Truth, trust, and sanctions: on institutional selection in sender-receiver games

Ronald Peeters, Marc Vorsatz, Markus Walzl

Truth, trust, and sanctions: On institutional selection in sender-receiver games

RM/07/034

JEL code: A13, C72, Z13

Maastricht research school of **E**conomics
of **TE**chnology and **OR**ganizations

Universiteit Maastricht
Faculty of Economics and Business Administration
P.O. Box 616
NL - 6200 MD Maastricht

phone : ++31 43 388 3830
fax : ++31 43 388 4873

Truth, trust, and sanctions: On institutional selection in sender-receiver games*

Ronald Peeters† Marc Vorsatz†,‡ Markus Walzl†

September 4, 2007

Abstract

This paper reports on a laboratory experiment which investigates the impact of institutions and institutional choice in constant-sum sender-receiver games. We compare individual sender and receiver behavior in two different institutions: A sanction-free institution which is given by the bare sender-receiver game and a sanctioning institution which in addition offers the receiver the opportunity to (costly) sanction the sender after receiving feedback on the senders private information. We conduct the experiment in two phases: First, individuals are randomly assigned to an institution, and second they can choose the institution themselves.

We find that sanctioning takes place predominantly after the receiver has trusted a lie by the sender. Those who are responsible for sanctioning are also responsible for truth-telling in excess with respect to models of rational payoff-maximizing agents. Thereby, the sanctioning institution exhibits more truth-telling. Most importantly, agents who sanction reveal preference for the sanctioning institution while the other subjects almost exclusively opt for the sanction-free institution. As a consequence, both institutions typically coexist in the second phase of the experiment and the sanctioning institution exhibits a higher level of truth-telling and lower aggregate material payoffs.

To offer an explanation of our experimental findings, we formalize preferences for truth-telling as psychological payoffs and analyze the sender-receiver game as a dynamic psychological game *à la* Battigalli and Dufwenberg (2006). We demonstrate that standard models of social preferences are not able to explain observed sanctioning behavior and excessive truth-telling. Explicit psychological costs of lying and the exposition to a lie, however, are able to fill this gap. To this end, we model deontological and consequentialistic preferences for truth-telling and evaluate their respective explanatory power.

JEL Classification: A13, C72, Z13.

Keywords: Experiment; Sender-receiver games; Strategic information transmission; Institutional selection; Dynamic psychological games; Social norms.

*Financial support by Meteor and NWO is gratefully acknowledged.

†Department of Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Email: {R.Peeters, M.Vorsatz, M.Walzl}@algec.unimaas.nl.

‡Corresponding author.

1 Introduction

Everyone faces the decision to tell the truth or to lie to someone else at countless occasions, ranging from social contexts ("Did you like my nephew's performance as a solo-soprano in the boy's choir?") to economic settings ("Shall I reveal the potential pitfalls of this investment project?"). As diverse as the possible answers to these questions, are the various reflections upon the concept of "truth" in the history of philosophical thought. Following a deontological concept, some scholars view truth as an indispensable duty of mankind.[1] This perspective is contested by consequentialists who regard the decision to tell the truth as the result of a maximization of individual or societal welfare.[2] In short, a deontological perspective puts more emphasis on the means than on the goals while the reversed holds for consequentialists.[3] These different views on truth-telling are, however, not only a feature of discussions among philosophical scholars, but represent an ostinato of recent discussions on truth-telling to medical patients,[4] whistle-blowing programs for government officials,[5] or the debate on work ethics for managers (see *e.g.* The Economist (2007)).

But there is not only heterogeneity of individuals and philosophical systems if it comes to an attitude or appreciation of truth-telling. We can also observe a high degree of heterogeneity if we look at the way real life institutions (contracts, trade rules, or social norms *etc.*) actually enforce truth-telling. As an example consider, for instance, the various arrangements with which sports clubs want to induce truthful reports on the usage of sporting facilities by their members. Some clubs use costly monitoring or electronic registration techniques together with a menu of sanctioning opportunities for misbehavior, while other clubs just rely on their member's compliance. In a similar way, corporate cultures differ in the implementation of truthful reports by a CEO to the board of governors (see *e.g.* Frey and Jegen (2001)). While some firms use incentive contracts to establish truth-telling as a payoff-maximizing strategy for the CEO, other firms rely on the intrinsic motivation or the reputation-based development

[1] This view is shared by most religious ethical systems. The most pronounced exposition in philosophy is perhaps given by Immanuel Kant's categorial imperative (see Kant (1999, p.45)) according to which everyone should act in such a way that he also wants his actions to become a universal law ("Handle nur nach derjenigen Maxime, durch die du zugleich wollen kannst dass sie ein allgemeines Gesetz werde."). This idea has been applied by by the linguist Paul Grice who formulates a (descriptive) cooperation principle of communication (see Grice (1975)) according to which everyone contributes to "talk-exchange" in an expected way. Grice thereby emphasizes the importance of truth-telling as an essential tool to make communication feasible.

[2] The most prominent representatives of this perspective are utilitarians starting with Joseph Priestley and Jeremy Bentham. A pronounced presentation of this view can be found in Bentham (1996, ch.1) "... the greatest happiness of the greatest number is the foundation of morals and legislation". For an example for a modern theory of philosophical thought that picks up on this view see *e.g.* Nyberg (1993). An extreme form of consequentialism is clearly represented by the homo economicus who's only goal is the maximization of individual profit—regardless of the means or the well-being of others.

[3] Of course, there is no such dichotomy of philosophical schools in this respect. *E.g.,* Grice's work—though of deontological flavor—clearly pursues an important goal also familiar to consequentialists: The feasibility of communication.

[4] See, for instance, http://depts.washington.edu/bioethx/topics/truth.html.

[5] See, for instance, http://www.truthtellingproject.org.

of "conventions" within the profession. Other (economic) examples range from tax compliance to markets with incomplete information about product characteristics and to contracts for delegated expertise (with *e.g.,* lawyers, doctors, financial advisors, or external auditors).

Our experiment analyzes constant-sum sender-receiver games and thereby tries to capture the feature of strategic information transmission inherent to all these applications. We investigate how truth-telling is influenced by institutional details (*i.e.,* sanctioning opportunities) and institutional choice. In particular, we ask in how far truth-telling in a constant-sum sender-receiver game can be interpreted as a social norm which deserves enforcement through costly sanctions and whether self-selection of individuals into institutions creates sub-societies of distinct properties (*i.e.,* aggregate payoffs and norm-compliance through truth-telling).

The game played throughout the experiment is a sender-receiver game with two types of senders (each type being equally likely). Senders submit a (not necessarily truthful) message about their type to a receiver. If the receiver's action matches with the sender's type, the receiver gets a payoff of five and the sender a payoff of one; in case of a mismatch, the payoffs are opposite. Hence, the game is constant-sum, sender's and receiver's payoffs are antagonistic, and messages have no influence on the players' payoffs (cheap-talk).

For our experiment, we adopt a two-by-two within-subjects design. Firstly, we have two kinds of institutions. In the sanction-free institution individuals just play the sender-receiver game—being randomly assigned the role of sender and receiver, respectively. In the sanctioning institution, after having played the game and having observed the true type of the sender, the receiver is given the option to costly sanction the sender by reducing both payoffs to zero. Secondly, there are two phases. In the random assignment phase, which is the first phase, subjects are randomly assigned to an institution. In the following selection phase, subjects can select the institution (before knowing whether to play the role of sender or receiver).

Standard theory with rational, payoff-maximizing agents predicts that no information transmission takes place in any sequential equilibrium of the sender-receiver game (see Crawford and Sobel (1982)). That is, senders report the true state (*i.e.* tell the truth) half of the time and receivers believe in their reports (*i.e.* trust) half of the time.[6] Moreover, a receiver will never choose to sanction the sender if this reduces his own material payoff ex post (which is the case in our setting). Hence, senders and receivers would prefer to be in an institution without such a sanctioning opportunity.[7]

Sender-receiver games have been investigated in a few recent laboratory studies. Dickhaut et al. (1995) demonstrate that information transmission decreases in the degree of conflict

[6]Throughout the paper we will refer to a report by the sender that reveals the true type as *truth-telling* and a choice by the receiver that coincides with the sender's report as *trust*.

[7]Strictly speaking, players *weakly* prefer the sanction-free institution if they anticipate sequential equilibrium behavior. The expectation of some "accidental" sanctioning, however, would lead to a strict superiority of the sanction-free institution.

between sender and receiver in line with Crawford and Sobel (1982)'s predictions. However, experiments by Cai and Wang (2006) show that, on the aggregate level, senders overcommunicate in the sense that they reveal more private information than what standard theory predicts. A laboratory study by Gneezy (2005) indicates that the probability of lying is increasing in the potential gains from deception to the sender and decreasing in the potential losses to the receiver. Sánchez-Pagés and Vorsatz (forthcoming) investigate the relation between excess truth-telling and the willingness to (costly) sanction lies on the individual level. They show that those who sanction as receivers—in particular after they trusted a sender who deceived them—tell the truth excessively and are responsible for almost all excess truth-telling on the aggregate level. Our design allows us to extend this result by relating sanctioning behavior in the sanctioning institution to truth-telling and trust in the sanction-free institution.

From the relatively young but growing literature on institutional selection, the recent paper by Guererk et al. (2006) seems closest to our contribution. In their experiment, individuals choose between a sanctioning and a non-sanctioning institution to play a public good game. They find that a small number of individuals who prefer to contribute and to (costly) sanction free-riders—and therefore choose the sanctioning institution—is sufficient to establish higher aggregate payoffs in the sanctioning institution thereby attracting successively all participants. Hence, it is the higher material payoff which establishes the sanctioning institution as sole "winner" of institutional competition. However, it is not clear that in our setting the sanctioning institution will be as successful, since the joint material payoff can at best be as high as in the sanction-free institution.

In our experiment, we observe excessive truth-telling and trust compared to predictions by standard theory, and identify *sanctioners* as individuals who sanction after they have trusted a lie. We find that sanctioners are mainly responsible for excessive truth-telling in both and for excessive trust in the sanction-free institution (*i.e.,* if we withdraw the sanctioners from our data, individuals tell the truth and trust on average in half of the cases as predicted by sequential equilibrium behavior). With respect to the institutional choice in the selection phase, we observe that sanctioners predominately choose the sanctioning institution while the vast majority of other subjects opts for the sanction-free institution. Both institutions typically coexist in the selection phase, and the sanctioning institution exhibits more truth-telling and trust. As we observe sanctions in the sanctioning institution throughout the selection phase, one can conclude that there are individuals (sanctioners) who deliberately choose an institution with lower material payoffs but a higher level of truth-telling (and trust). Hence, the members of our experimental society self-select themselves into sub-societies of distinct economic performance (*i.e.,* payoff generation) and (normative) behavior (*i.e.,* truth-telling).

Our findings offer several implications. First, the robust observation of excess truth-telling (in line with previous studies) suggests that truth-telling is easier to implement than indicated by models with rational, payoff-maximizing agents. In particular, details of institutional design (such as opportunities for costly sanctions) that are irrelevant in these models have a systematic impact on individual behavior and aggregate institutional performance.

Second, sequential equilibrium (assuming rational, payoff-maximizing agents) fails to explain the behavior of sanctioners and the choice for and performance of the sanctioning institution. We demonstrate that models of distributional preferences (such as *e.g.,* inequity version) can only provide a partial explanation of these observations. While these models are able to motivate the observation of sanctioning, they are unable to explain significantly different sanctioning rates after a sender trusted a lie and after he did not trust the truth (both histories leading to the same payoff distribution). Towards a better understanding of actual behavior, we investigate the explanatory power of deontological and consequentialistic preferences for truth-telling. In the deontological model, we assign psychological costs to lies (and benefits to truth-telling) regardless of the consequences while in the consequentialistic model, we restrict costs from lying and benefits from telling the truth to situations where the sender believes the receiver to trust him in more than half of the cases (*i.e.,* psychological costs are not independent of the consequences and depend on beliefs). The deontological model proves to be a fruitful shortcut to explain the history-dependence of sanctioning behavior, but fails to describe actual excessive truth-telling on the individual level. The consequentialistic model in turn is rich enough to give a coherent explanation of our experimental findings.

Third, if institutional choice is endogenized, individuals self-select into sub-societies with distinct aggregate payoffs and levels of truth-telling and trust. In particular, self-selection leads to societies that are accurately described by standard theory next to societies that require a more complex modelling (*e.g.,* explicit preferences for truth-telling) for a coherent explanation. Hence, our experiment does not only demonstrate stable institutional diversity, but also emphasizes the impact of self-selection on accurate descriptions of truth-telling decisions by economic models.

The remainder of the paper is organized as follows. In Section 2, we present the experimental setting, design and procedures. Section 3 lists all results. In Section 4, we provide intuitions, implications and explanations. Section 5 concludes. The theoretical analysis and the instruction of the experiment are relegated to the Appendices.

2 Experimental set-up

2.1 Setting

We consider a sender-receiver game with payoffs as depicted in Figure 1. In this game, there

Figure 1: Sender-receiver game.

are two players: the sender and the receiver. The sender is either of type A or of type B. The actual type is chosen by nature and it is only known to the sender. It is common knowledge that nature selects each type with equal probability. The receiver decides whether to take action A or to take action B. In case the action matches with the type, the receiver gets a payoff of 5 and leaves the sender with a payoff of 1. Payoffs are reversed in case the action does not match with the type.

Before the receiver has to decide upon the action to take, but after the sender has learnt her type, the sender sends one of the following two messages to the receiver: message A ("the type selected by nature is type A") and message B ("the type selected by nature is type B"). Throughout this paper we say that the sender tells the 'truth' if her message matches with her type, otherwise we say she tells a 'lie'. Moreover, we say that the receiver does 'trust' the message if his action matches with the message received, otherwise we say he does 'distrust' the message.[8] Hence, the combinations truth–trust and lie–distrust lead to a payoff of 5 to the receiver and only 1 to the sender and the combinations truth–distrust and lie–trust lead to the reversed payoffs.

We consider this game in two different institutions, the sanction-free institution and the sanctioning institution. In the sanction-free institution, the sender and the receiver just play the sender-receiver game presented above. In the sanctioning institution, the receiver has additionally the option to sanction after the sender-receiver game has been played and after he has observed the real type of the sender. If the receiver sanctions, both players' payoffs are reduced to zero, otherwise the payoffs remain unchanged.

2.2 Design and procedures

The experiment was conducted with the help of the z-Tree toolbox (Fischbacher (2007)) in the experimental computer laboratory at Maastricht University in October and November 2006. All students of the Faculty of Economics and Business Administration were invited via email to register for the experiment. In total, we had 8 sessions with 20 subjects. Subjects received written and context-free instructions (see Appendix B) that they could study at their own pace. Eventual clarifying questions were dealt with privately. Before the experiment started, every subject had to answer some control questions correctly (see Appendix B).

[8]Note that we use 'truth', 'lie', 'trust', and 'distrust' as mere labels of (combinations of) actions that allow for a reduced form representation of the game (see Appendix A). A connection to the respective philosophical concepts is discussed in Section 4.

We conducted the experiment in two phases. The first phase is referred to as *random assignment phase* and it lasts 60 rounds. In each round, the 20 subjects are randomly divided in such a way that 6 subjects are assigned to the sanction-free institution and the remaining 14 are assigned to the sanctioning institution. Next, each subject is randomly matched to another subject within the same institution to form a pair. Within each pair, one subject is randomly chosen to be the sender, the other subject is the receiver. After all subjects are informed about the institution assigned to and the their role, the respective game is played.

The second phase of each session is referred to as *selection phase* and it lasts 40 rounds. At the beginning of each round, subjects decide in which institution to play. After all subjects have made their decision, each subject is randomly matched to another subject which has chosen the same institution to form a pair. In case of an odd number of subjects in an institution, one randomly chosen subject stays single and receives a fixed payoff of 3. Again, in each pair, one subject is randomly chosen to be the sender, the other subject is the receiver. After all subjects are informed about the selected institution and their role, the respective game is played.

After each round subjects were informed about all decisions taken within the respective pair, the resulting payoffs, and the individual accumulated payoff. Feedback on the identity of the subject they were matched to was never given.

After the experiment, subjects were paid off privately in cash. The average payment to the subjects was € 16.23 with the average session lasting 105 minutes.

3 Results

In any sequential equilibrium for the standard assumptions of (expected) payoff maximizing individuals and common belief of rationality, the sender tells the truth half of the time and the receiver trusts half of the time in both institutions. Moreover, the receiver will never choose to sanction and both players would weakly prefer the sanction-free institution (a strict preference for the sanction-free institution could be due to the anticipation of some "accidental" sanctioning). In the remainder of the paper, we will call truth-telling and trust rates *excessive* if they exceed these standard equilibrium predictions. Moreover, standard equilibrium predictions will serve as the Null-hypotheses for our tests if not stated otherwise.

This section is divided into four subsections, each of which is devoted to one of the decisions: Sanctioning, institutional selection, truth-telling and trust. The first subsection reports on observed sanctioning behavior throughout the experiment. In particular, we classify individuals by means of their sanctioning behavior. This allows us to study institutional selection, truth-telling and trust separately for individuals with distinct sanctioning behavior.

3.1 Sanctioning

Figure 2 illustrates the development of sanctioning behavior after the histories lie–trust and truth–distrust over rounds (clustered per 5 rounds). Sanctioning after the other histories has taken place only once for each history.

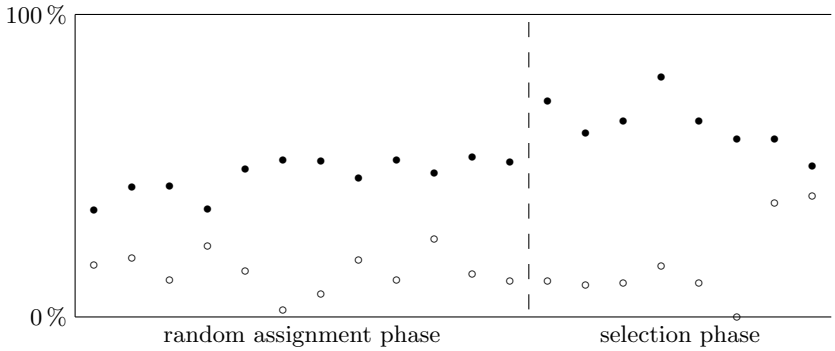

Figure 2: Sanctioning rates after truth–distrust (circles) and lie–trust (bullets) over rounds (5-round averages).

First of all, the occurrence of sanctioning (in both phases) is obviously in sharp contrast to what standard theory predicts. Figure 2 also indicates that there is more sanctioning after lie–trust than after truth–distrust, an observation that is confirmed by our statistical analysis (the one-sided p-value of the corresponding Wilcoxon signed-rank test is 0.0072 for the random assignment and 0.0173 for the selection phase). Moreover, we see that the transition to the selection phase increases sanctioning after lie–trust ($p = 0.0150$, one-sided), but not after truth–distrust ($p = 0.3363$, one-sided). The difference between the two trends narrows down at the end of the experiment, but the sanctioning rate after the history truth–distrust over the last ten rounds is there based on only 14 observations.

In our experiment, individuals take decisions in four different dimensions (neglecting phases, institutions and histories): truth-telling, trust, sanctioning, and institutional choice. Decisions regarding truth-telling, trust, and institutional selection can be driven by the individual's preferences *and* their beliefs over action choices of the other individuals in the respective institution. In contrast, the decision to sanction appears rather independent of beliefs but a more straightforward expression of preferences. Therefore, we distinguish individuals according to their sanctioning behavior. Subjects with a sanctioning rate of more than 80% in the random assignment phase as a receiver after having trusted a lie by a sender are classified as *sanctioners*, provided that the number of observations is at least 4. All subjects that are not classified as sanctioners, are classified as *others*.[9] Based on this procedure, 51

[9] A separate analysis of individuals who sanction after lie–trust *and* after truth–distrust (*i.e.,* sanctioning contingent on the payoff distribution) and of individuals who only sanction after lie–trust but not after truth–distrust (*i.e.,* sanctioning contingent on a specific history) is impossible due to the small number of observations of the history truth–distrust. The requirement of at least four lie–trust observations may well lead to the

out of the 160 participants in the experiment are classified as sanctioners.

Table 1 summarizes the average rates of sanctioning after lie–trust and truth–distrust in the two different phases for three different groups of individuals: all, sanctioners, and others. Moreover, this table presents the p-values of one-sided Wilcoxon signed-rank tests for differences in sanctioning rates between the random assignment and the selection phase. Because of the limited number of observations of the history truth–distrust in the selection phase, the data does not allow for meaningful tests on sanctioning behavior after this history and, therefore, the table only reports on the test results for sanctioning after lie–trust. However, we can observe from simple inspection that the sanctioning rates after truth–distrust do not change across phases. Most importantly, there is a significant increase in overall sanctioning, but not for the two different groups of individuals.

	lie–trust		truth–distrust	
	random assignment phase	selection phase	random assignment phase	selection phase
all	47 %	64 % [0.0150]	15 %	15 % [0.3363]
sanctioners	95 %	89 % [0.1468]	19 %	14 % [0.4276]
others	16 %	32 % [0.0917]	15 %	17 % [0.5000]

Table 1: Average sanctioning rates after lie–trust and truth–distrust. Between squared brackets, the p-values of a one-sided Wilcoxon signed-rank test on the difference in sanctioning between phases.

Result 1 (**Sanctioning**). *The sanctioning rate after lie–trust is higher than after truth–distrust and selection increases the sanctioning rate after lie–trust.*

3.2 Institutional Selection

The higher sanctioning rate after lie–trust in the selection phase (see Result 1) strongly suggests that subjects with different attitudes towards sanctioning prefer different institutions. Table 2 supports this interpretation. It can be seen that, in more than two-thirds of the cases, individuals have selected the sanction-free institution. However, the sanctioners have chosen the sanctioning institution in more than half of the cases, whereas the others selected this institution in one out of five cases only. The sanctioning institution broke down (because no pair of subjects opted for it) in only one out of eight sessions (the first one). The difference between the rates by which sanctioners and others select the sanctioning institution is found to be significant ($p = 0.0072$, one-sided Wilcoxon signed-rank test).

misidentification of sanctioners as others due to a lack of the history lie–trust for the respective individual. Hence, the number of sanctioners may be underestimated in our analysis.

	sanction-free institution	sanctioning institution
all	68 %	32 %
sanctioners	48 %	52 %
others	80 %	20 %

Table 2: Institutional selection.

Result 2 (**Institutional selection**). *Both institutions co-exist. Sanctioners choose more often the sanctioning institution than the others.*

3.3 Truth-telling

Figure 3 displays the development of the average rate of truth-telling during the sessions. The

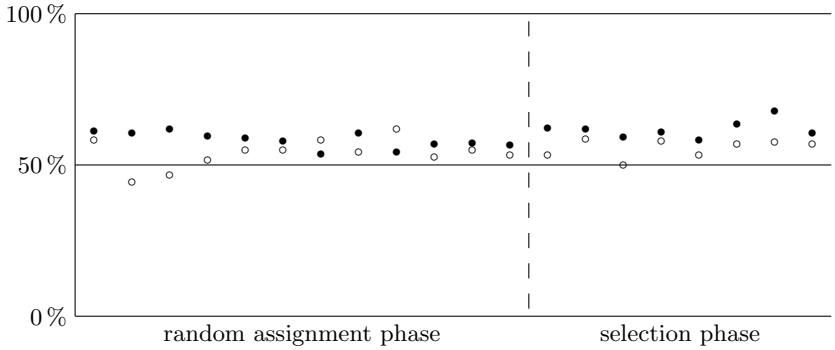

Figure 3: Truth-telling in the sanction-free institution (circles) and in the sanctioning institution (bullets) over rounds (5-round averages).

figure reveals that subjects tend to tell the truth excessively in both institutions and both phases. Moreover, excessive truth-telling seems more prominent in the sanctioning institution in the selection phase while the difference between the two institutions is less visible in the random assignment phase. Finally, the opportunity to select an institution seems to (slightly) increase the probability of truth-telling in the sanctioning institution, but has no obvious impact in the sanction-free institution.

Table 3 summarizes average rates of truth-telling in both institutions and phases for the three different groups of individuals and the p-values of the respective one-sided Wilcoxon signed-rank tests on excessive truth-telling. In addition, the table displays the p-values of the one-sided Wilcoxon signed-rank tests on the difference between truth-telling rates in the two phases.

Except for the sanctioning institution during the selection phase, we find significant excess truth-telling on the overall population. If we leave out the session where the sanctioning institution broke down, excess truth-telling in the sanctioning institution during the selection

	sanction-free institution		sanctioning institution	
	random assignment phase	selection phase	random assignment phase	selection phase
all	54 %	55 %	58 %	62 %
	(0.0209)	(0.0105)	(0.0105)	(0.0708)
	[0.2643]		[0.4168]	
sanctioners	64 %	64 %	74 %	79 %
	(0.0072)	(0.0212)	(0.0072)	(0.0212)
	[0.4721]		[0.1170]	
others	49 %	53 %	51 %	42 %
	(0.3121)	(0.1170)	(0.3121)	(0.0537)
	[0.0537]		[0.0212]	
difference	15 %	11 %	23 %	37 %
	(0.0072)	(0.0401)	(0.0072)	(0.0072)

Table 3: Average truth-telling rates for the overall population, the sanctioners and the others, and the difference between the latter two sub-populations. Between brackets, we display the p-values of the one-sided Wilcoxon signed-rank tests for excess truth-telling (first three rows) and the difference in truth-telling between sanctioners and others (last row). Between squared brackets, we display the p-values of the one-sided Wilcoxon signed-rank tests on the difference in truth-telling between the two phases.

phase becomes significant as well.[10] Excess truth-telling is caused by the sanctioners, who in both institution and phases told the truth excessively and significantly more often than the others. Excess truth-telling among others is nowhere found to be significant. Moreover, the data indicates that the others tend to lie excessively in the sanctioning institution throughout the selection phase ($p = 0.0537$, one-sided).

Aggregate data reveals a significant difference in the truth-telling rates between the two institutions. One-sided Wilcoxon signed-rank tests on this data lead to a p-value of 0.0517 for the random assignment phase and a p-value of 0.1355 for the selection phase. With a disregard of the first session, the latter p-value decreases to 0.0485.

On the overall level, the transition to the selection phase did not lead to a significant change of the truth-telling rate. The only significant result is that the others start to tell the truth less often in the sanctioning institution. For the sanction-free institution, the increase in the truth-telling rate from the random assignment to the selection phase for the others is significant with a p-value of 0.0537.

Result 3 (**Truth-telling**). *There is more excessive truth-telling in the sanctioning institution than in the sanction-free institution in both phases. Sanctioners are responsible for the excessive truth-telling everywhere. Institutional selection has no significant overall effect on truth-telling, but leads to less truth-telling among the others in the sanctioning institution.*

[10] In the first session (and only in this session), the sanctioning institution broke down. The insignificance is caused by this session.

3.4 Trust

Figure 4 displays the development of the average rate of trust during the sessions. In the

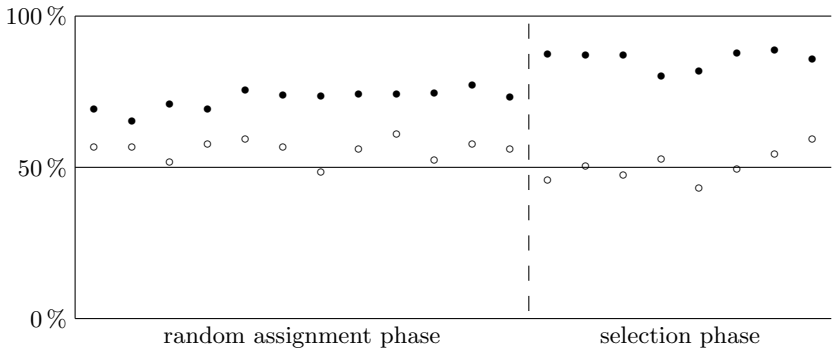

Figure 4: Trust in the sanction-free institution (circles) and in the sanctioning institution (bullets) over rounds (5-round averages).

sanctioning institution, receivers seem to trust excessively and there seems to be more trust when the institution is element of choice. In the sanction-free institution, subjects seem to trust excessively when randomly assigned to an institution. In the selection phase, however, the average rate of trust seems to be in line with what standard theory predicts.

Table 4 summarizes all average rates of trust in both institutions and phases for the three different groups of individuals and the p-values of the respective one-sided Wilcoxon signed-rank tests on excessive trust. In addition, the table displays the p-values of the one-sided

	sanction-free institution		sanctioning institution	
	random assignment phase	selection phase	random assignment phase	selection phase
all	56 % (0.0071)	50 % (0.4721) [0.0072]	73 % (0.0072)	86 % (0.0072) [0.0072]
sanctioners	55 % (0.1180)	48 % (0.3999) [0.0917]	89 % (0.0072)	92 % (0.0072) [0.4721]
others	56 % (0.0536)	51 % (0.4232) [0.0537]	65 % (0.0072)	78 % (0.0072) [0.0072]
difference	−1 % (0.3121)	−3 % (0.2643)	24 % (0.0072)	14 % (0.0754)

Table 4: Average trust rates for the overall population, the sanctioners and the others, and the difference between the latter two sub-populations. Between brackets, we display the p-values of the one-sided Wilcoxon signed-rank tests for excess trust (first three rows) and the difference in trust between sanctioners and others (last row). Between squared brackets, we display the p-values of the one-sided Wilcoxon signed-rank tests on the difference in trust between the two phases.

Wilcoxon signed-rank tests on the difference between trust rates in the two phases.

In the random assignment phase, we find excess trust in both institutions, whereas in the selection phase, there is excess trust only in the sanctioning institution. In the sanctioning institution, excess trust is caused by both types of individuals. Throughout the random assignment phase, sanctioners trust significantly more often than the others. In general trust rates seem to be higher in the sanctioning institution. Indeed, one-sided Wilcoxon signed-rank tests lead to a p-value of 0.0004 for both phases.

Finally, as for truth-telling, we close this subsection with an account of the difference between average trust rates in the random assignment and the selection phase. We find that trust increases for the sanctioning institution, but decreases for the sanction-free institution. For both institutions these shifts are mainly caused by the others (though only significant with a p-value of 0.0537 for the sanction-free institution). We summarize these results as follows.

Result 4 (**Trust**). *There is excessive trust in both institutions in the random assignment phase and in the sanctioning institution in the selection phase. In both phases, there is more trust in the sanctioning institution than in the sanction-free institution. For the two subpopulations, excess trust is only found in the sanctioning institution. Sanctioners are only found to trust significantly more than the others when randomly assigned to the sanctioning institution. Institutional selection increases trust in the sanctioning institution and reduces trust in the sanction-free institution.*

4 Implications and explanations

In this section, we aim at theoretical explanations of our experimental findings. In particular, we will investigate in how far observed behavior is in line with the hypothesis of payoff maximization, social preferences, or explicit preferences for truth-telling.

In our experiment, we do not find convincing evidence to reject the hypothesis that subjects classified as "others" are rational payoff-maximizers. Others do not tell the truth or trust excessively—the only exceptions are excessive trust in the sanctioning institution and the tendency to lie in the sanctioning institution during the selection phase. One is tempted to explain these deviations as a response to excessive aggregate trust and truth-telling levels. It should be kept in mind, however, that the best response of a rational, payoff-maximizing subject to an aggregate trust-level of more than 50 % is to tell the truth in 100 % of the cases (and likewise for a best-response to excessive truth-telling levels). Hence, our results suggest that others fail to eventually learn aggregate truth-telling and trust rates and that their behavior seems to reflect a (boundedly rational) learning process. More importantly,

the others nearly sanction[11] and their occasional choice of the sanctioning institution can well be explained through the opportunity to free-ride on anticipated high truth-telling and trust levels. Hence, the hypothesis that others are payoff maximizers cannot be rejected.

On the other hand, the behavior of "sanctioners" is in sharp contrast to the theoretical predictions based on the assumption of payoff maximization. They (i) tell the truth excessively (*i.e.,* they report the true state of the world with more than 50 %—but also in less than 100 %—of the cases), (ii) sanction predominately after the history lie–trust, and (iii) choose predominately the sanctioning institution. In what follows, we will investigate the predictive power of models of distributional preferences (*e.g.,* inequity aversion, maximin-preferences, or preferences for efficiency) and deontological and consequentialistic preferences for truth-telling in these respects.

4.1 Distributional preferences

As a benchmark, we start with a discussion of models of distributional preferences (*e.g.,* utility functions as proposed by Levine (1998), Fehr and Schmidt (1999), Bolton and Ockenfels (2000), or Charness and Rabin (2002))[12] that can all be expressed by a utility function that assigns to a given history of the game a utility that depends on the respective payoff distribution. A society of rational agents each of them endowed with such a (not necessarily identical) utility function (and with common knowledge thereof) would lead to the following sequential equilibrium behavior.[13]

(i) No information transmission. In any sequential equilibrium and independently of the institution, senders tell the truth half of the time and receivers trust half of the time. Hence, distributional preferences cannot explain excessive truth-telling and trust (in contrast to Result 3 and 4). In fact, the best-response correspondences for distributional preferences are identical to the payoff maximizing case.[14]

(ii) History-independent sanctioning. Receivers may sanction but their sanctioning behavior after histories that lead to the same payoff distribution is identical. This indicates that distributional preferences allow for an explanation of sanctioning as such (see Result 1). The reason is that receivers may prefer the material payoff distribution $(0, 0)$ to $(5, 1)$ if, *e.g.,*

[11] And, moreover, the sanctioning behavior of the others after the histories truth–distrust and lie–trust do not differ (the one-sided p-value of the corresponding Wilcoxon signed-rank test is 0.3121 for the random assignment and 0.1042 for the selection phase)

[12] For a recent overview of models of distributional preferences see *e.g.,* Sobel (2005).

[13] For a detailed analysis see Subsection A.5.

[14] Extreme altruism of one player could of course support a sequential equilibrium with 100 % truth-telling and trust. However, to admit such a preference structure would make it difficult to explain sanctioning behavior (by the same individuals).

they are sufficiently inequity averse, spiteful, or negatively reciprocal.[15] However, this implies the same sanctioning behavior after the history lie–trust and truth–distrust in contrast to Result 1.

(iii) Choice of the sanctioning institution. Agents may strictly prefer the sanctioning institution. In Appendix A (Proposition 2(iii)) we demonstrate that in a society with inequity averse agents as proposed by Fehr and Schmidt (1999), agents sanction after the histories lie–trust and truth–distrust and they strictly prefer the sanctioning institution. However, this explanation of institutional choice leaves the distinct truth-telling and trust levels across institutions in the selection phase unexplained (see Results 3 and 4).

Hence, even though the assumption of distributional preferences is able to motivate sanctioning, it fails to explain excessive truth-telling or trust and the history dependence of sanctions. To introduce history dependence, we continue with an explicit modelling of preferences for truth-telling in the next two subsections.

4.2 Deontological preferences

From a deontological point of view, a lie is bad *per se* regardless of intentions or consequences. This can be captured by a preference structure with psychological costs of lying for senders and psychological costs of the exposition to a lie for receivers and respective psychological benefits from truth-telling and the exposition to it. Clearly, this includes the extreme case of individuals who are "programmed" to tell the truth. As we discuss in more detail in Subsection A.6, there are configurations of agents with deontological preferences (and with common knowledge thereof) which lead to the following sequential equilibrium behavior.

(i) Information transmission. Senders always tell the truth and receivers always trust in both institutions. This indicates that deontological preferences can explain excessive truth-telling. However, it follows from Proposition 4(ii) and (iii) in Subsection A.6 that there are no configurations of agents with deontological preferences such that senders tell the truth in more than 50 % of the cases *and* in less than 100 % of the cases—which is what we observe (see Result 3).

(ii) History-dependent sanctioning. Receivers sanction after and only after the history lie–trust. This demonstrates that deontological preferences are able to explain history dependent sanctioning behavior. However, the history lie–trust would never be on an equilibrium path. Moreover, note that sanctioning after lie–trust and no sanctioning after lie–distrust is

[15] The reciprocity explanation has to be taken with care. According to recent reciprocity models such as Dufwenberg and Kirchsteiger (2004) or Falk and Fischbacher (2006), receivers sanction in our setting if they regard the sender's action as unkind. In this respect, it is hard to imagine that receivers regard a lie as unkind if it is the result of randomization by the sender. Therefore, the receiver's psychological costs from a lie have to depend on the beliefs about the sender's actions. We report on a model along these lines in Subsection 4.3.

not due to different psychological costs for the receiver but due to different material payoffs.

(iii) Choice of the sanctioning institution. Agents prefer the sanctioning institution and the sanctioning institution is characterized by more truth-telling and trust compared to the sanction-free institution. Hence, deontological preferences can explain the choice of the sanctioning institution (due to higher truth-telling probability and a correspondingly higher psychological payoff), but they fail to predict the exact pattern of truth-telling (between 50 % and 100 %) and sanctions (on the equilibrium path).

In sum, deontological preferences are indeed able to explain the history-dependence of sanctions and the choice of an institution that guarantees a higher level of truth-telling, but they fail to explain truth-telling probabilities between 50 % and 100 %. This does not come as a surprise. By assumption, deontological preferences ignore the consequences of a lie (*e.g.,* psychological costs are identical after lie–trust and lie–distrust). The respective psychological costs are therefore independent of the receiver's choice. Hence, psychological benefits and costs are just an offset to the receiver's utility and do not alter the best-response correspondence. But with an unaltered best-response for the receiver, the sender's best response is either to always tell the truth (if the corresponding benefits or the probability of a sanction are sufficiently high), or to tell the truth with probability 50 % (if psychological benefits and the probability of a sanction are sufficiently low). As a next step, we therefore introduce a consequentialistic notion of preferences which models a relationship between actions and psychological costs.

4.3 Consequentialistic preferences

To model consequentialistic preferences for truth-telling, we assume that senders perceive psychological benefits from telling the truth and suffer from a lie only if they believe that the receiver expects them to be honest in more than 50 % of the cases. In particular, the higher the probability with which they expect the receiver to believe in their honesty, the higher the psychological costs from lying and the psychological benefits from telling the truth. Whenever senders expect receivers to believe in their honesty with a probability of 50 % or less, however, they do not perceive any psychological costs and benefits. Likewise, receivers perceive benefits from truth-telling and suffer from a lie whenever they formed the belief that the sender tells the truth in more than 50 % of the cases. Again, psychological costs from lying and benefits from telling the truth increase in the probability with which the receiver expects the sender to tell the truth and psychological payoffs are zero whenever the anticipated probability of truth-telling is less or equal to 50 %. There are configurations of agents with such consequentialistic preferences (and with common knowledge thereof) which lead to the following

sequential equilibrium behavior.[16]

(i) Information transmission. Senders (receivers) tell the truth (trust) in more than 50 % and less than 100 % of the cases. Hence, the fact that a receiver's psychological payoff is no longer independent of the actions (via the respective equilibrium beliefs), induces the existence of sequential equilibria with truth-telling probabilities between 50 % and 100 %.

(ii) History-dependent sanctioning. Receivers sanction after and only after the history lie–trust. In particular, the absence of sanctions after lie–distrust is not only due to the higher material payoffs (as for deontological preferences) but also driven by lower psychological costs.

(iii) Choice of the sanctioning institution. Agents strictly prefer the sanctioning institution and the sanctioning institution is characterized by more truth-telling and trust compared to the sanction-free institution but also lower aggregate material payoff.

To summarize, distributional preferences fail to explain essential observations of our experiment like the history-dependence of sanctions and excessive truth-telling, while the explicit encounter of psychological costs of lying and benefits of truth-telling provide a coherent explanation of the data. Overall, both ad hoc costs and benefits of a deontological preference structure as well as belief-dependent costs and benefits of a consequentialistic model prove to be powerful in this respect. We thereby regard consequentialistic modelling as in some sense superior—it provides a more coherent explanation of the data and its description of receiver behavior seems to have more intuitive appeal. A closer look at the nature of preferences for truth-telling (*e.g.*, by eliciting the beliefs of the subjects or by introducing identical costs of sanctions after lie–trust and lie–distrust)—though obviously suggested by our present work—is nonetheless beyond the scope of the present paper.

5 Concluding remarks

For the last decade, one of the most active fields of research in experimental economics has been the analysis of situations with unobservable or uncontractable actions.[17] As a bottom line, this research has shown that standard economic theory with rational, payoff maximizing agents does a poor job in describing action choices for a given game or contract *and* the choice of contracts. *E.g.,* in experimental labor markets an agent's effort choice tends to be increasing in a fixed wage which is paid *ex ante* (see Fehr et al. (1993)), agents sanction certain actions by other agents even if it reduces their own material payoff (see Güth et al. (1982)), and some agents deliberately choose institutions with the opportunity to sanction other agents' action

[16] For a more detailed analysis see Subsection A.7.

[17] Consider for example labor-market experiments such as Fehr et al. (2006) or experiments on public good provision such as Fehr and Gächter (2000).

choice even though these institutions lead to lower material payoffs (see Guererk et al. (2006)). Inspired by these experimental findings, powerful theories of distributional preferences (see Sobel (2005)), reciprocity (see Dufwenberg and Kirchsteiger (2003) or Falk and Fischbacher (2006)), and procedural justice (see Sen (1997) and Brandts and Charness (2003)) have been developed to give a coherent explanation of the data. Our contribution helps to translate this experimental and theoretical research from the field of unobservable or uncontractable actions (moral hazard or incomplete contracts) to situations of unobservable types, *i.e.,* problems of adverse selection such as Akerlof (1970)'s market for lemons, monopolistic screening, or signalling games. The following environment illustrates the connection.

Consider a seller (the sender) who is privately informed about the quality of a car (good or bad) and a buyer (receiver) who only knows that the car is good and bad with 50 % probability. The seller sends a (not necessarily truthful) message about the quality of the car and the receiver chooses one out of two contracts. Contract A (low price and no maintenance) is optimal for the buyer if the car is good (payoff 5 for the receiver and 1 for the seller), while contract B (high price and maintenance) is optimal for the buyer if the car is bad (again, payoff 5 for the receiver and 1 for the seller). The combination good car and contract B or bad car and contract A leads to a payoff distribution of 1 for the buyer and 5 for the seller. Clearly, the unique sequential equilibrium (see Proposition 1 in Subsection A.4) yields an expected payoff of 3 for seller and buyer (independent of the institution) and a market-breakdown whenever the buyer's reservation utility is above 3. Our experimental evidence suggests that a market-breakdown is less likely due to excessive truth-telling. Moreover, sanctioning opportunities lead to even higher truth-telling levels (and therefore fewer market-breakdowns). As institutional selection enhances truth-telling in the sanctioning institution while reducing it in the sanction-free institution, our experiment illustrates the importance of self-selection of individuals with different preferences for truth-telling into distinct institutions. In particular, it provides in a nutshell an example for the robust coexistence of different institutional environments in a hidden-type setting.

From a theoretical perspective, our experiment demonstrates that preference structures that have been developed to describe laboratory and field behavior in the case of moral hazard and incomplete contracts are of limited descriptive power if it comes to settings with hidden types. Only the explicit encounter of preferences for truth-telling allows for a coherent explanation of the data. The existence of such preferences, however, has important implications for the analysis of situations with hidden types. In particular, our paper suggest the existence of an "intrinsic motivation to tell the truth" that comes along with the willingness to (costly) punish lies. This relaxes the importance of truth-telling constraints that are a corner-stone of any standard analysis of screening or signalling games and emphasizes the role of institutional (self-)selection driven by non-profit motives.

Finally, our findings suggest to model these preferences with great care. Deontological preference structures that assign psychological costs to a lie and benefits to truth-telling regardless of intentions and consequences (and in particular, models which assume that individuals are programmed to tell the truth) only provide an incomplete description of our data. A consequentialistic model that accounts for psychological costs and benefits that depend on beliefs (and thereby on intentions and consequences), however, can achieve a satisfactory explanation of the data. This shows some similarity with the limits of models of distributional preferences when explaining sanctioning behavior in settings with moral hazard or incomplete contracts (see Falk et al. (2005)). Only the explicit consideration of intentions in models of sequential reciprocity (see *e.g.* Dufwenberg and Kirchsteiger (2003)) draws a coherent picture. In this sense, our experiment provides a nice illustration of the importance of dynamic psychological games as a modelling structure that allows for an easy capture of belief-hierarchies as a primitive of the agent's utility function.

References

1. Akerlof G (1970). The market for 'lemons': Quality uncertainty and the market mechanism. Quarterly Journal of Economics 84 (3): 488-500.

2. Aumann R and A Branderburger (1995). Epistemic conditions for Nash equilibrium. Econometrica 63 (5): 1161-1180.

3. Battigalli P and M Dufwenberg (2006). Dynamic psychological games. mimeo.

4. Bentham J (1996). An introduction to the principles of morals and legislation (ed. J Burns and H Hart). New York: Oxford University Press.

5. Bolton P and A Ockenfels (2000). ERC: A theory of equity, reciprocity, and competition. American Economic Review 90 (1): 166-193.

6. Brandts J and G Charness (2003). Truth or consequence. Management Science 49: 116-130.

7. Cai H and J Wang (2006). Overcommunication in strategic information transmission games. Games and Economic Behavior 56 (1): 7-36.

8. Charness G and M Rabin (2002). Understanding social preferences with simple tests. Quarterly Journal of Economics 117 (3): 817-869.

9. Crawford V and J Sobel (1982). Strategic information transmission. Econometrica 50 (6): 1431-1451.

10. Dickhaut J, K McCabe and A Mukherji (1995). An experimental study on strategic information transmission. Economic Theory 6 (3): 389-403.

11. Dufwenberg M and G Kirchsteiger (2004). A theory of social reciprocity. Games and Economic Behavior 47 (2): 268-298.

12. Falk A, E Fehr and U Fischbacher (2005). Driving forces behind informal sanctions. Econometrica 73 (6): 2017-2030.

13. Falk A and U Fischbacher (2006). A theory of reciprocity. Games and Economic Behavior 54 (2): 293-315.

14. Fehr E and S Gächter (2000). Cooperation and punishment in public goods experiments. American Economic Review 90 (4): 980-994.

15. Fehr E, G Kichsteiger and A Riedl (1993). Does fairness prevent market clearing? An experimental investigation. Quarterly Journal of Economics 108 (2): 437-459.

16. Fehr E, A Klein and K Schmidt (2006). Fairness and contract design. Econometrica 75 (1): 121-154.

17. Fehr E and K Schmidt (1999). A theory of fairness, competition, and cooperation. Quarterly Journal of Economics 114 (3): 817-868.

18. Fischbacher U (2007). zTree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10 (2): 171-178.

19. Frey B and R Jegen (2001). Motivation crowding theory. Journal of Economic Surveys 15 (5): 589-611.

20. Geanakoplos J, D Pearce and E Stacchetti (1989). Psychological games and sequential rationality. Games and Economic Behavior 1 (1): 60-79.

21. Gneezy U (2005). Deception: The role of consequences. American Economic Review 95 (1): 384-394.

22. Grice P (1975). Logic and conversation. In: D Davidson and G Harman (eds.) The Logic of Grammar. Encino: Dickenson. pp. 64-75.

23. Guererk Ö, B Irlenbusch and B Rockenbach (2006). The competitive advantage of sanctioning institutions. Science 312 (5770): 108-111.

24. Güth W, R Schmittberger and B Schwarze (1982). An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization 3 (3): 367-388.

25. Kant I (1999). Grundlegung zur Metaphysik der Sitten (ed. B Kraft and D Schönecker). Hamburg: Meiner.

26. Levine D (1998). Modeling altruism and spitefulness in experiments. Review of Economic Dynamics 1 (3): 593-622.

27. Nyberg D (1993). The Varnished Truth: Truth Telling and Deceiving in Ordinary Life. Chicago: University of Chicago Press.

28. Sánchez-Pagés S and M Vorsatz (forthcoming). An experimental study of truth-telling in a sender-receiver game. Games and Economic Behavior. doj:10.1016/ j.geb.2006.10.014.

29. Sen A (1997). Maximization and the act of choice. Econometrica 65 (4): 745-779.

30. Sobel J (2005). Interdependent preferences and reciprocity. Journal of Economic Literature 43 (2): 392-436.

31. The Economist May 10th (2007). New Graduation Skills.

A Theoretical analysis

In the following, we provide a detailed equilibrium analysis of the institutions introduced in Section 2 with the preference notions discussed in Section 4. The analysis of distributional and deontological preferences amounts to a straightforward exercise of sequential equilibrium computation. In contrast, consequentialistic preferences require a more involved framework. Here, the agent's utility is a function that does not only depend on terminal histories and material payoffs, but also on belief hierarchies (a second-order belief of the sender and a first-order belief of the receiver). Such a payoff structure can be analyzed within the framework of *dynamic psychological games* as introduced by Battigalli and Dufwenberg (2006).[18] In what follows, we will provide a unified treatment of the preferences discussed in Section 4 within this framework.

We start (Subsection A.1) with some additional notation that complements the presentation of the two institutions in Section 2. In line with the analyzes in Section 3, we will work with a reduced form of the sender-receiver game with actions of the sender and receiver labelled as "truth-telling" and "trust". We continue (Subsection A.2) with a general introduction to the dynamic psychological game. Finally, after having presented three lemmas (Subsection A.3) on best-response behavior regarding sanctioning, truth-telling and trust, Subsections A.4–A.7 present the four preference structures and resulting propositions that are referred to in Section 4.

[18]This paper, henceforth BD(2006), generalizes *psychological* games as introduced by Geanakoplos et al. (1989), henceforth GPS(1989). While GPS(1989) allow the utility function of a player to depend on a hierarchy of his own beliefs, BD(2006) admit utility functions that depend on hierarchies of beliefs of all players.

A.1 Notation

As mentioned in the introductory text to this appendix we will consider the sender-receiver game of Section 2 in its reduced form. Instead of considering the game as a dynamic game of incomplete but perfect information, we will approach it as a dynamic game with complete but imperfect information. This is done by considering the strategies of the players as meta-strategies on the level of truth-telling and trust.

The reduced sender-receiver game consists therefore of a set of players $N = \{S, R\}$ (Sender and Receiver, generic element i) and for each player a set of pure strategies: $A_S = \{t, l\}$ ($t$ruth and $l$ie, generic element a_S) and $A_R = \{t, d\}$ ($t$rust and $d$istrust, generic element a_R). Mixed strategies are a probability distributions over A_S and A_R, parameterized by σ_S and σ_R—the probabilities to tell the truth and to trust, respectively. As all strategy spaces have cardinality two, we can abbreviate $-a_i = A_i \backslash \{a_i\}$ and $A_{-i} = A_{N \backslash i}$. A combination of strategies induces a probability distribution over the set of histories $H = \{(t,t), (t,d), (l,t), (l,d)\}$ (generic element h) that are composed by the sender's and the receiver's pure strategies. Finally, the payoffs $x_i(h)$ of player i for history h are given by

$$x_S(h) = \begin{cases} 1 & \text{if } h = (t,t) \text{ or } h = (l,d) \\ 5 & \text{if } h = (t,d) \text{ or } h = (l,t) \end{cases}$$

and

$$x_R(h) = \begin{cases} 5 & \text{if } h = (t,t) \text{ or } h = (l,d) \\ 1 & \text{if } h = (t,d) \text{ or } h = (l,t). \end{cases}$$

In the sanction-free institution, the players just played the sender-receiver game. For this institution the set of terminal histories Z (generic element z) equals the set H and the payoffs in these terminal histories are defined by $x_i(z) = x_i(h)$ for $i = S, R$.

In the sanctioning institution, the receiver has an additional set of pure actions $A'_R = \{s, n\}$ ($s$anction or $n$ot, generic element a'_R). After having observed the history h, the receiver decides to sanction or not. Therefore a pure sanctioning strategy is a mapping $a' : H \to A'$. We denote a mixed sanctioning strategy by σ'_R such that $\sigma'_R(h)$ denotes the probability that the receiver sanctions in response to the history h. For this institution the set of terminal histories Z (generic element z) equals the set $\{h \times A'_R \,|\, h \in H\}$. The payoffs in each terminal history $z = (h, a'_R)$ are given by

$$x_i(z) = \begin{cases} x_i(h) & \text{if } a'_R = n \\ 0 & \text{if } a'_R = s \end{cases}$$

for $i = S, R$. Let $x(z)$ denote the distribution of material payoffs for a terminal history z.

A.2 Dynamic psychological games

Consider a set X, a Borel sigma-algebra $\mathcal{B}$ on X, and a collection of events $\mathcal{C} \subseteq \mathcal{B}$. A conditional probability system (cps) on $(X, \mathcal{B}, \mathcal{C})$ is a function $\mu(\cdot|\cdot) : \mathcal{B} \times \mathcal{C} \to [0,1]$ such

that (i) $\mu(\cdot|F)$ is a probability measure over X for every $F \in \mathcal{C}$, (ii) $\mu(F|F) = 1$, and (iii) $E \subseteq F' \subseteq F$ implies $\mu(E|F) = \mu(E|F')\mu(F'|F)$ for all $E \in \mathcal{B}$ and $F, F' \in \mathcal{C}$.

Define for every player i and $k = 0, 1, 2, \ldots$ the set X_{-i}^k recursively as follows. $X_{-i}^0 = A_{-i}$ and $X_{-i}^k = X_{-i}^{k-1} \times \Delta^H(X_{-i}^{k-1})$ where $\Delta^H(X)$ denotes the set of all cps for X for a given set of histories H. A cps $\mu_i^k \in \Delta^H(X_{-i}^{k-1})$ is called a kth order cps. A *hierarchy* of *cps* for player i is a countably infinite sequence of *cps*' $\mu_i = (\mu_i^1, \mu_i^2, \ldots)$. A hierarchy is *coherent* if the cps of distinct orders assign the same conditional probabilities to lower-order events. A hierarchy is collectively coherent if each of its cps assigns probability one to the coherence of the hierarchy of the other players. We denote the set of collectively coherent hierarchies for player i by M_i and $\bigtimes_{i \in N} M_i \equiv M$ (generic element μ).

A *dynamic psychological game* is a structure $\Gamma = (N, H, (u_i)_{i \in N})$ where

$$u_i : Z \times M \times A_{-i} \to \mathbb{R} \tag{1}$$

is player i's *psychological payoff function.* So, psychological payoffs are defined on the set of terminal histories (Z) and on player $-i$'s strategy space. Moreover, psychological payoffs may depend on elements of M, *e.g.,* the sender's psychological payoff may depend on her second-order cps μ_S^2—*i.e.,* on her belief on the receiver's belief in truth-telling—and thereby introducing a cost for the sender to "fool" a receiver who believes in her honesty.

The primitive for the equilibrium analysis are behavior strategies denoted by σ_i. The sender's behavior strategy is fully determined by σ_S, the receiver's behavior strategy in the sanction-free institution is fully determined by σ_R and in the sanctioning institution by the pair (σ_R, σ_R').[19]

An assessment for player i is a pair (σ_i, μ_i) where σ_i is a behavior strategy and μ_i is a hierarchy of cps' for player i. In the two player case such an assessment is called *consistent* if first-order cps' are stochastically independent and a player's higher-order beliefs in μ assigns probability one to the respective lower-order beliefs in μ.[20] An assessment profile $(\sigma_i, \mu_i)_{i \in N}$ will be denoted by (σ, μ).

Let us denote the expected utility for player i with psychological payoff function u_i from choosing pure strategy a_i for a consistent belief hierarchy μ by $\mathbb{E}_\mu[u_i|a_i]$ and denote the expected utility from a behavior strategy profile σ by $\mathbb{E}_\mu[u_i|\sigma]$. An assessment profile (σ, μ) is a *sequential equilibrium* (SE) if it is consistent and

$$\text{Supp}\,(\sigma_i) \subseteq \underset{a_i \in A_i}{\text{argmax}}\ \mathbb{E}_\mu[u_i|a_i].$$

We denote the set of sequential equilibria of a dynamic psychological game Γ by $SE(\Gamma)$.

[19] Following BD(2006), we interpret a mixed strategy σ_i of player i as the common (first-order)-belief of the other player(s) about the strategy choice of i (see Aumann and Branderburger (1995)).

[20] For more than 2 players consistency additionally requires the coincidence of first-order beliefs of any two players about a third player's strategies (see Assumption 6 in BD(2006)).

The following result guarantees the existence of a sequential equilibrium in a dynamic psychological game.

Theorem 1. ***Battigalli and Dufwenberg (2006)*** *For any dynamic psychological game* Γ *a sequential equilibrium exists* ($SE(\Gamma) \neq \emptyset$).

The psychological payoff function $u_i : Z \times M \times A_{-i} \to \mathbb{R}$ is flexible enough to capture all preference structures discussed in Section 4. For the sake of simplicity and following the main contributions to the literature on non-standard preferences (see Sobel (2005)), we restrict ourselves to risk-neutral players and (in case of deontological and consequentialistic preferences) to an additively separable structure of material and psychological payoffs. Furthermore, we adopt as a convention that a player's utility in case of sanctioning is zero. Relaxing these assumptions does not alter the main conclusions regarding the descriptive power of the different preference structures expressed in Section 4 but drives the analysis much more cumbersome. Finally, we exclude extreme forms of altruism (such that players strictly prefer a payoff distribution of 1 for themselves and 5 for the other player) and summarize as follows.

Assumption 1. *(i) Consider the sanctioning institution and let* $a'_R = s$. *Then* $u_i(z, \mu) = 0$ *for all* i. *(ii) Fix* μ, *let* $x(z) = (5, 1)$ *and* $x(z') = (1, 5)$. *Then,* $u_S(z, \mu) > u_S(z', \mu)$ *and* $u_R(z, \mu) < u_R(z', \mu)$.

A.3 Three lemmas

We first analyze the receiver's optimal sanctioning decision (in the sanctioning institution) and thereby adopt the convention that the receiver sanctions ($a'_R(h) = s$) if and only if he is strictly better off by doing so.

Lemma 1. *Fix a belief hierarchy* μ *and a history* $h = (a_S, a_R)$. *Then,* $\sigma'_R(h) = 1$ *if and only if* $0 = u_R((h, s), \mu) > u_R((h, n), \mu)$ *and* $\sigma'_R(h) = 0$ *otherwise in any sequential equilibrium.*

Proof. Obvious. □

Lemma 1 implies that only a pure action at the punishment stage can be part of an equilibrium. We will refer to this action as a^*_R.

To elicit the best response correspondences for the receiver and the sender in the sender-receiver game, we henceforth assume that the receiver indeed chooses a^*_R in the sanctioning stage (of the sanctioning institution). The following lemma fully determines the best response correspondence of the receiver.[21]

[21] For expositional ease we suppress a^*_R and μ as arguments of the psychological payoff function from now on. Hence, we continue with $u_i(a_S, a_R)$. Moreover, we slightly abuse notation and denote by $\mu^1_{R(S)}$ the probability with which the receiver (sender) believes the sender (receiver) to tell the truth (trust), *i.e.,* μ^1_i denotes player i's first-order belief on σ_{-i}.

Lemma 2. *Let $u_i(\cdot)$ satisfy Assumption 1. The best response of the receiver is given by*

$$\sigma_R = \begin{cases} 1 & \text{if } \mu_R^1 > \bar{\mu}_R \\ [0,1] & \text{if } \mu_R^1 = \bar{\mu}_R \\ 0 & \text{if } \mu_R^1 < \bar{\mu}_R \end{cases}$$

with

$$\bar{\mu}_R = \frac{u_R(l,d) - u_R(l,t)}{u_R(l,d) - u_R(l,t) + u_R(t,t) - u_R(t,d)}.$$

Proof. The receiver's best response correspondence results from the following maximization program.

$$\max_{\sigma_R \in [0,1]} \mu_R^1 \left(\sigma_R u_R(t,t) + (1-\sigma_R) u_R(t,d)\right) + (1-\mu_R^1)\left(\sigma_R u_R(l,t) + (1-\sigma_R) u_R(l,d)\right)$$

which is equivalent to the problem

$$\max_{\sigma_R \in [0,1]} \sigma_R(\mu_R^1 \left(u_R(t,t) - u_R(t,d)\right) + (1-\mu_R^1)\left(u_R(l,t) - u_R(l,d)\right)) + C_1 \tag{2}$$

where C_1 is a constant that does not depend on σ_R. Any feasible σ_R solves this latter problem if $\mu_R^1 = \bar{\mu}_R$ (as defined in the lemma). If $\mu_R^1 > \bar{\mu}_R$, (2) is solved by $\sigma_R = 1$; and, if $\mu_R^1 < \bar{\mu}_R$, (2) is solved by $\sigma_R = 0$. □

Analogously, the sender's best response correspondence is determined as follows.

Lemma 3. *Let $u_i(\cdot)$ satisfy Assumption 1. The best response of the sender is given by*

$$\sigma_S = \begin{cases} 1 & \text{if } \mu_S^1 < \bar{\mu}_S \\ [0,1] & \text{if } \mu_S^1 = \bar{\mu}_S \\ 0 & \text{if } \mu_S^1 > \bar{\mu}_S \end{cases}$$

with

$$\bar{\mu}_S = \frac{u_S(l,d) - u_S(t,d)}{u_S(l,d) - u_S(l,t) + u_S(t,t) - u_S(t,d)}.$$

Proof. The sender's best response correspondence results from the following maximization program.

$$\max_{\sigma_S} \sigma_S \left(\mu_S^1 u_S(t,t) + (1-\mu_S^1) u_S(t,d)\right) + (1-\sigma_S)\left(\mu_S^1 u_S(l,t) + (1-\mu_S^1) u_S(l,d)\right)$$

which is equivalent to the problem

$$\max_{\sigma_S} \sigma_S[\mu_S^1(u_S(t,t) - u_S(l,t)) + (1-\mu_S^1)(u_S(t,d) - u_S(l,d))] + C_2 \tag{3}$$

where C_2 is a constant that does not depend on σ_S. Any feasible σ_S solves this problem if $\mu_S^1 = \bar{\mu}_S$ (as defined in the lemma). If $\mu_S^1 > \bar{\mu}_S$, (3) is solved by $\sigma_S = 0$; and, if $\mu_S^1 < \bar{\mu}_S$, (3) is solved by $\sigma_S = 1$. □

A.4 Payoff maximization

In the case of payoff maximization, the player's utility coincides with his individual (material) payoff. This is a special case of the psychological payoff function in (1)—the player's utility only depends on the terminal node $z \in Z$, and the player's individual payoff in this terminal node.[22]

Assumption 2. *Let $x_i(z)$ be player i's material payoff for a terminal history z. Then, player i's utility for z is given by $u_i(z) = x_i(z)$.*

This assumption leads to the benchmark result discussed at the beginning of Section 3 (for a proof see *e.g.,* Sánchez-Pagés and Vorsatz (forthcoming)).

Proposition 1. *Let $u_i(\cdot)$ satisfy Assumption 1 and 2.*

(i) In the sanctioning institution, in any SE, $\sigma'_R(h) = 0$ for all $h \in H$.

(ii) In both institutions, in any SE, $\sigma_S = \sigma_R = \frac{1}{2}$.

(iii) Players weakly prefer the sanction-free institution.

A.5 Distributional preferences

For distributional preferences, we assume that a player's utility is a continuous function of the payoff-distribution in a terminal node and formalize as follows.

Assumption 3. *Player i's utility for z is given by*

$$u_i(z) = f_i(x(z)).$$

Proposition 2. *Let $u_i(\cdot)$ satisfy Assumption 1 and 3.*

(i) Let $h, h' \in H$ and suppose $x(h) = x(h')$. Then, $\sigma'_R(h) = \sigma'_R(h')$ in any SE (in the sanctioning institution). Suppose that $u_R(h) < 0$, then $\sigma'_R(h) = 1$ in any SE.

(ii) In both institutions, in any SE, $\sigma_S = \sigma_R = \frac{1}{2}$.

(iii) Suppose $u_i(z) = x_i(z) - \kappa \max(x_{-i}(z) - x_i(z), 0) - \lambda \max(x_i(z) - x_{-i}(z), 0)$ for all $i \in \{S, R\}$ with $\kappa = 2$ and $\lambda = 0.6$.[23] Then, $\sigma'_R(h) = 0$ if $x(h) = (1, 5)$ and $\sigma'_P(h) = 1$ if $x(h) = (5, 1)$. Moreover, players prefer the sanctioning institution.

[22] Recall that, throughout this appendix, z refers to both institutions simultaneously and that $z = h$ for the sanction-free institution and $z = (h, a'_R)$ for the sanctioning institution.

[23] Model and specification taken from Fehr and Schmidt (1999).

Proof. (i) With Assumption 3, $x(h) = x(h')$ implies $u_R(h) = u_R(h')$. But then Lemma 1 indicates that $\sigma'_R(h) = \sigma'_R(h')$ in any SE.

(ii) From Assumption 3 and Part (i) it follows that $u_R(t,t) = u_R(l,d)$ and $u_R(l,t) = u_R(t,d)$. This implies by Lemma 2 that $\bar{\mu}_R = \frac{1}{2}$ which fixes the receiver's best response correspondence. Now suppose in contradiction to the proposition that $\mu_R^1 \neq \frac{1}{2}$ and let without loss of generality be $\mu_R^1 > \frac{1}{2}$. Then, Lemma 2 indicates that $\sigma_R = 1$. But then $\mu_S^1 = 1$ and Lemma 3 imply that the sender's best response is $\sigma_S = 0$. But $\sigma_S = 0$ contradicts $\mu_R^1 > \frac{1}{2}$. $\mu_S^1 = \frac{1}{2}$ follows analogously.

(iii) Suppose $u_i(z) = x_i(z) - \kappa \max(x_{-i}(z) - x_i(z), 0) - \lambda \max(x_i(z) - x_{-i}(z), 0)$ for all $i \in \{S, R\}$ with $\kappa = 2$ and $\lambda = 0.6$. Then, $u_R(t,t) = u_R(l,d) = 5 - 0.6 \cdot 4 = 3.6$ and $u_R(l,t) = u_R(t,d) = 1 - 2 \cdot 4 = -7$. Hence, $\sigma'_R(h) = 1$ if $x(h) = (5,1)$ and $\sigma'_R(h) = 0$ if $x(h) = (1,5)$. Given Part (ii), the sender's and receiver's expected utility in the sanction-free institutions is $\frac{1}{2}3.6 - \frac{1}{2}7 = -1.7$. In the sanctioning institution, the sender's expected utility is $\frac{1}{2}0 - \frac{1}{2}7 = -3.5$ and the receiver's expected utility is $\frac{1}{2}3.6 - \frac{1}{2}0 = 1.8$. Hence, with random assignment of the role of a sender and a receiver, each player expects a utility of $-0.85 > -1.7$ and therefore opts for the sanctioning institution. □

A.6 Deontological preferences

For deontological preferences, we assume that players regard lying as *per se* bad and therefore incur psychological costs from lying and being exposed to a lie, respectively. A player's utility therefore depends on her/his material payoff *and* the strategy of the sender as follows.

Assumption 4. *Let $x_i(z)$ be player i's payoff for a terminal history z. Then, player i's utility for z is given by*

$$u_i(z) = x_i(z) + g_i(z)$$

with $g_i(h) = c_i^+ > 0$ if $a_S = t$ and $g_i(h) = c_i^- < 0$ if $a_S = l$. For the sanction-free institution $g_i(z) = g_i(h)$ and for the sanctioning institution $g_i(z) = g_i(h)$ if $a'_R = n$ and $g_i(z) = 0$ if $a'_R = s$.

Proposition 3. *Let $u_i(\cdot)$ satisfy Assumption 1 and 4.*

(i) Suppose $a_S = t$, then $\sigma'_R(h) = 0$. Suppose that $a_S = l$ and $x_R(h) < |c_R^-|$, then $\sigma'_R(h) = 1$.

(ii) Consider the sanctioning institution and suppose that $|c_R^-| \in (1,5)$. Then in any SE, $\sigma_S = \sigma_R = 1$, and $\sigma'_R(h) = 1$ for $h = (l,t)$ and $\sigma'_R(h) = 0$ otherwise.

(iii) Consider the sanction-free institution and suppose that $c_S^+ - c_S^- < 4$. Then, $\sigma_S = \frac{1}{2}$ and $\sigma_R \in (\frac{1}{2}, 1)$ in any SE. Suppose $c_S^+ - c_S^- \geq 4$. Then, $\sigma_S = \sigma_R = 1$ in any SE.

(iv) Suppose $c_S^+ = c_R^+ = 0$ and $c_S^- = c_R^- = 2$. Then, players strictly prefer the sanctioning institution.

Proof. (i) Follows from Assumption 4 and Lemma 1.

(ii) With Assumption 4 and Lemma 1 it follows that $\sigma'_R(h) = 1$ if $x_R(h) + g_R(h) < 0$ and $\sigma'_R(h) = 0$ otherwise. Then, $|c_i^-| \in (1, 5)$ implies $u_S(h) = 0$ for $h = (l, t)$, and Lemma 3 implies that $\sigma_S = 1$ is the best response for the sender (independent of μ_S^1). The respective unique best response of the receiver is $\sigma_R = 1$ (see Lemma 2).

(iii) With Assumption 4 it follows that $u_R(l, d) - u_R(l, t) = u_R(t, t) - u_R(t, d)$. This implies (by Lemma 2) that $\bar{\mu}_R = \frac{1}{2}$ which fixes the receiver's best response correspondence. Moreover (see Lemma 3), $\bar{\mu}_S = \frac{1}{2} + \frac{1}{8}(c_S^+ - c_S^-)$. For $c_S^+ - c_S^- < 4$, this implies $\bar{\mu}_S \in (\frac{1}{2}, 1)$ which fixes the sender's best response correspondence and implies the first part of the result. Finally, suppose $c_S^+ - c_S^- \geq 4$. Then, the sender's best response is $\sigma_S = 1$ (see Lemma 3) such that $\mu_R^1 = 1$ and the receiver's best response becomes $\sigma_R = 1$ (see Lemma 2).

(iv) With $c_R^- = 2$, Part (ii) implies that $\sigma'_R(l, t) = 1$ and $\sigma'_R(h) = 0$ for all other $h \in H$. Moreover, $\sigma_R = \sigma_S = 1$ in any SE. Hence, the sender's expected utility in the sanctioning institution is $u_S(t, t) = 1$ and the receiver's expected utility is $u_R(t, t) = 5$ (recall that $c_S^+ = c_R^+ = 0$). Hence, random assignment induces an expected payoff of 3 from the sanctioning institution. For the sanction-free institution, $c_i^+ + |c_i^-| = 2$ induces $\mu_R^1 = \frac{1}{2}$ and $\mu_S^1 = \frac{3}{4}$ (see Part (iii)). This yields an expected utility for the sender and receiver of 2. Hence, they both prefer the sanctioning institution. □

A.7 Consequentialistic preferences

For consequentialistic preferences, we denote by α the sender's (second-order) belief about the receiver's belief on σ_S (the probability of truth-telling). And we denote by $\beta = \mu_R^1$ the receiver's first-order belief about σ_S. A player's utility now depends on her/his material payoff *and* the elements of belief-hierarchies α and β as follows.

Assumption 5. *Let $x_i(z)$ be player i's payoff for a terminal history z. Then, the sender's utility for z and α is given by*

$$u_S(z, \alpha) = x_S(z) + g_S(z, \alpha)$$

with $g_S(h, \alpha)$ continuous, $\frac{\partial g_S(h,\alpha)}{\partial \alpha} > 0$ if $a_S = t$, $a_R = t$, and $\alpha > \frac{1}{2}$, $\frac{\partial g_S(h,\alpha)}{\partial \alpha} < 0$ if $a_S = l$, $a_R = t$, and $\alpha > \frac{1}{2}$, and $g_S(h, \alpha) = 0$ otherwise. The receiver's utility for z and β is given by

$$u_R(z, \beta) = x_R(z) + g_R(z, \beta)$$

with $g_R(h, \beta)$ continuous, $\frac{\partial g_R(h,\beta)}{\partial \beta} > 0$ if $a_S = t$, $a_R = t$, and $\beta > \frac{1}{2}$, $\frac{\partial g_R(h,\beta)}{\partial \beta} < 0$ if $a_S = l$, $a_R = t$, and $\beta > \frac{1}{2}$, and $g_R(h, \beta) = 0$ otherwise. For the sanction-free institution

$g_i(z,\cdot) = g_i(h,\cdot)$ *and for the sanctioning institution* $g_i(z,\cdot) = g_i(h,\cdot)$ *if* $a'_R = n$ *and* $g_i(z,\cdot) = 0$ *if* $a'_R = s$.

Proposition 4. *Let* $u_i(z,\mu)$ *satisfy Assumption 1 and 5.*

(i) There is always a SE with $\sigma_R = \sigma_S = \frac{1}{2}$ *in the sanction-free institution and there is always a SE with* $\sigma_S = \sigma_R = \frac{1}{2}$ *and* $\sigma'_R(h) = 0$ *for all* $h \in H$ *in the sanctioning institution.*

(ii) Consider the sanctioning institution and suppose that $|g_R((l,t),1)| \in (1,5)$. *Then there is a SE with* $\sigma_R = \sigma_S = 1$, $\sigma'_R(z) = 1$ *for* $h = (l,t)$ *and* $\sigma'_R = 0$ *otherwise.*

(iii) Consider the sanction-free institution and suppose that $|g_S((t,t),\alpha)| < |g_S((l,t),\alpha)|$ *for any* $\alpha > \frac{1}{2}$ *and* $|g_R((t,t),\beta)| < |g_R((l,t),\beta)|$ *for any* $\beta > \frac{1}{2}$. *Moreover, suppose* $|g_S((t,t),1)| + |g_S((l,t),1)| < 4$. *Then, there is a SE with* $\sigma_R \in (\frac{1}{2},1)$ *and* $\sigma_S \in (\frac{1}{2},1)$.

(iv) Suppose that $g_i((t,t),\frac{6}{11}) = 1$ *and* $g_i((l,t),\frac{6}{11}) = -2$. *Then, players prefer the sanctioning institution.*

Proof. (i) For a belief-system with $\mu_S^1 = \mu_R^1 = \frac{1}{2}$ (and therefore $\alpha = \beta = \frac{1}{2}$), Assumption 5 implies that the player's utility is given by their material payoff $x_i(z)$. Hence, it follows analogously to Proposition 1 that $\mu_S^1 = \mu_R^1 = \frac{1}{2}$ indeed constitutes a SE.

(ii) With Assumption 5 and Lemma 1 it follows that $\sigma'_R(h) = 1$ if $|g_R(h,\alpha)| > x_R(h)$ and $\sigma'_R = 0$ if $|g_R(h,\alpha)| \leq x_R(h)$. Then, $|g_R(h,\alpha)| \in (1,5)$ implies $u_S(h) = 0$ for $h = (l,t)$, and Lemma 3 implies that $\sigma_S = 1$ is the best response for the sender (independent of μ_S^1). The respective best response of the receiver is $\sigma_R = 1$ (see Lemma 2) consistent with $\alpha = \beta = 1$.

(iii) With the $|g_S((t,t),\alpha)| < |g_S((l,t),\alpha)|$ for any $\alpha > \frac{1}{2}$, $|g_R((t,t),\beta)| < |g_R((l,t),\beta)|$ for any $\beta > \frac{1}{2}$, and $|g_S((t,t),1)| + |g_S((l,t),1)| < 4$, it follows that $\bar{\mu}_S \in (\frac{1}{2},1)$ and $\bar{\mu}_R \in (\frac{1}{2},1)$. Then, the result follows analogously to the proof of Proposition 1 (for details see *e.g.,* the Appendix in Sánchez-Pagés and Vorsatz (forthcoming)).

(iv) Obviously, player's are indifferent between both institutions if players coordinate on the equilibrium characterized in Part(i). Assume from now on that players coordinate on the equilibria characterized in Part (ii) and (iii), respectively. With $g_i((t,t),\frac{6}{11}) = 1$ and $g_i((l,t),\frac{6}{11}) = -2$, Part (ii) implies an expected utility for the sender lager than $u_S(t,t) = 2$ (recall that $g_i(h,\cdot)$ is monotone increasing in α and β, respectively) and for the receiver larger than $u_R(t,t) = 6$. Hence, random assignment induces an expected payoff of at least 4 from the sanctioning institution. For the sanction-free institution institution, $g_i((t,t),\frac{6}{11}) = 1$ and $g_i((l,t),\frac{6}{11}) = -2$ induces $\bar{\mu}_R = \frac{6}{11}$ and $\bar{\mu}_S = \frac{4}{5}$. This yields an expected utility for the sender of $\frac{63}{55}$ and the receiver of $\frac{235}{55}$. Hence, random assignment induces an expected payoff of $\frac{149}{55}$ which is less than 2.71 and therefore does not outperform the sanctioning institution. Hence, both players prefer the sanctioning institution. □

B Instructions

Welcome

Dear participant,

thank you for taking part in this experiment! It will last about 2 hours. You will be compensated according to your performance during the experiment. In order to ensure that the experiment takes place in an optimal setting, we would like to ask you to follow the general rules during the whole experiment:

- please do not communicate with your fellow students!
- please do not forget to switch off your mobile phone!
- read the instructions carefully. If something is not well explained or any question turns up now or at any time later in the experiment, then ask one of the experimenters. Do, however, not ask out loud, but raise your hand! We will clarify questions privately.
- you may take notes on this instruction sheet if you wish.
- after the experiment, please remain seated till we paid you off.
- if you do not obey the rules, the data becomes useless for us. Therefore we will have to exclude you from this experiment and you will not receive any compensation.

Your decisions are anonymous. None of your fellow students nor anybody else will ever learn them from us.

Environment 1

The central situation of the experiment is the situation depicted in Figure 5 with the following underlying story.

A	B
1 ; 5	5 ; 1

Table A

A	B
5 ; 1	1 ; 5

Table B

Figure 5: Central situation of the experiment

There are two players, a *sender* and a *receiver*. In the beginning, the computer randomly selects one of the payoff tables A and B, each with equal probability. Only the sender will be (correctly) informed which table has been selected. Next, the sender transmits either the message *"Table A has been selected"* or the message *"Table B has been selected"* to the receiver. Please, observe that the sender can transmit whatever message he prefers. After

observing the sender's message, the receiver decides whether to take *action A* (that is to select column A) or to take *action B* (that is to select column B). The interpretation of the actions is that the receiver says either *I believe the actual payoff table is A* or *I believe the actual payoff table is B*. The payoffs to the sender and the receiver, which are given by the numbers in the corresponding cell, depend only on the table actually chosen by the computer and the action selected by the receiver. The first number in the cell corresponds to the payoff of the sender, the second number to the payoff of the receiver. In short, if the receiver's action matches with the actual table she receives 5 ECU (Experimental Currency Units) and the sender 1 ECU. Otherwise, payoffs are the opposite. For example, if the computer chooses table A, the tells the receiver that table A has been selected, and the receiver takes action A, then the sender gets 1 ECU and the receiver 5 ECU.

Environment 2

The second environment extends the first environment. After receiving feedback on the table chosen by the computer and the decisions of the sender and the receiver, the receiver has to make a final decision. She has to decide whether to *accept* the payoffs for both participants or whether to *reduce* the payoff of both participants to zero.

Matching

This experimental session consists of 100 rounds. In total, 20 subjects participate in this experiment. In every of the first 60 rounds, the computer assigns you randomly to one of the two environments. With 70 % probability you will be assigned to the second environment. Next, you are randomly matched with another participant from the same environment to form a pair. In each pair, one participant is randomly chosen to be the sender, and one to be the receiver. This process is random. Your profile may change every round with respect to three variables: the environment you are assigned to (1 or 2), the participant you are matched with (some subject from the same environment), and the role you have (sender or receiver). The matching is anonymous, so you will never learn with whom you formed a pair. After every round you receive a complete feedback of the decisions of both players, the payoffs from the round, and your accumulate payoff.

In the second phase of the experiment, the last 40 rounds, you can decide whether you want to be in environment 1 or in environment 2. This decision is taken every round anew. Given your decision for the current round, you are again randomly matched with another participant from the same environment to form a pair. In each pair, one participant is randomly chosen to be the sender, and one to be the receiver. Observe that if an odd number of participants choose an environment it becomes impossible to divide all players into pairs. In this situation, the participant that stays single does not have to make decisions and gets a fixed payoff of

3 ECU. The matching is anonymous, so you will never learn with whom you formed a pair. After every round you receive a complete feedback of the decisions of both players, the payoffs from the round, and your accumulate payoff.

Payment

The points that you accumulate in course of the experiment will determine your payment. The exchange rate ECU/Euros is such that every ECU in the experiments is equal to 5 Eurocents.

Closing

At the end of the experiment, we would like to ask you to complete a short on-screen questionnaire. But, before we start, we would like to ask you to answer the control questions on the bottom of this page. Once ready, please raise your hand, and one of the experimenters will check your answers. The software will be started as soon as *all* answers have been checked. So, please, be patient.

Thank you again and good luck with the experiment! And, please, make your decisions carefully—your reward depends on your performance during the experiment.

Control questions

Please, answer the following questions! One of the experimenters will go round, check the answers and discuss any problems.

Please fill in your subject id: _____

Statement	**True**	**False**
In the 43th round of the experiment, I will be able to select my favorite environment.		
If I am playing the role of sender this round, I can be sure to be playing the role as receiver next round.		
I never know whom of the other participants I am matched with.		
As a sender I can be sure that the receiver regards my message as credible.		
In the second environment, before making the decision of whether or not to reduce the payoffs of both participants, I am informed about the selected table and the payoffs resulting from my choice as a receiver.		
My decisions in the first phase do not influence my payoffs.		