Social surplus determines cooperation rates in the one-shot Prisoner's Dilemma

Gary Charness, Luca Rigotti, and Aldo Rustichini

June 5, 2016
Social surplus determines cooperation rates in the one-shot Prisoner’s Dilemma

Gary Charness, Luca Rigotti, and Aldo Rustichini

June 5, 2016

Abstract: We provide evidence on how cooperation rates vary across payoff parameters in the Prisoner’s Dilemma (PD), using four one-shot games that differ only in the payoffs from mutual cooperation. In our experiment, participants play only the PD game, and play the game once and only once, so there are no potential confounds or methodological issues. Our results show that higher monetary payoffs from cooperation are associated with substantially higher cooperation rates, which increase monotonically from 23% to 60%. Participants’ beliefs about cooperation rates track closely actual cooperation rates: higher cooperation is expected from others when mutual cooperation payoffs are higher. This is true also for participants who, in a follow-up experiment, only make guesses about the choices of others.

Keywords: Prisoner’s Dilemma, cooperation rates, beliefs, experiment

JEL Codes: A13, C70, C91, D03

Contact: Gary Charness (corresponding author), University of California, Santa Barbara, 2127 North Hall, Santa Barbara, California, 93106, Luca Rigotti, University of Pittsburgh, 4901 Posvar Hall, Pittsburgh, Pennsylvania, 15260, Aldo Rustichini, University of Minnesota – Twin Cities, 4-101 Hanson Hall, Minneapolis, Minnesota, 55455. We thank the editor and referees of an earlier draft of the paper for very helpful comments that have led to a substantial revision of the analysis and the exposition.
1. Introduction

The Prisoner’s Dilemma (PD) is perhaps the most well-known game in the social sciences. The game was formulated by Merrill Flood and Melvin Dresher at RAND in 1950 and was later formalized by Albert Tucker. Figure 1 presents the canonical representation of a symmetric PD, with $T > R > P > S$.

**Figure 1: Canonical symmetric Prisoner’s Dilemma**

<table>
<thead>
<tr>
<th></th>
<th>Cooperate</th>
<th>Defect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperate</td>
<td>R, R</td>
<td>S, T</td>
</tr>
<tr>
<td>Defect</td>
<td>T, S</td>
<td>P, P</td>
</tr>
</tbody>
</table>

A great deal of research has been conducted on the PD, much of it focused on how to induce mutual cooperation and achieve a higher social payoff. Much of this research has centered on the Prisoner’s Dilemma as a repeated game. Robert Axelrod conducted two famous tournaments around 1980, with Anatol Rapoport first to introduce the Tit-for-Tat strategy (cooperate until one’s counterpart defects and then make the same choice as was made by one’s counterpart in the previous period) that proved to be the most successful strategy. Kreps, Milgrom, Roberts, and Wilson (1982) showed that if either player has even a very small probability of being committed to playing tit-for-tat, in equilibrium both players cooperate until the last few periods. This is because even an uncommitted player has an incentive to “build a reputation” for being committed to Tit-for-Tat, as doing so induces the other player to cooperate.
Less research has been done in one-shot (non-iterated) PD games.¹ Rabin (1993) showed that mutual cooperation is an equilibrium of the one-shot game if both players care sufficiently about the “kindness” of their counterpart. In experiments with multiple rounds with re-matching, Andreoni and Varian (1999) and Charness, Fréchette, and Qin (2007) demonstrated that mutual cooperation is sustainable if the players credibly pre-commit to making some range of positive transfer payments in the event that one’s counterpart has chosen cooperation. However, to our surprise, it appears that there are no studies that consider how the cooperation rate is affected by the parameters in a pure one-shot PD, where a participant makes only one choice in a session.²

![Figure 2: Our Prisoner’s Dilemma Games](image)

If we assume selfish preferences and identify utils with monetary payoffs, the prediction is that we should never see cooperation. But numerous experimental results falsify this assumption, so it is important to study the relationship between cooperation rates and monetary outcome profiles. We do so with a test involving the symmetric one-shot Prisoner’s Dilemma shown below: specifically we vary the payoff from mutual cooperation, denoted as $x$, in four experimental treatments. Figure 2 illustrates the implementation of the Prisoner’s dilemma,

---

¹ Note that the iterated Prisoner’s Dilemma is not the same game as the one-shot version. This fact has been well known since the early days of the Prisoner’s Dilemma literature, as demonstrated by John Nash’s comment in the footnote on p. 16 of Flood (1952).

² As discussed in section 2, the few papers that deal with one-shot games either feature multiple periods with re-matching of some form or environments in which one is not paid with certainty for the choice made. To our knowledge, this is the only paper in which subjects make only one decision in the laboratory.
where $x \in \{3,4,5,6\}$ takes on one value in each session, that we use to test the predictions that different theories (linking the monetary outcome profile and utility) provide in our environment.

A number of models formalize the intuitive prediction that the rate of cooperation shall increase as $x$ increases, by postulating utility of players over monetary joint payoffs. Any model of simple altruism in monetary payoffs where one has a fixed positive coefficient on the payoff of the other player, delivers this prediction. This idea goes back to Edgeworth’s Mathematical Physics (page 53, footnote 1), where he introduced the coefficient of effective sympathy, a positive number that weights the utility of the other and adds the result to one’s utility. Alternatively, Rapoport (1967) suggests that the rate of cooperation will depend on the $(R-P)/(T-S)$, from Figure 1. This represents the gain from mutual cooperation normalized by the full range of one’s possible payoffs. We vary $R$ and hold all else constant, so that the predicted rate of cooperation increases with $x$.

Rabin (1993) makes a similar prediction based purely on reciprocity considerations, using a “kindness” function. If one believes that the other person is being kind in the choice made, one may be willing (depending on one’s sensitivity to fairness considerations) to sacrifice money to be kind in return. Since cooperation is fully kind regardless of the value of $x$ (see p. 1287) and since one must have a larger fairness sensitivity to give up more money, there should be less cooperation as $x$ decreases. Charness and Rabin (2002) present a model (in their Appendix) in which one cares both about the total payoff for the players and the lowest payoff for any player.

---

3 Given the abundant evidence of heterogeneity in social preferences (see, e.g., Charness and Rabin, 2002), such models do not predict a single threshold at which point every individual switches from defection to cooperation.

4 The equilibrium predictions of a linear-preference model such as Edgeworth's are not unambiguous for a substantial range of the coefficient of effective sympathy, which is natural to assume is smaller than 1. For a large interval of this parameter values either (e.g. when $x = 3$, values between 0 and 1/5) the unique equilibrium is the joint defection, or (when $x = 3$, values between 1/5 and 1/2) the game has three equilibria, with the only symmetric one being the mixed. For larger values of the coefficients cooperating becomes a dominant strategy, and the prediction of increasing frequency of cooperation with $x$ is correct.
in addition to one’s own material payoff. Since we hold the minimum payoff constant in our design and social efficiency increases with $x$, the cooperation rate is predicted to increase with $x$.

A similar prediction is made by the Daley and Sadowski (2016) theory of magical thinking; their monotonicity axiom implies that weakening the motives for defection will induce less defection among players. Finally, a monotonically-increasing relation between cooperation rates and mutual-cooperation payoffs follows if participants might see the one-shot game in the light of repeated interactions, and adopt in the one-shot interaction norms of behavior that are learned in the more familiar environment of repeated interactions. In a repeated game, a higher mutual cooperation payoff decreases the relative gain from defection, $7-x$; so cooperation is easier to sustain when $x$ is larger. Experimental evidence on repeated games (Dal Bo and Fréchette, 2011) shows that cooperation rates do increase when $x$ increases in a repeated game.

We discuss the evidence supporting this hypothesis in more detail in the discussion section.

All of these theoretical views have a directional prediction in common about behavior.6

**Hypothesis 1**: The rate of cooperation increases as the payoffs from mutual cooperation increase in our games.

We find strong support for this hypothesis, as the cooperation rates increase steadily as the payoff from mutual cooperation increases. The cooperation rate is 23 percent, 34 percent, 51 percent, and 60 percent when $x$ is respectively 3, 4, 5, and 6.

We also elicited beliefs about the probability of cooperation from players after they have made their choices. Participants were not aware at the time of their PD choices that they would be asked to make a guess about the overall cooperation rate. We expected a positive relationship

---

5 For example, the cooperation outcome can be sustained by a grim-trigger strategy if $x$ is larger than $7 - 5\delta$.

6 We note that these hypotheses are directional and pertain to individual behavior. We do not invoke equilibrium reasoning in our one-shot game, where there is considerable heterogeneity (particularly regarding beliefs).
between the payoffs from mutual cooperation and beliefs about the overall cooperation rate. This can originate in different ways: for example, because individuals tend to overestimate how others are similar to them, as they cooperate more they expect others to do the same; another possibility is that individuals form beliefs and then choose what to do, in which case they may correctly anticipate more cooperation from others when $x$ is higher.\footnote{Costa-Gomes, Huck, and Weizsäcker (2014) and Rubinstein and Salant (2016) find that beliefs and actions are highly correlated in P.D. games; the latter paper suggests that this correlation may reflect strategic justification. The first paper in fact establishes a causal link between beliefs and actions, using an approach involving instrumental variables. Fehr et al. (2003), Bellemare and Kröger (2007), Sapienza, Toldra-Simats, and Zingales (2013), and Naef and Schupp (2009) provide evidence on the relationships between beliefs and action in trust and dictator games. It is not our purpose to establish causal relationships in this paper; we simply note correlations. Our own view is that beliefs are likely to affect actions and actions are likely to affect beliefs.} We expected a positive relationship between the payoffs from mutual cooperation and beliefs about the overall cooperation rate, since participants are, at least to some extent, likely to share the preferences, norms or motivations of others. We also anticipated that individual differences in preferences, norms or motivations affecting behavior would affect in the same direction the belief in the behavior of others, and thus specifically that cooperators and defectors would have different beliefs about the overall cooperation rate. We find strong support for each of the hypotheses below.

**Hypothesis 2:** Participants’ beliefs on overall cooperation rate increase as the payoffs from mutual cooperation increase in our games.

**Hypothesis 3:** Beliefs about the overall cooperation rate are higher for people who choose cooperation than for those who choose defection.

We also conduct additional sessions in which new participants make no PD choices, but instead are shown the instructions for the game and are incentivized to provide their beliefs about overall cooperation rates in the previous sessions. In these sessions we then also incentivize them for their beliefs about the beliefs of the people in their session about these overall cooperation
Here we also see that these beliefs increase significantly and steadily with increases in mutual-cooperation payoffs. We also see that beliefs about the beliefs of others about the overall cooperation rate are significantly different than one’s own beliefs. This suggests that people do not automatically presume that the beliefs of others are the same as their own (self-similarity).

In view of the existing results on one-shot and repeated games (Jones, 2008; Burks et al., 2009; Fehr and Huck, 2015, Gill and Prowse, 2015, Proto, Rustichini, and Sofiamos, 2016) we also expect cognitive skills to play a role in the choice of cooperation. In this experimental literature, two important regularities emerge. First, subjects with higher cognitive skills behave closer to the equilibrium predictions when the games are strictly competitive, or more generally when there is little possible gain from cooperation. Second, when gains from cooperation are possible, players with higher cognitive skills are more likely to achieve a constrained efficient outcome. Our PD belongs to the second class, so we formulate our hypothesis as:

**Hypothesis 4:** Cooperation rates are higher for participants with higher cognitive skills.

Our study makes four main contributions: First, we provide evidence that systematically increasing the material payoff from mutual cooperation (holding everything else constant) leads to a monotonic increase in cooperation rates in PD games in which one makes only one choice. Second, elicited beliefs about the cooperation rates among other players also respond to our treatment variable: people expect more cooperation when the payoff from mutual cooperation is higher. Third, more cooperation is expected by those players who have chosen cooperation. Fourth, cooperation rates are higher for participants with higher cognitive skills, in spite of the fact that those participants expect a comparatively lower cooperation rate from others.

The remainder of the paper is as follows. We present a literature review in section 2. Our experimental design and results are in section 3, and section 4 concludes.
2. Related literature

To our knowledge, the earliest systematic research involving the Prisoner’s Dilemma is reported in Flood (1952). He considered different payoff matrices with 100 iterations of the game. A well-known study is that of Rapoport and Chammah (1965), where (in the so-called pure matrix condition), each pair of participants plays the same game throughout for 300 plays; 10 pairs each play one of seven games.

Ahn et al. (2001) find that payoff values associated with fear (P-S) and greed (T-R) are important determinants of behavior in four PD games in which both T and S are varied. Participants made successive choices in each game, with feedback provided about the resulting payoff after each game, so that history and updating most likely affect choices. Engel and Zhurakhovska (2012) manipulate the payoff from mutual defection in a one-shot Prisoner’s Dilemma, finding that cooperation rates increase as this payoff decreases. However, each person makes 11 contingent choices in a within-subject, strategy-method design.\(^8\)

Our design involves a pure one-shot symmetric game.\(^9\) Having a test of behavior in a one-shot environment is useful to better understand the recent literature examining behavior in repeated PDs (Dal Bó and Fréchette, 2011, 2013; Fudenberg, Rand, and Dreber, 2012; Proto, Rustichini, and Sofianos, 2014). The fundamental finding of the repeated PD game is that

\(^8\) While there is some evidence that the use of the strategy method (Selten, 1967) preserves treatment effects, although not levels (see Brandts and Charness, 2011), within-subject designs may generate spurious effects (see Charness, Gneezy, and Kuhn, 2012); further, there is scope for confusion in this task. In addition, only one of the 11 choices was selected (at random) for payoff implementation. This payment device involves an additional layer of uncertainty and may or may not generate the same results as definite payment.

\(^9\) Of course asymmetric Prisoner’s Dilemma games have also been studied, as in Andreoni and Varian (1999) and Charness, Fréchette, and Qin (2007). We chose symmetric games to ease comprehension, to simplify the analysis, and to provide greater statistical power (twice as many relevant observations with symmetry). We would expect the same principles to apply with asymmetric games, although fairness concerns could then enter the picture.
cooperation increases with the discount rate, a result consistent with standard predictions.\textsuperscript{10} Our finding that the fraction of cooperation increases with the payoff to \((C, C)\) is in the same spirit since in both cases the frequency of cooperation increases with the payoff from cooperation.

3. Experimental Design and Results

3.1. Design

We conducted 12 experimental sessions at UCSB with the game in Figure 2, where the payoff represented dollar amounts to be received at the end of the session. There were three sessions for each of the four values (3, 4, 5, and 6) for the payoff from mutual cooperation. The number of people in each session ranged from 22 to 34, with a total of 348 people participating, each in only one session). Sessions were conducted in a large classroom, with space around each person for privacy. The game matrix was drawn on the board and the instructions were handed out and read aloud. Each possible combination of outcomes was discussed to aid comprehension.

Each person made a choice of “A” or “B” in the game matrix and was anonymously and randomly paired for payoff purposes with another person in the room. At the end of the session, a questionnaire was handed out to collect demographic data.\textsuperscript{11} This included a question in which the individual was asked to guess the percentage of people who had chosen cooperation (“A”) and received $3 if the guessed percentage was within five percentage points of the actual realized percentage.\textsuperscript{12} The experimental instructions and the questionnaire are presented in the Appendix.

\textsuperscript{10} This is in a weak sense, that is the set of equilibrium payoff outcomes increases in the direction of cooperation.

\textsuperscript{11} Costa-Gomes and Weizsacker (2008) show that asking about beliefs before or after actions are chosen does not induce a different distribution of actions among players. In that paper, each player had to guess the behavior of the player with whom she is paired, rather than the overall behavior.

\textsuperscript{12} This elicitation approach, perhaps first used in Charness and Dufwenberg (2006), has the advantage of being easy to understand. In addition, it is immune to risk-preference considerations. A minor drawback is that the
Each person made one choice and one guess. The experiment took about 40 minutes in all, with average earnings (including a $4 payment for showing up on time) of approximately $11.

We followed up these sessions with others in which the participants were shown the instructions for the game and asked to guess the overall cooperation rate in the original experiment. Guesses within five percentage points of the actual percentage earned $3. After we collected these guesses, we asked participants for their guesses about the average guess made by the other participants in their session, with guesses within five percentage points of the actual value earning $3. They also completed demographic questionnaires. Participants received a $6 show-up fee and earned an average of about $8 for a 30-minute session. There were 199 participants: 45 participants who made guesses about behavior when $x = 6$, 49 when $x = 5$, 51 when $x = 4$, and 54 when $x = 3$. There were two sessions for each treatment. The mutual-cooperation payoffs for a particular session were randomly-determined on the day of the session.

3.2. Results

**Choices:** Table 1 shows the aggregate choices made in each of our four treatments. We use the test of proportions (Glasmapp and Poggio, 1985), as the choice of whether to cooperate or defect is binary. A Wilcoxon-Mann-Whitney rank-sum test gives very similar results.

We see a very clear pattern: the cooperation rate increases steadily as the payoff from mutual cooperation increases. A Spearman correlation test confirms the strong relationship between the payoffs from mutual cooperation and the cooperation rate, with $\rho = 0.286$ and $p = 0.000$. The difference across treatments is always significant at least at the 10 per cent level for sensible range of guesses is reduced to [5,95] rather than [0,100]. There were almost no guesses outside this sensible range, indicating that guesses were not frivolous and that some degree of comprehension was present.

$^{13}$ All $p$-values are rounded to three decimal places and reflect one-tailed tests, except where otherwise indicated.
adjacent values of the payoffs from mutual cooperation, and it is strongly significant for all other pairwise comparisons. The difference in the cooperation rates at the extremes of the range of $x$ is large and highly significant.\textsuperscript{14} The cooperation rates are roughly linear in the cooperation payoffs: in a probit regression where the dependent variable is the choice to cooperate and the independent variable is the mutual cooperation payoff, each additional dollar paid for the joint cooperation outcome increases the probability of cooperating by about 12%.

\textbf{Table 1: Cooperation rates and tests of proportions, by treatment}

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Cooperation rate</th>
<th>Test for $x = 5$</th>
<th>Tests for $x = 4$</th>
<th>Tests for $x = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = 6$ (86)</td>
<td>60.47%</td>
<td>-1.290*</td>
<td>-3.637***</td>
<td>-4.835***</td>
</tr>
<tr>
<td>$x = 5$ (83)</td>
<td>50.60%</td>
<td>-</td>
<td>-2.304**</td>
<td>-3.596***</td>
</tr>
<tr>
<td>$x = 4$ (98)</td>
<td>33.67%</td>
<td>-</td>
<td>-</td>
<td>-1.499*</td>
</tr>
<tr>
<td>$x = 3$ (81)</td>
<td>23.46%</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

\textbf{Notes}: The number of observations is shown in parentheses. $Z$-statistics are given for each pairwise test of proportions; *, **, and *** indicate statistical significance at $p = 0.10$, 0.05, and 0.01, respectively (one-tailed tests, as per our directional hypothesis). $x$ represents the payoffs from mutual cooperation.

**Beliefs** As mentioned before, we elicited incentivized beliefs about aggregate cooperation rates in each session by asking players to guess (after making the game choice and not knowing that this question was coming) what percentage of other players had chosen a particular action.\textsuperscript{15} Table 2 shows the beliefs in each treatment. The two rightmost columns will be discussed later.

A Wilcoxon-Mann-Whitney test indicates significance for almost every pairwise

\textsuperscript{14} Indeed, one can imagine considerably more cooperation with a cooperation payoff of 6.75 and considerably less cooperation with a cooperation payoff of 2.25.

\textsuperscript{15} Notice that this does not ask one to predict the behavior of the person with whom one is matched. These beliefs about others’ play correspond to the “population frame” in the Rubinstein and Salant (2016) terminology.
difference in overall beliefs across treatment. A simple Spearman correlation test confirms this very strong relationship between the payoffs from mutual cooperation and beliefs about cooperation, with $\rho = 0.551$ and $p = 0.000$.

### Table 2: Beliefs about cooperation rates by other players.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Actual rate</th>
<th>Overall beliefs</th>
<th>Cooperator beliefs</th>
<th>Defector beliefs</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = 6$</td>
<td>60.47%</td>
<td>57.92% (2.77)</td>
<td>67.77% (3.00)</td>
<td>42.85% (4.15)</td>
</tr>
<tr>
<td>$x = 5$</td>
<td>50.60%</td>
<td>47.70% (2.88)</td>
<td>64.63% (3.13)</td>
<td>30.37% (3.05)</td>
</tr>
<tr>
<td>$x = 4$</td>
<td>33.67%</td>
<td>42.78% (2.35)</td>
<td>59.19% (3.30)</td>
<td>34.71% (2.59)</td>
</tr>
<tr>
<td>$x = 3$</td>
<td>23.46%</td>
<td>32.46% (2.47)</td>
<td>46.92% (5.38)</td>
<td>28.03% (2.55)</td>
</tr>
</tbody>
</table>

**Note:** Standard errors are shown in parentheses.

Since there is considerable heterogeneity among subjects, in Figure 3 we show the cumulative distribution of beliefs across the population in each of our treatments, where reported beliefs about other players’ cooperation rates are grouped in bins of length 0.1 (from 0 to 0.1, from 0.1 to 0.2, and so on). Beliefs differ substantially across treatments, and as we increase the payoff from joint cooperation, players guess that others will cooperate more.

Figure 3 clearly illustrates that beliefs move with monetary payoffs in the same way in which cooperation rates move with changes in monetary payoffs. This co-movement might have other explanations: for example, Rubinstein and Salant (2016) elicit beliefs about others’ play in

---

16 Comparing $x = 6$ to $x = 5$, 4, and 3 gives, respectively, $Z = -2.671$, -4.069, and -6.138, with $p = 0.004$, 0.000, and 0.000; comparing $x = 5$ to $x = 4$ and 3 gives, respectively, $Z = -1.131$ and -3.640, with $p = 0.129$ and 0.000; comparing $x = 4$ to $x = 3$ gives $Z = -2.998$, with $p = 0.001$. 

11
the game of chicken. They find that players tend to report beliefs that are biased toward rationalizing their action and that are biased toward thinking others behave as they do.

**Figure 3 – Cumulative Beliefs by Mutual Cooperation Payoff**

Each line represents the cumulative distribution of beliefs for a particular value of the mutual-cooperation payoff.

**Regression analysis** For an overall view of the different potential explanatory variables, we present below the estimation of an OLS robust regression model in which beliefs depend on the payoff from mutual cooperation (x), the SAT score as a proxy of cognitive skills, and sex and age as control variables. We perform this regression for all sessions, and we present separate regressions for the sessions with subjects who played the PD and those with non-players. The SAT score in our sample is approximately normally distributed, with mean 1923.8 and SD 187.23. Some subjects could not or did not provide a score.

Variables are expressed in their natural units to make an intuitive assessment of effect size easier; the SAT score is in units of 100 (about half of the SD of 187.23). The mutual-cooperation payoff is significant; an increase in $1 increases the probability of cooperation expected from others by 6 per cent (the standardized coefficient (beta) is 0.285). SAT is also significant, and negative: an increase by 100 points in the score induces a reduction by 1.3 per cent of the expected cooperation (beta is -0.101). Male and age are not significant. In a similar
regression with the absolute value of the difference of the guess of the participant from the mean of the belief in the same class of payoff shows no significant effect of SAT score. This confirms that higher SAT is not affecting the belief by improving the ability to guess the correct value.

**Table 3: OLS for beliefs, all subjects.**

<table>
<thead>
<tr>
<th>Category</th>
<th>All participants</th>
<th>Players</th>
<th>Non Players</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperation payoff</td>
<td>6.635*** (0.99)</td>
<td>7.606*** (1.27)</td>
<td>5.194*** (1.61)</td>
</tr>
<tr>
<td>SAT</td>
<td>-1.387** (0.60)</td>
<td>-1.308* (0.77)</td>
<td>-1.606* (0.91)</td>
</tr>
<tr>
<td>Male</td>
<td>0.011 (2.26)</td>
<td>-0.899 (2.82)</td>
<td>1.935 (3.61)</td>
</tr>
<tr>
<td>Age</td>
<td>-0.313 (0.52)</td>
<td>-0.314 (0.58)</td>
<td>-0.003 (0.81)</td>
</tr>
<tr>
<td>Constant</td>
<td>51.630*** (16.42)</td>
<td>42.354*** (19.39)</td>
<td>61.848*** (24.37)</td>
</tr>
<tr>
<td>N</td>
<td>487</td>
<td>308</td>
<td>179</td>
</tr>
</tbody>
</table>

**Notes:** Standard errors are in parentheses. ***, **,* denote significance at \( p = 0.01, 0.05, \) and 0.10, respectively (two-tailed). The sample includes subjects who chose and those who only made guesses.

A regression (not shown) with the same independent variables on the guess on the guess of others in the subsample of the participants who did not play the game shows the same pattern, with the coefficient of the mutual-cooperation payoff equal to 3.65 \( (p = 0.005, \beta = 0.21) \) and of SAT equal to -1.39 \( (p = 0.054, \beta = -0.14) \).\(^{17}\)

In summary, the effect of the mutual-cooperation payoff confirms the monotonic relation we described earlier; as the mutual cooperation payoff increases players think there will be more cooperation, and they also think others think there will be more cooperation. As the SAT score

\(^{17}\) For comparison, the coefficients of the guess in the regression with the same independent variables are 5.14 for \( x \) \( (p = 0.001, \beta = 0.23) \) and -1.605 \( (p = 0.086, \beta = -0.12) \). A regression using second-order beliefs as the dependent variable is not qualitatively different from the regression for first-order beliefs, further confirming that SAT does not simply improve the ability to guess the right value.
increases, players think cooperation rates will be lower, and they think that others think that cooperation rates will be lower. In view of our findings on the regressions of the distance of the guess from the average guess and from the guess of others, the channel from SAT to belief is not due to a better ability to predict the action of others.

**Beliefs and choices** We now document some features of the data that may provide an insight for the co-movement of behavior and beliefs. First, we show that, in the aggregate, the players’ guesses about the behavior of others match their own behavior. Second, we show that behavior is related to the expected payoff of the game, where the expectation is computed using beliefs about others’ play. Finally, we use payoff and beliefs to predict cooperation choices in our data.

Regarding Hypothesis 3, Table 2 shows that beliefs of cooperators about the cooperation are higher than the beliefs of defectors, with the difference ranging from 18 to 34 percentage points. Wilcoxon-Mann-Whitney rank-sum tests confirm this difference.

We can compare the elicited beliefs about the rate at which others have cooperated with the choices of players who have those beliefs by looking at the frequency of cooperators among all players who assign a given probability to their opponent having cooperated. In other words, we look at all players who say that the probability of cooperation is $y\%$ and compute the fraction among them that choose cooperation. Figure 4 presents this comparison. On the horizontal axis, the first bin includes those who think the fraction of cooperators is between 0 and 10 percent, the second bin includes those who think the fraction of cooperators is between 11 and 20 percent, and so on. For each of these bins, a point in the diagram is determined by the fraction of players who actually cooperated on the vertical dimension and the average belief about how many would

---

18 The Z-statistics for comparisons for $x = 3, 4, 5,$ and 6 are 3.002, 4.858, 5.938, and 4.446, respectively, with corresponding p-values of 0.001, 0.000, 0.000, and 0.000.
cooperate on the horizontal one. The Figure shows that beliefs and action frequency have a strong correspondence, and the correlation between the two series is 0.98.

**Figure 4: Beliefs and actions.**

![Figure 4: Beliefs and actions.](image)

<table>
<thead>
<tr>
<th>Bin</th>
<th>N</th>
<th>Average beliefs about cooperation</th>
<th>Actual frequency of cooperation</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0,10%)</td>
<td>24</td>
<td>0.054</td>
<td>0.042</td>
</tr>
<tr>
<td>(10%,20%)</td>
<td>39</td>
<td>0.127</td>
<td>0.077</td>
</tr>
<tr>
<td>(20%,30%)</td>
<td>46</td>
<td>0.223</td>
<td>0.239</td>
</tr>
<tr>
<td>(30%,40%)</td>
<td>44</td>
<td>0.328</td>
<td>0.227</td>
</tr>
<tr>
<td>(40%,50%)</td>
<td>30</td>
<td>0.429</td>
<td>0.233</td>
</tr>
<tr>
<td>(50%,60%)</td>
<td>33</td>
<td>0.525</td>
<td>0.515</td>
</tr>
<tr>
<td>(60%,70%)</td>
<td>58</td>
<td>0.636</td>
<td>0.603</td>
</tr>
<tr>
<td>(70%,80%)</td>
<td>29</td>
<td>0.729</td>
<td>0.724</td>
</tr>
<tr>
<td>(80%,90%)</td>
<td>32</td>
<td>0.834</td>
<td>0.875</td>
</tr>
<tr>
<td>[90%,100%]</td>
<td>12</td>
<td>0.942</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Left panel: scatterplot and regression line of actual cooperation rate (vertical axis) and belief (horizontal axis). The table on the right has the corresponding data.

The last two columns in Figure 4 show that the rate of cooperation increases with the belief about the rate of cooperation. Can this observation possibly be consistent with a model of behavior in which players best respond to their beliefs? In order to answer that question we need a bit of notation. Let \( u(C,C) \) denote the player’s utility when she cooperates and so does her opponent, let \( u(C,D) \) denote the player’s utility when she cooperates and her opponent defects, and so on for \( u(D,C) \) and \( u(D,D) \). Let \( q \) be a player’s belief about the likelihood her opponent cooperates. In choosing an action, the player compares two lotteries: Cooperate has expected payoff \( qu(C,C) + (1-q)u(C,D) \), and Defect has expected payoff \( qu(D,C) + (1-q)u(D,D) \). If she is an expected utility maximizer, the player will cooperate if and only if:

\[
qu(C,C) + (1-q)u(C,D) > qu(D,C) + (1-q)u(D,D) \quad \text{or} \\
q[u(C,C) - u(D,C)] > (1-q)[u(D,D) - u(C,D)].
\]
Consider two players who have the same utility function over outcomes, assume beliefs do not depend on actions, and assume the first player is more pessimistic about the chances her opponent cooperates than the second player is. Then, one can easily verify that if the first player prefers to cooperate, so does the second. So, a model in which players best respond to beliefs could produce the data we observe. This does not mean the model above is the only one consistent with our data; for example, Table 2 above provides some evidence that beliefs may not be independent of actions and that dependence is also consistent with our data.

Table 4: Cooperation, payoff and beliefs Odds ratios reported.
Left columns: Logistic regression, Right column: OLS

<table>
<thead>
<tr>
<th>Category</th>
<th>Logit Odds Ratio</th>
<th>Logit Marginal Effect</th>
<th>Linear regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperation payoff</td>
<td>1.261* (0.168)</td>
<td>0.038* (0.022)</td>
<td>0.030* (0.022)</td>
</tr>
<tr>
<td>SAT</td>
<td>1.183** (0.095)</td>
<td>0.028** (0.013)</td>
<td>0.027** (0.013)</td>
</tr>
<tr>
<td>Belief</td>
<td>1.054*** (0.007)</td>
<td>0.009*** (0.001)</td>
<td>0.009*** (0.001)</td>
</tr>
<tr>
<td>Male</td>
<td>0.763 (0.219)</td>
<td>-0.045 (0.047)</td>
<td>-0.388 (0.047)</td>
</tr>
<tr>
<td>Age</td>
<td>1.111* (0.067)</td>
<td>0.017 (0.010)</td>
<td>0.015 (0.009)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.0001*** (0.0002)</td>
<td>-</td>
<td>-1.015*** (0.332)</td>
</tr>
<tr>
<td>N</td>
<td>308</td>
<td>308</td>
<td>308</td>
</tr>
</tbody>
</table>

Notes: Standard errors are in parentheses. ***, **, * denote significance at $p = 0.01$, 0.05, and 0.10, respectively (two-tailed tests).

Table 4 reports the results of regression analysis of the choice of cooperation. The mutual cooperation payoff, after we condition for beliefs, is associated with a significant increase in cooperation rates: in the linear model, an increase of one dollar increases the cooperation rate by

Rubinstein and Salant (2016) note that this “single-crossing” property does not rest on expected utility theory.
3 per cent. The SAT score increases cooperation rates: an increase of 100 points is associated with a 2.7 per cent increase. A one per cent increase in the expected cooperation rate of others is associated with an almost one to one increase in cooperation (0.9 per cent).

**Self-similarity** The surprisingly close relationship between actions and beliefs displayed in Table 2, Figure 4 and Table 4 could potentially be explained by self-similarity. Rubinstein and Salant (2016) define self-similarity as “a player tends to think that other players behave similarly to him, and thus will report beliefs that are biased toward his own action.” That paper reports that self-similarity is present in a game of chicken. This theory suggests in our case that making a choice in the PD colors one’s beliefs about the actions taken by others. In other words, all else equal, cooperators assess a higher probability that others have also cooperated. We see clear evidence of this in Table 2, where the beliefs of cooperators about the cooperation of others are always higher than the beliefs of defectors. Yet, this evidence can still be consistent with our result that as the mutual-cooperation payoff increases beliefs about how much others cooperate also increase. In order to quantify the presence of self-similarity, and understand whether or not this effect is solely responsible for the relationship between beliefs and mutual cooperation payoff, we run a regression of beliefs on all the variables presented in Table 3 above plus variables that measures subjects’ behavior. The results are in Table 5, and they include a specification where actions and mutual cooperation payoffs are also interacted with each other.

Having chosen to cooperate on average increases the predicted cooperation rate by about 25 percentage points, an effect consistent with the results presented in Table 2. Despite accounting for this, the payoff from mutual cooperation is significant in explaining subjects’ beliefs about cooperation rates. This significance is only slightly decreased if we allow the mutual-cooperation payoff and cooperation rate to interact, in specification (2). Taken together,
these two suggest that even when self-similarity is accounted for, beliefs are also explained by the payoff subjects receive for mutual cooperation. Subjects anticipate the effect of monetary payoff on the actions of others even after accounting for their own action.

**Table 5: OLS for beliefs about cooperation rates**

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperation payoff</td>
<td>4.678***</td>
<td>3.734**</td>
</tr>
<tr>
<td></td>
<td>(1.14)</td>
<td>(1.47)</td>
</tr>
<tr>
<td>Cooperate</td>
<td>25.641***</td>
<td>15.077</td>
</tr>
<tr>
<td></td>
<td>(2.53)</td>
<td>(10.90)</td>
</tr>
<tr>
<td>Coop. x Coop. Payoff</td>
<td>-</td>
<td>2.296</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2.30)</td>
</tr>
<tr>
<td>SAT</td>
<td>-1.676**</td>
<td>-1.700**</td>
</tr>
<tr>
<td></td>
<td>(0.67)</td>
<td>(0.67)</td>
</tr>
<tr>
<td>Male</td>
<td>0.325</td>
<td>0.355</td>
</tr>
<tr>
<td></td>
<td>(2.44)</td>
<td>(2.44)</td>
</tr>
<tr>
<td>Age</td>
<td>-0.621</td>
<td>-0.616</td>
</tr>
<tr>
<td></td>
<td>(0.50)</td>
<td>(0.50)</td>
</tr>
<tr>
<td>Constant</td>
<td>57.630***</td>
<td>61.948***</td>
</tr>
<tr>
<td></td>
<td>(16.85)</td>
<td>(17.40)</td>
</tr>
<tr>
<td>N</td>
<td>308</td>
<td>308</td>
</tr>
</tbody>
</table>

**Notes:** Standard errors are in parentheses. ***, **, * denote significance at $p = 0.01$, $0.05$, and $0.10$, respectively (two-tailed tests).

**Beliefs without choice** To better isolate the effect of payoffs on beliefs we conducted sessions in which participants made no choice in the PD, but only guessed the average cooperation rate in our original sessions. Since they took no action in the game, their beliefs could not be colored by their choices. A strong relationship between mutual-cooperation payoffs and beliefs here would confirm that the increase in the belief in the cooperation of others as $x$ increases is not entirely driven by the action taken in the game. Figure 5 presents the cumulative distribution of subject’s predicted cooperation rates by mutual cooperation payoffs for both players and non-players.

One can see that even for subjects who do not play in the game there is evidence of the
close relationship between beliefs about the cooperation rate and mutual-cooperation payoffs. Beliefs increase monotonically with the payoffs from mutual cooperation, although the distinctions across values of our treatment variable are not as stark as they are in Figure 3. So, although one can attribute the difference between solid and dotted lines for a given \( x \) to the influence of a subject’s action on their beliefs, the difference across colors is still substantial.

**Figure 5 – Cumulative Beliefs by Mutual Cooperation Payoff, Players and Non Players**

![Cumulative Beliefs by Mutual Cooperation Payoff, Players and Non Players](image)

Each line represents the cumulative distribution of beliefs for a particular value of the mutual cooperation payoff. Dotted lines are for subjects who have played the game, while solid lines are for those who did not play.

Rank-sum tests on pairwise comparisons indicate significance in most cases.\(^{20}\) A simple Spearman correlation test confirms a strong relationship between the payoffs from mutual cooperation and beliefs about cooperation, with \( \rho = 0.238 \) and \( p = 0.000 \). So the relationship found for the players is robust to whether one has taken an action or not.

As mentioned, to test for self-similarity we also elicited guesses about the guesses made by the other participants in the same session as the participant. Table 6 presents the first- and

---

\(^{20}\) Comparing \( x = 6 \) to \( x = 5, 4, \) and 3 gives, respectively, \( Z = -1.762, -1.838, \) and -3.454, with \( p = 0.039, 0.033, \) and 0.000; comparing \( x = 5 \) to \( x = 4 \) and 3 gives, respectively, \( Z = -0.338 \) and -1.917, with \( p = 0.368 \) and 0.028; comparing \( x = 4 \) to \( x = 3 \) gives \( Z = -1.233, \) with \( p = 0.109 \). These are all one-tailed tests.
second-order beliefs of non-players, in comparison to the actual cooperation rate, while Figure 6 shows cumulative first- and second-order beliefs for non-players.

Table 6: Beliefs by non-players, by treatment

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Actual rate</th>
<th>Beliefs about rate</th>
<th>Beliefs about beliefs</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = 6$</td>
<td>60.47%</td>
<td>63.42% (3.25)</td>
<td>63.07% (2.67)</td>
</tr>
<tr>
<td>$x = 5$</td>
<td>50.60%</td>
<td>56.16% (3.13)</td>
<td>53.09% (2.56)</td>
</tr>
<tr>
<td>$x = 4$</td>
<td>33.67%</td>
<td>52.65% (3.78)</td>
<td>54.08% (2.60)</td>
</tr>
<tr>
<td>$x = 3$</td>
<td>23.46%</td>
<td>46.35% (3.21)</td>
<td>49.17% (2.87)</td>
</tr>
</tbody>
</table>

Notes: Standard errors are shown in parentheses.

Figure 6 – First- and Second-order Beliefs of Non Players, by Mutual Cooperation Payoff

Each line represents the cumulative distribution of beliefs for a particular value of the mutual cooperation payoff. For subjects who did not play, dotted lines are the first order beliefs, while solid lines are the second order beliefs.

Examining second-order beliefs, there is a positive (albeit weaker) relationship between these beliefs and mutual-cooperation payoffs. This is not so surprising, since we are considering a more difficult cognitive task when one has not even made a choice in the underlying game. In any case, while the only significant pairwise significance is found for comparisons between $x = 6$ and $x = 3, 4, 5$, the Spearman test gives $\rho = 0.222$ and $p = 0.001$. A simple OLS regression with the second-order guess as the dependent variable and the payoffs from mutual cooperation as the independent variable has a coefficient of 4.001 on the payoff term, with $p = 0.001$. 
Figure 6 also highlights differences between one’s own guess about the cooperation rate and one’s guess about the average guess of others about this cooperation rate. If self-similarity were the key driving force, we should expect little difference between these two guesses. While Table 6 indicates that the average second-order guess differs little from the average first-order guess, there is a great deal of heterogeneity (the first-order guess exceeded the second-order guess in 99 cases and the reverse was true in 87 cases; the guesses were the same in only 13 cases). The average absolute value of the difference in these guesses is 13.6; the difference between the guesses was less than 10 in only 46.7% of the cases (93 of 199). A Kolmogorov-Smirnov test rejects that the first- and second-order guesses are the same ($\chi^2 = 7.329$, $p = 0.026$, two-tailed test). So people do not just assume that everyone else is the same as they are.

It is interesting to compare the guesses made by non-players with the guesses made by players, considering also the players’ behavior (see Table 3). Guessed cooperation rates are higher for those who didn’t play the game, are similar to those of the cooperators in the game, but are different from the non-cooperators. Comparing the rates of cooperators for $x = 6, 5, 4$ and $3$ gives, respectively, $Z = -1.195, -1.499, -0.698$ and $0.000$, with $p = 0.232, 0.134, 0.485$ and $1.000$, two-tailed tests. The same comparisons between non-cooperators and non-players give $Z = 3.624, 5.015, 3.583$ and $4.192$, with $p < 0.001$ in all cases, two-tailed tests.

**Comparison with Repeated-Games Data** Our test of behavior in a one-shot environment offers some insight into behavior of our participants in repeated PDs, which is an active topic of current research. Just as in our one-shot experiment, in cooperation in repeated games increases as the payoff from mutual cooperation increases (see Table 3, page 417, of Dal Bó and Fréchette,
This hints at the natural question of what might be the relation between these observed results. One can compare the cooperation rates in Table 3 of Dal Bó and Fréchette (2011) with those in our Table 1, by utilizing the Rapoport (1967) ratio \((R-P)/(T-S)\).

**Table 7: Comparison of PD cooperation rates in repeated games and one-shot games**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Rapoport ratio</th>
<th>(\delta = \frac{3}{4})</th>
<th>(\delta = \frac{1}{2})</th>
<th>Avg. of (\delta)’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x = 3)</td>
<td>0.167</td>
<td>23.46%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(R = 32) (DF)</td>
<td>0.184</td>
<td>20.25%</td>
<td>9.82%</td>
<td>15.04%</td>
</tr>
<tr>
<td>(x = 4)</td>
<td>0.333</td>
<td>33.67%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(R = 40) (DF)</td>
<td>0.395</td>
<td>58.71%</td>
<td>17.98%</td>
<td>38.34%</td>
</tr>
<tr>
<td>(x = 5)</td>
<td>0.500</td>
<td>50.60%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(R = 48) (DF)</td>
<td>0.605</td>
<td>76.42%</td>
<td>35.29%</td>
<td>55.86%</td>
</tr>
<tr>
<td>(R = 32) (DF)</td>
<td>0.667</td>
<td>60.47%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: DF refers to repeated games in Dal Bó and Fréchette (2011). \(\delta\) only applies to DF conditions.

In our data, at \(x = (3, 4, 5, 6)\) we have Rapoport ratios (0.167, 0.333, 0.500, 0.667), and we get cooperation rates (23.46%, 33.67%, 50.6%, 60.47%). In Dal Bó and Fréchette (2011), \(R = (32, 40, 48)\) gives Rapoport ratios (0.184, 0.395, 0.605) and the cooperation rates are (20.25%, 58.71%, 76.42%) when the discount rate is \(\frac{3}{4}\) and (9.82%, 17.98%, and 35.29%) when the discount rate is \(\frac{1}{2}\). Their data with \(\delta = \frac{3}{4}\) fits rather well with our data: a simple OLS regression of the ratio against the observed cooperation rate gives a coefficient of 0.837 \((t = 4.62, p = 0.006,\) two-tailed test) and an adjusted \(R^2\) of 0.772. The choice of \(\delta = \frac{3}{4}\) is more natural than \(\delta = \frac{1}{2}\); the latter describes a repeated environment with very few interactions, while \(\delta = \frac{3}{4}\) seems a better approximation to the field environment. A simple averaging of the cooperation rates with their

\[\begin{array}{c|c|c}
R, R & 12, 50 \\
50, 12 & 25, 25 \\
\end{array}\]

21 The payoff matrices in their games are of the form (where \(R \in \{32, 40, 48\}\):
two discount rates yields a substantially better fit with our data, as the OLS regression gives a coefficient of $1.128 (t = 11.56, p = 0.000)$ and an adjusted $R^2$ of 0.957 in this case.

This analysis shows the Rapoport (1967) ratio effectively predicts cooperation rates in both one-shot games and repeated games. To our knowledge, we are the first to note this result.

4. Conclusion

We provide a clean test of how cooperation rates vary across several one-shot Prisoner’s Dilemma games, which differ only in the material payoffs from mutual cooperation; we call the test clean because participants’ only activity in the session is playing this game, they play one and only one instance of it, and are paid for their choice with certainty.

Our data provide support for the four hypotheses we formulated. Cooperation rates increase monotonically in the payoff from joint cooperation: within the range of our parameter choices, these rates change substantially, from 23% to 60%. Participants also expect this behavior from others: the belief in the cooperation rates of others respond strongly and significantly to our treatment variable and higher cooperation is expected when the material payoff from mutual cooperation is higher. In the experiment where participants do not choose, the monotonicity of beliefs on payoffs extends to the second-order beliefs over the guess of others. Our data confirms a general idea on the role of cognitive skills (Jones, 2008; Burks et al., 2009; Proto et al., 2015): in non-zero sum games, when gains from cooperation are possible, higher cognitive skills are associated with a larger cooperation rate, and a larger distance from the game-theoretic predictions that one can formulate taking own monetary payoffs as utility.

An interesting direction of future research would be to make the quantitative comparison (like the one we attempted in Table 7) more precise, bringing into the analysis the role of personality traits and particularly cognitive skills, and compare the estimates with those one can obtain from the other theoretical viewpoints recalled in section 1.
References


Pellegrini, Anthony, Cary Roseth, Shanna Milner, Catherine Bohn, Mark Van Ryzin, Natalie Vance, Carol Cheatham, and Amanda Tarullo (2001), “Social Dominance in Preschool Classrooms,” Journal of Comparative Psychology, 121(1), 54-64.


Rubinstein Ariel, and Yuval Salant (2016), “’Isn’t everyone like me?’ On the presence of self-similarity in strategic interactions,” Judgment and Decision Making, forthcoming.


Appendix

Sample Instructions ($x = 5$)

Thank you for participating in this experiment. You will receive $4 for your participation, in addition to other money to be paid as a result of decisions made in the experiment.

You will be paired anonymously with another person in the room, using a simple numbering scheme based on your ID number in this experiment. Each person will be making a simultaneous choice between A and B in the following decision matrix:

<table>
<thead>
<tr>
<th>Other person</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5, 5</td>
<td>1, 7</td>
</tr>
<tr>
<td>B</td>
<td>7, 1</td>
<td>2, 2</td>
</tr>
</tbody>
</table>

In each cell, the first number represents the outcome for you and the second number represents the outcome for the other person. Each unit represents $1.

Thus, if both people choose A, you would receive 5 and the other person would receive 5. If both people choose B, you would receive 2 and the other person would receive 2. If you choose A and the other person chooses B, you would receive 1 and the other person would receive 7. If you choose B and the other person chooses A, you would receive 7 and the other person would receive 1.

Please feel free to ask questions.