The Game Prisoners Really Play: Preference Elicitation and the Impact of Communication

Michael Kosfeld
University of Zurich

Ernst Fehr
University of Zurich

Jörgen W. Weibull
Boston University

October 10, 2003
Unfinished version: Please do not distribute!

Abstract

To be written.

1 Introduction

The Prisoners' Dilemma (PD) is surely the classic among all games studied in the social and behavioral sciences. Created in 1950 by Al Tucker to illustrate the possible non-social-desirability of Nash equilibrium, it has since then inspired a vast literature in fields as diverse as biology, economics, philosophy, political science, psychology, and sociology.

In its original form the PD describes the interaction between two prisoners, each of whom has to decide whether to confess or not confess. If no one confesses, both go to jail for a short period of time; if only one confesses and the other does not, the confessor goes free and the other goes to jail for a long time; if both confess, both go to jail for an intermediate period of time. *Assuming that each prisoner minimizes the time he has to spend in jail*, the game has a unique Pareto-inferior Nash equilibrium in dominant strategies, where both prisoners confess.

Since its creation the dilemma of the prisoners who cannot succeed in cooperating (i.e., not confessing), has served as a game-theoretic model for literally hundreds of applications capturing different forms of economic and social cooperation. At the same time, it has been the object of intensive empirical, in particular experimental investigation, as well. The typical finding in these experiments concerning the one-shot game stands in stark contrast to the theoretical prediction given before. While the actual level of cooperation reacts readily to various kinds of experimental manipulation, it is bounded well above zero percent (Reference).

How do we explain this conflict between the game theoretic analysis and the empirical observation? Before we come up with an answer we should be aware of two important facts. The first is that in the formal definition of a *game*, the preferences of the players constitute an essential element. The other integral parts are the set of players, the players' strategies and information sets, and the outcomes or consequences that are determined by the different combinations of strategies of the players. These latter parts constitute the so-called *game form*, or *game protocol* that together with the players' preferences make up the game. The second important fact is that in an experiment, it is relatively easy to implement a game protocol but it is much harder to implement a game. The latter requires that the experimenter knows, or is able to control, the preferences of the players. Until recently, the common solution to this problem — at least in economic experiments — is to use monetary rewards and assume that players' preferences are equivalent to the maximization of individual monetary rewards. Whether this assumption is correct, however, is an empirical question and by now, a large number of results suggest that indeed, in many situations individuals may have preferences that are *not* equivalent to the maximization of individual monetary rewards (Reference). Note, however, that most of these findings offer only an indication, rather than clear evidence, for the existence of alternative preferences. The reason is that in these studies subjects do not reveal their complete preferences but only their behavior in some of all possible game situations, e.g., in the situation when the opponent cooperates.[1] In consequence, observations can sometimes equally well be explained by players having wrong beliefs or exhibiting certain features of boundedly-rational behavior (Example).

The purpose of this study is to shed some new light on this issue. Our main goal is to elicit individuals' *true* preferences in the one-shot PD protocol. There have been earlier attempts to address this problem, all of them relying on questionnaire methods where individuals are asked to rank the different outcome in the PD protocol (Watabe *et al.* 1996, Kollock 1997, Gibbons and Van Boven 2001). We deviate from this methodology and take an approach that is more in line with an old economic tradition of regarding preferences as being something that is revealed by behavior (Samuelson 1938). The key methodological idea of our approach is that, if we see an individual choosing outcome ω from a set of outcomes $\{\omega, \tilde{\omega}\}$, we can interpret this as revealing the individual's preference $\omega \succ \tilde{\omega}$. By applying a variant of the so-called strategy method (Selten 1967) to the PD protocol we are able to elicit sufficient information about players' preferences to construct the *true PD game*.

Our second goal is to use our method of preference elicitation to analyze the impact of communication in the PD protocol. (...)

[1] An exception is Fischbacher *et al.* (2001).

2 Preference Elicitation

2.1 General Approach

In the classic two-player Prisoners' Dilemma protocol two players simultaneously and independently of each other have to decide between two pure strategies: to "cooperate" (C) or to "defect" (D). A combination of these strategies leads to a unique outcome determining the monetary consequences for both players. Let us denote by $\Omega = \{CC, CD, DC, DD\}$ the set of outcomes in the PD protocol, where we first denote the strategy chosen by player 1 and second the strategy chosen by player 2. By definition, the payoff function of player 1, mapping outcomes into monetary consequences, satisfies the condition $\pi_1(DC) > \pi_1(CC) > \pi_1(DD) > \pi_1(CD)$. Interchanging the strategies of the two players the same condition holds for the payoff function of player 2, i.e. $\pi_2(CD) > \pi_2(CC) > \pi_2(DD) > \pi_2(DC)$. Figure 1 gives an example of a PD protocol in normal form, comprising the different outcomes and monetary consequences.

		Player 2	
		C	D
Player 1	C	$30, 30$	$5, 50$
	D	$50, 5$	$10, 10$

Figure 1: PD protocol

We are interested in the players' preferences over the four possible outcomes and, of course, the corresponding monetary consequences in the PD protocol. One such preference is the "standard economic" preference that is based on the maximization of a player's individual material payoff, i.e., player i strictly prefers outcome ω to outcome $\tilde{\omega}$ if and only if $\pi_i(\omega) > \pi_i(\tilde{\omega})$. However, there exist other possible preferences, and the latter may only be one plausible candidate for the real preferences human beings actually have. In order to get a complete picture of the players' preferences, clearly a lot of information is needed — in particular if we think in terms of von-Neumann-Morgenstern preferences. However, relatively little information is needed if we want to solve the game with the concept of strict Nash equilibrium. In that case ordinal preferences are sufficient. Precisely, the only information we need to know is: *what outcome does a player prefer given the other player chooses C, and what outcome does he prefer given the other player chooses D.* This is achieved by applying a variant of the so-called strategy method (Selten 1967), where a players makes a *conditional choice* between C and D both for the case that the other player chooses C and for the case that the other player chooses D. Thus, for example, player 1 chooses an outcome from the set $\{CC, DC\}$ as well as an outcome from the set $\{CD, DD\}$. Similarly, player 2 chooses an outcome from the set $\{CC, CD\}$ as well as an outcome from the set $\{DC, DD\}$. The observed choice behavior reveals exactly the information about players' preferences we need. Moreover, the elicitation procedure

can be made incentive compatible, by letting both players make a conditional *and* an unconditional choice between C and D. If with positive probability a player's conditional choice and the opponent's unconditional choice becomes payoff relevant, revealing one's preference truthfully is a weakly dominant strategy.

What type of preferences do we elicit by this procedure? Without loss of generality we can restrict attention to the role of player 1. Note that player 1 makes two choices, each from a set of two outcomes. In consequence, there are four different types of preferences player 1 can reveal. The first type of preference corresponds to choosing D independent of what the other player does revealing the preference $DC \succ CC$ and $DD \succ CD$. Since this preference coincides with the maximization of player 1's material payoff, we call this preference a *selfish* preference.[2] The second type of preference corresponds to a choice of C in case the other player chooses C and a choice of D in case the other player chooses D. This behavior reveals the preference $CC \succ DC$ and $DD \succ CD$, which is a preference for reciprocity. Therefore a player revealing such a preference is called a *reciprocator*. The third type of preference reflects the choice of C independent of what the other player does, i.e., the preference $CC \succ DC$ and $CD \succ DD$. Since this preference is equivalent to maximizing the opponent's material payoff, we call this preference an *altruistic* preference. Finally, an individual who chooses D if the other player chooses C and chooses C if the other player chooses D, i.e. reveals the preference $DC \succ CC$ and $CD \succ DD$, is called an *anti-reciprocator*. Table 1 summarizes the four possible types of preferences we elicit by our procedure.[3]

	A player's		
type of preference	choice if other player chooses C	choice if other player chooses D	revealed preference (as player 1)
selfish	D	D	$CC \prec DC, DD \succ CD$
reciprocal	C	D	$CC \succ DC, DD \succ DD$
altruistic	C	C	$CC \succ DC, DD \prec CD$
anti-reciprocal	D	C	$CC \prec DC, DD \prec CD$

Table 1: Revealed preferences in the PD protocol

2.2 Experimental Design

In the experiment, we analyzed the PD protocol with payoff matrix as given in Figure 1. While we chose a neutral frame in the experiment labelling strategies A and B rather

[2]This preference is identical to the standard economic preference described above.

[3]Note that the different types of preferences we elicit by our procedure correspond to the pure strategies of the second player in a *sequential* PD protocol. In this sense our procedure is identical to the strategy method used in sequential game experiments.

than "cooperate" and "defect", we will use the conventional PD terminology in this paper. Altogether we implemented three treatments: a communication treatment, a no-communication treatment and an information treatment. Each treatment consisted of two parts, where in each part we elicited subjects' preference in the PD protocol and also subjects' belief about the preference of their opponent. The first part was the same in each treatment: preferences were elicited using both the behavioral approach described above and a questionnaire method, beliefs were elicited using the quadratic scoring rule (details given below). The second part differed across the treatments. In the communication treatment, before subjects' preferences and beliefs were elicited a second time, subjects could communicate with their (newly assigned) interaction partner. In the no-communication treatment, no communication was allowed and the second part of the experiment was identical to the first part. In the information treatment, subjects learned the preference of their new interaction partner before we elicited their preference and belief. As a basic principle, subjects always learned that there was a second part only after the first part was over. (Note: In this paper, we will not discuss the results from the second part of the information treatment but focus only on observations from the first part of this treatment, which was identical to the first part of the other treatments.)

In each part of a treatment, the elicitation of subjects' preferences proceeded as follows. All subjects were randomly matched into pairs. With the exception of the second part in the communication treatment, interaction was always anonymous, i.e., subjects did not know the identity of their opponent. A subject's preference was elicited using first the behavioral procedure described above and second a questionnaire method, where subjects had to rank the four different outcomes in the PD protocol on a scale from 1 to 4. Precisely, subjects were asked to assign rank 1 to the outcome they liked best, rank 2 to the outcome they liked second best, rank 3 to the outcome they liked third best, and, finally, rank 4 to the outcome they liked the least.[4] After preferences were elicited in these ways, a random device determined whether with respect to the behavioral procedure a subject's unconditional or conditional decision became payoff relevant. With probability 0.5 a subject's unconditional decision became relevant and the opponent's conditional decision became relevant, and with the same probability it was the other way round. In that way we ensured that our behavioral elicitation procedure was incentive compatible, as well as that in the questionnaire procedure subjects were not influenced by the actual outcome realized in their case.

[4]Note that we asked subjects to give us their subjective ranking over the different *outcomes* in the game and not only over the different monetary consequences. Thus, for instance, a subject had to compare outcome CC, where she and the other player cooperate, with outcome CD, where she cooperates and the other player defects, rather than only compare payoff profile $(30, 30)$ and payoff profile $(5, 50)$.

2.3 Procedural Details

At the beginning of the experiment subjects were given written instructions containing all the details of the experiment. To ensure the understanding of the experimental procedures all subjects had to answer several control questions and the experiment did not start until all subjects had answered all questions correctly. In addition all key aspects of the experiment were orally summarized. Then the first part began. Half of the subjects made the unconditional decision first and the conditional decision second, the other half made the conditional decision first and the unconditional decision second.[5] Before subjects ranked the four outcomes, each subject gave a belief about the preference of their opponent. Precisely, we asked each subject for her estimation of the probability that the other player will cooperate if she cooperates herself and if she defects herself, respectively. The quadratic scoring rule was used to make the elicitation of beliefs incentive compatible.[6] Subsequently, subjects ranked the four outcomes in the PD protocol from 1 to 4. After all decisions were made, one subject threw a die to determine for whom of the subjects the unconditional decision and for whom the conditional decision was payoff relevant. Subjects did not learn the outcome immediately, however, but were informed that there would be a second part of the experiment first. In the no-communication treatment, the second part was exactly the same as the first part with subjects randomly matched into new pairs. In the communication treatment, before the second part began subjects had the opportunity to communicate with their newly assigned interaction partner. For this, subjects were led to an extra room, where they could communicate face-to-face for about 5 minutes. After the communication phase subjects were asked to summarize the content of the communication on a prepared sheet and were led to their seat in the laboratory again. At their place, subjects individually answered 13 questions measuring specific interpersonal trust and sympathy between interaction partners.[7] The second part began, which from then on was identical to the second part of the no-communication treatment. At the end of the experiment, all subjects were informed about their and their opponent's payoff relevant decision in the first and in the second part of the experiment and the resulting payoff they earned in the experiment.

In total 96 subjects participated in the experiment, 24 in the communication treatment, 24 in the no-communication treatment and 48 in the information treatment. This implies that with respect to the elicitation of subjects' preferences in the first part we have 96 independent observations. All subjects were students from the University of Zurich and the Swiss Federal Institute of Technology Zurich. No subject participated in more than one session. With the exception of the questionnaire ranking, all decisions had monetary consequences where 10 points in the experiment represented 3 Swiss Francs (1 Swiss

[5] No differences were found.

[6] In each case a subject earned a payoff equal to $4(1-(\theta-s)^2)$, where $\theta \in [0,1]$ is the subject's stated probability belief that the other player cooperates and s takes the value 1 if the other player cooperates and 0 otherwise. Given this rule, subjects maximize their expected payoff by stating their beliefs correctly.

[7] Questions were taken from the specific interpersonal trust scale (Johnson-George and Swap 1982).

Franc ≈ .59 $US at the time of the experiment). On average subjects received 27.06 Swiss Francs including a show-up fee of 10 Swiss Francs. A session lasted about 60 minutes on average. All decisions were made on a computer screen. We used the experimental software z-Tree (Fischbacher 1999) to run the experiment.

3 Results

In this section we analyze the decisions made in the first part of any treatment in the experiment (96 observations). We first present our results concerning all behaviorally revealed preferences in PD protocol. Given these preferences we can construct the distribution of different PD *games* that are represented by the PD *protocol*. Our next result analyzes the unconditional decisions made, connecting a subject's unconditional decision to her revealed preference and her belief about the preference, i.e. conditional decision, of her opponent. Finally, we compare the results from the questionnaire method to the behaviorally revealed preferences.

3.1 Revealed PD Preferences

Our first main result concerns the empirical distribution of preferences in the PD protocol.

Result 1 *The large majority (85 percent) of the subjects have either selfish or reciprocal preferences in the Prisoners' Dilemma protocol.*

Support for Result 1 is given by Figure 2, which shows the relative frequencies of the different preferences in the PD. The figure nicely demonstrates that the large majority of the subjects reveal either selfish or reciprocal preferences. The precise numbers are 47 percent selfish, 38 percent reciprocal, 9 percent altruistic and 6 percent anti-reciprocal preferences.

3.2 PD Games

Given the empirical distribution of preferences, we can provide an answer to the question, what game individuals *really* play when they are confronted with the PD protocol. To answer this question we calculate the probability that, given the above distribution of revealed preferences, a pair of preferences is randomly drawn to play the game. As a result we also obtain the probability that the PD protocol is actually perceived as the classic PD game, i.e., both individuals have selfish preferences.

Result 2 *The simultaneous Prisoners' Dilemma protocol is perceived as the classic Prisoners' Dilemma game in only 22 percent of the cases. In 14 percent of the cases the Prisoners' Dilemma protocol is seen as a coordination game, i.e., both individuals have*

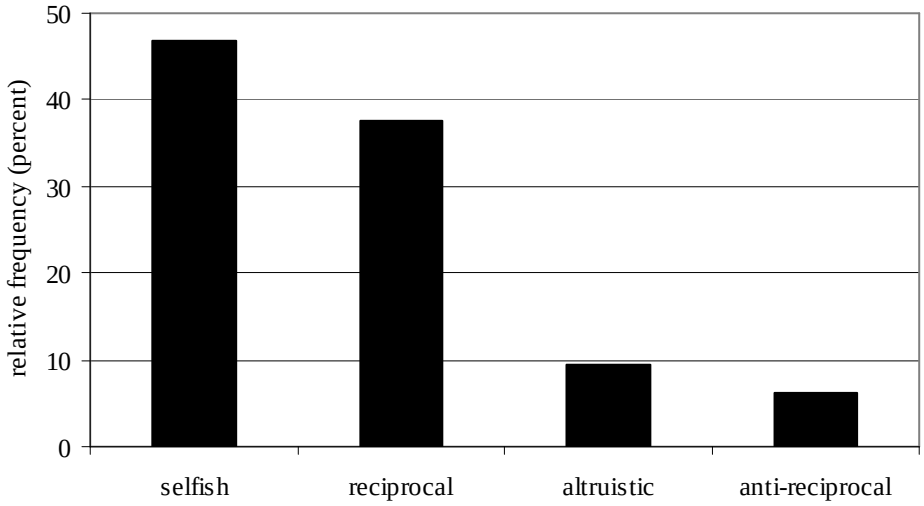

Figure 2: Distribution of revealed PD preferences

reciprocal preferences. Most likely (35 percent) the game is played by one reciprocal and one selfish individual.

Figure 3 shows the relative frequencies of the most interesting games, which involve either selfish, reciprocal, or altruistic preferences.

The largest frequency (35 percent) have those games that involve one selfish and one reciprocal player. In these games the unique Nash equilibrium is where both players defect, i.e., choose (D, D). The second largest class of games, which occur with a relative frequency of 22 percent, is the classic PD game, where both players have selfish preferences, i.e., maximize their individual material payoff. In this game the unique Nash equilibrium is again (D, D). With probability of 14 percent two reciprocal individuals meet to play the game. In that case the game is in fact a coordination game, which has two strict Nash equilibria, (C, C) and (D, D). A selfish and an altruistic individual meet with probability of 9 percent, leading to a unique Nash equilibrium, where the altruistic player cooperates and the selfish player defects. With probability of 7 percent an altruist is confronted with a reciprocal opponent, a game which has a unique Nash equilibrium of (C, C). Finally, in 1 percent of the cases two altruists meet, leading, again, to a unique Nash equilibrium of (C, C).

Thus, in only 71 percent of the cases, when two randomly selected individuals are called to play the PD protocol, the strategy profile (D, D) constitutes a Nash equilibrium of the resulting game. In 22 percent of the cases strategy profile (C, C) is a Nash equilibrium of the resulting game.

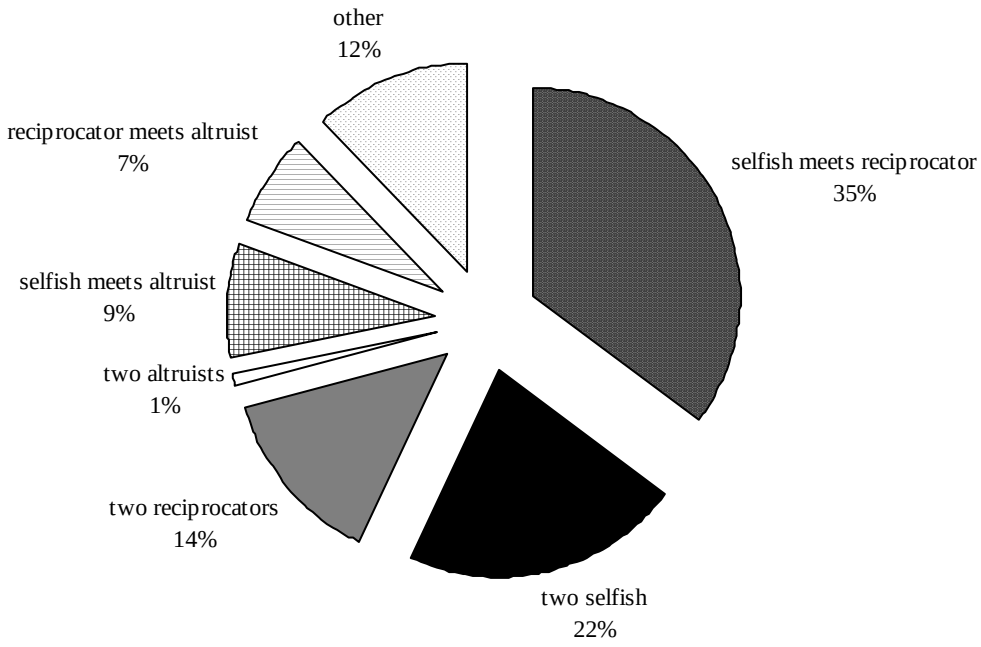

Figure 3: Distribution of PD games.

3.3 Unconditional Choices

Our experimental design allows us to analyze what unconditional decisions subjects made, in particular given the information we have about their revealed preference and their belief about the opponent's conditional decision.

Result 3 *Reciprocal and altruistic subjects show the highest probability for unconditional cooperation. However, also selfish subjects cooperate with probability of about* 0.3*. The average unconditional cooperation rate across all subjects is* 0.56*.*

Unconditional cooperation rates are shown in Figure 4 for the four different types of preferences. As can be seen, reciprocal subjects cooperate with the highest probability, which is equal to 0.92. Altruists cooperate with the second highest probability equal to 0.72. Most interestingly, also individuals who reveal a selfish preference cooperate roughly in one out of three cases, the probability is 0.29. The cooperation rate for an anti-reciprocal subject is 0.17.[8]

Here something on the different beliefs and optimality of individual behavior.

[8]The perhaps somewhat surprising observation that the cooperation rate of an altruist is lower than the one of a reciprocator seems to be due to the small number of altruists we see in the experiment. In total, 9 subjects reveal an altruistic preference and from these 7 subjects cooperate unconditionally.

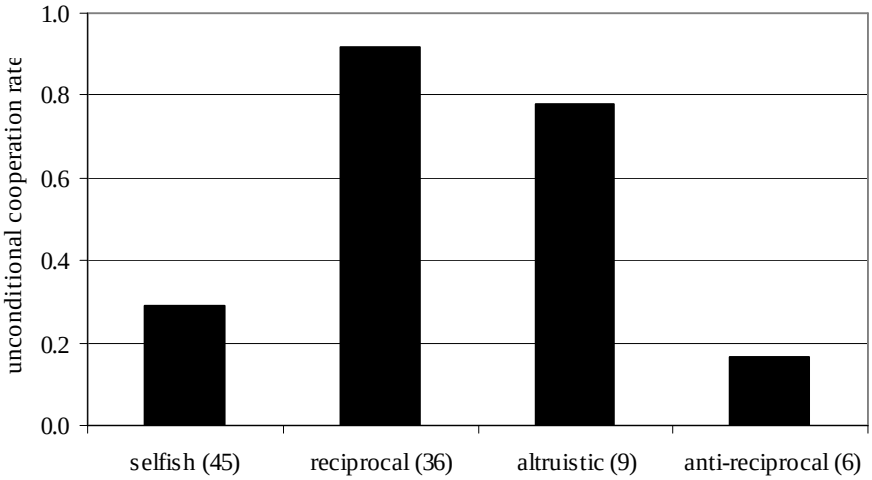

Figure 4: Unconditional cooperation (number of observations in parentheses)

3.4 Questionnaire Data

The next result compares the preferences subjects indicated by their ranking of the four outcomes in the PD protocol to their behaviorally revealed preferences. This result is interesting since all preference elicitation studies so far have used questionnaire data.

Result 4 *Subjects' preferences based on the questionnaire method are more cooperative than the behaviorally revealed preferences.*

Support for Result 4 is given by Figure 5, which compares the distribution of preferences for both elicitation methods.

While according to the behavioral method 47 percent of the subjects have selfish preferences, according to the questionnaire only 33 percent have such preferences. At the same time, 46 percent indicate a reciprocal preference if they have to rank the four outcomes in the PD, compared to 38 percent if they reveal their preference by behavior. With respect to altruistic preferences the relative difference between both elicitation methods is largest: We find twice as many altruists if we use the questionnaire (18 percent) than if use the behavior-based approach (9 percent). For anti-reciprocal preferences it is the other way round. Whereas 6 percent of the subjects show an anti-reciprocal preference based on behavior, 3 percent show such preference in the questionnaire. A marginal homogeneity test confirms that the two preference distributions are statistically different at the 5-percent level ($p = 0.0158$).

Result 4 seems intuitive. At least, it goes into the direction one might have suspected. However, the reason for the significant difference between both elicitation methods is not immediately clear. One plausible explanation could be that subjects' preferences are

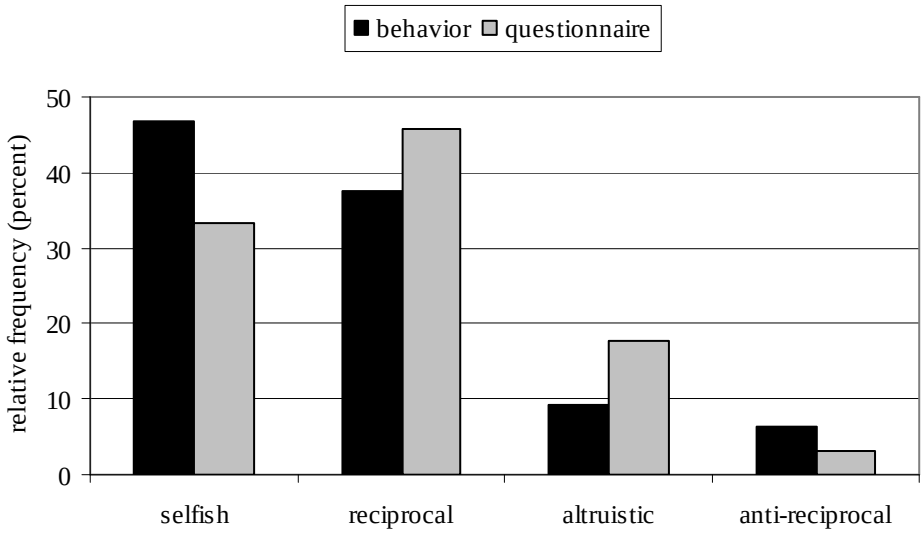

Figure 5: Distribution of preferences elicited by behavior and by questionnaire.

more cooperative in the questionnaire data because (1) the ranking of the four outcomes was not made incentive compatible and (2) subjects have an incentive to appear "nice". However, they may well be other reasons for our finding. Further studies are needed to fully answer this problem.

We can use the questionnaire data to find out about subjects' most and least preferred outcome in the PD. The result is given in Figure 6.

	C	D
C	CC 52	CD 7
D	DC 25	DD 16

rank 1

	C	D
C	CC 10	CD 63
D	DC 18	DD 9

rank 4

Figure 6: Relative frequency (in percent) of being the most (rank 1) and least (rank 4) preferred outcome, seen from the viewpoint of the row player.

Seen from the viewpoint of the row player, overall 52 percent of the subjects rank outcome CC highest, where both players cooperate, while only 25 percent of the subjects rank outcome DC highest, where the subject defects and the other player cooperates, although this outcome generates the highest material payoff to the subject. For a large majority of the subjects (63 percent) the least preferred outcome is outcome CD, where

the subject cooperates and the other player defects. Yet, still 18 percent give outcome DC the lowest rank.

It seems bizarre that 7 percent of the subjects rank outcome CD highest and that 10 percent rank outcome CC lowest. This suggests that the questionnaire method may indeed generate quite some amount of noise in the elicitation of subjects' preferences. Again, we believe that a thorough comparison of the two elicitation procedures is a worthwhile undertaking.

4 The Effect of Pre-play Communication

In this section we apply our method of behavior-based preference elicitation to one of the most frequently discussed phenomena in social dilemma games: the effect of pre-play communication. (Something on this phenomenon and the relevant literature.)

In principle, there are two possible explanations for the effect of pre-play communication. First, communication may change the players' *preferences*. In particular, face-to-face communication between the two players can easily be imagined to make players' preferences more cooperative. The basic idea is that if players have seen and have talked to each other they evaluate the different outcomes in the PD protocol differently. This explanation seems closely related to the argument that players develop a form of group identity when communicating with each other. Second, communication may change players' *beliefs* about the preference of their opponent. If a player has a preference $CC \succ DD$ and believes that with some sufficiently positive probability the opponent is a reciprocator, it is optimal for him to cooperate, independent of whether the player himself is a reciprocator or selfish (if he is an altruist he should cooperate anyway).

A major feature of our experimental design is that we can distinguish between these two explanations, because we measure all information that is needed: We elicit subjects' unconditional behavior, their preferences, and their beliefs. Therefore we are able to examine the explanatory validity of both arguments and we can answer the question what it really is, that causes the effect of pre-play communication.

To study the effect of pre-play communication we used the following experimental design. As a basis we took again the Prisoners' Dilemma protocol given in Figure **??**. We implemented two treatments, a *communication* treatment and a *control* treatment, where each treatment consisted of two parts. In the first part, which was identical for both treatments, we elicited subjects preferences and beliefs exactly as described before, i.e., subjects were randomly and anonymously matched into pairs, each subject made a decision as player 1 and as player 2, expressed his or her beliefs in the role of player 1 about the behavior of player 2, and gave a subjective ranking over the four different outcomes in the PD protocol. Then, before subjects learned what decision was relevant and what payoff they earned, the second part of the experiment started, which differed between the two treatments.

In the second part of the control treatment, subjects were again randomly and anonymously matched into pairs and we elicited subjects' preferences and beliefs using again the same procedure as in the first part of the experiment.

In the second part of the communication treatment, subjects were also randomly matched into new pairs. However, before we elicited subjects' preferences and beliefs each pair of subjects had the opportunity to communicate with each other face-to-face for about five minutes. Communication was free, i.e., we did not force subjects to follow any fixed protocol.[9] After the five minutes were over subjects were led to their PC's and we elicited subjects's preferences and beliefs using the same procedure as in the first part of the experiment.

In both treatments subjects did not know that there would be a second part until the first part of the experiment was over. At the end of both treatments subjects learned which of their decisions were relevant in each of the two parts of the experiment and what payoff they earned. Finally, all subjects were paid separately from each other.[10] In total, 24 subjects participated in each of the two treatments.

4.1 Preferences

We first report the effect of pre-play communication on subjects' preferences in the PD.

Result 5 *Pre-play communication causes subjects' preferences in the PD to become significantly more cooperative.*

Support for Result 5 comes from Figure 7 and Figure 8 showing the distributions of revealed preferences in the first and the second part of the communication treatment and the control treatment, respectively.

Figure 7 clearly shows that pre-play communication has a strong effect on players' preferences. While before communication 38 percent of the subjects reveal a selfish preference, after communication only 4 percent, i.e., one single subject, show a selfish preference in the PD. At the same time, the relative frequency of reciprocal and altruistic preferences increases from 50 to 67 percent and from 8 to 29 percent. The share of anti-reciprocators decreases from 4 percent to zero. A marginal homogeneity test confirms the statistically significant difference between the two distributions ($p = 0.0174$).

In the control treatment, on the other hand, we find no statistically significant difference between the two distributions of preferences we elicited in the first and in the second part of the experiment. A marginal homogeneity test produces a value of $p = 0.2839$. If anything, Figure 8 shows that the relative frequency of selfish preferences increases while

[9]Subjects were asked to summarize in a few written words the content of their conversation.

[10]This was particularly important in the communication treatment, where we told subjects at the beginning of the second part that at the end of the experiment everyone had to wait at his desk until he or she was called to be paid and leave the building. Only after a subject had left the building the next subject was called and paid.

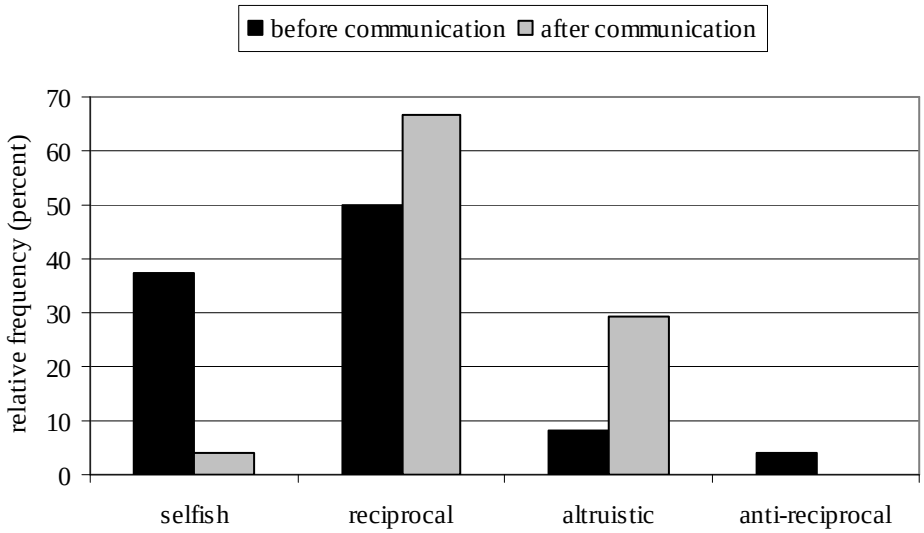

Figure 7: Distribution of preferences in the communication treatment.

the share of reciprocal and altruistic preferences decreases, i.e., preferences become *less* cooperative in the second part of the control treatment. Relative frequencies are 46 and 63 percent for selfish preferences, 33 and 29 percent for reciprocal preferences, 17 and 4 percent for altruistic preferences, and 4 percent for anti-reciprocal preferences in the first and the second part of the control treatment, respectively.

It is interesting to see how preferences change on the individual level in the communication treatment, i.e., *who* becomes altruistic and *who* becomes reciprocal. Our data show that selfish individuals become reciprocal or altruistic each with probability of roughly 1/2. Three out of four reciprocators remain reciprocators, while the other become altruistic. Finally, all altruists become reciprocators and all anti-reciprocators become selfish.[11]

4.2 Beliefs

So far it has become clear that pre-play communication does affect players' preferences. The next question is: what is the effect on players' beliefs?

Result 6 *Pre-play communication raises subjects' belief that the opponent cooperates after cooperation, and to some degree also after defection.*

Result 6 is based on Table 2 showing the average belief in the role of player 1 about the cooperative behavior of player 2. The upper part reports beliefs in the communication treatment, the lower part in the control treatment.

[11] Assuming this flow of preferences to form a stationary process, one can see immediately that reciprocal and altruistic preferences form an absorbing set.

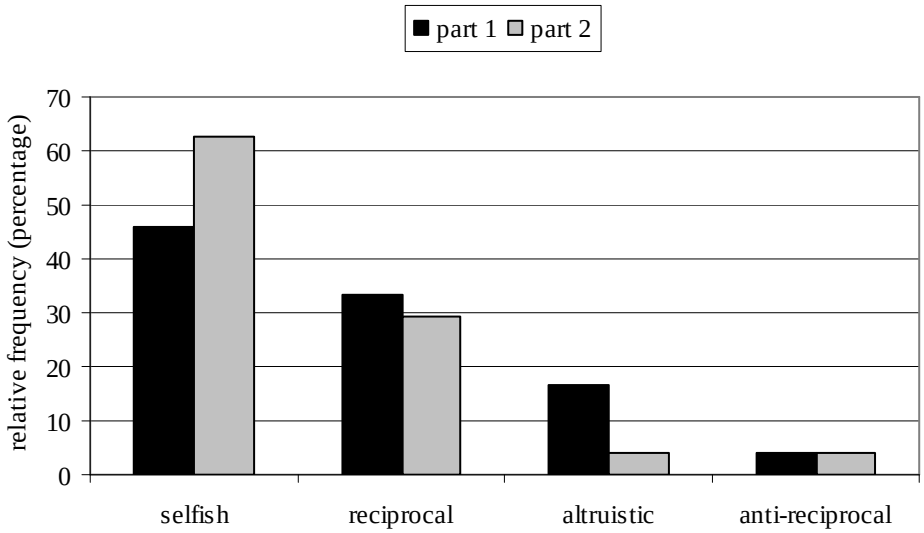

Figure 8: Distribution of preferences in the control treatment.

Before looking at the implication of pre-play communication, note that in the first part of the communication treatment ("before communication") and in the first part of the control treatment ("part 1"), i.e., where there was no difference between the two treatments, also the average belief is almost the same. Roughly, subjects believe that the opponent cooperates after cooperation with probability of one half, and cooperates after defection with probability of almost 0.2.

As can be seen from Table 2, pre-play communication affects beliefs in two ways. First, it strongly increases the average belief that player 2 cooperates if player 1 cooperates, from 0.54 to 0.88. Second, it slightly increases also the belief that player 2 cooperates if player 1 defects, from 0.18 to 0.3.[12] In the control treatment no such effect is observed. To the contrary, in this treatment the average belief decreases from 0.53 to 0.36 if player 1 cooperates and decreases from 0.18 to 0.15 if player 1 defects.[13] Hence, overall we find that pre-play communication has a significant effect on subjects' beliefs: it makes subjects's belief about the cooperative behavior of the opponent more optimistic.

4.3 Unconditional choices

The effect on preferences and beliefs caused by pre-play communication suggests that unconditional choices, i.e., decisions in the role of player 1 should become more cooperative

[12]According to a Wilcoxon signed-rank test, the first difference is significant at the 1 percent level ($p = 0.0037$), whereas the second increase is not statistically significant.

[13]The first decrease is significant at the 5 percent level ($p = 0.0235$), the second decrease is, again, not statistically significant.

Communication treatment	Probability that player 2 cooperates	
	if player 1 cooperates	if player 1 defects
before communication	0.54	0.18
after communication	0.88	0.30
Control treatment	**Probability that player 2 cooperates**	
	if player 1 cooperates	if player 1 defects
part 1	0.53	0.18
part 2	0.36	0.15

Table 2: Average belief in the role of player 1 that player 2 cooperates.

as well. Our next result shows that this is indeed the case.

Result 7 *With pre-play communication all subjects cooperate in the role of player 1.*

Table 3 reports subjects' rate of cooperation in the role of player 1 in the communication and in the control treatment. We see that in the communication treatment cooperation rates rise substantially after subjects had the possibility to communicate with each other. In fact, with pre-play communication *all* subjects cooperate as player 1, independent of their individual preference. In contrast, in the control treatment cooperation rates in part 2 decrease or remain rather close to those in part 1. On average, cooperation rates are roughly 1/2 in both parts of this treatment.

Communication treatment	Probability to cooperate as player 1			
	selfish	reciprocal	altruistic	anti-reciprocal
before communication	0.44	0.92	1	0
after communication	1	1	1	—
Control treatment	**Probability to cooperate as player 1**			
	selfish	reciprocal	altruistic	anti-reciprocal
part 1	0.18	1	0.75	0
part 2	0.27	1	0	0

Table 3: Cooperation rates of player 1 in both treatments.

References

Gibbon, R. and L. Van Boven (2001) "Contingent social utility in the prisoners' dilemma," *Journal of Economic Behavior & Organization*, 45, 1-17.

Kollock, P. (1997) "Transforming social dilemmas: group identity and cooperation," in: P. Danielson (ed.), *Modeling rational and moral agents*, pp. 186-210, Oxford: Oxford University Press.

Watabe, M., S. Terai, N. Hayashi, and T. Yamagishi (1996) "Cooperation in the one-shot prisoner's dilemma based on expectations of reciprocity," *Japanese Journal of Experimental Social Psychology*, 36, 183-196.

Weibull, J.W. (2001) "Testing game theory," mimeo, Stockholm School of Economics.