR, and that this is driven by the reporting of good agents who do not commit to punish. The QRL-k model correctly predicts that Good agents issue threats at lower rates

<sup>46</sup>It 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.

under RR and HG than DE, but fails to predict differences in the reporting of Good agents by monitors.

One possible explanation is that monitors may sometimes have antagonistic feelings against agents as a whole, and that those antagonistic feelings are more frequent under DE because DE shifts payoffs from monitors towards agents. Table B.2 shows that although on average, monitors tend to do better than agents, this varies across treatments. Under both RR and HG the average monitor payoff is roughly 60 points ahead of the average agent payoff, and the share of rounds in which the agent comes out ahead is under 15%. In contrast, under DE the average monitor payoff is only 12 points ahead the average agent payoff, and the agent comes out ahead in 42% of rounds. It is therefore plausible that monitors would be more likely to harbor antagonistic feeling towards agents under DE than either RR or HG.

Table B.2: Agent and monitor payoffs and behavior across treatments

|   | DE    | HG    | RR     |
|---|-------|-------|--------|
| Average Agent Payoff                          | 50.13 | 15.07 | 39.47  |
| Average Monitor Payoff                        | 62.0  | 74.2  | 100.87 |
| Share of Rounds Agent Ahead                   | 0.42  | 0.15  | 0.10   |
| Share of Good Non-Threatening Agents Reported | 0.13  | 0.06  | 0.06   |

In turn, reporting Good, non-threatening, agents is an efficient way for spiteful monitors to act on their feelings. Reporting a Good non-threatening agent causes a 10 points direct loss to the monitor, versus a 100 points direct loss to the agent. In contrast, reporting a Good agent that commits to punish causes a 200 points direct loss to the monitor, versus a 200 points direct loss to the agent, while reporting a Bad agent that commits to punish causes a 170 points direct loss to the monitor, versus a 200 points direct loss to the agent.

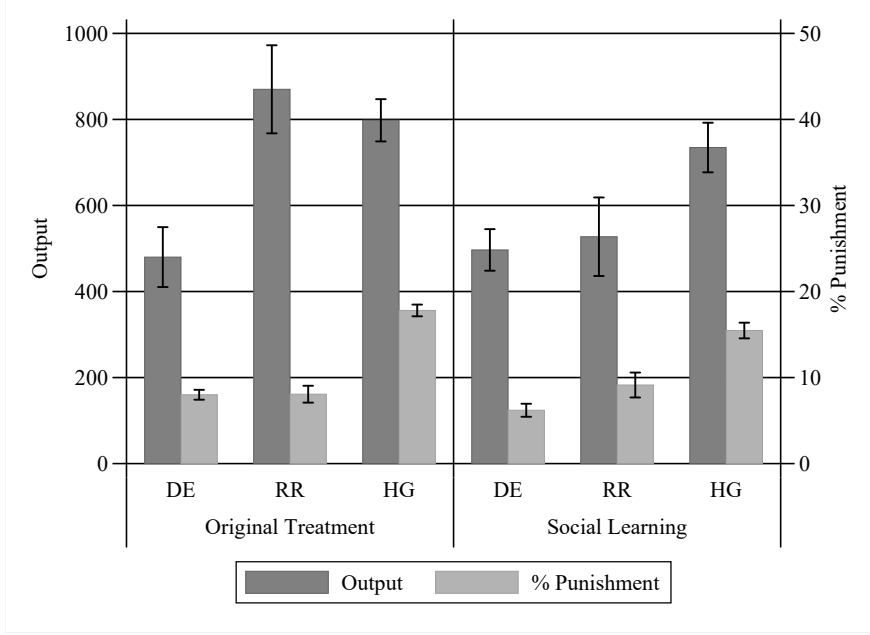
Reflecting such social preferences in the model of Section VII seems plausibly doable but goes beyond the scope of the paper.

### B.3 Social Learning Treatments

In this section we describe in greater detail the results from our second wave of experiments. In those treatments we provide participants with conditional payoff information elicited in

previous sessions of the same experiment.<sup>47</sup> We describe how conditional payoff information affects outcomes under our three elicitation mechanisms (DE, RR, HG).

Figure B.3: 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  $\pm 1$  standard error from the mean (clustered at the session level).

**Output and punishment.** Figure B.3 displays average output and punishment frequencies for all elicitation mechanisms in our original treatments and the social learning treatments. The figure reveals that our finding that randomized response gets the best of both survey procedures (see section VI.A) no longer holds once social learning is possible. While there is no significant impact on average output under DE and HG<sup>48</sup>, the availability of conditional payoff information reduces average output under RR from 870 to 527 points (OLS:  $p = 0.018$ , RS:  $p = 0.046$ ). As a consequence, there is no longer a significant difference in average output between DE (497) and RR (527) under social learning (OLS:  $p = 0.769$ , RS:

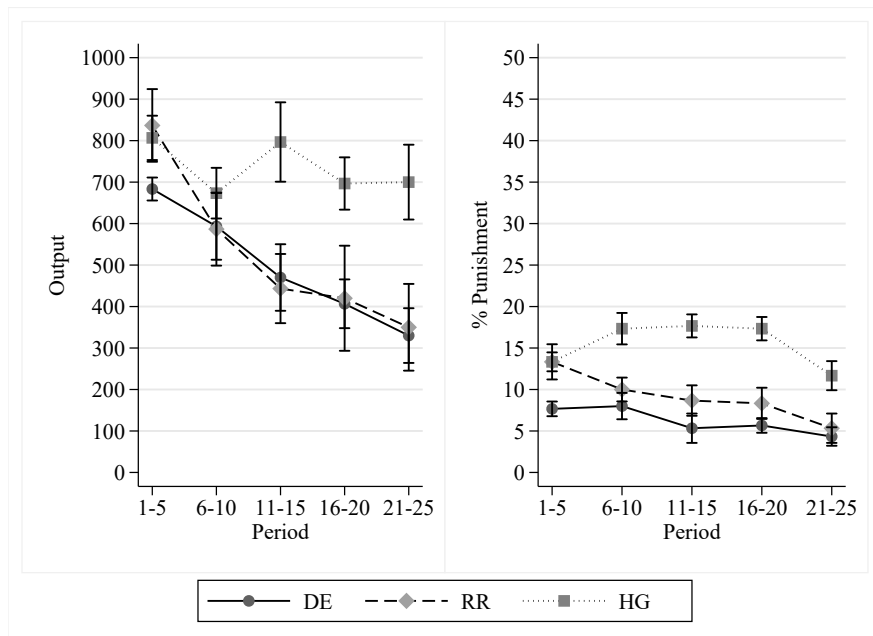
<sup>47</sup>Agents are informed about sample averages of agent profits conditional on agent type and commitment to punish. Monitors learn sample averages of monitor profits conditional on agent quality, agent's commitment to punish, and the reporting decision.

<sup>48</sup>Under DE average output slightly increases from 480 to 497 points (OLS:  $p = 0.846$ , RS:  $p = 1.000$ ) and under HG average output slightly decreases from 798 to 735 points (OLS:  $p = 0.410$ , RS:  $p = 0.699$ ).



$p = 0.873$ ).<sup>49</sup> 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:  $p = 0.143$ ) and HG (17.8% vs. 15.5%, OLS:  $p = 0.048$ , RS:  $p = 0.035$ ), the punishment frequency under RR increases insignificantly (8.1% vs. 9.1%, OLS:  $p = 0.546$ , RS:  $p = 0.563$ ).

Figure B.4: 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  $\pm 1$  standard error from the mean (clustered at the session level).

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 B.4 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

<sup>49</sup>In the presence of social learning average output under HG is significantly higher than under DE (OLS:  $p = 0.003$ , RS:  $p = 0.024$ ) and under RR (OLS:  $p = 0.064$ , RS:  $p = 0.097$ ).

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:  $p = 0.022$ ), and 350 points under RR (OLS:  $p = 0.017$ , RS:  $p = 0.071$ ). 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 B.5 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:  $p = 0.028$ ). 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.<sup>50</sup> 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 (77%) and DE (77%, OLS:  $p = 0.880$ , RS:  $p = 0.784$ ). 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:  $p = 0.017$ , HG vs. RR: OLS:  $p = 0.029$ , RS:  $p = 0.058$ ).<sup>51</sup>

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:  $p = 0.565$ ) and HG (53% vs. 44%, OLS:  $p = 0.086$ , RS:  $p = 0.292$ ), 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:  $p = 0.041$ ). As a consequence, the rate of reporting intents against Bad agents is no longer different between RR and DE (OLS:  $p = 0.611$ , RS:  $p = 0.855$ ). With respect to reporting

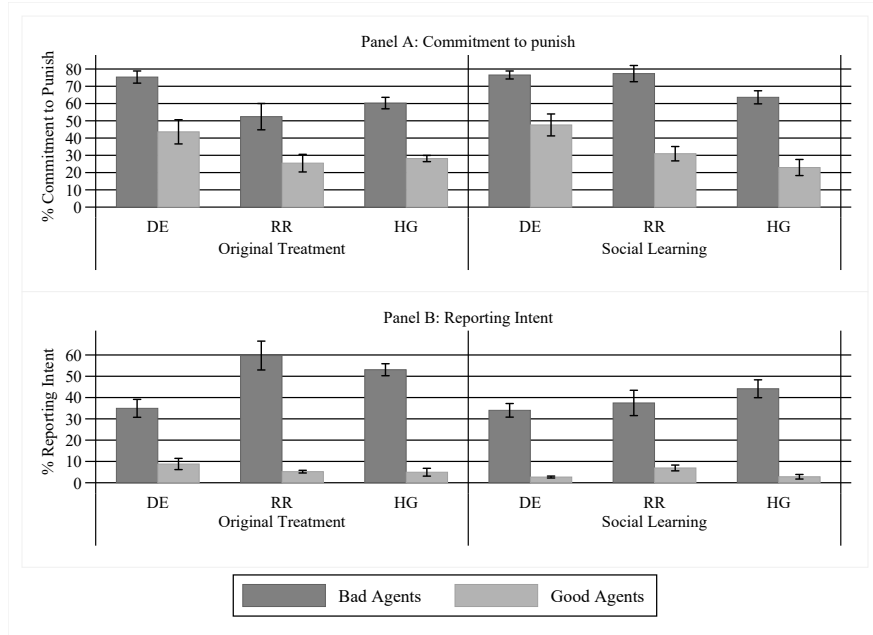
---

<sup>50</sup>In 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:  $p = 0.613$ ), Good agents (44% vs. 48%, OLS:  $p = 0.675$ , RS:  $p = 0.699$ ). RR: Good agents (25% vs. 31%, OLS:  $p = 0.412$ , RS:  $p = 0.686$ ). HG: Bad agents (60% vs. 64%, OLS:  $p = 0.512$ , RS:  $p = 0.619$ ), Good agents (28% vs. 23%, OLS:  $p = 0.305$ , RS:  $p = 0.619$ ).

<sup>51</sup>Note 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:  $p = 0.065$ ) and HG (23%, OLS:  $p = 0.004$ , RS:  $p = 0.015$ ). The commitment rates of Good agents between RR and HG are not significantly different (OLS:  $p = 0.209$ , RS:  $p = 0.619$ ).

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:  $p = 0.290$ ) and HG (5% vs. 3%, OLS:  $p = 0.324$ , RS:  $p = 0.463$ ), but slightly increases under RR (5% vs. 7%, OLS:  $p = 0.256$ , RS:  $p = 0.197$ ). This implies 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:  $p = 0.054$ ).<sup>52</sup>

Figure B.5: impact of social learning on commitment to punish and reporting intent



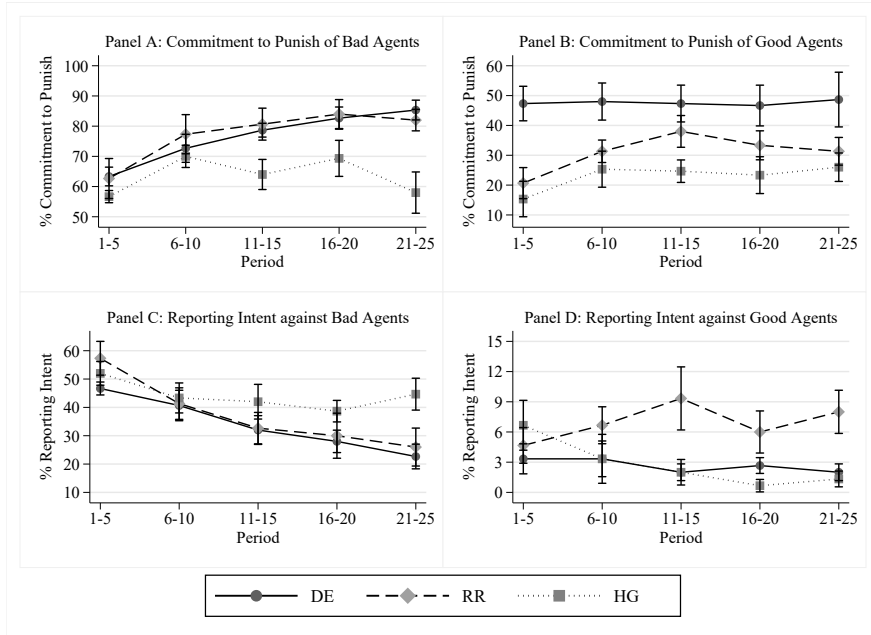
**Note:** The figure shows the observed frequencies of reporting intent and commitment to punish in all treatments. The variable *% Commitment to Punish* measures the within-type percentage of agents who commit to punish. The variable *% Reporting Intent* measures the frequency with which monitors intend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

Figure B.6 illustrates the dynamics of commitment to punish and reporting intents in the

<sup>52</sup>Under 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:  $p = 0.394$ ). In case of the unrelated question, conditional payoff information reduces reporting intents against Bad agents from 64% to 45% (OLS:  $p = 0.042$ , RS:  $p = 0.084$ ), but increases reporting intents against Bad agents from 12% to 21% (OLS:  $p = 0.035$ , RS:  $p = 0.048$ ).

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

Figure B.6: 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  $\pm 1$  standard error from the mean (clustered at the session level).

**Estimating treatment effects.** Figure B.7 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 shrinks 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:  $p = 0.896$ ). The 95% confidence interval for the difference is  $[-0.01, 0.01]$ . As a consequence, the estimated treatment effect is essentially 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:  $p = 0.699$ ), 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:  $p = 0.240$ ).<sup>53</sup>

## C Structural Investigation

The QRL- $k$  model is identified in experimental data: parameters  $(\alpha, \sigma, \nu, \rho)$  can be recovered from  $\mu^{\text{QRL}}$ , which can be estimated using the sample distribution.

**Proposition C.1** (model identification). *For all survey games, HG, DE, and RR, parameters  $\alpha, \sigma, \nu$  and  $\rho$  are identified from the following moments:*

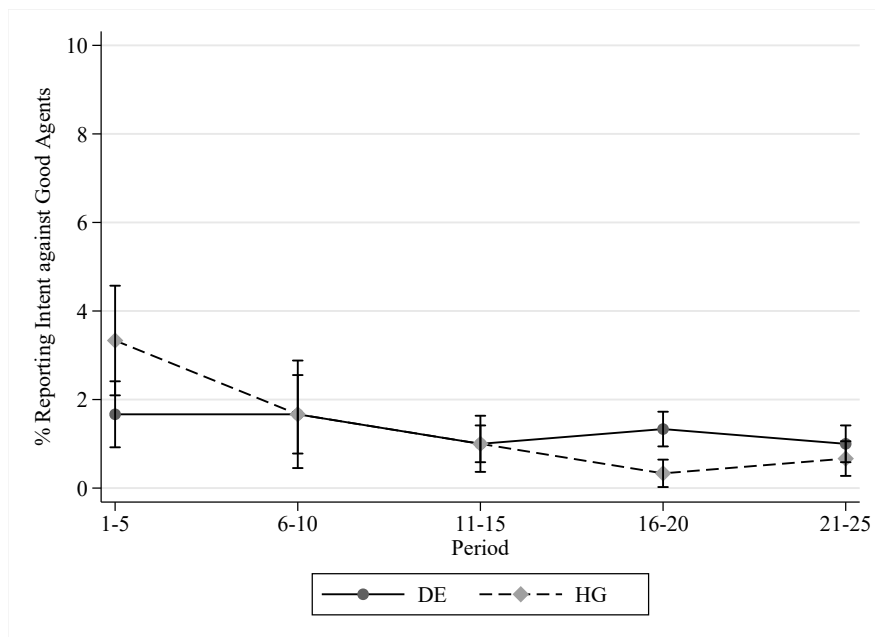
- *play by the monitor at all histories,  $(\mu^{\text{QRL}}(r = 1|c, \tau))_{\substack{c \in \{0,1\}, \\ \tau \in \{G,B\}}}$ , and,*
- *play by the agent conditional on her type,  $(\mu^{\text{QRL}}(c = 1|\tau))_{\tau \in \{G,B\}}$ .*

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 sensitivity of findings to different specifications.

---

<sup>53</sup>Under RR the reporting of Good agents remains somewhat higher in the presence of social learning:  $R_G^{RR} = 3.5\%$ . This rate is significantly higher than under DE (OLS:  $p = 0.005$ , RS:  $p = 0.054$ ) and HG (OLS:  $p = 0.023$ , RS:  $p = 0.037$ ). This effect is driven by responses to the unrelated question. If we exclude cases in which the monitor answered the unrelated question, the reporting rate for Good agents under RR drops to 1.2%. This rate is not significantly different from those under DE (OLS:  $p = 0.719$ , RS:  $p = 0.234$ ) or HG (OLS:  $p = 0.728$ , RS:  $p = 0.667$ ).

Figure B.7: 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  $\pm 1$  standard error from the mean (clustered at the session level).

**In-sample fit.** For each treatment HG, DE, and RR in our first wave of experiments, we estimate model parameters  $\rho$  (share of level 2 players),  $\alpha$  (monitor altruism),  $\sigma$  (the scale of payoff-responsive shocks), and  $\nu$  (the mass of payoff-non-responsive shocks). While parameters are identified analytically (Proposition C.1), we note that the model is overidentified. Given the potential for misspecification this leads us to estimate parameters using the simulated method of moments (McFadden, 1989), specifically by minimizing the relative distance between sample and simulated moments.

Table 3, in the main text, 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  $\rho$  is lower under RR than under DE. This does not impact the behavior of monitors (given agent's type and commitment decision) but it 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 payoff-non-responsive perturbations  $\nu$  is large,

especially under DE. As Table C.1 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. These contradictory facts end up being rationalized through a high rate of payoff-non-responsive errors.

Table C.1: in-sample fit of empirical (E) and simulated (S) moments; “report” refers to an intended report  $r = 1$

| moment                      | Emp./Sim. | HG    | DE    | RR    |
|-----------------------------|-----------|-------|-------|-------|
| threat given bad            | E         | 0.603 | 0.753 | 0.524 |
|                             | S         | 0.443 | 0.619 | 0.416 |
| threat given good           | E         | 0.281 | 0.436 | 0.255 |
|                             | S         | 0.200 | 0.516 | 0.306 |
| report bad given no threat  | E         | 0.893 | 0.832 | 0.933 |
|                             | S         | 0.976 | 0.958 | 0.966 |
| report bad given threat     | E         | 0.292 | 0.191 | 0.293 |
|                             | S         | 0.347 | 0.211 | 0.294 |
| report good given no threat | E         | 0.056 | 0.128 | 0.059 |
|                             | S         | 0.071 | 0.106 | 0.048 |
| report good given threat    | E         | 0.033 | 0.037 | 0.031 |
|                             | S         | 0.024 | 0.042 | 0.034 |

## D Proofs

**Proof of Proposition 1.** Monitors always submit report  $r = 0$  conditional on the agent type being Good, since this maximizes social surplus  $Y$  while minimizing expected potential costs  $\mathbb{E}[d\tilde{r}K_M|r]$ . In turn, whenever  $\pi > 0$  a Good agent maximizes her payoff by not committing to punish.

Consider the case where the agent is Bad. If the agent commits to punish, the monitor

sends report  $r = 1$  if and only if

$$\begin{aligned} \left(\frac{1}{n} + \alpha \frac{n-1}{n}\right) \gamma L_B - K_M &\geq \pi \left[ \left(\frac{1}{n} + \alpha \frac{n-1}{n}\right) \gamma L_B - K_M \right] \\ \iff \alpha &\geq \frac{nK_M - \gamma L_B}{(n-1)\gamma L_B} = \alpha^*. \end{aligned}$$

If instead the agent commits not to punish, since  $\alpha \geq 0$ , the monitor reports the agent with probability 1.

Altogether, recalling that  $p = \text{prob}(r = 1)$ , this implies that it is optimal for the agent to commit to punish if and only if

$$\begin{aligned} -(\pi + (1 - \pi)p) \times (D + K_A) &\geq -D \\ \iff \pi + (1 - \pi)p &\leq \frac{D}{D + K_A}. \end{aligned}$$

This concludes the proof.  $\blacksquare$

**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 distributions of realized reports  $\tilde{r}$  conditional on any configuration  $(\tau, c)$  coincide under oRR, and HG. As a result, any joint distribution of outcomes  $(\tau, c, \tilde{r})$  supported by equilibrium play in one game is supported by equilibrium play in the other game.  $\blacksquare$

**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,  $\mu$ -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 (4) 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

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

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

**Proof of Proposition 4.** 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 5.** Consider the problem of a Good agent:

- A Good agent of level 1 believes that the monitor has level 0 and 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  $\pi$ . As a result, 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) of Proposition 4 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.

In turn, the behavior of monitors is the same across treatments RR and DE conditional on the agent's commitment to retaliation. Hence, a lower probability of commitment to retaliate under RR than DE translates into a higher aver share of positive reports against Good agents under RR than DE. ■

**Proof of Proposition C.1.** We first consider the hard-garbling game HG. Consider the behavior of a monitor after history  $(\tau, c)$ . 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

$$(9) \quad \frac{\log \frac{\bar{r}(G, 0) - \nu/2}{1 - \bar{r}(G, 0) - \nu/2}}{\log \frac{\bar{r}(B, 0) - \nu/2}{1 - \bar{r}(B, 0) - \nu/2}} = -\frac{L_G}{\gamma L_B}.$$

It follows from the assumption that  $\bar{r}(G, 0) < .5 < \bar{r}(B, 0)$  that the left-hand side of (9) is strictly decreasing in  $\nu$ . Hence (9) has at most one solution. Given  $\nu$ , values  $\tilde{r}(G, c = 0)$

and  $\tilde{r}(G, c = 1)$  pin-down  $\sigma$ :

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

Given  $\sigma$ , parameter  $\alpha$  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  $\rho_{\text{HG}}$ , 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  $\rho_{\text{HG}}$  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  $\rho_{HG}$ .

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

## E 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**

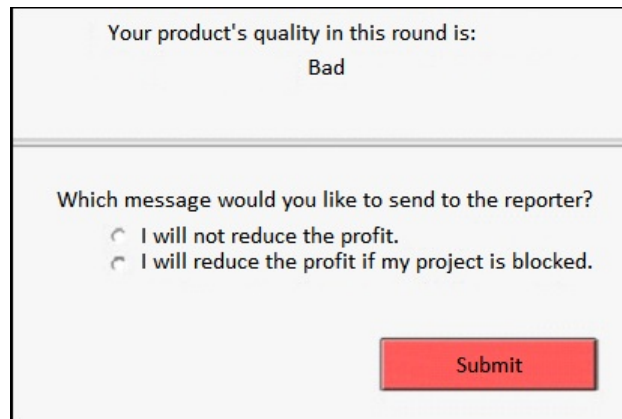
Each sender is assigned a new project in every round. This project can be of good or bad quality. Each quality is realized with a probability of 50%. The sender cannot influence the quality of the project. The quality of the project determines the impact of the project on the total return that is distributed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

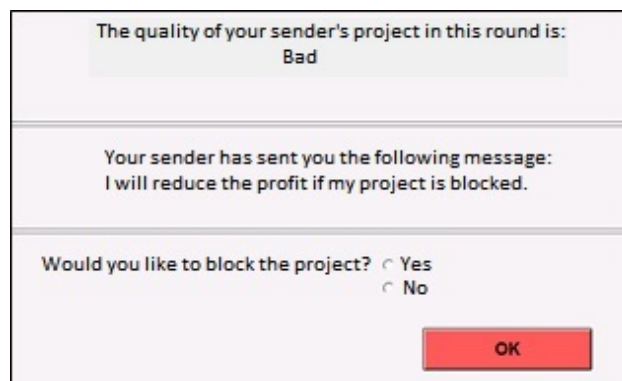
### 3) **Evaluation of the project by the reporter**

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.

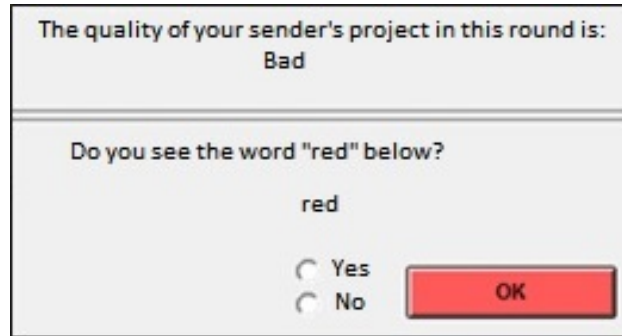
The computer screen that asks the reporter whether or not he/she wants to block the project looks as follows:



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 quality of your sender's project in this round is:  
Bad

---

Do you see the word "red" below?

red

☐ Yes ☐ No

OK

**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:<br>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 four types of projects: "Implemented good projects", "Blocked good projects", "Implemented bad projects", and "Blocked bad projects", each with "xxx" in both the "Number" and "Return" columns. Below the table is a row for "Total Return" with "xxx" in the "Return" column. Below the table is a section titled "Your profit" with a calculation: "Your profit = Endowment + Return / 10 - Reduction = xxx". To the right of this is a section titled "Sender's profit" with a calculation: "Sender's profit = Endowment + Bonus - Cost of reduction = xxx". Below the "Sender's profit" section is a table with two columns: "Sender's project" and "Implemented". The table lists three rows: "Bonus", "Profit Reduction", and "Cost of reduction", each with "xxx" in the "Implemented" column. At the bottom right of the interface is an "OK" button.

| Projects in this period:  |        |        |
|---------------------------|--------|--------|
|                           | Number | Return |
| Implemented good projects | xxx    | xxx    |
| Blocked good projects     | xxx    | xxx    |
| Implemented bad projects  | xxx    | xxx    |
| Blocked bad projects      | xxx    | xxx    |
| Total Return              |        | xxx    |

**Your profit**

Total Return xxx  
Profit Reduction xxx  
Your profit = Endowment + Return / 10 - Reduction = xxx

**Sender's profit**

Sender's project Implemented  
Bonus: xxx  
Profit Reduction xxx  
Cost of reduction xxx  
Sender's profit = Endowment + Bonus - Cost of reduction = xxx

OK

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<br>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.