**Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2012/2**

Conditional Cooperation With Negative Externalities – An Experiment

Christoph Engel
Lilia Zhurakhovska

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2012/2

Conditional Cooperation With Negative Externalities – An Experiment

Christoph Engel / Lilia Zhurakhovska

February 2012

revised version August 2014

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

Conditional Cooperation With Negative Externalities – An Experiment

Christoph Engel[*] / Lilia Zhurakhovska[#+]

Abstract

Empirically, the commons are not as tragic as standard theory predicts. The predominant explanation for this finding is conditional cooperation. Yet many real life situations involve insiders, who are directly affected by a dilemma, and outsiders, who may be harmed if the insiders overcome the dilemma. The quintessential illustration is oligopoly. If insiders overcome their dilemma and collude, this inflicts harm on the opposite market side. In our experiment, harm on outsiders significantly reduces conditional cooperation of insiders. We can exclude that this result is driven by inequity aversion, reciprocity or efficiency seeking. Only guilt aversion can rationalize our findings, with guilt being most pronounced if the active insiders not only inflict harm on the outsider, but increase their own payoff at the expense of the outsider.

Keywords: conditional cooperation, negative externalities, prisoner's dilemma, beliefs, inequity aversion, efficiency, guilt aversion.

* Max Planck Institute for Research on Collective Goods, Bonn
\# University of Erlangen-Nuremberg
\+ University of Cologne

I. Introduction

Many economic and social situations can be modelled as problems of cooperation. The quintessential cooperation problem is a prisoner's dilemma. Already Rapoport and Chammah (1965) refute the textbook prediction and demonstrate that many people are willing to cooperate in a prisoner's dilemma. This finding has been supported by a wide range of economic studies using laboratory experiments (e.g. Poundstone, 1992; Andreoni and Miller, 1993) and field studies (e.g. Ostrom, 1999; Ostrom, Dietz et al., 2002). The predominant explanation for this behavior is conditional cooperation (Keser and van Winden, 2000; Fischbacher, Gächter et al., 2001; Kocher, Cherry et al., 2008; Fischbacher and Gächter, 2010). Conditionally cooperative individuals cooperate if they know, experience, or believe that others are unlikely to exploit them. Under these conditions, they are willing to forego the possibility to exploit others.

Now, many dilemmas of life are more complex than a two-person prisoner's dilemma. Specifically, co-operation problems are frequently embedded in a wider social context. If insiders successfully overcome the dilemma, some outsiders suffer. In such a situation, the moral balance becomes complicated. Insiders must decide whether to let down each other in order to avoid harming outsiders, or to accept harm on outsiders in the interest of being loyal to their in-group fellows.

In the field, this conflict between kindness at the interior and meanness at the exterior is not uncommon. Sometimes, being mean is the very purpose of cooperation, as in a military coalition or in a trade union. At other instances, the harm is more a side-effect which is deliberately taken into account. Those closer to the source of a river build a dam, knowing that this deprives those closer to the estuary of the benefits of the river. A municipality builds a landfill to keep garbage off its streets, knowing that this puts the groundwater of neighboring municipalities at risk.

The most obvious motivation of our paper, however, is oligopoly. Viewed from inside the supply side of the market, competition may be interpreted as a prisoner's dilemma. In this perspective, collusion is the equivalent of cooperation, competitive behavior is defection. Individually, each supplier is best off if the other suppliers are faithful to the cartel, and she undercuts the collusive price or, for that matter, surpasses her quota. Yet if they cooperate, suppliers impose a distributional loss on the demand side, and they generate a deadweight loss, to the detriment of society.

Harm to an outsider may affect cooperation in two ways: insiders may generally become less willing to cooperate, and they may become less sensitive to the experienced or expected level of cooperation among insiders. Harm to an outsider might thus reduce unconditional and conditional cooperation. To test whether this is the case, we experimentally investigate a dilemma with a passive outsider. We deliberately use a very simple setting. A simultaneous symmetric one-shot two-person prisoner's dilemma with binary action space serves as our *Baseline*. In

three treatments we add a passive third participant. Whenever at least one of the active players chooses to cooperate, the passive participant suffers harm. The three treatments differ in the severity of the harm.

We find that harm to a passive outsider affects unconditional and conditional cooperation. We find significant main effects: participants cooperate less if cooperation inflicts harm on an outsider, and they cooperate the more, the more they are optimistic about the cooperativeness of other active participants. Yet we find a significant negative interaction between the level of harm and subjects' expectations about the choices of other active participants. If cooperation inflicts harm on a passive outsider, optimism about the cooperativeness of other active participants is less likely to tilt the balance in favor of cooperation, compared with the *Baseline*.

This is not only an interesting finding in and of itself. It also helps us better understand what motivates conditional cooperation. None of the standard explanations for conditional cooperation predicts this negative interaction effect: inequity aversion, reciprocity, or an efficiency motive; nor does the intuitive aversion against inflicting harm on passive outsiders. The only way to rationalize the robust interaction effect requires a utility function with guilt aversion, and guilt most pronounced if the two active players take advantage of the passive outsider. The more they are optimistic that their active counterpart cooperates, the more it becomes likely that this happens. This result suggests that guilt aversion is the most plausible motive driving conditional cooperation, even if there is no outsider.

The remainder of the paper is organized as follows: Section 2 relates the paper to the existing literature. Section 3 introduces the design. Section 4 makes theoretical predictions. Section 5 presents and discusses results. Section 6 concludes.

II. Related Literature

The effects of externalities on passive outsiders have only rarely been studied. To the best of our knowledge, they have not been tested in a standard prisoner's dilemma. Most related is a paper by one of us with another co-author. Engel and Rockenbach (2011) study a standard repeated four-person linear public good game with three passive outsiders. They vary the direction of the externality and the endowment of the outsiders. Insiders do not cooperate more if this has the additional advantage of making outsiders better off, and they do not cooperate less if this has the additional disadvantage of making outsiders worse off. Rather results are in line with insiders trying to increase the payoff gap between themselves and outsiders. We build on this design, but focus on the most interesting effect, the apparent absence of reticence to impose harm on passive outsiders. Our design differs in the following respects: we implement a one-shot game. This excludes the shadow of the future as a potential confounding factor. We use two-person games. This excludes expectations and experiences about heterogeneity as a possible explanation. We use various levels of harm. This way we do not only see whether any level of harm categorically influences choices, but can investigate whether more

pronounced harm has a more pronounced effect. Finally, and most importantly, we elicit beliefs. That way we can disentangle cognitive and motivational effects of imposing harm on passive outsiders.

Other relevant studies are for example Güth and van Damme (1998). They present an ultimatum game with an externality on an inactive third player. The proposer decides how to divide the pie between three players. The division is executed if and only if the responder accepts. Otherwise, all three players receive nothing. In this game, the outsider receives very little. If the responder only learns the fraction the proposer wants to give the outsider, proposers keep almost everything for themselves. In anticipation, responders are very likely to reject the (mostly unknown) offer. Bolton and Ockenfels (2010) study lottery choice tasks in which the actor's choice also influences the payoff of a non-acting second player. This induces participants to take larger risks, provided the safe option yields unequal payoffs. Abbink (2005) plays a two-person bribery game in which corruption negatively affects passive workers. He concludes that reciprocity between briber and official overrules concerns about distributive fairness towards other members of the society. Ellman and Pezanis-Christou (2010) study how a firm's organizational structure influences ethical behavior towards passive outsiders. A firm of two players decides on its production strategy, which influences a passive third player. They find that horizontally organized firms in which the firm's decision corresponds to the average of both individual decisions are less likely to harm the outsider than consensus-based firms or firms in which one of both members is the boss. There is a rich experimental literature on oligopoly (see the meta-study by Engel 2007), yet it does not focus on the fact that oligopoly is socially embedded.

III. Design

In our experiment, we have a *Baseline* with just two active players, and three treatments with an additional passive outsider who is negatively affected by insiders choosing a cooperative move.

III.1. The Game

Table 1
Payoff Matrix

	C	D
C	R€, R€, $-h$€	S€, T€, $-h$€
D	T€, S€, $-h$€	P€, P€,0€

C cooperative move, D defective move
In each cell, the left payoff is for the row player, the middle payoff is for the column player, the right payoff is for the outsider (if there is one)

Our game is a standard symmetric two-choices prisoner's dilemma with two active players and a passive player, as in Table 1. If both players cooperate, each of them earns *R€*, and the passive player earns $-h$*€*. If one cooperates and the other defects, the cooperator earns *S€*, while the defector earns *T€*, and the passive player earns $-h$*€*. If both defect, each of them earns *P€*, and the passive player earns *0€*. Following the labels originally introduced by Rapoport and Chammah (1965), *R* stands for "reward", *S* for "sucker", *T* for "temptation" and *P* for "punishment".

We choose the following parameters: *R=5, S=0, T=10, P=2.45*. In the *Baseline* there is no outsider (and thus no harm on passive players is implemented: *h=0*) while we add an outsider in the three *Treatments*. We vary the level of harm. In treatment *Small* the level of harm is *h=.3*; in *Middle h=2.1*; in *High h=4.8*.

III.2. Considerations Motivating the Design

In a stylized way, our game captures a one-shot Bertrand market with constant marginal cost where two firms individually decide whether to set the collusive price *(C)* or to engage in a price war *(D)*. Our introduction of harm on a passive outsider is meant to capture the loss in consumer welfare, and in total welfare, inherent in anticompetitive behavior. If both firms engage in (tacit or explicit) collusion, both set the monopoly price and split the monopoly profit evenly *(R=T/2)*. If only one of them starts a price war, it undercuts the collusive price by the smallest possible decrement. As is standard in the theoretical literature, in this interpretation of our design we assume the decrement to be infinitesimally small. This implies that the aggressive firm cashes in the entire monopoly profit *(T)*, while the firm that is faithful to the cartel receives nothing *(S)*. Therefore, in the experiment, we do not confine harm to the situation where both active players cooperate. Yet if both firms start fighting, they end up in the Nash equilibrium. This removes harm on the opposite market side, and the deadweight loss resulting from the fact that some demand is not served although marginal cost is below marginal willingness to pay.

We deliberately avoid a market frame. This not only makes sure that our results are not driven by the frame. It is also necessary to isolate the effect of externalities. In a market setting, from their world knowledge subjects would know that collusion is illegal and might be motivated by this social and legal norm, rather than by their reticence to impose harm.

Our choice of parameters is primarily driven by experimental concerns. We create the maximum difference between the sucker payoff *S=0* and the temptation payoff *T=10*. That way, both the premium for beating one's opponent and the penalty for losing in competition are largest. By contrast, the payoff in case both players defect *(P)* almost holds the middle between the reward for cooperation *(R)* and the payoff when being the sucker *(S)*. For this payoff, we deliberately have not chosen either extreme. If participants earn *0€* in case both defect, cooperation is no longer strictly dominated. Strictly speaking, the game is no longer a prison-

er's dilemma. At the opposite extreme, the equilibrium is not affected. But if participants earn *5€* in case both defect, gains from cooperation are *0€*. The situation is no longer a dilemma.

In a repeated game, the effects of optimism and reticence to impose harm would be overshadowed by reputation effects. We therefore test our subjects on a one-shot game. That way, we also need not be concerned that players might take turns. There is no room for an equilibrium in iterations.

III.3. Beliefs

After the prisoner's dilemma we elicit beliefs about the cooperativeness of active players in the game. I.e., each player is asked to estimate the number of active players in her session who chose the cooperative move (labeled neutrally). If a participant guesses the number exactly right, she earns an additional 2€. If her estimate is within a range of +/- 2 around the true number, she earns an additional 1€.[2]

III.4. Procedures

Subjects know that the experiment has several parts,[3] but receive specific information about the content of each part only immediately before playing the relevant part. Group composition varies between the parts. No information about other participants' decisions and therefore about any earnings is given to the subjects before the end of the entire experiment, so that independence is preserved. All instructions are read aloud by the experimenter immediately before the relevant part to achieve common knowledge about the procedure.[4]

The experiment was run in February 2014 at the University of Bonn with a computerized interaction using z-Tree (Fischbacher, 2007). ORSEE (Greiner, 2004) was used to invite subjects from a subject pool of approximately 8000 subjects. Each subject participated only in one session. We collected 48 independent observations from active players in the prisoners' dilemma in almost each condition (we have only 44 independent observations in treatment *Middle*[5]). In total 258 subjects participated in the experiment (70 in the role of passive players). In each session of the treatments, at the beginning of the experiment the active and the passive players were randomly picked from the pool of participants present in the laboratory. Subjects were on average 23.54 year old (range 17-55). 56.57% were female. Almost all of them were students, with various majors. Each session lasted about one and a half hours. Participants in each session received a show-up fee of 10€ that suffices to cover potential losses.

2 Subjects were informed on their computer screens how many active players participated in total in their session.

3 We ensure that the number of parts is the same in all treatments to exclude any behavioral changes caused simply by differences in expectations about the duration of the experiment. The data from subsequent stages of the experiment are not relevant for the present paper and are therefore not presented here.

4 See the Appendix for an English translation of the instructions.

5 In two sessions, not all invited participants showed up, so that we could not fill one group of three.

Subjects earned on average 20.61€ (equivalent to $28.02 on the last day of the experiment, range 5.2€-45€).[6]

IV. Predictions

Our game is a one-shot prisoner's dilemma. If participants hold standard preferences they defect, irrespective of their beliefs about the behavior of other active participants, and irrespective of the harm cooperative moves inflict on outsiders.

Empirically, many experimental participants have been found to be conditional cooperators (Fischbacher, Gächter et al., 2001; Fischbacher and Gächter, 2010). Pure conditional cooperators prefer cooperation over defection if they expect their counterpart to cooperate with certainty. This implies that they resist the temptation to exploit their counterpart. In line with previous experiments, we expect conditional cooperation to be more prevalent than outright selfishness. Yet we expect participants to be less than perfectly optimistic. Then conditional cooperators run the risk of not getting gains from cooperation. If they are neutral to risk and losses, they compare the expected utility from cooperation with the expected utility from defection. In which ways this comparison is moderated by the fact that cooperation inflicts harm on the outsider depends on the motive that drives (conditional) cooperation.

Two otherwise plausible motives are not meaningful in our setting. We implement a one-shot game, which is why the prospect of future gains cannot sustain cooperation (cf. Kreps, Milgrom et al., 1982). Outsiders are passive, which is why active participants cannot reciprocate their expected kindness (cf. Rabin, 1993; Charness and Haruvy, 2002). Consequently possible treatment effects cannot be explained in terms of reciprocity. We discuss three reasonable motives for lower cooperation in the *Treatments* than in the *Baseline*: utility from increasing efficiency, disutility from inequity aversion, and disutility from guilt aversion. In this section, we explain the intuition behind our theoretical predictions. For a formal approach please see the Appendix. As we will explain in the following paragraphs, for each of these motives beliefs about the cooperativeness of active co-players play an important role. Thus, we derive predictions, based on the three competing motives, about (1) how the degree of harm on a passive outsider influences own cooperativeness; (2) how optimism about the cooperation of co-players is correlated with own cooperativeness, and (3) how both effects interact.

If conditional cooperation is driven by efficiency motives, as advocated by Charness and Haruvy (2002) and Engelmann and Strobel (2004), cooperators derive extra utility from the entire society's total payoff. The larger the harm for the outsider, the smaller society's payoff

6 The average payoff was 23.25€ in the *Baseline*. In treatment *Small* subjects earned on average 20.70€ (15.95€ for passive players), in *Middle* 20.34€ (15.06€ for passive players), and in *High* 19.02€ (15.30€ for passive players).

if a player cooperates. We therefore expect less cooperation the larger the harm.[7] In the *Baseline,* the more a player is optimistic that the other player will cooperate, the more she is likely to gain the temptation payoff of 10 units for herself if she defects. Since in our game, $T+S=2R$, there is no efficiency loss either, provided the other active player indeed cooperates. We therefore also expect less cooperation the more an efficiency minded participant is optimistic. Yet the more pronounced the harm, the smaller the efficiency gap between cooperation and defection. In efficiency terms, optimism is less important for the choice between cooperation and defection if harm can only be avoided by joint defection.

Thus, if efficiency seeking drives conditional cooperation this leads to:

H1a: There is less cooperation the more severe the harm.

H2a: The more a participant is optimistic, the less she cooperates in the Baseline.

H3a: The more severe the harm, the less optimism is critical for the decision to defect.[8]

Conditional cooperation may also follow from inequity-aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). If a participant is perfectly optimistic about the cooperativeness of others, she runs no risk of being exploited herself. Her aversion against disadvantageous inequity is immaterial. She cooperates if she is sufficiently averse against being an exploiter herself. Her degree of aversion against advantageous inequity is critical for her decision to cooperate. If she is less than perfectly optimistic, she has to balance potential disutility from being exploited (since her counterpart defects) against disutility from exploiting her (since her counterpart cooperates). Consequently if inequity aversion drives conditional cooperation, we expect a positive main effect of beliefs.

The difference in payoffs between the two active players is not affected by the level of harm. Thus, inequity aversion towards the other active player does not lead to a treatment main effect. Yet arguably, active players also experience inequity aversion vis-à-vis the passive bystander. By our choice of parameters, an active participant is always better off than the passive participant, and the difference is the larger the more severe the harm. If the other active participant cooperates, the passive participant suffers harm with certainty. When she decides whether to cooperate or to defect herself, the inequity with respect to the bystander cancels out. The active player's choice is only critical for the bystander's harm if the other player defects. It is the larger the more severe the harm. Hence we predict a negative main effect of *Treatment*, the more so the more pronounced the harm, and therefore the additional advantageous inequity. The more an active player expects the other active player to cooperate any-

7 Note, however, that, with the parameters of the experiment, defection is never efficient. Even with $h = 4.8$, the society loses a higher amount. If at least one player cooperates, the society earns $(2R = S + T = 10) - 4.8 = 5.2$. If both players defect, the society earns $2P = 4.9$. The hypothesis thus requires that active players are less motivated to cooperate if efficiency gains are smaller, even if they remain positive.

8 Note that mathematically this statement is identical to **H3b** and **H3c1**. We just use different words because they are more intuitive with efficiency seeking.

way, the less her own cooperative choice is likely to be the cause of harm to the outsider. Therefore the more severe the harm, the bigger the effect of optimism.

Thus, if inequity-aversion drives conditional cooperation, we expect:

H_{1b}: There is less cooperation the more severe the harm.

H_{2b}: The more a participant is optimistic, the more she cooperates in the *Baseline*.

H_{3b}: The more severe the harm, the more optimism is critical for the decision to cooperate.

A third motive that may support conditional cooperation is guilt aversion (Battigalli and Dufwenberg, 2007; Dufwenberg, Gächter et al., 2011). Guilt aversion reacts to factual or normative beliefs. Guilt is the more pronounced the more one disappoints the expectations of one's peers. We assume that individuals experience guilt if they let down another active player who cooperates herself. This is the more likely the more they expect the other active participant to cooperate. We therefore expect a positive main effect of beliefs.

There are different ways how guilt aversion can be modeled in the *Treatments*. The first option is for the active participant to experience guilt with respect to the outsider whenever she makes a cooperative move herself: irrespective of the decision of her co-player the outsider suffers harm. With this specification of guilt, we expect less cooperation the more severe the harm: the larger the harm, the more pronounced the disutility inherent in a cooperative move. Yet with this definition of guilt aversion we do not expect an interaction between the level of harm and the degree of optimism: any cooperative move increases guilt, whether or not the other active participant cooperates.

Alternatively, the active player might only feel guilty if she is responsible for the harm. This is only the case if she alone has made the cooperative move. We again expect the negative main effect of *Treatment*: the more pronounced the harm, the more disutility the participant suffers if she is the only one to cooperate. We further expect the interaction between optimism and harm to be positive. The more the participant believes that her counterpart will cooperate anyway, the less she is likely to suffer additional guilt from being responsible for harming the outsider.

Contrast this with a third definition of guilt. Now the active participant experiences additional guilt if she not only has made a cooperative move, and thereby inflicted harm on the outsider, but if she also has benefitted from this action, since the other active player has cooperated as well. With this definition of guilt, we again expect a negative main effect of *Treatment*. The reason is the same as with guilt resulting from a unilateral cooperative move. Even if the active participant is alone to cooperate, the outsider still suffers. The main effect of optimism now not only depends on the general strength of guilt aversion but, additionally, on the ratio between general guilt aversion and aversion against jointly exploiting the outsider. With this

specification of guilt, the interaction effect changes sign. The more a participant expects the other active participants to cooperate, the less she will be inclined to cooperate herself. This is the only way how to avoid the additional guilt disutility from jointly taking advantage of the passive outsider.

Depending on the specification of guilt aversion, we thus predict:

H_{1c}: There is less cooperation the more severe the harm.

H_{2c}: The more a participant is optimistic, the more she cooperates in the *Baseline*.

H_{3c1}: The more severe the harm, the more optimism is critical for the decision to cooperate.

or

H_{3c2}: The more severe the harm, the less optimism is critical for the decision to cooperate.

Table 3 sums up predictions. With standard preferences there should be no treatment difference and beliefs are immaterial, irrespective of the level of harm. If (a sufficient fraction of) active participants are conditional cooperators, there is less cooperation the more severe the harm. This treatment main effect is to be expected whatever the motive that drives conditional cooperation. If conditional cooperation follows from efficiency seeking, the more a participant is optimistic, the less she cooperates. If conditional cooperation follows from inequity aversion or guilt aversion, the more a participant is optimistic, the more she cooperates. If participants are motivated by efficiency seeking, inequity aversion or guilt aversion from being responsible for harm, in the *Treatments* they cooperate the more the more they expect their counterpart to cooperate as well. If subjects experience guilt whenever they make a cooperative move and thereby harm the outsider, they are neither sensitive to beliefs nor to the harm they inflict on an outsider. If conditional cooperation results from guilt aversion and is most pronounced if the two active participants jointly take advantage of the passive outsider, in the *Treatments* they cooperate *less* the more they expect their counterpart to cooperate as well.

V. Results

In this section, we report the data and test our hypotheses. In particular, we examine which of the motives discussed in the previous section explains our results.

Figure 1 collects choices in the *Baseline* and the *Treatments*. The degree of cooperation monotonically decreases in the harm inflicted on a passive outsider. Yet, non-parametrically we do not find a significant difference between the *Baseline* and the *Small* (h=.3€) treatment (chi^2: $p = 0.149$). By contrast, the difference between the *Baseline* and the two remaining treatments

is significant (chi^2: $p \leq 0.003$)[9]. With this qualification we support hypotheses $\mathbf{H_{1a}}$, $\mathbf{H_{1b}}$ and $\mathbf{H_{1c}}$ that expected harm to outsiders to dampen cooperation. This is a first hint against choices being driven by standard preferences.

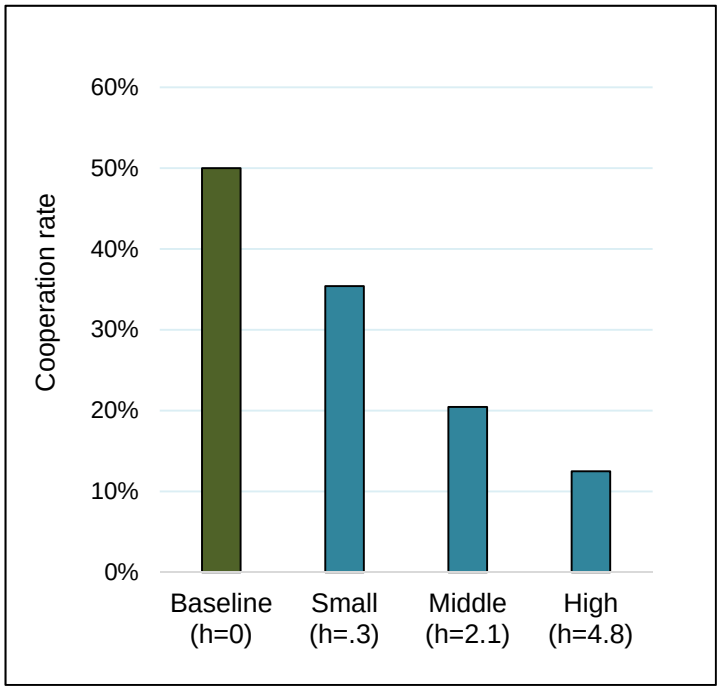

Figure 1

Degree of Cooperation in Prisoner's Dilemma with Harm

On the vertical axis, one can see the cooperation rate in percent of active players. The harm imposed on outsider (in €) is presented on the horizontal axis. The scale goes from 0 (green bar: *Baseline*) to 4.8 (blue bars: treatments with cooperation leading to harm).

If we run a linear probability model (LPM)[10] (Table 3, Model 1), we replicate the non-parametric result, i.e., we do not find a negative significant effect on cooperation rates in *Small* (h=.3€), but we do find this effect in *Middle* (h=2.1€) and in *High* (h=4.8€). Table 3, Model 2 establishes a significant positive main effect of beliefs. We thus support hypotheses $\mathbf{H_{2b}}$ and $\mathbf{H_{2c}}$. This is in line with choices based on inequity aversion or guilt aversion. It is further evidence against choices based on standard preferences, and even more so against choices based on efficiency seeking.

9 Comparing the treatments with each other, the only statistically significant difference is between *Small* and *High* (*Small* vs. *Middle*: chi2: p = 0.111; *Middle* vs. *High*: chi2: p = 0.302; *Small* vs. *High*: chi2: p = 0.009).

10 We estimate linear probability models since we are mainly interested in the interaction of treatment (defined by specific levels of harm on the passive outsider) and beliefs; with a non-linear logit or probit model, interaction terms could not be interpreted directly (Ai and Norton, 2003).

Table 3
Explaining cooperation rate – comparison Baseline and Treatments (h=.3; 2.1; 4.8)

linear probability model (LPM)

Dependent variable: cooperation rate

	Model 1	Model 2	Model 3	Model 4
Small (.3€)	-.146	-.079	.157*	
	(.101)	(.076)	(.093)	
Middle (2.1€)	-.296***	-.211***	.026	
	(.095)	(.074)	(.115)	
High (4.8€)	-.375***	-.324***	.132	
	(.087)	(.070)	(.085)	
Harm				.010
				(.015)
Belief		.971***	1.506***	1.312***
		(.093)	(.118)	(.090)
*Small (.3€)** Belief			-.468***	
			(.178)	
*Middle (2.1€)** Belief			-.467*	
			(.272)	
*High (4.8€)** Belief			-.966***	
			(.216)	
Harm*Belief				-.160***
				(.043)
Constant	.500***	.0189	-.247***	-.179***
	(.073)	(.067)	(.070)	(.045)
N	188	188	188	188
P model	<.001	<.001	<.0001	<.0001
R^2	0.1000	0.4303	.4769	.4660

Linear Probability Model. Robust standard errors are presented in parentheses. *Small*, *Middle* and *High* are treatment dummies that equal 1 for observations in these treatments. *Baseline* is the reference category. Significance at the 10%, 5% and 1% by *, ** and ***.

In the final step (Model 3), we interact treatments with beliefs. For all three treatments, we find a significant negative interaction[11] We also find a highly significant negative interaction effect if we interpret harm as a continuous variable (Model 4). We thus support hypothesis $\mathbf{H_{3c2}}$ and reject the competing hypothesis $\mathbf{H_{3c1}}$ as well as $\mathbf{H_{1c}}$ and $\mathbf{H_{2c}}$. This effect is only predicted by guilt aversion, and only if participants experience particularly pronounced guilt if they jointly take advantage of the outsider. Figure 2 collects average marginal effects of treatment conditional on the individual degree of optimism. As one sees, irrespective of the severity of harm, the treatment effects are driven by optimistic participants. While such participants are very inclined to cooperate if this has no adverse effects on outsiders, their sensitivity with respect to optimism is substantially dampened if the price for cooperation is harming

11 With harm of 2.1 €, it is only weakly significant, p = .087.

an outsider. The negative externality not only reduces unconditional but also conditional cooperation.

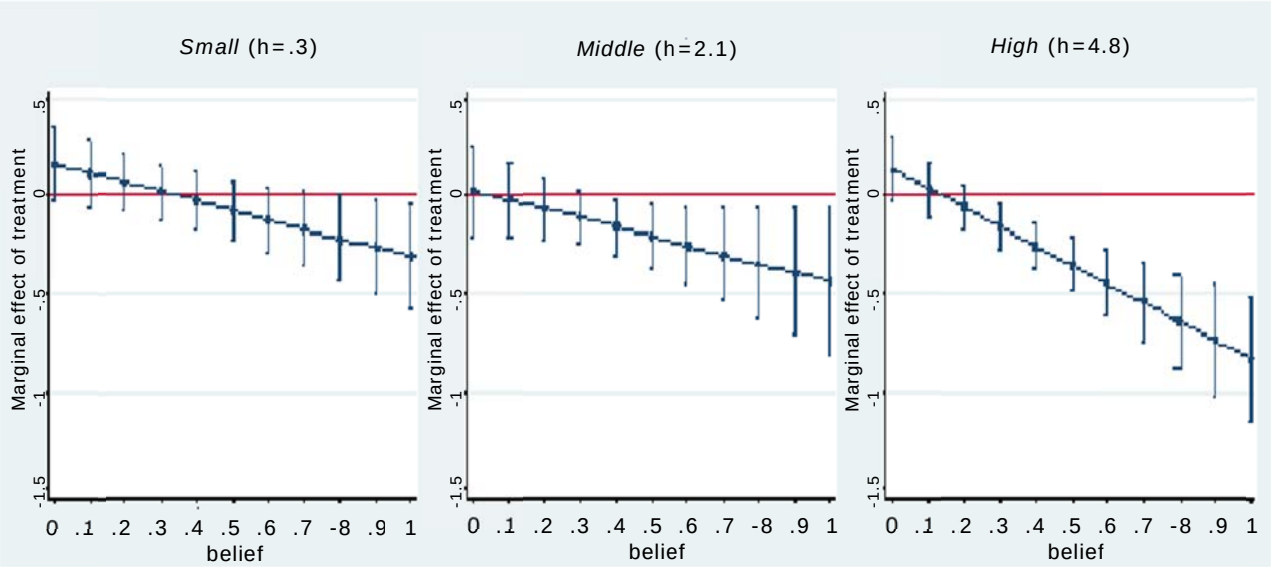

Figure 2

Marginal Effects of Treatments Conditional on Optimism

On the horizontal axis, one can see the expected cooperation rate of active players in percent (belief). The marginal effect of treatments on the cooperation rate (Table 3, Model 3 with level = 0, i.e., *Baseline* as the reference category) is presented on the vertical axis.

We therefore conclude

> **Result:** If making a cooperative move in a simultaneous symmetric two-person two-action prisoner's dilemma imposes harm on a passive outsider, active participants are averse against guilt resulting from jointly taking advantage of the outsider.

VI. Conclusions

Many dilemma situations that individuals face in real life are embedded in a wider social context. Not so rarely, if individuals overcome their dilemma, outsiders suffer. Often these outsiders cannot protect themselves. The canonical illustration of this situation is oligopoly. For firms, collusion is a dilemma. If all firms collude, all gain their share of monopoly profit. Yet each individual firm is best off if all others set the collusive price while this firm undercuts and reaps the entire profit. Now if all firms resist this temptation, they are all better off. But by this very fact they exploit the opposite market side, and society suffers from a deadweight loss.

We have studied such an embedded dilemma in the laboratory. If harm to the outsider is small, we do not find a significant reduction of cooperation, compared with the same dilemma with no outsider. Yet with more severe harm, the difference is significant. We find that coop-

eration is less likely in a symmetric one-shot-two-person prisoner's dilemma if cooperation inflicts harm on a passive outsider. More importantly, we find that active subjects cooperate more the more they expect others to cooperate but that this positive correlation is the less pronounced the more severe the resulting harm on the outsider.

One should always be cautious when extrapolating from the choices of students in a context-free laboratory to analogous situations in the field. That said our findings suggest that those who have a chance to act are sensitive to the fact that a local public good could be a global public bad. Knowing that internal cooperation entails external harm reduces their willingness to take action. From a policy perspective, the negative interaction effect is even more important. In the field, those who decide to contribute their fair share to a public good are next to never sure about the actions of others who decide simultaneously. Our results suggest: the more pronounced this strategic uncertainty, the more insiders become sensitive to the risk of harming outsiders. If policy makers are concerned about the global, rather than the local public good, heavy-handed intervention may not be necessary to protect those who have no chance to protect themselves.

This result also contributes to the understanding of conditional cooperation in general. Since we study one-shot experiments, cooperation cannot be motivated by the shadow of the future. Since the outsiders are passive by design, a treatment main effect of cooperation cannot be motivated by reciprocity. The fact that we find a positive main effect of beliefs is evidence against efficiency seeking (which predicts a negative effect of beliefs). Yet the most important indicator is the robust negative interaction between harm and beliefs. It is only predicted by one specific version of guilt aversion: participants experience disutility from letting down the other active player and from harming the passive player; but their disutility is even more pronounced if they jointly take advantage of the passive participant. Since guilt aversion is the only plausible candidate for rationalizing our findings, our experiment suggests that guilt aversion is the most plausible candidate for explaining conditional cooperation in general. In the future, it would be very interesting to develop experimental designs that test this motive even if there is no outsider.

References

ABBINK, KLAUS (2005). Fair Salaries and the Moral Costs of Corruption. *Advances in Cognitive Economics*. B. N. Kokinov. Sofia, NBU Press**.**

AI, CHUNRONG AND EDWARD C. NORTON (2003). "Interaction Terms in Logit and Probit Models." *Economics Letters* **80**: 123-129.

ANDREONI, JAMES AND JOHN MILLER (1993). "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma. Experimental Evidence." *Economic Journal* **103**(418): 570-585.

BATTIGALLI, PIERPAOLO AND MARTIN DUFWENBERG (2007). "Guilt in Games." *American Economic Review* **97**(2): 170-176.

BOLTON, GARY E. AND AXEL OCKENFELS (2000). "ERC: A Theory of Equity, Reciprocity and Competition." *American Economic Review* **90**: 166-193.

BOLTON, GARY E. AND AXEL OCKENFELS (2010). "Betrayal Aversion. Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. Comment." *American Economic Review* **100**: 628-633.

CHARNESS, GARY AND ERNAN HARUVY (2002). "Altruism, Equity, and Reciprocity in a Gift-Exchange Experiment. An Encompassing Approach." *Games and Economic Behavior* **40**: 203-231.

DUFWENBERG, MARTIN, SIMON GÄCHTER AND HEIKE HENNIG-SCHMIDT (2011). "The Framing of Games and the Psychology of Play." *Games and Economic Behavior* **73**: 459-478.

ELLMAN, MATTHEW AND PAUL PEZANIS-CHRISTOU (2010). "Organisational Structure, Communication and Group Ethics." *American Economic Review* **100**(5): 2478-91.

ENGEL, CHRISTOPH (2007). "How Much Collusion? A Meta-Analysis on Oligopoly Experiments." *Journal of Competition Law and Economics* **3**: 491-549.

ENGEL, CHRISTOPH AND BETTINA ROCKENBACH (2011). We Are Not Alone. The Impact of Externalities on Public Good Provision, MPI Collective Goods Preprint No. 29.

ENGELMANN, DIRK AND MARTIN STROBEL (2004). "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments." *American Economic Review* **94**: 857-869.

FEHR, ERNST AND KLAUS M. SCHMIDT (1999). "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* **114**: 817-868.

FISCHBACHER, URS AND SIMON GÄCHTER (2010). "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Good Experiments." *American Economic Review* **100**: 541-556.

FISCHBACHER, URS, SIMON GÄCHTER AND ERNST FEHR (2001). "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment." *Economics Letters* **71**: 397-404.

GREINER, BEN (2004) "An Online Recruitment System for Economic Experiments". In K. Kremer, and V. Macho (eds.), *Forschung und Wissenschaftliches Rechnen*. Gesellschaft für Wissenschaftliche Datenverarbeitung Bericht, Göttingen: Datenverarbeitung **63**: 79–93.GÜTH,

WERNER AND ERIC VAN DAMME (1998). "Information, Strategic Behavior, and Fairness in Ultimatum Bargaining. An Experimental Study." *Journal of Mathematical Psychology* **42**: 227-247.

KESER, CLAUDIA AND FRANS VAN WINDEN (2000). "Conditional Cooperation and Voluntary Contributions to Public Goods." *Scandinavian Journal of Economics* **102**: 23-39.

KOCHER, MARTIN, TODD L. CHERRY, STEPHAN KROLL, ROBERT J. NETZER AND MATTHIAS SUTTER (2008). "Conditional Cooperation on Three Continents." *Economics Letters* **101**(3): 175-178.

KREPS, DAVID M., PAUL R. MILGROM, JOHN ROBERTS AND ROBERT B. WILSON (1982). "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma." *Journal of Economic Theory* **27**: 245-252.

OSTROM, ELINOR (1999). "Coping with the Tragedies of the Commons." *Annual Review of Political Science* **2**: 493-535.

OSTROM, ELINOR, THOMAS DIETZ, NIVES DOLSAK, PAUL C. STERN, SUSAN STONICH AND ELKE U. WEBER, Eds. (2002). *The Drama of the Commons*. Washington, National Academy Press.

POUNDSTONE, WILLIAM (1992). *Prisoner's Dilemma*. New York, Doubleday.

RABIN, MATTHEW (1993). "Incorporating Fairness into Game Theory and Economics." *American Economic Review* **83**: 1281-1302.

RAPOPORT, ANATOL AND ALBERT M. CHAMMAH (1965). *Prisoner's Dilemma. A Study in Conflict and Cooperation*. Ann Arbor, University of Michigan Press.

Appendix

Appendix 1. Theoretical Predictions

If participants hold standard preference, they only care about their payoff and maximize

$\pi = pR + (1-p)S - pT + (1-p)P$	(1)

By the definition of the dilemma $T > R, P > S$. Cooperation is dominated. Participants are not sensitive to the expected size of p. The amount of harm resulting from making a cooperative move does not enter their utility, and is therefore also immaterial.

If participants have positive utility from efficiency, their utility is given by (2).

$u = p(R+(2R-h)e)+(1-p)(S+(T+S-h)e)$ $-p(T+(T+S-h)e)-(1-p)(P+2Pe)$	(2)

where $e > 0$ captures the strength of the efficiency motive. The first derivative wrt h is given by $-(1-p)e$ which is negative whenever $p < 1$. Harm has a negative effect on cooperation. The first derivative wrt p is given by

$\frac{\partial u}{\partial p} = R+(2R-h)e-(S+(T+S-h)e)-(T+(T+S-h)e)+(P+2Pe)$	(3)

which, with the parameters of the experiment, boils down to $-2.55 - 5.1e + eh$. This is negative, given $h \leq 4.8$. The cross-derivative of (3) wrt h is $e > 0$.

If participants are inequity averse, utility is defined by (4)

$u = p(R-(R+h)\beta)+(1-p)(S-(T-S)\alpha-(S+h)\beta)$ $-p(T-(T-S+T+h)\beta)-(1-p)(P-P\beta)$	(4)

The first derivative wrt h is $(1-p)\beta$ which is negative for all $p < 1$. The first derivative wrt p is given by (5)

$\frac{\partial u}{\partial p} = R-(R+h)\beta-(S-(T-S)\alpha-(S+h)\beta)-(T-(T-S+T+h)\beta)+(P-P\beta)$	(5)

which, with the parameters of the experiment, boils down to $10\alpha + (12.55+h)\beta - 2.55$. This term is positive for sufficiently large α and/or β. The first derivative of (5) wrt h is simply $\beta > 0$.

If participants experience guilt whenever they let down an insider or an outsider, their utility follows from (6)

$u = p(R-\gamma h)+(1-p)(S-\gamma h)-p(T-(R\text{-}S)\gamma)-(1-p)P$	(6)

where $\gamma > 0$ captures the strength of guilt aversion. The first derivative wrt h is $-\gamma < 0$. The first derivative wrt p is given by $R - S - T + \gamma(R - S) + P$. With the parameters of the experiment this is $5\gamma - 2.55$, which is positive provided $\gamma > .51$. Since the first derivative wrt p does not depend on h, the cross-derivative is 0.

If participants only experience guilt wrt to the outsider if they are personally responsible for the harm, (6) is replaced by (7)

$u = pR+(1-p)(S-\gamma h)-p(T-(R\text{-}S)\gamma)-(1-p)P$	(7)

The first derivative wrt h is $-\gamma + p\gamma$ which is negative for all $p < 1$. The first derivative wrt p is given by $R + \gamma h - S - T + \gamma(R - S) + P$ or, with the parameters of the experiment, $(5 + h)\gamma - 2.55$, which defines under which conditions the effect of optimism is positive. The cross derivative of this term wrt h is $\gamma > 0$.

Finally if participants feel more guilt for jointly taking advantage of the outsider, their utility is given by (8)

$u = p(R-\delta\gamma h)+(1-p)(S-\gamma h)-p(T-(R\text{-}S)\gamma)-(1-p)P$	(8)

where $\delta > 1$ captures this qualification of guilt. The first derivative wrt h is $p(\gamma - \delta g) - \gamma$. This is negative irrespective of the size of p, given $\delta > 1$. The first derivative wrt p is given by $-\delta\gamma h + R + \gamma h - S - T + \gamma(R - S) + P$. With the parameters of the experiment this is $5\gamma + (1 - \delta)\gamma h - 2.55$. The cross derivative wrt h is $(1 - \delta)\gamma < 0$, since we assume $\delta > 1$.

Appendix 2. Instructions

In each session subjects participated in an experiment, which included 4 parts. Parts 2, 3 and 4 were post-experimental tests. We use neither test for this paper. Recall that subjects were not aware of the content of the subsequent parts when making their choices in the first part (the prisoners' dilemma). We only analyze subjects' behavior in Part 1 of the experiment. Therefore, we report first the full instructions of the *Baseline* and afterwards only Part 1 of the treatment *Small*. The difference between *Small*, *Middle* and *High* lies only in the level of harm for the passive player.

2.1. Baseline

Welcome to our experiment. Please remain quiet and do not talk to the other participants during the experiment. If you have any questions, please give us a signal. We will answer your queries individually.

Course of Events

The experiment is divided into four parts. We will distribute separate instructions for each of the four parts of the experiment. Please read these instructions carefully and make your decisions only after taking an appropriate amount of time to reflect on the situations, and after we have fully answered any questions you may have. Only when all participants have decided will we move on to the next part of the experiment. All of your decisions will be treated anonymously.

Your Payoff

At the end of the experiment, we will give you your payoff in cash. Each of you will receive the earnings resulting from the decisions you will have made in the course of the experiment. It is possible to make a loss in one part of the experiment. These losses will be subtracted from the earnings in the other parts and from your show-up fee.

Thus:

Total payment =

+ Earnings from Part 1

+ Earnings from Part 1a

+ Earnings from Part 2

+ Earnings from Part 3

+ Earnings from Part 4

+ 10 €

In Part 2, however, losses are possible, too. Should you incur losses, these will be deducted from your earnings from Part 1, Part 3, or Part 4 and from your show-up fee of 10€.

We will explain the details of how your payoff is made up for each of the four parts separately. In each of the four parts, possible payoffs are given in Euro, which is the currency you will be paid in.

Part 1

The basic idea of this part of the experiment is as follows: you are anonymously paired by us with another participant. You and the other participant will make one decision.

We will show you one tables that look as follows:

		Type B	
		Above	*Below*
Type A	*Above*	5€, 5€	0€, 10€
	Below	10€, 0€	2.45€, 2.45€

We will let you know at the start whether you are a Type A or a Type B participant. (You will probably notice that the payments given to both types are symmetrical; the distinction between Type A and Type B is solely for the purpose of explaining the experiment.)

The decisions *Above* or *Below* determine the payoffs to you and the other participant. In each of the four cells of the table, the figure on the left denotes A's profit, while the figure on the right denotes B's profit.

For instance, if Type A chooses the option *Above* and Type B chooses the option *Above*, then both receive a payment of 5€. If Type A chooses *Above* and Type B chooses *Below*, then Type A receives zero profit and Type B gets 10€. The same is valid for a *Below*/*Above* constellation. Finally, if Type A chooses *Below* and Type B chooses *Below*, then both receive a payment of 2.45€.

Let us first begin with some test questions. (The aim of these questions is merely to verify whether all participants have fully understood the instructions. Neither the questions nor the answers have anything to do with your final payment.) Then the screen on which your actual decisions are marked will appear.

Do you have any further questions?

Part 1a

This part of the experiment refers to the previous part where you made eleven decisions, "Above" or "Below". The number of participants in the roles A and B who participated in this task will be presented to you on the screen. We ask you to estimate how many participants of the experiment selected "Above". In case you make a precise estimation, you can gain 2€ in addition. If your estimation deviates by +/-2, you still gain 1€ in addition. Otherwise, you gain nothing in addition.

Part 2

The basic idea of this part of the experiment is as follows. In the following, you will be requested to make six decisions. In this part of the experiment, no other participant is paired with you. The payoffs therefore relate only to you. In each of your six decisions, you may therefore choose to play a "lottery" or decline.

What are these "lotteries" then? In these lotteries, a computer-simulated random toss of a coin determines whether you win or lose money. If the coin shows "tails" (i.e., a number), you win 6€; if it is "heads", you lose. How much you lose depends on the particular lottery. Losses vary between 2€ and 7€. If losses occur, they are subtracted from the earnings from the other parts of the experiment at the end of the experiment.

You can accept or refuse these lotteries on an individual basis, just as you can accept or refuse all. If you refuse, you will make no profit and lose nothing, i.e., your payoff will be zero. If you accept, the toss of the coin determines your payoff, as described above.

In the end, one of the six lotteries is randomly chosen, and then the payment is determined according to your decision and the coin throw for this particular lottery. Thus, once again the lot decides twice in a row: first, one of the lotteries is drawn by lot, and then the toss of a coin decides whether or not you win in this lottery – on condition that you have decided to go for the lottery.

Let us first begin with some test questions. (The aim of these questions is merely to verify whether all participants have fully understood the instructions. Neither the questions nor the answers have anything to do with your final payment.) Then the screen on which your actual decisions are marked will appear.

Part 3

This part of the experiment is as follows: one Type X participant has to decide between two situations (1 or 2). His decision influences his own payoff, and the payoff of one other randomly paired Type Y participant, as follows:

Situation 1: Type X receives a payoff, determined by lot, of 5€ or 10€, Type Y receives a payoff of zero Euro. The likelihood with which Type X either receives 5€ or 10€ is systematically varied in the following table. Type X must make a decision for each of the eleven constellations (a total of 11 decisions).

Situation 2 remains the same for all 11 constellations: Type X and Type Y both receive 5€.

In this part, all participants must initially make their decisions in the role of Type X.

We will proceed with the payoff as follows:

- The lot is drawn to determine whether your payments, following your own decisions, classify you as a Type X or a (passive) Type Y. We will draw one half of the group as Type X and the other as Type Y.
- The next draw pairs each Type Y participant with a Type X participant.
- Finally, the third draw determines one single payoff-relevant situation out of the total of eleven situations.

Therefore, one out of the eleven decisions emerges as the basis for payoff. With a probability of ½, it will be your own decision, and with the same likelihood it will be another participant's decision.

Example for Part 3

	Profit	With likelihood of
You	10€	30%
	5€	70%
Other participant	0€	100%
	1	
Your decision		
	2	
Both	5€	100%

As stated above, all participants will make eleven decisions of this kind. Please mark your decision by clicking on the appropriate box.

Part 4

In this part of the experiment, no other participant is paired with you. The payoffs therefore relate only to you. The decisions of the other participants only have an influence on their own respective payoffs.

In this part of the experiment, you are asked to decide in 10 different situations (lotteries) between option A and B. These situations will be presented to you on consecutive screens. The two lotteries each comprise 2 possible monetary payoffs, one high and one low, which will be paid to you with different probabilities.

The options A and B will be presented to you on the screen as in the following example:

The computer uses a random draw program, which assigns you payments exactly according to the denoted probabilities.

For the above example, this means:

Option A obtains a payoff of 2 Euro with a probability of 10% and a payoff of 1.60 Euro with a probability of 90%.

Option B obtains a payoff of 3.85 Euro with a probability of 10% and a payoff of 0.10 Euro with a probability of 90%.

Now you have to click on the particular option you decide for.

Please note that at the end of the experiment only one of the 10 situations will eventually be paid. Yet, each of the situations can be randomly chosen with equal probability to be the payoff-relevant one.

After this, a draw will determine whether for the payoff-relevant situation the high payoff (2.00 Euro or 3.85 Euro) or the low payoff (1.60 Euro or 0.10 Euro) will be paid.

2.2. Treatment Small

Part 1

The basic idea of this part of the experiment is as follows: you are anonymously paired by us with two other participants. There exist Type A, Type B and Type C players. Type C is passive in that experiment. If you are not Type C, you and one other participant will make one decision.

We will show you eleven tables that look as follows:

		Type B	
		Above	Below
Type A	Above	5€, 5€, -.3€	0€, 10€, -.3€
	Below	10€, 0€, -.3€	2.45€, 2.45€, 0€

We will let you know at the start whether you are a Type A or a Type B participant. (You will probably notice that the payments given to both types are symmetrical; the distinction between Type A and Type B is solely for the purpose of explaining the experiment.)

The decisions Above or Below determine the payoffs to you and the other participants. In each of the four cells of the table, the figure on the left denotes A's profit, while the figure on the right denotes B's profit. Type C receives either –.3€ or 0€, depending on the decisions of Type A and B.

For instance, if Type A chooses the option Above and Type B chooses the option Above, then both receive a payment of 5€ and Type C receives –.3€. If Type A chooses Above and Type B chooses Below, then Type A receives zero profit, Type B gets 10€, and Type C receives –.3€. The same is valid for a Below/Above constellation. Finally, if Type A chooses Below and Type B chooses Below, then both receive a payment of 2.45€ and Type C receives 0€.

Let us first begin with some test questions. (The aim of these questions is merely to verify whether all participants have fully understood the instructions. Neither the questions nor the answers have anything to do with your final payment.) Then the screen on which your actual decisions are marked will appear.