Enjoy the Silence: An Experiment on Truth-telling*

Santiago Sánchez-Pagés† and Marc Vorsatz‡

January 19, 2007

Abstract

We analyze experimentally two sender-receiver games with conflictive preferences. In the first game, the sender can choose to tell the truth, to lie, or to remain silent. The latter strategy is costly and similar to an outside option. If sent, the receiver can either trust or distrust the sender's message. In the second game, the receiver must decide additionally whether or not to costly punish the sender after having observed the history of the game. We investigate the existence of two kinds of social preferences: Lie-aversion and preference for truth-telling. In the first game, senders tell the truth more often than predicted by the sequential equilibrium concept, they remain silent frequently, and there exists a positive correlation between the probability of being truthful and the probability of remaining silent. Our main experimental result for the extended game shows that those subjects who punish the sender with a high probability after being deceived are precisely those who send fewer but more truthful messages. We then explore two formal models of the baseline game that can account for our experimental results. First, we fit the data to the logit agent quantal response equilibrium; secondly, we solve for the Perfect Bayesian Nash equilibria of a stylized version of the baseline game with two types of senders. The equilibrium predictions obtained in both cases are consistent with both preferences for truth-telling and lie-aversion although the latter seems to be more pronounced.

Keywords: Experiment, Lie-Aversion, Social Preferences, Strategic Information Transmission, Truth-Telling.

JEL-Numbers: C72, C73, D83.

*We thank Alvin Roth for his suggestions at the early stage of this work and we are very grateful to Jordi Brandts, Sjaak Hurkens, and seminar audiences at Maastricht and Aberdeen for their valuable comments. This research has been possible thanks to the financial support of the Development Research Trust Fund of the University of Edinburgh.

†Address: Economics, University of Edinburgh, 50 George Square, EH8 9JY, Edinburgh, U.K., E-Mail: ssanchez@staffmail.ed.ac.uk

‡Address: Corresponding author. Department of Economics (AE1), University Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands, E-Mail: m.vorsatz@algec.unimaas.nl

1 Introduction

Strategic information transmission games, introduced by Crawford and Sobel [7], are characterized by the following structure: A *sender* with private information regarding a state variable transmits a message about the actual state to the *receiver* who, in turn, takes a subsequent action that is payoff-relevant to both players. The main theoretical result of Crawford and Sobel [7] states that less information about the true state is revealed by the sender's message as the preferences of the sender and the receiver become less aligned.

Several laboratory experiments have analyzed individual behavior in strategic information transmission games. In the first experimental analysis of this kind, Dickhaut et. al [8] corroborated the main insight of Crawford and Sobel [7]. More recently, Cai and Wang [3] have provided evidence showing that senders reveal more private information than predicted by the most informative sequential equilibrium and explained this finding by means of the quantal response equilibrium (McKelvey and Palfrey [15] and [14]) and a behavioral type analysis (see, among others, Stahl and Wilson [19], Nagel [16], Costa-Gomes et al. [5], and Crawford [6]). Other experimental works have further investigated the importance of moral behavior in these settings. Gneezy [11] has shown that if preferences are conflictive (the better the outcome fot the receiver is, the worse for the sender and vice versa), the probability of lying is increasing in the potential gains to the sender and decreasing in the potential loss to the receiver. Sutter [20] has established the existence of a second-order deception because telling the truth in Gneezy's experiment can also be regarded as a lie if the sender expects the receiver not to trust the message. Hurkens and Kartik [12] have found that Gneezy's data cannot reject the hypothesis that subjects are of one of the following two behavioral types: either an individual does not lie at all or s/he lies whenever the outcome obtained from lying is preferred to the outcome obtained from telling the truth. Finally, Sánchez-Pagés and Vorsatz [17] have shown that moral behavior plays a decisive role in explaining the reported excessive truth-telling with respect to the standard sequential equilibrium analysis. In particular, if (a) the sender can only choose between the messages *truth* and *lie*, (b) the receiver's strategy set contains only the actions *trust* and *distrust*, and (c) preferences are conflictive but the final payoff distribution is not too unequal, senders tell the truth significantly more often than predicted by the sequential equilibrium analysis with purely selfish players. If this baseline game is extended by allowing the receiver to costly punish the sender once the outcome of the

baseline game is observed, the punishment rate after the history (*lie,trust*) turns out to be considerably higher then the one after the history (*truth,distrust*) although the implemented payoff distributions are identical. The authors also show that those subjects who do punish very often after the history (*lie,trust*) tell also the truth very often in the extended game, whereas the rest of the subjects play equilibrium strategies on the aggregate. Since the overall percentage of truth-telling in the baseline game is not significantly different from the one in the extended game, there is strong evidence that the excessive truth-telling in the original game is also linked to some kind of moral behavior.

The present work continues the analysis of social preferences in strategic information transmission games along this line. One may think of two distinct types of social preferences that can explain excessive truth-telling, *preference for truth-telling* and *lie-aversion*. According to the first one, individuals gain some extra utility from being honest whereas the latter implies that they loose some utility when they deceive others. Since it is impossible to distinguish between these two forms of preferences if the message space M of the sender is $M = \{truth, lie\}$, it is necessary to construct a new experiment in order to study which of them is prominent and how they relate to each other. One possible way to achieve this goal is to extend our previous analysis by introducing a third strategy for the sender: *Remaining silent*. By giving subjects this option we can discriminate between the two types of moral behavior outlined above. For example, subjects may have a strong preference for truth-telling that would make *truth* a dominant strategy. Hence, senders would always send a message. Note that this would not be any longer the case if subjects display a symmetric strong lie-aversion because then *lie* becomes a dominated strategy and since telling the truth under conflictive preferences would give the sender only a low material payoff, she would prefer to remain silent in the first place. On the other hand, remaining silent may in our framework be due to other two reasons: First, subjects may have a weaker preference for truth-telling and recognize that if required to send a message they will tend to be truthful. If, additionally, receivers trust often, the sender will get only a low material payoff and may thus prefer to remain silent. Alternatively, subjects may be lie-averse and since material incentives favor the use of deception, remaining silent becomes a way to avoid lying.

In our experimental analysis we consider two different games. In the *Benchmark Game* the sender chooses among the actions *truth, lie,* and *remain silent*. If the sender transmits

a message the receiver can either *trust* or *distrust* the message, if the sender remains silent the receiver can just take a random action. Since preferences are not aligned, a rational and self-interested sender will optimally not transmit any information. There is a large set of strategies to achieve this but the percentage of truthful messages must be one-half in any of them. Since remaining silent does not change the receiver's beliefs either but comes at a strictly positive cost c, a rational and self-interested sender would never take this action and, therefore, all sequential equilibria of the baseline game are characterized by the fact that the percentage of truth-telling and the percentage of lies is one-half and messages are always sent.[1] In our parametrization of the baseline game we find that the sequential equilibrium prediction does not hold on two grounds: the percentage of truthful messages is greater than fifty percent (it is 51.67% for $c = 0.1$ and 53.9% for $c = 0.5$ but only the latter difference is significant) and subjects remain silent rather often (the corresponding percentages are 10.17% for $c = 0.1$ and 10.25% for $c = 0.5$). The former result hints into the direction of a weak preference for truth-telling whereas the latter suggests the existence of lie-aversion. Moreover, since subjects who send more truthful messages are also more likely to remain silent, it seems to be the case that both types of social preferences are positively correlated.

To analyze this point further we proceed in the following way. First, the *Punishment Game* extends the baseline treatment by giving the receiver the option of reducing the payoff distribution to (0,0) once he observes the potential outcome of the baseline game. A rational and self-interested receiver would never reduce payoffs and, consequently, the standard sequential equilibrium prediction with respect to the percentage of truth-telling and the percentage of lies is the same as in the baseline game. We find however that the punishment rate after history (*lie,trust*) is above 42% for both c. Note that, the payoff distribution associated to the history (*truth,distrust*) is the same as the one after history (*lie,trust*). But since the corresponding punishment rates are "only" equal to 18% for $c = 0.1$ and 4% for $c = 0.5$, a lie is clearly perceived as a deception if the receiver trusts the message. Moreover, due to the specific role rotation mechanism we implement, we are able to track down the behavior of every subject in both roles. It turns out that those subjects who punish very often after history (*lie,trust*) behave in a morally consistent way; that is, they are responsible for the

[1] In order to avoid misunderstandings we explain the terminology *percentage of truthful messages* and *percentage of truth-telling*. The former is the probability of telling the truth conditional on a message being sent whereas the latter refers to the unconditional probability.

excessive truth-telling and remain silent more often. In fact, they tell the truth in 87.65% (76.59%) of the occasions for $c = 0.1$ ($c = 0.5$) whereas the remaining subjects do so in only 52.08% (51.41%) of all occasions. Similarly, the morally consistent individuals remain silent in 30.83% (33.25%) of all observations for $c = 0.1$ ($c = 0.5$), but the remaining group of individuals does so only in 16.53% (13.25%) of all occasions. Moreover, they trust a 25These results corroborate the central role of social preferences in the Punishment Game.

In a second step, we formulate the logit agent quantal response equilibrium of McKelvey and Palfrey [15] for the Benchmark Game when individuals have social preferences. We introduce into the utility specification of the sender one parameter that measures preference for truth-telling and one parameter that measures the degree of lie-aversion. The receiver on the other hand is supposed to face a utility loss whenever he gets the lower payoff, after history (*truth,distrust*) because he is inequity averse and after (*lie,trust*) because he feels deceived. If we now take the punishment rates after these histories as estimates for the utility loss faced by the receiver and fit the data of the Benchmark Game to the equilibrium predictions, it is possible to determine the two parameters of social preferences uniquely. We obtain that truth-telling enhances the senders payoff by roughly 4% for $c = 0.1$ and 12% for $c = 0.5$. On the other, lying decreases the payoff around 10% for both c. Hence, the data of the Benchmark Game is consistent with both types of social preferences in this indirect equilibrium approach, although lie-aversion seems to be slightly more prominent.

One drawback of the former quantification is that we have to consider aggregate data. But individual behavior is diverse and, therefore, it is necessary to explain the structural features of our data by means of a theoretical model with different types of players. We consider a stylized version of the Benchmark Game in which remaining silent becomes an outside option. Moreover, the sender is assumed to have either a high or a low cost of lying and the receiver prefers the history (*truth,distrust*) over (*lie,trust*). For the parameter values of interest, there is one pooling and one separating Perfect Bayesian Nash equilibrium. The percentages of truthful and trusted messages are greater than fifty percent in both equilibria, the difference is that the sender with a high cost of lying tells always the truth in the pooling equilibrium whereas she remains silent in the separating equilibrium. This prediction can therefore account for the positive correlation between the percentage of truthful messages and the probability of remaining silent simply because subjects may coordinate sometimes on

the pooling and sometimes on the separating equilibrium.

We proceed as follows. In the next section, we introduce the games formally and describe our experimental design and procedures. Afterwards, we present the results of the experiments. In Section 4, we perform the theoretical analysis. Finally, we conclude. Complementary material (proofs, the instructions corresponding to the Punishment Game, and a description of the computer programs and the raw data) is relegated to the Appendix.

2 Experimental Design and Procedures

In this section, we introduce the two games we compare experimentally and provide a detailed overview of the experimental procedures. We start with the baseline treatment.

The Benchmark Game

Let $P = \{\text{sender,receiver}\}$ be the set of players. At the beginning of the game, the payoff tables A and B are selected randomly with equal probability, *i.e.* $p(A) = p(B) = 0.5$, and only the sender is informed about the payoff table actually chosen. Selecting table $\theta \in \Theta = \{A, B\}$ means that final payoffs are realized according to θ. Both tables depend only on the action U or D taken by the receiver later on.

Table A	Sender	Receiver
Action U	5	1
Action D	1	5

Table B	Sender	Receiver
Action U	1	5
Action D	5	1

Table 1: Payoff Tables

After the sender has been informed, she chooses a mixed strategy with support on her action space $M = \{A, B, N\}$. That is, if table θ is actually selected, the sender communicates with probability $p(m|\theta)$, $m \in \Theta$, that table m determines payoffs and with probability $p(N|\theta)$ she remains silent and pays the strictly positive cost c. If the sender takes action $m \in M$, the receiver believes with probability $\mu(\theta|m)$ that the actual payoff scheme is represented by table θ. Taking into account these beliefs, the receiver chooses a mixed strategy with support on her action set $\mathcal{A} = \{U, D\}$. Formally, given $m \in M$, the receiver takes action $a \in \mathcal{A}$ with probability $q(a|m)$. Finally, both players receive their payoff. We denote by $x_s(a, m, \theta)$ and

$x_r(a, m, \theta)$ the payoffs of the sender and the receiver when the receiver takes action a upon message m and the true state is θ. ■

In the Benchmark Game, the sender chooses a mixed strategy over the set of actions telling the truth, lying, and remaining silent. In the two former cases the receiver has two actions at hand: He can trust the message (*i.e.*, take action D upon message A and action U upon message B) or he can distrust it (*i.e.*, take action U upon message A and action D upon message B).[2] If the sender remains silent, the receiver has to randomize between U and D.[3] Self-interested and fully rational senders will never remain silent in the Benchmark Game. This is because any strategy in which the sender tells the truth as often as she lies leaves the receiver's prior beliefs about the underlying payoff table unchanged, which is optimal given that preferences are opposed. Since remaining silent does not change prior beliefs either but comes at a strictly positive cost, the sender will never remain silent and the optimal percentages of truth-telling and lying must both be equal to fifty percent. A self-interested and fully rational receiver will foresee this and trust the sender with probability one-half.

Remark 1: *The set of sequential equilibria of the Benchmark Game is such that the sender tells the truth and lies with probability one-half each and the receiver trusts the sender's message with probability one-half.*

To analyze the effect the cost c has on the propensity to remain silent we conducted two experimental series, one with $c = 0.1$ and one with $c = 0.5$. The results of these two treatments are compared to the ones of an extended game in which the receiver can punish the sender at a strictly positive cost once he observes the outcome of the original game.

The Punishment Game

The Punishment Game extends the Benchmark Game. Let H be the set of all histories of the Benchmark Game. After observing a particular history $h = h(a, m, \theta) \in H$, the receiver reduces the payoffs of both participants to zero with probability $r(h) \in [0, 1]$ and he accepts

[2] Here we assume implicitly that the receivers trust (distrust) can be equated with receivers believing the payoff table to be the one (opposite) of the sender's message and playing a best response to this belief. We controlled for this assumption in our earlier paper (see [17]) by asking explicitly the receiver's beliefs before taking an action.

[3] Theoretically it is possible that remaining silent conveys information with respect to the payoff table. But this is not the case because A was the underlying payoff table in 48% of all cases when the sender remained silent and with probability 0.515 the receiver chose action U afterwards.

the payoff distribution induced by the Benchmark Game with probability $1 - r(h)$. Then, both players receive their payoff. ■

It is straightforward to determine the set of sequential equilibria of the Punishment Game because a self-interested and fully rational receiver would never punish the sender.

Remark 2: *The set of sequential equilibria of the Punishment Game is such that the receiver does never punish the sender, the sender tells the truth and lies with probability one-half each, and the receiver trusts the sender's message with probability one-half.*

We conducted our experiments at Maastricht University in May and June 2006. Students from the economics and business faculty were able to register online using their matriculation number. By doing this we ensured that every student participated in only one session. In total 240 undergraduate students took part in the experiment. We organized a total of 20 sessions, for both cost factors we had five sessions on the Benchmark and the Punishment Game. Twelve subjects participated in every experimental session.

To perform the experiment we employed the computer software Z-Tree developed by Fischbacher [10]. At the beginning of a session, students met outside the laboratory. We prepared cards with the numbers from one to twelve and let each student draw one card. If more than twelve students showed up for a particular session, we offered three Euros in case somebody was willing to register for a different session. If an insufficient number of students decided to leave, we put additional empty cards into the deck and determined the participating students randomly. Left-out students received also a compensation of three Euros. We also reminded everybody that any kind of communication inside the laboratory would lead to an immediate cancellation of the session. Afterwards, students entered the laboratory and took seat in front of the computer with the number corresponding to their card. The computers were placed in such a way that subjects could not see each other and we put next to each computer an envelope containing the instructions and an official payment receipt (see the Appendix for the instructions corresponding to the Punishment Game).

Before the first round of a session, the computer randomly divided subjects into groups of six without revealing the actual matching. Thus, the students, who were all anonymous to each other, did not know who else was in their group. We informed every subject that s/he would only play against subjects belonging to the same group. So we implicitly divided

our subject pool into a total of 40 groups of six subjects, ten groups playing each of the four treatments. In each of the fifty rounds of an experimental session the computer matched the subjects belonging to the same group into three new pairs and assigned different roles (sender or receiver) within pairs. The matchings were balanced so that after fifty rounds every subject played the game exactly ten times against each of her/his five opponents. Moreover, every subject met every opponent five times in each role. The pre-determined order of the matchings was unknown to the subjects. Thus, dynamic strategies should not matter.

In every round, after pairs had been formed and roles had been assigned, the sender was informed of whether table A or B had been selected. Then, the sender chose an action from the set $M = \{A, B, N\}$ telling the receiver which table corresponds to the actual payoff table or remaining silent at cost c. Afterwards, the receiver took either action U or action D. This constituted the end of the round in the Benchmark Game. In the sessions corresponding to the Punishment Game, the receiver was further informed about the induced payoffs of her/his action. Finally, s/he had to decide between accepting these payoffs or reducing the payoff of both participants to zero.

At the end of a session, we called subjects one by one to step forward to the control desk for payment. Subjects received 6.7 cents per point in the payoff table. As a result, for $c = 0.1$ ($c = 0.5$) the average payment in the one hour session corresponding to the Benchmark and the Punishment Game was equal to 10.1 (9.9) Euros and 8.6 (8.5) Euros, respectively.

3 Results

We describe in this section the results of our experiments. Before we do so, one comment regarding our statistical analysis is worth mentioning. First, we only consider data from the last forty rounds in order to account for possible learning effects. We calculate then the percentages of the variables of interest (*i.e.*, truthful messages and trust) for all groups. In this way, we obtain ten independent observations for each of the four treatments. Then, we perform Wilcoxon signed rank and Mann-Whitney U tests on these observations.

3.1 The Benchmark Game

The first descriptive analysis regards the behavior of the senders in the Benchmark Game. We investigate the existence of social preferences by looking both at the percentage of truthful

messages and the frequency of silence. The relevant information is graphically presented in the four panels of Figure 1. The top row corresponds to the case $c = 0.1$ the bottom one to $c = 0.5$. In the panels to the right we present for all rounds the percentage of truthful messages and the percentage with which subjects remain silent. The latter percentage belongs always to the lower of the two lines in a given graph. The scatterplots on the left hand side show the strategy of every subject and include an OLS regression on the percentage with which a subject remains silent using a constant and the percentage of truthful messages as regressors.

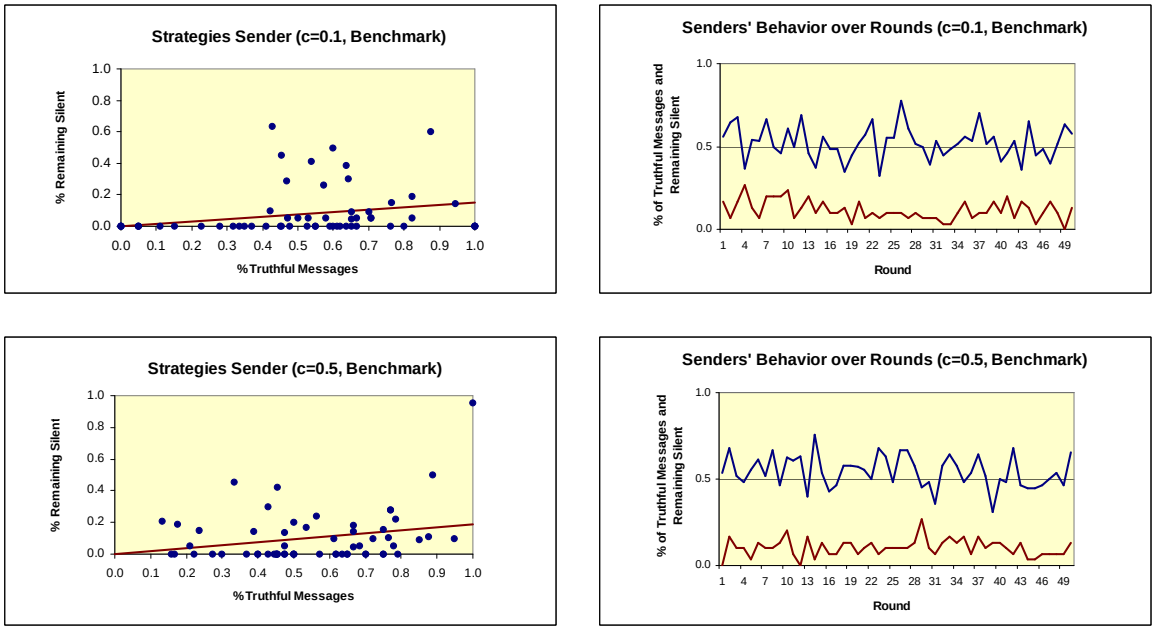


Figure 1: Senders' Behavior (Benchmark Game)

In the top right panel of Figure 1 it can be seen that the percentage of truthful messages per round oscillates around the equilibrium prediction for $c = 0.1$ with some more values above fifty percent than below. The percentage with which the senders remain silent lies around 10% in nearly all rounds. The data looks quite similar for $c = 0.5$ only there seem to be less cases when the percentage of truthful messages is below fifty percent. The overall percentage of truthful messages is 51.67% for $c = 0.1$ and 53.9% for $c = 0.5$. According to our first null hypothesis, which follows from the sequential equilibrium predictions, the percentage of truthful messages should be equal to fifty percent. The corresponding alternative hypothesis states that social preferences are prevalent and therefore the percentage of truthful messages

will be strictly greater than fifty percent. We obtain that the null hypothesis cannot be rejected for $c = 0.1$ although the median observation is 52.85% (the one-sided p-value of the Wilcoxon signed rank test is 0.2377) but it can be rejected for $c = 0.5$ (the one-sided p-value of the Wilcoxon signed rank test is 0.0095). The median observation is 53.15% in the latter case. Moreover, the medians are not significantly different from each other (the two-sided p-value of the Mann-Whitney U test is 0.8798).[4]

Result 1: *In the Benchmark Game, the percentage of truthful messages is greater than fifty percent for both c but the difference is only significant for* $c = 0.5$*. Moreover, the percentage of truthful messages for* $c = 0.5$ *is not significantly different from the one for* $c = 0.1$*.*

The strategies of every subject can be found in the two left panels of Figure 1. On average a subject remains silent with probability 0.1017 when $c = 0.1$ and with probability 0.1025 when $c = 0.5$, but we also see also that a lot of subjects never followed that strategy. We obtain from a simple OLS regression that subjects who tell the truth more often are also more likely to remain silent. The estimated slopes for $c = 0.1$ and $c = 0.5$ are equal to 0.1532 and 0.1913, respectively, with the intercept not being significant at the five percent significance level in either regression.[5] Since the estimated equations are very similar to each other and the overall percentages are also very close, we deduce that subjects remain silent equally often across the two specifications. This intuition is confirmed by the statistical analysis, because the two-sided p-value of the corresponding Mann-Whitney U test is 0.7613. Note that the median observation is 5% for $c = 0.1$ and 9.2% for $c = 0.5$.

Result 2: *In the Benchmark Game, the probability of remaining silent is significantly greater than zero for both c but not significantly different from each other. Moreover, those subjects who send more truthful messages are more likely to remain silent.*

Now we analyze the receivers behavior. In the four panels of Figure 2 we present a

[4]In our earlier paper (see [17]) where the sender could only choose among the strategies *truth* and *lie*, we obtained that the percentage of truthful messages was significantly greater than fifty percent if the implemented payoff distribution was either (1,2) or (2,1) but not if it was either (9,1) or (1,9). Our findings here seem to be consistent with our former results since the present specification can be thought off as an intermediate case.

[5]One word of caution with respect to the OLS regressions. The more often an individual remains silent the fewer observations we can have with respect to her/his honesty. This can become a major problem if a lot of subjects decide to remain silent very often. But this is not the case as in total only three subjects remain silent in more than sixty percent. Moreover, we can obtain a similar result by dividing the subject pool into two groups: In one group we put all subjects that remain silent at least once and in the other group the rest of the subjects. The probability that a given message is true for the former group is 0.6364 for $c = 0.1$ and 0.5903 for $c = 0.5$. The corresponding probabilities for the rest of the subjects are only 0.4536 and 0.4975.

histogram of the percentage of messages trusted (the two panels on the left hand side) and show how this percentage evolves over rounds (the two panels on the right hand side). We see that the level of trust lies above fifty percent in nearly all rounds, independently of the value of c, reaching sometimes even considerably high values. The empirical distribution is rather uniform for $c = 0.1$ and seems to be closer to the normal for $c = 0.5$. Moreover, there are considerably more subjects who trust with a probability greater than 0.5 in both distributions. The overall percentage of trust is 55.25% for $c = 0.1$ and 56.60% for $c = 0.5$.

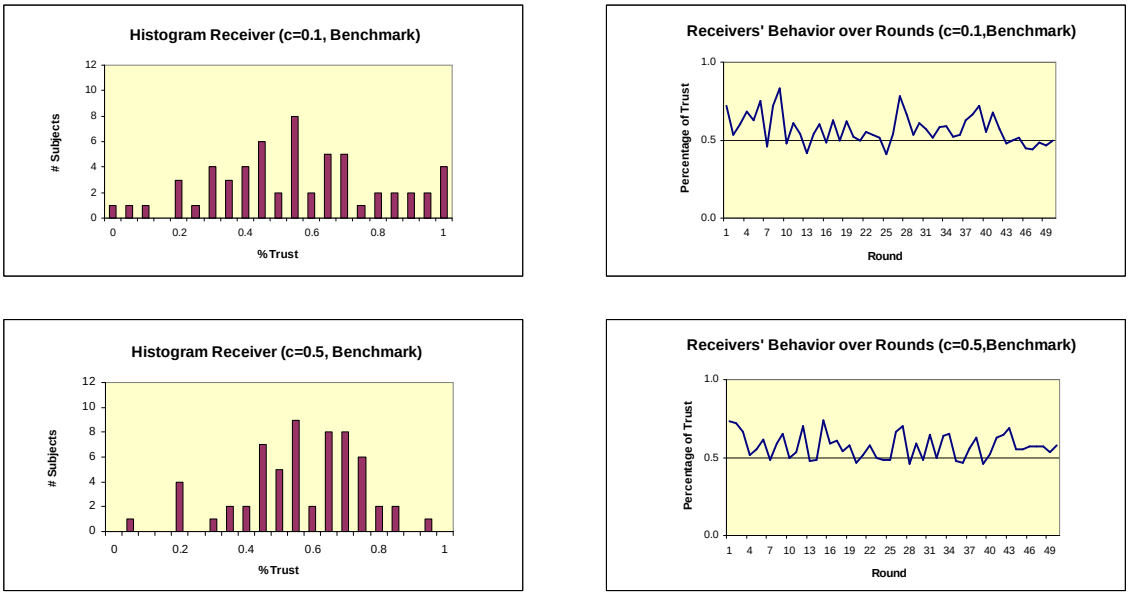


Figure 2: Receivers' Behavior (Benchmark Game)

In our statistical analysis we check whether the receivers adjust their behavior in response to the senders strategies in the correct direction by testing the null hypothesis that the receivers trust with probability 0.5 against the alternative hypothesis that this probability is greater than 0.5. The median observation is 52.7% for $c = 0.1$ and 57.8% for $c = 0.5$. We obtain by means of a one-sided Wilcoxon signed rank test that the difference is in both cases significant, the corresponding one-sided p-values are 0.0050 for $c = 0.1$ and 0.0072 for $c = 0.5$. Additionally, we cannot reject the hypothesis that the percentage of trust for $c = 0.1$ is the same as the one for $c = 0.5$ (the two-sided p-value of the Mann-Whitney U test is 0.7913).

Result 3: *In the Benchmark Game, the percentage of messages trusted is significantly greater than fifty percent for both c and the two values are not significantly different from each other.*

We sum up our findings for the Benchmark Game. Since the percentage of truthful

messages is slightly above fifty percent in the Benchmark Game, we have found evidence in favor of weak preferences for truth-telling. Moreover, the receivers seem to predict the senders behavior correctly because they trust more often than they distrust. Finally, the fact that the percentage with which subjects remain silent is considerably bounded away from zero but smaller than one hints in the direction of a weak form of lie-aversion. The OLS regression revealed additionally that the two types of social preferences seem to be positively correlated. None of these results depends on the value of c.

3.2 The Punishment Game

From the previous analysis, it is evident that the observed excessive truth-telling and the considerable frequency of silence are likely to be driven by social preferences. But of what type? In this section, we investigate this point further with the help of the Punishment Game. In particular, we are going to identify a fraction of subjects who behave *morally* in the role of the receiver and show that they behave in a *consistent* way; that is, as senders they tell the truth very often and remain silent frequently.

Models of inequity aversion such as the ones of Fehr and Schmidt [9] and Bolton and Ockenfels [1] incorporate experimental evidence showing that individuals take into account in the evaluation of an outcome not only their own payoff but also the payoff of others. But preferences may even be more complex because, as it has been suggested by Sen [18], individuals may also care about how a particular payoff distribution has been reached; that is, they may have concerns for *procedural justice*. The experimental findings of Brandts and Charness [2] and Sánchez-Pagés and Vorsatz [17] support this conjecture.

c=0.1	Trust	Distrust
Truth	0	0.1795
Lie	0.4658	0

c=0.5	Trust	Distrust
Truth	0.0025	0.0398
Lie	0.4239	0

Figure 3: Punishment Behavior

Figure 3 provides further evidence in favor of the notion of procedural justice. First we see that the punishment rates after the histories (*truth,trust*) and (*lie,distrust*) are virtually zero. This is explained by the fact that these histories are the most favorable ones to the receiver. Since the potential payoff distribution after the histories (*lie,trust*) and (*truth,distrust*) is

(5,1), a sufficient inequity averse receiver may prefer the payoff distribution (0,0) instead, and consequently punish the sender. However, according to the models of inequity aversion the punishment rate should be the same after both histories. This is clearly not the case and the extra punishments after the history (*lie,trust*), 28.63% for $c = 0.1$ and 38.41% for $c = 0.5$, should be attributed to the deceit the sender commits. On the contrary, in the history (*truth,distrust*) the sender conforms to social norms and ultimately it is the receiver's fault why he gets the lower payoff. In case the sender remains silent the punishment rates are 10.49% (6.19%) for $c = 0.1$ ($c = 0.5$) if the receiver gets the lower payoff and zero for both c if the receiver gets the higher payoff. Since these values compare to the punishment rates after history (*truth,distrust*) in Figure 3, we refer from now on to the punishment rate after this history as the baseline utility loss induced by inequity aversion.

Result 4: *Procedural justice matters. The punishment rate after the history (lie,trust) is significantly higher than the one after history (truth,distrust).*

Result 4 will be important in the next section where we quantify social preferences for the Benchmark Game but it should also have crucial consequences in the Punishment Game itself. This is because the senders' incentive structure changes to some extent. A subject who aims uniquely at maximizing her/his payoff should always lie and always trust in the Benchmark Game but a purely self-interested sender should now avoid to lie because the risk of being punished and ending up with nothing is too high. This is similar to what happens in ultimatum games, where substantial offers can be explained as responses to the threat of rejection. Here, the alternatives are either to tell the truth or to remain silent. The former strategy bears a big risk of being easily outguessed by the receiver which results in a low payoff. By remaining silent the sender incurs into a cost but since the associated punishment rates are rather small, this strategy looks far more promising. On the basis of our data it yields in fact the highest expected payoff, i.e. 2.64 (2.36) for $c = 0.1$ ($c = 0.5$) versus 1.73 (2.07) from telling the truth and 2.28 (2.35) from lying.

Now, we turn our attention to the behavior of the sender. It can be seen in the top right panel of Figure 4 that the percentage of truthful messages for $c = 0.1$ is in almost all rounds greater than fifty percent, leading to an overall percentage of 64.74%. We obtain by means of a one-sided Wilcoxon signed rank test that the median observation of 56.9% is significantly greater than the equilibrium prediction of fifty percent (the corresponding one-sided p-value

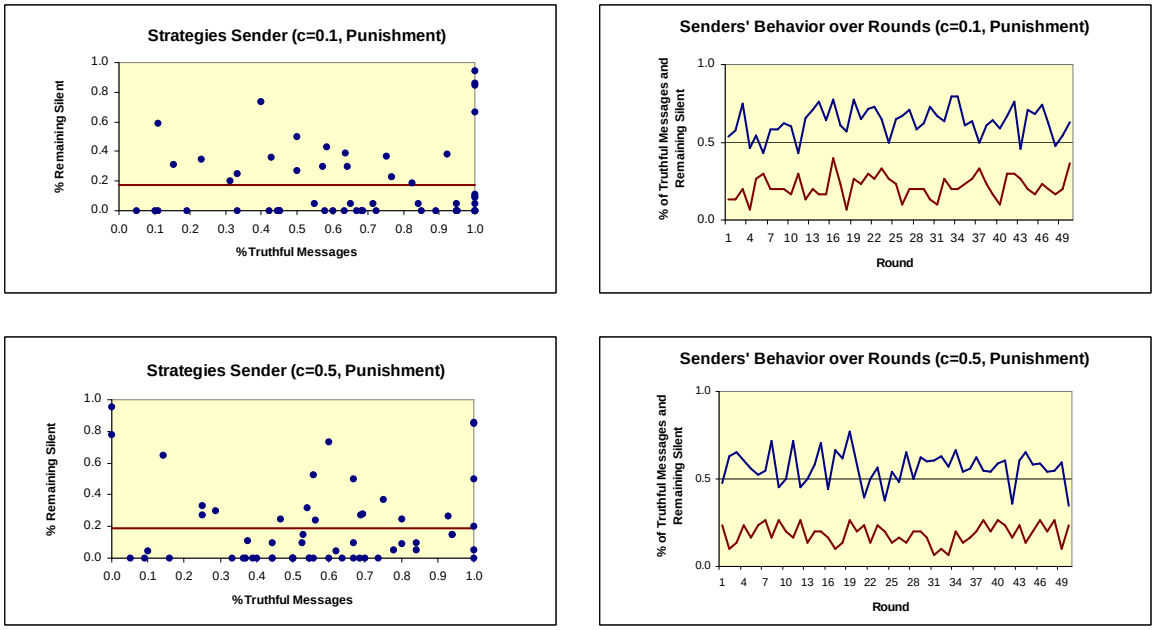


Figure 4: Senders' Behavior (Punishment Game)

is 0.0499). In the lower right panel, where we plot the senders' behavior over rounds for $c = 0.5$, it can be observed that the percentage of truthful messages is very high for a lot of rounds, specially for the first twenty. But the overall percentage of truthful messages is "only" 56.53% because there are also quite a few rounds where the percentage is low. The one-sided Wilcoxon signed rank test reveals nevertheless that the median observation of 54.5% is significantly greater than fifty percent (the corresponding one-sided p-value is $= 0.0329$). Moreover, there is no difference in the percentage of truthful messages across the two costs (the two-sided p-value of the Mann-Whitney U test is 0.6495).

Result 5: *In the Punishment Game, the percentage of truthful messages is significantly greater than fifty percent for both c and the percentage of truthful messages for $c = 0.1$ is not significantly different from the one for $c = 0.5$.*

The two right panels of Figure 4 reveal also that the percentage with which subjects remain silent increases slightly during the course of the sessions for $c = 0.1$ whereas it stays rather constant, at a high level though, for $c = 0.5$. This is consistent with our earlier finding that this action yields the highest payoff. The overall percentage of silence in the last forty rounds of the experiment is equal to 22.33% for $c = 0.1$ and 18.33% for $c = 0.5$ whereas the median observations, 12.50% for $c = 0.1$ and 16.65% for $c = 0.5$, are somewhat lower. The

difference between these two values is not significant at any conventional confidence level (the two-sided p-value of the Mann-Whitney U test is 0.8797). In the two left panels we see that individual behavior is very volatile, the only pattern that can perhaps be identified is that a lot of subjects do always transmit a message. This idea is supported by the OLS regression because the percentage of truthful messages is no longer significant as an explanatory variable in either case. The estimated constant is 0.1767 for $c = 0.1$ and 0.1853 for $c = 0.5$.[6]

Result 6: *In the Punishment Game, the probability of remaining silent for* $c = 0.1$ *is not significantly different from the one for* $c = 0.5$ *but higher than in the Benchmark Game. The probability of remaining silent is also independent of the percentage of truthful messages.*

Our final descriptive analysis in this section regards the receivers' behavior in the Punishment Game. The power the receivers have due to their punishment opportunity and the importance of procedural justice according to Result 4 made the senders tell the truth often. The best response for the receivers would thus be to always trust the senders' messages. We see in the two left panels of Figure 5 that a lot of subjects trust almost always and that only a few trust less than the equilibrium prediction of 0.5. This leads to a very skewed distribution. In the two right panels of the same figure we observe additionally that the percentage of trust is in all rounds far higher than the equilibrium prediction.

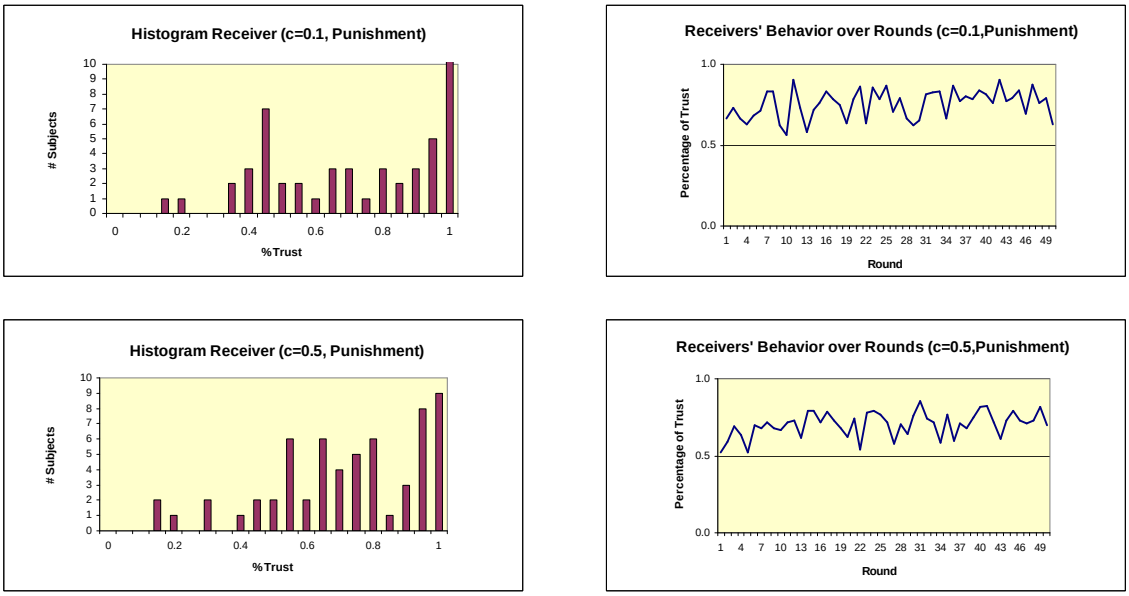


Figure 5: Receivers' Behavior (Punishment Game)

[6]Seven subjects remained always silent in the Punishment Game. Since their behavior cannot be represented in the left panels of Figure 4, we excluded them also from the statistical analysis.

The results of the statistical analysis are therefore straightforward. The overall percentage of trust in the Punishment Game is 76.84% for $c = 0.1$ and 71.98% for $c = 0.5$. The median observations are 76.5% for $c = 0.1$ and 70.35% for $c = 0.5$. The latter two values are obviously significantly greater than the equilibrium prediction (the one-sided p-value of the Wilcoxon signed rank test is in both case 0.0030) but not significantly different from each other (the two-sided p-value of the Mann-Whitney U test is 0.7337).

Result 7: *In the Punishment Game, the percentage of messages trusted is significantly greater than fifty percent for both c and the two values are not significantly different from each other.*

To sum up our findings for the Punishment Game so far. Most importantly, we have established the central role of procedural justice by showing that the punishment rates after the histories (*truth,distrust*) and (*lie,trust*) are very distinct from each other. This result does not only contradict the standard sequential equilibrium hypothesis but also more general models of inequity aversion. Moreover, we have found that the percentage of silence is independent of the percentage of truthful messages at the individual level. This result comes up in sharp contrast to the one obtained in the Benchmark Game. The reason for this is that since remaining silent avoids the potential receivers' outrage when they are deceived, some senders who would have lied in the Benchmark Game may now prefer to abstain from sending any message. As we will see next, these subjects do not display substantial moral concerns and this makes the correlation vanish. Finally, note that the results are again very robust with respect to the cost factor c.

3.3 The Separability Hypothesis

So far we have evidenced that moral considerations enter into the subjects decisions when they have to decide whether or not to punish the sender. We perform now a final consistency test to show that social preferences are also important when subjects act in the role of the sender. Thanks to the role rotation mechanism all subjects play the Punishment Game equally often in both roles. If we now divide the whole subject pool into two groups, one group containing those subjects who punish the senders frequently after history (*lie,trust*) and one group consisting of the rest of the subjects, is it true that those subjects who care about procedural justice as a receiver send less frequent but more truthful messages? In other words, is this moral behavior consistent across roles?

Since the punishment rate after history (*lie,trust*) lies between forty and fifty percent for both c, we classify all subjects that punish the sender after this history in more than eighty percent of all occasions as the ones with real concerns for procedural justice.[7] This criterium is met by 23 out of 60 subjects for $c = 0.1$ and by 17 subjects for $c = 0.5$. In the former (latter) case, these subjects send a truthful message with probability 0.8765 (0.7659) and remain silent with probability 0.3083 (0.3355). The corresponding probabilities for the remaining subjects are 0.5208 (0.5141) and 0.1653 (0.1324) for $c = 0.1$ ($c = 0.5$), respectively. Moreover, the morally consistent subjects also trust around 25% more than the others and receive a roughly 10% lower material payoff.[8]

We can conclude from this result that moral concerns are also crucial in understanding the senders' behavior in the Punishment Game because if all individuals that punish the sender a lot after history (*lie,trust*) are taken away from the subject pool, the observed excessive truth-telling nearly vanishes and the percentage of remaining silent declines considerably. Moreover, we have also re-established for the morally consistent subjects the positive correlation between the probability of sending a truthful message and the probability of remaining silent we had already found in the Benchmark Game.

4 Modelling and Estimating Social Preferences

In the former sections we have established two main results: First that subjects' behavior in the Benchmark Game is on the aggregate consistent with weak forms of both lie-aversion and preferences for truth-telling and second, that a fraction of subjects, identified by means of the Punishment Game, indeed display them. In the present section we will develop two formal models that can account for these observations.

Our desiderata for these models is that they must generate a percentage of truthful and trusted message above 50% and, at the same time, they must display a positive correlation between truth-telling and the probability of remaining silent across subjects. In the first

[7] In expectation every subject plays the Punishment Game in the last forty rounds of the experiment 20 times in the role of the receiver. In this 20 repetitions the subject will play the history (*lie,trust*) about 6 times. Hence, we selected in expectation those subjects who punished the sender in at least five out of six occasions.

[8] We come to a similar conclusion if we put into one group all subjects that never punish the sender after history (*lie,trust*) and into the other group all remaining subjects. For $c = 0.1$ ($c = 0.5$) 13 (28) individuals satisfy this criterium. The non-punishers send a truthful message with probability 0.4689 (0.4990) and remain silent with probability 0.0731 (0.1375). The corresponding probabilities for the remaining individuals are 0.7095 (0.6285) and 0.2638 (0.2219), respectively.

model we modify slightly the utility function for senders and receivers and then apply the logit agent quantal response equilibrium (logit-AQRE) concept proposed by McKelvey and Palfrey [15]. In the second model, we build on the idea of a fixed cost of lying proposed by Gneezy [11] and formalized by Kartik [13], and solve a simplified version of the Benchmark Game with two types of senders. The results obtained in both cases are consistent with our desiderata and can explain the observed data as reasonably as formal models can.

4.1 Agent Quantal Response Equilibrium

Since it is impossible to extract social preferences directly from the Benchmark Game, we apply now an indirect equilibrium approach. To do so, we suppose that the utility functions u_s of the sender and u_r of the receiver take the following form:

$$u_s(h) = \begin{cases} \alpha_T \cdot x_s & \text{if } m = \{\textit{truth}\} \\ \alpha_L \cdot x_s & \text{if } m = \{\textit{lie}\} \\ 3 - c & \text{if } m = \{\textit{remain silent}\} \end{cases} \qquad u_r(h) = \begin{cases} \beta_T & \text{if } h{=}(\textit{truth,distrust}) \\ \beta_L & \text{if } h{=}(\textit{lie,trust}) \\ x_r & \text{otherwise} \end{cases} .$$

According to this specification the utility of the sender does not only depend on her payoff x_s but also on the type of message she sends. In particular, if she tells the truth her payoff is enhanced by factor $\alpha_T \geq 1$ and if she lies a loss of factor $1 - \alpha_L \geq 0$ is suffered. Moreover, if the sender remains silent, the almost fair payoff distribution $(3 - c, 3)$ is implemented in expectation and we assume for simplicity that the sender's utility is in this case equal to her payoff. On the other hand, the receiver takes into account the notion of procedural justice which can be seen by the fact that his utility does depend on the whole history of the Benchmark Game. Following our findings in the Punishment Game the receiver feels deceived after history (*lie,trust*) and his payoff is downgraded to $\beta_L \leq 1$. The corresponding utility associated to (*truth,distrust*) is β_T, where $1 \geq \beta_T \geq \beta_L$. This specification incorporates both the receiver's inequity aversion and concerns for procedural justice. Finally, we assume again for simplicity that the receiver's utility is equal to her payoff in all other cases.

In the next step, we define the logit agent quantal response equilibrium (logit-AQRE) proposed McKelvey and Palfrey [15] for the given utility functions. The central idea of this model of bounded rationality, which encompasses the sequential equilibrium concept as a special case, is that individuals make mistakes when they try to maximize their utility but have correct beliefs about their opponents actions. According to the logit-AQRE, which is

parameterized by $\lambda \in [0, \infty)$, the sender transmits message m in state θ with probability $p(m|\theta) = \frac{e^{\lambda u_s(m|\theta)}}{\sum_{i \in M} e^{\lambda u_s(i|\theta)}}$, where $u_s(i|\theta) = \sum_{a \in \mathcal{A}} q(a|i) \cdot u_s(a, m, \theta)$ denotes the expected utility from sending message i in state θ. Similarly, the receiver chooses action a after observing message m with probability $q(a|m) = \frac{e^{\lambda u_r(a|m)}}{\sum_{j \in A} e^{\lambda u_r(j|m)}}$. Here, $u_r(j|m)$ corresponds to the expected payoff of taking action j upon message m; that is, $u_r(j|m) = \sum_{\theta \in \Theta} \mu(\theta|m) \cdot u_r(j, m, \theta)$, where $\mu(\theta|m) = \frac{p(m|\theta)}{\sum_{i \in \Theta} p(i|\theta)}$ is the receiver's posterior belief about the state θ when he observes message m. Observe that the logit-AQRE selects one sequential equilibrium for $\lambda \to \infty$ and $\alpha_T = \alpha_L = \beta_T = \beta_L = 1$.

Due to the symmetry of the Benchmark Game it will be the case that $p(A|A) = p(B|B)$, $p(B|A) = p(A|B)$, and $q(D|A) = q(U|B)$; that is, the probabilities of telling the truth, lying, and trust are independent of the selected payoff table. Hence, from now on, let p_T and p_L be the probabilities that the sender tells the truth and lies, respectively. Consequently, the sender remains silent with probability $1 - p_T - p_L$. The probability that a given message is truthful is denoted by $\bar{p}_T$. Now, if q_T is the probability that the receiver trusts the sender's message, the logit-AQRE for the Benchmark Game with cost c is fully specified by the following equations:

$$\begin{aligned} q_T &= \frac{e^{\lambda(\bar{p}_T(5-\beta_L)+\beta_L)}}{e^{\lambda(\bar{p}_T(5-\beta_L)+\beta_L)} + e^{\lambda(5-\bar{p}_T(5-\beta_T))}} \\ p_T &= \frac{e^{\lambda \cdot \alpha_T \cdot (5-4 \cdot q_T)}}{e^{\lambda \cdot \alpha_T \cdot (5-4 \cdot q_T)} + e^{\lambda \cdot \alpha_L \cdot (1+4 \cdot q_T)} + e^{\lambda(3-c)}} \\ p_L &= \frac{e^{\lambda \cdot \alpha_L \cdot (1+4 \cdot q_T)}}{e^{\lambda \cdot \alpha_T \cdot (5-4 \cdot q_T)} + e^{\lambda \cdot \alpha_L \cdot (1+4 \cdot q_T)} + e^{\lambda(3-c)}} \end{aligned} \quad . \tag{1}$$

Observe that given q_T, $\alpha_T(5-4q_T)$ is the expected utility of telling the truth while $\alpha_L(1+4q_T)$ is the expected utility from lying. Similarly, given $\bar{p}_T$, the receiver gets an expected utility of $\bar{p}_T(5 - \beta_L) + \beta_L$ from trusting the sender's message and of $5 - \bar{p}_T(5 - \beta_T)$ from distrusting it. So far, it is not possible to determine the unobservable parameters α_T and α_L from the set of equations (1) because even if we use aggregate data for q_T, p_T, p_L, and $\bar{p}_T$, we are still left with five variables ($\alpha_T, \alpha_L, \beta_T, \beta_L$, and λ) and only three equations. One way solve this problem is to set the punishment rates after the histories (*lie,trust*) and (*truth,distrust*) equal to $1 - \beta_L$ and $1 - \beta_T$. This is appropriate because β_L and β_T are then simply the empirical payoffs of the receiver after these histories in the Punishment Game and, most importantly, these estimates are derived from the decisions of the receivers themselves. If we rewrite the

set of equations (1) in terms of α_T, α_L, and λ, we obtain that

$$\lambda = \frac{\ln\left(\frac{q_T}{1-q_T}\right)}{\bar{p}_T(10-\beta_T-\beta_L)-5+\beta_L}, \quad \alpha_T = \frac{\ln\left(\frac{p_T}{1-p_T-p_L}\right)+\lambda(3-c)}{\lambda(5-4q_T)}, \text{ and } \quad \alpha_L = \frac{\ln\left(\frac{p_L}{1-p_T-p_L}\right)+\lambda(3-c)}{\lambda(1+4q_T)}. \tag{2}$$

We are now able to determine the parameters α_T and α_L for our data. Additionally, we also calculate the parameter λ which must take a non-negative value by definition of the logit-AQRE. Typically λ is taken to be a measure of rationality, because individuals play completely random if $\lambda = 0$ and are fully rational if $\lambda \to \infty$. So, higher values indicate a more rational behavior.

Parameters	$c = 0.1$	$c = 0.5$	**Estimates**	$c = 0.1$	$c = 0.5$
q_T	0.5525	0.5660	λ	176.50	2.6377
p_T	0.4633	0.4833	α_L	0.9060	0.9281
p_L	0.4350	0.4142	α_T	1.0424	1.1286
$\bar{p}_T$	0.5167	0.5385			
β_L	0.5342	0.5761			
β_T	0.8205	0.9602			

Table 2: Estimates of Social Preferences

Our first observation is that all estimates take values in the desired range. The rationality parameter λ is positive and α_L (α_T) is smaller (bigger) than one. The value of α_T suggests that telling the truth gains the sender an extra of four percent of her payoff when $c = 0.1$ and twelve percent when $c = 0.5$. This estimation is obviously driven by the fact that the sender tells the truth more often in the latter case. The value of α_L is nearly the same in both cases. This makes also sense given that the percentage with which the senders remain silent was rather independent of c. Surprising is only the big difference in the values of λ, which is easily explained. If we have a look at the first equation in the set of equations (2), we see that λ is an increasing function in β_T. But we observe in Table 2 that β_T changes considerably across cost treatments, implying that the expected payoff of distrust is higher for $c = 0.1$. Since there is nearly no change in the probability of trust across cost treatments, this is a sign of a more rational behavior when $c = 0.1$, so λ must increase considerably.

4.2 A Model of Lie-Aversion

Our experimental results are consistent with the idea that individual behavior is diverse, which is for example clearly documented by our finding that only a subset of individuals remains

silent in the Benchmark Game. We will now take the variety of individual preferences into account and introduce different types of players into a model inspired by the one of Kartik [13]. This approach puts the burden of the proof on the *consequences* of the act of lying and is thus different from the one proposed by Charness and Dufwenberg [4] who claim that the potential guilt generated by deception is relative to the *expectations* held by the party who is lied to. In particular, we transform the Benchmark Game into a simple simultaneous move game with an outside option for the sender; that is, we assume implicitly that a message is decomposed into two parts: First, whether or not the sender remains silent and second, whether or not a sent message is truthful. By doing so we can abstract from the beliefs held by the receivers about the truthfulness of a message thereby concentrating on the behavior of the sender and, in particular, on her propensity to remain silent.

Simplified Benchmark Game

Let $P = \{\text{sender,receiver}\}$ be the set of players. In the beginning of the game, the type of the sender is randomly determined. We assume that the sender can be either of type h or of type l. The probability that the sender is of type h is ε, where $0 < \varepsilon < 0.5$. Only the sender is informed about her type, indexed by i. The first decision is made by the sender; she must decide whether or not to remain silent (opt out of the game). For every type $i = l, h$, we write $\sigma_1^i = 0$ if she remains silent and $\sigma_1^i = 1$ otherwise. In case the sender remains silent, the utility distribution $(3 - c, 3)$ is implemented. If the sender does not remain silent, players enter into a second stage in which the sender of type i chooses a probability distribution over the set of actions $\{truth, lie\}$. We denote the probability that the sender tells the truth by σ_2^i. Similarly, the receiver selects a probability distribution over the set of actions $\{trust, distrust\}$. The probability that the receiver takes the action *trust* is denoted by ϕ. The actions in the second stage are taken simultaneously and independently. Moreover, let $\mu^i = 1 - \mu^j$ be the posterior belief the receiver holds about the sender's type being i after observing the opting out decision of the sender.

Under the fixed (or belief-independent) cost of lying, subjects get a disutility from the act of sending a false message. In particular, we assume that a sender of type l faces a cost from lying of $k^l > 0$ whereas a sender of type h suffers an utility loss of $k^h > k^l$ if she is not honest. Finally, we maintain our assumption of the previous model that the utility of the receiver decreases whenever he gets the low payoff, either because of inequity aversion or because of

deception. In particular, we suppose that $\beta_L < \beta_T$. ■

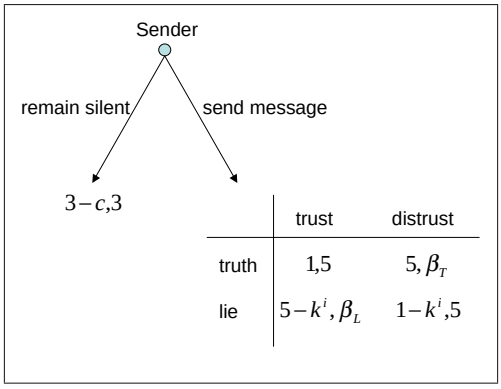


Figure 6: Simplified Benchmark Game for Type i

Consider the Simplified Benchmark Game with two types of senders. This is a sequential move game in which the decision of whether or not to remain silent may convey information regarding the sender's type. Hence, both pooling and separating Perfect Bayesian Nash equilibria can emerge. Formally, a *Perfect Bayesian Equilibrium (PBE)* of the Simplified Benchmark Game is a strategy profile $(\sigma^l, \sigma^h, \phi)$ together with the posterior belief $\mu^i = 1 - \mu^j$ such that σ^i maximizes the sender's expected utility given ϕ, and ϕ maximizes the receiver's expected utility given σ^i and μ^i, where the beliefs are consistent with σ^i via Bayes' rule. The PBE of the Simplified Benchmark Game are summarized in the following proposition.

Proposition 1 *In the Simplified Benchmark Game,*

*(i) for $k^l \leq 2c$ there exists a **Pooling** PBE in which no type remains silent ($\sigma_1^l = \sigma_1^h = 1$), type l tells the truth with with probability $\sigma_2^l = \frac{5-\beta_L}{10-\beta_T-\beta_L} - \frac{\varepsilon}{1-\varepsilon}\frac{5-\beta_T}{10-\beta_T-\beta_L}$, type h tells always the truth, and the receiver trusts with probability $\phi = \frac{1}{2} + \frac{k^l}{8}$.*

*(ii) for $k^l \leq 2c \leq k^h$ there exists a **Separating** PBE in which only type l sends messages ($\sigma_1^l = 1$ and $\sigma_1^h = 0$); she tells the truth with probability $\sigma_2^l = \frac{5-\beta_L}{10-\beta_T-\beta_L}$ and the receiver trusts with probability $\phi = \frac{1}{2} + \frac{k^l}{8}$.*

*(iii) for $k^l \geq 2c$ there exists a **Pooling** PBE in which both types of senders remain silent.*

Proof: See the Appendix. □

The percentage of truthful messages in the equilibria described in (i) and (ii) is equal to $\frac{5-\beta_L}{10-\beta_T-\beta_L} > 0.5$. In addition, the percentage of trusted message is as well above 50%. These

results are consistent with our desiderata. Moreover, note that if $k^l \leq 2c \leq k^h$, we have two equilibria: In one h remains silent, in the other she tells the truth for sure. Hence, this case may account also for the observed positive correlation between the percentage of truthful messages and the probability of remaining silent, simply by considering that sometimes subjects coordinated on the pooling and sometimes on the separating PBE. The pooling PBE in (i) is also consistent with the alternative interpretation of Gneezy's [11] results by Hurkens and Kartik [12] who postulate the existence of two behavioral types, one who always tells the truth and the other who tells the truth only if payoff is maximized by doing so. Proposition 1 also points in the direction of lie-aversion as the driving force behind our experimental observations because in an alternative version in which agents get an extra utility from being honest it would be the subjects with less intense preferences for truth-telling who either tell the truth less frequently (in a pooling PBE) or do remain silent (in a separating PBE). Hence, there would be a negative correlation between remaining silent and the percentage of truthful messages.[9]

5 Conclusion

A fast-growing body of the economic literature has established that people have "social preferences" and hold moral standards. In the present paper, we have studied in particular the existence of procedural concerns in the context of a strategic information transmission game with conflictive preferences; the better the outcome for the sender is, the worse for the receiver and vice versa. In a sense, this is the most hostile scenario for the prevalence of moral concerns since lies pay off from a purely materialistic point of view and this could lead to them being expected and therefore "forgiven". We incorporated the possibility of remaining silent with the aim of understanding the actual type(s) of moral behavior that underlie the excessive truth-telling observed in previous experiments of this kind. In a nutshell, the results show that people definitely care about lies and deception: They tend to be more honest and less opportunistic than what the standard theory would predict and dislike deception enough to forego material benefits in order to punish those who deceived them.

In the baseline version of our game, subjects tend to send more truthful messages than

[9] A model in which all moves are taken simultaneously cannot account for the positive correlation either. In that case both types remain silent or both send a message, but these equilibria do not coexist for the same range of parameters.

predicted by the sequential equilibrium concept, and consequently trust relatively often when they assume the role of receivers. They also remain silent rather frequently. Moreover, there is a positive correlation between silence and truth-telling. We then use punishments to substantiate the existence of procedural concerns in the population. Punishment rates were substantial when the receiver obtained the lower payoff, but were even higher when that distribution was achieved by means of deception. Both standard theory and inequity aversion models are unable to explain this result.

We also used this treatment as a way to identify a fraction of subjects who display a moral consistent behavior: Subjects who punish those who lie to them tell the truth more often, remain silent more frequently, and trust more than the rest of the population, who on average behaves according to the standard equilibrium prediction. Whereas the strategic value of truth-telling in the presence of punishments and procedural concerns made the positive correlation between silence and truth vanish in the overall, it remained strong within this group.

Finally we solve two formal models that can account for the observed data. First, we modified individual utilities to incorporate both preferences for truth-telling and lie-aversion, together with the dislike for deception, and then fit the data with the agent quantal response equilibrium of the baseline game. The estimates obtained were consistent with our intuitions. Secondly, we solved a simplified version of the Benchmark Game in which silence becomes an outside option and senders differ with respect to their cost of telling a lie. The Perfect Bayesian Nash equilibria of that game can replicate the aggregate observed data and can explain the positive correlation between truth-telling and silence.

We do not want to dismiss here the idea that the rate of lying is sensitive to the potential gains to the party who lies and to the potential loss to the party who is lied to. In a previous paper, we ourselves found evidence of this phenomenon by changing the distribution of payoffs in two different treatments similar to the one presented here. But it seems also unreasonable to dismiss the existence of two types of subjects who, let us emphasize this, do not differ on their thrust for utility maximization, but rather on the intensity of their moral concerns. We do not want to dismiss either other operationalizations of the concept of lie-aversion; namely Charness and Dufwenberg's [4] based on expectations. Indeed, it is possible to modify the game in section 4.2 to incorporate that notion. However, the results obtained in that case

cannot account for the observed data since in our experimental design the receiver may also suffer a let down when she is told the truth but she distrusts. Certainly, consequences and expectations do matter, but actions also carry moral "labels" that cannot be ignored. A lie is always a lie. Still, our experiment is not directly designed to test these distinct theories so we choose not to elaborate on this point any further.

The fact that at least a part of the population holds moral concerns in terms of preferences for truth-telling and, more importantly, in terms of lie-aversion, has clear implications for contract theory and mechanism design. The latter for instance assumes that people will lie as long us material incentives to do so exist. Incentive-compatibility must be fulfilled. This self-imposed assumptions on individual behavior render difficult and sometimes impossible to find mechanisms that succeed in the task of eliciting truthfully agents' private information. Without them, the set of mechanism capable of this would certainly enlarge.

References

[1] G. Bolton and A. Ockenfels, *ERC: A Theory of Equity, Reciprocity, and Competition*, American Economic Review **90** (2000), 166–193.

[2] J. Brandts and G. Charness, *Truth or Consequence: An Experiment*, Management Science **49** (2003), 116–130.

[3] H. Cai and J. Wang, *Overcommunication in Strategic Information Transmission*, Games and Economic Behavior **56** (2006), 7–36.

[4] G. Charness and M. Dufwenberg, *Promises and Partnership*, Econometrica **74** (2006), 1579–1601.

[5] M. Costa-Gomes, V. Crawford, and B. Broseta, *Cognition and Behavior in Normal Form Games: An Experimental Study*, Econometrica **69** (2001), 1193–1235.

[6] V. Crawford, *Lying for Strategic Advantages: Rational and Boundedly Rational Misrepresentations of Intentions*, American Economic Review **93** (2003), 133–149.

[7] V. Crawford and J. Sobel, *Strategic Information Transmission*, Econometrica **50** (1982), 1431–1451.

[8] J. Dickhaut, K. McCabe, and A. Mukherji, *An Experimental Study of Strategic Information Transmission*, Economic Theory **6** (1995), 389–403.

[9] E. Fehr and K. Schmidt, *A Theory of Fairness, Competition, and Cooperation*, Quaterly Journal of Economics **114** (1999), 817–864.

[10] U. Fischbacher, *Z-Tree - Zurich Toolbox for Readymade Economic Experiments - Experimenter's Manual*, Working Paper Nr. 21, Institute of Empirical Research in Economics, Zurich University (1999).

[11] U. Gneezy, *Deception: The Role of Consequences*, American Economic Review **95** (2005), 384–394.

[12] S. Hurkens and N. Kartik, *(When) Would I Lie to You? Comment on "Deception: The Role of Consequences"*, Mimeo (2006).

[13] N. Kartik, *Information Transmission with Almost Cheap Talk*, Mimeo (2005).

[14] R. McKelvey and T. Palfrey, *Quantal Response Equilibria in Normal Form Games*, Games and Economic Behavior **10** (1995), 6–38.

[15] ———, *Quantal Response Equilibria in Extensive Form Games*, Experimental Economics **1** (1998), 9–41.

[16] R. Nagel, *Unraveling in Guessing Games: An Experimental Study*, American Economic Review **85** (1995), 1313–1326.

[17] S. Sánchez-Pagés and M. Vorsatz, *An Experimental Study of Truth-Telling in a Sender-Receiver Game*, forthcoming Games and Economic Behavior.

[18] A. Sen, *Maximization and the Act of Choice*, Econometrica **65** (1997), 745–779.

[19] D. Stahl and P. Wilson, *On Player's Models of Other Players: Theory and Experimental Evidence*, Games and Economic Behavior **10** (1995), 218–245.

[20] M. Sutter, *Deception: The Role of Expectations. A Comment on Gneezy (2005)*, Mimeo (2006).

Proof of Proposition 1

Consider first the possibility of a separating PBE in which type i sends a message ($\sigma_1^i = 1$) but type j remains silent ($\sigma_1^j = 0$). Hence, $\mu^i = 1 - \mu^j = 1$. In that case, the game the sender i and the receiver play is represented by the matrix in Figure 6. If $k^i \geq 4$, *truth* becomes a dominant strategy and the only Nash equilibrium is (*truth,trust*). The utility of the sender is thus 1. Otherwise, there exists a unique equilibrium in mixed strategies where the sender tells the truth with probability $\sigma_2^i = \frac{5-\beta_L}{10-\beta_T-\beta_L}$ and receivers trust the sender's message with probability $\phi = \frac{1}{2} + \frac{k^i}{8}$. The sender's expected utility is then $3 - \frac{k^i}{2}$.

Suppose that the l type is the one who remains silent whereas the h type does not. Then, the sender gets in expectation either 1 if $k^l > 4$ and $3 - \frac{k^h}{2}$ otherwise. In this equilibrium the l type gets $3 - c$. This cannot be an equilibrium whenever $k^h \leq 2c$ since h will prefer to mimic. On the other hand, l will prefer to mimic if $2c > k^l$. Hence, for this to be an equilibrium it must be that $k^h \leq 2c \leq k^l$, which is a contradiction.

Suppose on the contrary that we are in a separating equilibrium in which the low type sends a message and the high type does not. The low type will do that as long as $k^l \leq 2c$. On the other hand, the high type will not feel tempted to mimic as long as $k^h \geq 2c$. Hence, such separating equilibrium can be supported if and only $k^l \leq 2c \leq k^h$.

Let us now discuss the existence of a pooling PBE in which both types of senders send a message ($\sigma_1^l = \sigma_1^h = 1$). In that case, the simultaneous move game played at the second stage can be represented by the following matrix:

	trust	*distrust*
truth, truth	$1, 5$	$5, \beta_T$
truth, lie	$1 + 4\varepsilon - \varepsilon k^h, 5 - (5 - \beta_L)\varepsilon$	$5 - 4\varepsilon - \varepsilon k^h, \beta_T + (5 - \beta_T)\varepsilon$
lie, truth	$5 - 4\varepsilon - (1 - \varepsilon)k^l, \beta_L + (5 - \beta_L)\varepsilon$	$1 + 4\varepsilon - (1 - \varepsilon)k^l, 5 - (5 - \beta_T)\varepsilon$
lie, lie	$5 - (1 - \varepsilon)k^l - \varepsilon k^h, \beta_L$	$1 - (1 - \varepsilon)k^l - \varepsilon k^h, 5$

In this matrix, the first action corresponds to a sender of type l and the second one to the type h sender. As it will become clear below two strategies at most will be played.

The receiver's best response to both (*lie,truth*) and (*lie,lie*) is *distrust* whereas it is *trust* to strategy (*truth,truth*). The best response to (*truth,lie*) depends upon the proportion of h type: It is *distrust* if and only if $\varepsilon > \frac{5-\beta_T}{10-\beta_L-\beta_T}$.

On the other hand, only the strategy (*truth,truth*) can be dominant, and this happens when $k^l \geq 4$. But in this case, the final utility of the sender is just 1. By deviating and remaining silent, the sender can get $3 - c$. So such pooling PBE cannot exist. When $k^l < 4$ the best response to *distrust* keeps being (*truth,truth*) but it is straightforward to show that the best response to *trust* is either (*lie,truth*) or (*lie,lie*). It is the former if and only if $k^h \geq 4$. In any case, this implies that there is no equilibrium in pure strategies of this simultaneous game if $k^l < 4$. Hence, we have to turn our attention towards the existence of mixed strategy equilibria.

Let us now compare each possible pair of strategies and find the probability ϕ with which the receiver must trust in order to support them as part of such equilibria.

(a) Suppose that two strategies played are (*truth,truth*) and (*truth,lie*). It must be that $k^h \leq 4$. The receiver would always play *trust* if $\varepsilon < \frac{5-\beta_T}{10-\beta_L-\beta_T}$ and such equilibrium could not exist since the sender would then only play (*truth,lie*). If $\varepsilon > \frac{5-\beta_T}{10-\beta_L-\beta_T}$, ϕ must be such that $\phi + 5(1-\phi) = \phi(1+4\varepsilon-\varepsilon k^h) + (1-\phi)(5-4\varepsilon-\varepsilon k^h)$ which is equivalent to $\phi = \frac{1}{2} + \frac{k^h}{8}$. This yields an expected utility of $3 - \frac{k^h}{2}$ for the sender. But observe that the expected utility from playing (*lie,lie*) is $1+4\phi-(1-\varepsilon)k^l-\varepsilon k^h = 3+\frac{k^h}{2}-(1-\varepsilon)k^l-\varepsilon k^h$. Since the latter is always higher than the former, there is no equilibrium in which only these two strategies are played.

(b) Suppose that the two strategies played are (*truth,truth*) and (*lie,truth*). The probability of playing *trust* must satisfy the equation $\phi + 5(1-\phi) = \phi(5 - 4\varepsilon - (1-\varepsilon)k^l) + (1-\phi)(1 + 4\varepsilon - (1-\varepsilon)k^l)$, which is equivalent to $\phi = \frac{1}{2} + \frac{k^l}{8}$. So, the expected utility of the sender is $3 - \frac{k^l}{2}$. We have compare it with the expected utility from the other two strategies: For strategy (*truth,lie*) it is $\phi(1 + 4\varepsilon - \varepsilon k^h) + (1-\phi)(5 - 4\varepsilon - \varepsilon k^h) = 3 - \varepsilon k^h - \frac{k^l}{2}(1 - 2\varepsilon)$, which is always smaller than $3 - \frac{k^l}{2}$. The expected utiltiy from (*lie,lie*) is $1+4\phi-(1-\varepsilon)k^l-\varepsilon k^h = 3+\frac{k^l}{2}-(1-\varepsilon)k^l-\varepsilon k^h$, which is also always smaller than $3 - \frac{k^l}{2}$.

Hence, we have to find the probabilities assigned to the strategies (*truth,truth*) and (*lie,truth*). In a mixed strategy equilibrium it must be the case that $5\sigma_{TT} + (\beta_L + (5 - \beta_L)\varepsilon)(1-\sigma_{TT}) = \beta_T\sigma_{TT} + (5-(5-\beta_T)\varepsilon)(1-\sigma_{TT})$, where σ_{TT} is the probability assigned to the strategy (*truth,truth*). This equation solves for $\sigma_{TT} = \frac{5-\beta_L}{10-\beta_T-\beta_L} - \frac{\varepsilon}{1-\varepsilon}\frac{5-\beta_T}{10-\beta_T-\beta_L}$.

The resulting expected utility for the sender is $3-\frac{k^l}{2}$. Hence, if $k^l \leq 2c$ this pooling PBE can be supported. Otherwise, the sender would prefer to deviate and remain silent.

(c) Suppose that the two strategies played are (*truth,truth*) and (*lie,lie*). The probability of playing *trust* must be such that $\phi + 5(1-\phi) = \phi(5-(1-\varepsilon)k^l - \varepsilon k^h) + (1-(1-\varepsilon)k^l - \varepsilon k^h)(1-\phi)$ which is equivalent to $\phi = \frac{1}{2} + \frac{(1-\varepsilon)k^l + \varepsilon k^h}{8}$. The resulting expected utility for the sender is $3 - \frac{(1-\varepsilon)k^l + \varepsilon k^h}{2}$. Again we have to compare this with the payoffs from the other two strategies: The expected utility from playing (*lie,truth*) is $3-(1-\varepsilon)k^l + \frac{(1-\varepsilon)k^l + \varepsilon k^h}{2}(1-2\varepsilon)$. Since the latter is never smaller, there is no pooling PBE in which only these two strategies are played.

(d) Suppose that the two strategies played are (*truth,lie*) and (*lie,truth*). The probability of playing *trust* must be such that $\phi(1+4\varepsilon-\varepsilon k^h) + (1-\phi)(5-4\varepsilon-\varepsilon k^h) = \phi(5-4\varepsilon-(1-\varepsilon)k^l) + (1-\phi)(1+4\varepsilon-(1-\varepsilon)k^l)$, which is equivalent to $\phi = \frac{1}{2} + \frac{(1-\varepsilon)k^l - \varepsilon k^h}{8(1-2\varepsilon)}$. The expected utility of the sender is then $3 - \frac{(1-\varepsilon)k^l + \varepsilon k^h}{2}$. If the strategy (*truth,truth*) was played instead, the sender would insure herself an expected utility of $3 - \frac{(1-\varepsilon)k^l - \varepsilon k^h}{2(1-2\varepsilon)}$. Since the latter is always greater than the former, there is no pooling PBE in which only these two strategies are played.

(e) Suppose that the two strategies played are (*truth,lie*) and (*lie,lie*). This cannot be an equilibrium if $\varepsilon > \frac{5-\beta_T}{10-\beta_L-\beta_T}$ because in that case the receiver would always play *distrust* in which case the sender would only play (*truth,lie*). So suppose $\varepsilon < \frac{5-\beta_T}{10-\beta_L-\beta_T}$. Then the probability of playing *trust* must be solve the equation $\phi(1+4\varepsilon-\varepsilon k^h) + (1-\phi)(5-4\varepsilon-\varepsilon k^h) = \phi(5-(1-\varepsilon)k^l - \varepsilon k^h + (1-(1-\varepsilon)k^l - \varepsilon k^h)(1-\phi)$. It solves for $\phi = \frac{1}{2} - \frac{k^l}{8}$ and in this case the expected utility to the sender is $3 - \frac{k^l}{2} - (1-\varepsilon)k^l - \varepsilon k^h$. The payoff from playing (*truth,truth*) is $3 + \frac{k^l}{2}$, and therefore, there is no equilibrium in which only these two strategies are played.

(f) Suppose that the two strategies played are (*lie,truth*) and (*lie,lie*). Moreover, let $k^h \leq 4$. In this case, the receiver would play *distrust* and the sender's best response would be to play (*lie,truth*) only. Hence, there cannot be an equilibrium in which the sender only plays these two strategies.

From the previous discussion it is evident that three or four strategies will not be played in

any mixed strategy equilibrium. Finally, consider the case of a pooling equilibrium in which both types remain silent. In that case, the off-the-equilibrium paths are undetermined and they can be made arbitrary. Suppose that if one of the types deviates and actually sends a message, the receiver believes that he is of type h with probability μ^h. This beliefs act exactly in the same way as ε in the analysis above. Then either the equilibrium described in (b) or in (d) arise. Notice that these utilities are independent of beliefs. Given the possible utilities, such pooling PBE can be supported only if $k^l > 2c$. □

Instructions of the Punishment Game ($c = 0.5$)

Welcome.

Thank you for coming. The purpose of this session is to study how people make decisions in a particular situation. If you have any questions, feel free to raise your hand and your question will be answered so everyone can hear. From now on until the end of the session, unauthorized communication of any nature with any other participant is prohibited. The experiment will be conducted through computers and all interactions between you will take place through them.

During the session you will play a game that gives you the opportunity to make money. What you earn depends partly on your decisions and partly on the decisions of others. At the end of the session, the amount you earned will be paid to you privately in cash.

We start with a brief instruction period. During the instruction period you will be given a description of the experiment. We are about to begin. Switch your phones off.

General Instructions.

Write down your name and your student ID number in the official receipt. You will need it to receive your payment at the end of the session. Next, we will go over a brief tutorial. Please interrupt at any time if you have a question.

In this session you will play a game, which is repeated for 50 rounds. Before the first round, the computer will randomly divide the participants into two groups of six. This division will last for the entire session. Participants within each group will play only among themselves. The assignment process is random and anonymous so you will not know who is in your group.

At the beginning of each round, you will be randomly joined with another participant from your group to form a pair. In each pair, one participant is randomly chosen to be the **sender**, and one to be the **receiver**. Remember that this process is random and the assignment changes every round. So, you will not know whom you are playing with.

Each round, after pairs have been formed and roles have been assigned, the computer selects one of the following two payoff tables. Final payoffs for both participants will essentially be determined according to the selected table and the action **U** or **D** to be taken later on by the receiver.

Table A	Sender	Receiver
Action U	5	1
Action D	1	5

Table B	Sender	Receiver
Action U	1	5
Action D	5	1

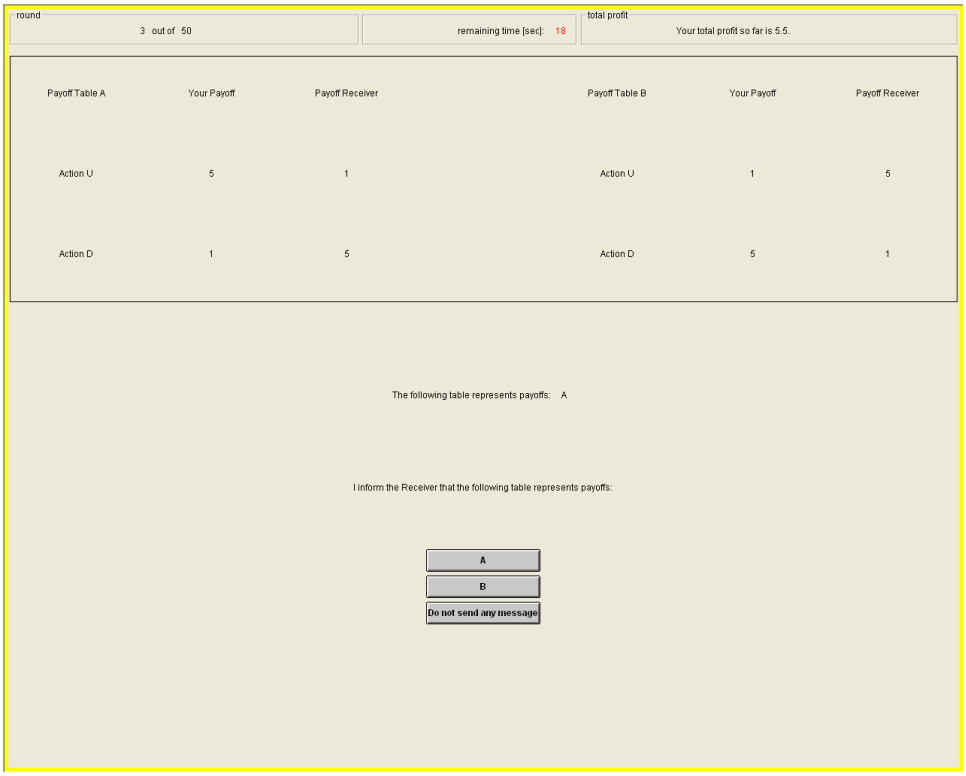
Sender's Instructions.

At the beginning of the round **only** the sender will be informed about the actual payoff table chosen by the computer. The sender is the first one to take a decision in the game. She/He has to select one of the following three options:

- Communicate to the receiver that table **A** represents payoffs.

- Communicate to the receiver that table **B** represents payoffs.
- Do not send any message to the receiver and pay a cost of 0.5 points.

The computer screen for the sender is as follows:



The two tables at the top of the screen represent payoffs according to the tables **A** and **B**. Below you find the information whether table **A** or table **B** was chosen by the computer. In our example it is table **A**. Additionally, you find three buttons labelled **A**, **B** and **Do not send any message**. By clicking on the buttons **A** or **B** you inform the receiver that you have observed the corresponding table. If you click on the button **Do not send any message** the receiver does not get any message and the sender has to pay a cost of 0.5 points. Please, take into account that the sender can select any option s/he wants.

The sender has 20 seconds to take this decision.

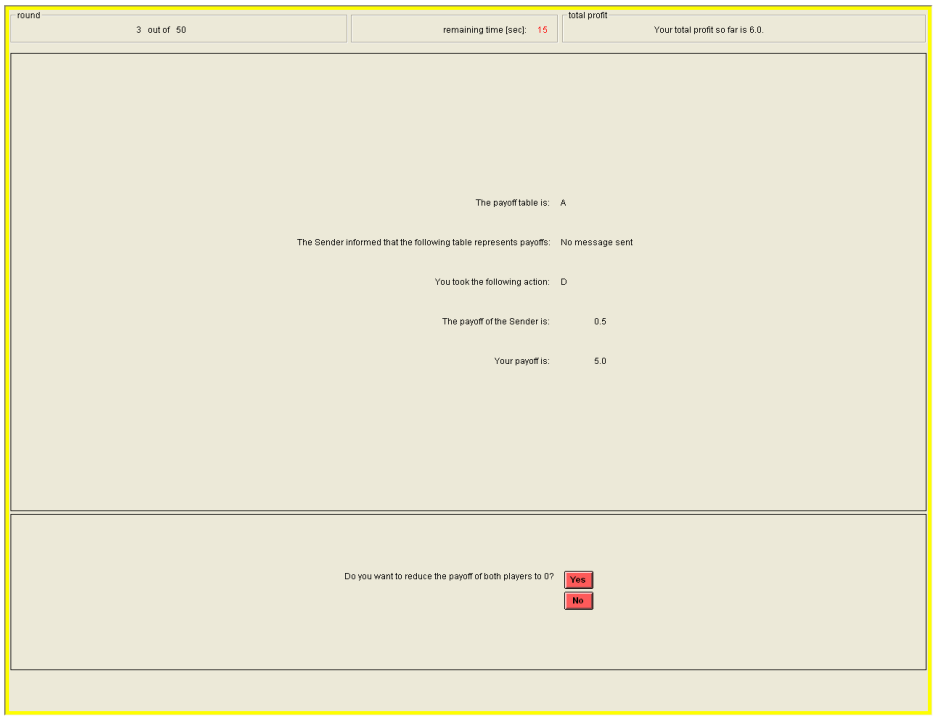
This is the only decision the sender has to make in this game.

Receiver's Instructions.

Once the receiver observes the message of the sender s/he is matched with, the receiver has to select between action **U** and **D**. The corresponding computer screen is as it can be seen on the top of the next page:

The two tables in the top of the screen represent payoffs according to the tables **A** and **B**. Below you find the decision of the sender. In our example the sender paid a cost of 0.5 points and did not send any message. On the right hand side, there are two buttons labelled **U** and **D**. By clicking on the buttons you take the corresponding actions.

The receiver has 20 seconds to take this decision. Once this action is taken, a new screen appears summarizing the outcome of the round so far.

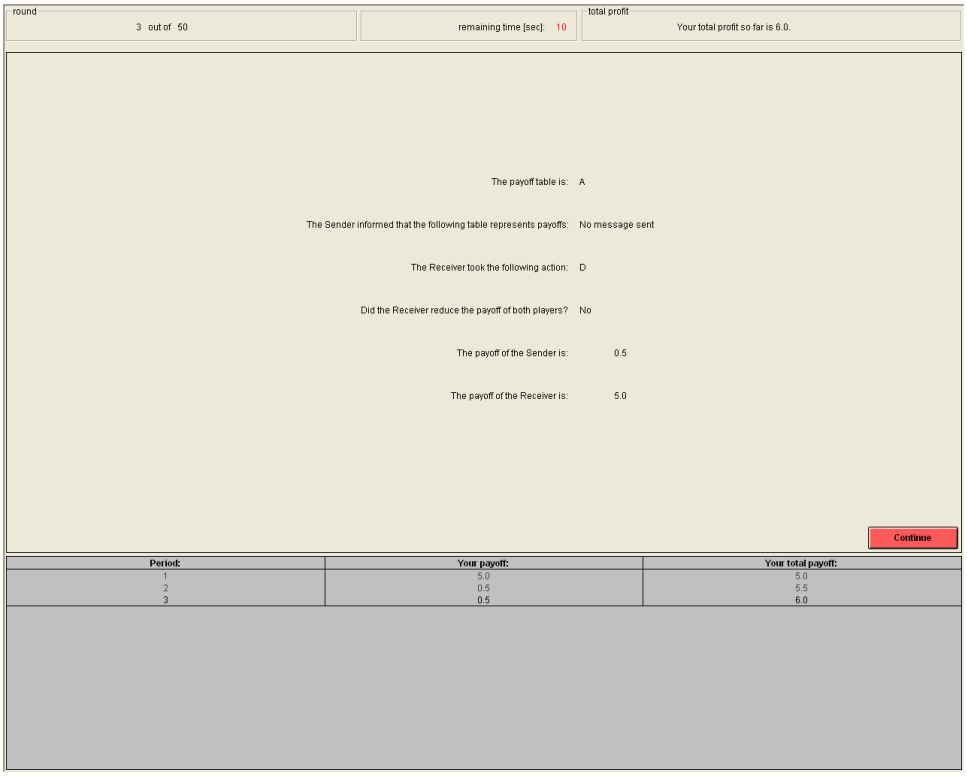


In the computer screen above you see that (a) the actual payoff table is represented by table **A**, (b) the sender **did not sent any message** and (c) the receiver took action **D**. As a result, the potential payoff of the sender and the receiver is 0.5 points and 5 points, respectively.

Now, the receiver is asked to take a final decision: She/He must either accept the current payoff distribution or reduce the payoff of both participants to zero. By clicking on the button **Yes**, the receiver reduces the payoffs of both participants. If you click on the button **No**, then the current payoff distribution is accepted. The receiver has 20 seconds to take this decision.

Summary of the Round.

The final screen is a summary of the round: It indicates the actual payoff table, the decisions of both the sender and the receiver, and the earnings of both participants in this round. Additionally, you are also informed about your total payoff up to this period.



The screen above is the sender's summary. It indicates that payoffs are represented by table **A** and that the sender **did not sent any message**. Moreover, since the receiver took action **D** and did not reduce payoffs, the sender gets 0.5 points and the receiver gets 5 points. At the end of a round, click on the button **Continue**.

Payment.

The points you accumulate during the course of the session will determine your payment. The exchange rate Points/Euros is such that every 15 points in the experiment are equal to 1 Euro; that is, if you total payoff after 50 rounds is equal to 165 points, then you get 11 Euros for your participation in this experiment.

Once the experiment has finished and you have answered the computerized questionnaire take the official receipt to the counter. Your personal data will be kept confidential and will be used only for statistical purposes. **Please, do not fill in your final payment into the official receipt. This will be done by us!!!** Once you are paid, you may leave.

Z-Tree Programs and Raw Data

Both the z-Tree programs and the raw data can be downloaded from the following website: http://www.personeel.unimaas.nl/m.vorsatz/

The z-Tree programs (finallieaversion1.ztt for the Benchmark Game and finallieaversion2.ztt for the Punishment Game) are ready to be implemented. However, if some question remains, please feel free to contact one of the authors. The raw data consists of ten Excel files. They are named in the following way: c-treatment-session#.xls. The variables follow also a rather straightforward denomination.

1. **Global Variables**

   - NumPeriods (number of repetitions)
   - nature (1=table A, 2=table B)
   - c (cost of remaining silent)
   - ratio (the higher of the two possible payoffs)

2. **Subject Variables**

   - player1 (1=message A, 2=message B, 3=remain silent, 0=subject was a receiver)
   - player2 (1=action U, 2=action D, 0=subject was a sender)
   - reduce (1=punish, 2=accept payoffs, 0=subject was a sender)
   - The variables Player1, Player2, and Reduce are equal to the actions taken by the participant the considered subject is matched with.

Finally, it should be mentioned that the subjects 1 to 6 form one independent observation and the subjects 7 to 12 another.