European Economic Review 166 (2024) 104753

Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/eer

The role of payoff parameters for cooperation in the one-shot Prisoner's Dilemma

Simon Gächter [a,b,c,*], Kyeongtae Lee [d], Martin Sefton [a], Till O. Weber [e]

[a] Centre for Decision Research and Experimental Economics (CeDEx), University of Nottingham, UK
[b] IZA Bonn, Germany
[c] CESifo Munich, Germany
[d] Economic Research Institute, Bank of Korea, South Korea
[e] Newcastle University Business School, Newcastle University, UK

ARTICLE INFO

JEL classification:
A13
C91

Keywords:
Prisoner's Dilemma
Cooperation
Payoff parameters
Temptation
Risk
Efficiency
Normalized gain
Normalized loss
K-index
Experiments

ABSTRACT

The Prisoner's Dilemma is arguably the most important model of social dilemmas, but our knowledge about how its material payoff structure affects cooperation is incomplete. We investigate the effect of variation in material payoffs on cooperation in one-shot Prisoner's Dilemma games. We report results from three experiments (N = 1,993): in a preliminary experiment, we vary the payoffs over a large range. In our first main experiment (Study 1), we present a novel design that varies payoffs orthogonally in a within-subjects design. Our second main experiment, Study 2, investigates the orthogonal variation of payoffs in a between-subjects design. In a complementary analysis we also study the closely related payoff indices of normalized loss and gain, and the K-index. A robust finding of our experiments is that cooperation increases with the gains of mutual cooperation over mutual defection.

1. Introduction

In many economic and social environments there is a conflict between individual and collective interests. The simplest model to represent such a conflict is the Prisoner's Dilemma (PD) and so it plays an important role in the behavioral sciences, where the PD is the topic of a vast literature in economics, sociology, political science, and social psychology. There is extensive evidence of cooperation in experimental PDs, and cooperation is observed even in carefully controlled anonymous one-shot interactions where participants have a real material incentive to defect (e.g., Cooper et al. (1996); Frank et al. (1993); Mengel (2018); Embrey et al. (2018); Dal Bó and Fréchette (2018)).[1] The literature has studied a wide variety of factors that affect cooperation (see, e.g., Balliet et al. (2009); Balliet (2010); Van Lange et al. (2014)), but perhaps from an economics perspective the most fundamental factor to consider is the material payoff structure. If players would be solely motivated by material payoffs, defecting would be a dominant strategy and the structure of

* Corresponding author at: Centre for Decision Research and Experimental Economics (CeDEx), School of Economics, University of Nottingham, UK.
E-mail address: simon.gaechter@nottingham.ac.uk (S. Gächter).

[1] Cooperation is also observed in repeated PD games that allow for strategic motives to cooperate (see, e.g., Embrey et al. (2018)). For a discussion of cooperation in finitely and infinitely repeated PD game experiments, see Mengel (2018) and Dal Bó and Fréchette (2018), respectively.

https://doi.org/10.1016/j.euroecorev.2024.104753
Received 12 December 2021; Received in revised form 4 May 2024; Accepted 8 May 2024
Available online 9 May 2024

payoffs – the relative size of payoffs in the PD – would not matter.

The fact that people sometimes cooperate in anonymous one-shot PDs violates the assumption that people always maximize material payoffs. Given this observation, we ask the most basic question, which we will make precise below: which features of the payoff structure explain cooperation? As we discuss in detail in Section 2, a surprisingly small literature has studied this question and a robust result has yet to emerge. Our contribution is to provide, across three experiments, a systematic analysis of the role of the material payoff structure for cooperation in one-shot PDs.

Our experiments are based on games in which two participants simultaneously choose to either 'Cooperate' or 'Defect' and their choices translate into money earnings as shown in Table 1.

We refer to the entries in Table 1 as payoffs, but to be clear they are the material payoffs resulting from their decisions and we make no claim about how they are related to utility more broadly construed. Following Rapoport and Chammah (1965) we choose the payoffs to satisfy the PD condition $T > R > P > S$. Thus, participants earn more from mutual cooperation than from mutual defection ($R > P$). However, cooperation is a '*risky*' choice that makes the participant vulnerable to being exploited by a defector ($P > S$). Additionally, each participant is '*tempted*' to choose defection as it increases her earnings against a cooperator ($T > R$). The PD condition ensures that the dominant strategy for money-maximizing participants is to defect. Rapoport and Chammah (1965) impose a second condition, $2R > T + S$, to ensure that mutual cooperation maximizes combined earnings. We focus on one-shot PDs that satisfy both conditions.

Our goal in this paper is to investigate three natural—*ceteris paribus*—payoff comparisons that capture the three sources of incentives alluded to in the previous paragraph:

First, a row player who assumes column player plays Cooperate gets a payoff increase of $T - R$ from defecting rather than cooperating. While a selfish player would defect since $T - R$ is positive, more generally a player who cares about own payoffs but trades this off against other considerations would have an increased incentive to defect as $T - R$ increases.[2] Thus, the higher T is relative to R, the higher the *temptation* to defect, holding all other payoffs constant.

Second, a row player who assumes column player plays Defect gets a payoff increase of $P - S$ from defecting rather than cooperating. Equivalently, a player who cooperates *risks* getting S rather than the payoff of P they would have got from defecting. Again, there is an increased incentive to defect and a decreased incentive to cooperate as $P - S$ increases. These two payoff comparisons are based on a player's interest in own payoff.

Third, it is also possible that players are motivated by collective interests, and so we consider a further payoff comparison, whereby players might also be more likely to cooperate the greater the payoff from mutual cooperation, R, is relative to the payoff from mutual defection, P, that is, the greater the *efficiency* of cooperation $R - P$.

We express these ceteris paribus payoff comparisons as percentage changes, using Mengel's (2018) payoff indices TEMPT $\equiv \frac{T-R}{T}$; RISK $\equiv \frac{P-S}{P}$; and EFF $\equiv \frac{R-P}{R}$. A property of the RISK, TEMPT, EFF indices is that they are invariant to multiplying the game's payoff matrix by a factor, for example, when using varying exchange rates across different subject pools. However, they are not invariant to adding a constant (e.g., in case of differing show-up fees). While this does not concern the within-subject pool investigation of the relative explanatory power of the three indices, a careful comparison across studies with varying show-up fees might warrant the use of normalized indices (see Section 5).

In the previous literature, several payoff indices have been proposed to predict the degree of cooperation in PDs (see Murnighan and Roth (1983)). Perhaps best known is Rapoport (1967)'s K-index ($\frac{R-P}{T-S}$) which is defined as the gains from mutual cooperation over mutual defection, ($R - P$), normalized by the payoff range ($T - S$). The K-index condenses a game's incentives into a single index based on all four elements of the payoff matrix. This can be viewed as a parsimonious prediction of how likely cooperation will be for a given payoff structure, but it has the disadvantage that PD games with very different incentives in terms of RISK, TEMPT and EFF may have the same K-index. In fact, several studies report varying rates of cooperation across PD games with different payoffs but the same K-index (e.g., Moisan et al. (2018)). Our approach, based on Mengel's indices, is not to predict the overall rate of cooperation in a game (that will no doubt depend on all four material payoffs, plus a host of other factors) but rather to examine the ceteris paribus effects of changes in particular incentives.

Our experiments are motivated by several observations about the previous literature and a preliminary experiment (which we will discuss in Section 2). First, the earliest studies and most of the subsequent research has examined payoff effects in the context of *repeated* PDs. Here, of course, players may have strategic reasons to cooperate, at least in early periods. This in turn complicates the interpretation of payoff indices as measuring incentives to defect. For example, for a given payoff matrix the incentive to defect differs according to whether a player is making a choice in the first or the last period.

Second, there are surprisingly few studies that have examined the effect of controlled payoff variation on cooperation in *one-shot* PDs and these offer an incomplete account of the role of material incentives for several reasons. Most of these studies vary more than one payoff index simultaneously across treatments and therefore cannot provide clear evidence on the relative effect size across the payoff indices.

Furthermore, most of these studies eliminate strategic reasons to cooperate by randomly matching participants across periods, but by allowing feedback between games they do allow for learning effects. For example, even if a participant plays against different participants across periods, the experience of being defected on in early periods may shape a participant's willingness to cooperate in later periods.

[2] These other considerations could, for example, reflect other-regarding concerns, such as utility derived from the payoffs of others.

Table 1
The Prisoner's Dilemma game.

	Cooperate	Defect
Cooperate	*R, R*	*S, T*
Defect	*T, S*	*P, P*

Notes: $T > R > P > S$. Row's payoff is given by the first entry in each cell.

Most relevant to our research is Mengel (2018). While few experiments examine controlled variation in payoffs, payoffs do differ considerably across studies and Mengel takes advantage of this variation to conduct a meta-analysis of the roles of RISK and TEMPT, controlling for EFF. For one-shot games Mengel finds that RISK best explains variation in cooperation rates and TEMPT has no explanatory power after controlling for RISK and EFF. However, her meta-analysis includes games that do not meet the Rapoport and Chammah (1965) PD conditions of $T > R > P > S$ and $2R > T + S$. As we show in Section 2.2, Mengel's result does not hold when imposing the PD conditions. In the restricted sample satisfying both conditions neither RISK nor TEMPT has a significant effect on cooperation after controlling for EFF. Moreover, Mengel's study is based on data from experiments that vary in many potentially important procedural variables, as well as in the payoffs they use, and so identifying the effect of payoff variation requires that these other procedural variables do not vary systematically with payoffs, or that they are adequately controlled for. In our experiments we vary payoffs systematically across treatments within a fixed design, offering an opportunity to corroborate (or not) Mengel's results via controlled experimental analysis.

We conduct a preliminary experiment and two new studies motivated by Mengel's results and those of our preliminary experiment. For our preliminary experiment, we run a lab experiment in which participants played 15 one-shot games with varying payoffs in a within-subject design. Payoffs were chosen to meet our two PD conditions while aiming for large variation in the RISK, TEMPT and EFF indices, resembling the variation across the studies that entered Mengel's meta-analysis. We find that cooperation is significantly higher when EFF is higher. However, this design includes only a few instances where one index varies while the other two indices are held constant.

In our first main experiment, Study 1, we vary RISK, TEMPT and EFF *orthogonally* across eight games that meet the two PD conditions. This allows us to conduct a clean test of the effect of changing one index while holding constant the remaining two. Again, we employ a within-subject design in which participants make decisions in all eight games. We recruit participants from two different subject pools. Our first subject pool is comprised of university student participants, as in most of the studies that motivated our experiment. Our second subject pool consists of workers on the Amazon Mechanical Turk (AMT) platform, which constitutes a more diverse subject pool regarding age, income, and education (e.g., Arechar et al. (2018); Snowberg and Yariv (2021)). Previous studies have found that cooperation varies systematically with demographic characteristics. For instance, older people tend to cooperate more than the young (e.g., Gächter and Herrmann (2011); List (2004); Matsumoto et al. (2016); Praxmarer et al. (2024)). Comparing subject pools allows us to test whether results based on student samples are transferable to a more diverse and, on average, older and presumably more cooperative population. In neither subject pool do we find any evidence that cooperation varies systematically with RISK. In contrast, cooperation decreases significantly with TEMPT and increases significantly with EFF in both subject pools.

A potential criticism of Study 1 is that the within-subject design allows for learning through enhanced experience in game play or induces an experimenter demand effect whereby participants might feel compelled to condition their action on the payoffs as these are the only things changing across the rounds. In our second main experiment, Study 2, we address this criticism by conducting a between-subject experiment using the same games as in Study 1 and, as far as possible, the same instructions and procedures. We recruit participants from the AMT platform. Participants play a single one-shot PD game, where the game is randomly drawn from one of the eight games used in Study 1. We find that cooperation is significantly higher when EFF is higher, whereas we do not find significant effects of RISK and TEMPT on cooperation.

Taken together, our experiments suggest that, in one-shot PDs where mutual cooperation maximizes social welfare, increasing EFF has a robust and positive impact on cooperation whereas decreasing RISK does not significantly enhance cooperation. Increasing TEMPT has the most detrimental effect on cooperation in our within-subject Study 1 but has an insignificant effect in our between-subject Study 2. Complementary analyses with the frequently used indices normalized loss, normalized gain, and the K-index, which are related to our indices RISK, TEMPT, and EFF, respectively, support our main conclusion: across all our experiments and subject pools, cooperation in the PD increases with the mutual gains from cooperation.

2. Related literature and some preliminary evidence

There is a vast experimental literature on PDs (for surveys see Balliet et al. (2009); Van Lange et al. (2014)). However, the very first published paper on PD experiments (Flood (1958)), the early work of Rapoport and Chammah (1965), and much of the subsequent experimental literature, has studied repeated PDs. The repeated PD offers a rich environment to study strategic behavior, but a complicated one in which to study the role of payoff structure for cooperation. Embrey et al. (2018) and Mengel (2018) discuss the effect of payoffs on cooperation in finitely repeated PDs. The role of incentives, unconfounded with strategic incentives, is laid bare in the one-shot PD. In the one-shot PD players have a dominant strategy to defect, but nevertheless cooperation is often observed. Many studies have investigated factors promoting cooperation (see, for example, Sally (1995) and Balliet (2010), which survey the role of communication) but there are surprisingly few studies that implement *controlled payoff variation* in the basic one-shot PD. We discuss

these in Section 2.1. Of course, payoffs vary greatly across studies, and so Mengel (2018) uses a meta-analysis to study the effect of payoff indices on cooperation. We discuss Mengel's study in Section 2.2.

2.1. Experiments varying payoff parameters

To our knowledge, seven experimental studies examined the effect of controlled payoff variation on cooperation in PDs. Charness et al. (2016) conducted a one-shot PD between-subject experiment varying R across four treatments. They found that average cooperation rates increase with R. However, note that both EFF and TEMPT change as R changes. Therefore, we cannot say whether increasing R increases cooperation because it increases efficiency, or decreases temptation, or both. Our experiments will allow us to separately identify the effects of EFF and TEMPT on cooperation.

Six studies implemented within-subject experiments where participants played multiple PDs with varying payoffs. Engel and Zhurakhovska (2016) studied 11 one-shot PDs where P varied across games and T, S and R were held constant. Each participant played all 11 PDs with no feedback between games. The authors found that cooperation decreases as P increases. Note, however, that this varies RISK and EFF simultaneously across games, and the observed decrease in cooperation may be due to either increasing RISK, or decreasing EFF, or both. Again, our experiments allow the separate identification of the effects of RISK and EFF.

Three studies used designs in which participants played a series of games against randomly changing opponents, with payoffs varying across games and feedback at the end of each game. Vlaev and Chater (2006) varied the K-index across games and found that the cooperation rate increased with the K-index. Schmidt et al. (2001) and Ahn et al. (2001) examined the impact of variations in 'greed' $\left(\frac{T-R}{T-S}\right)$ and 'fear' $\left(\frac{P-S}{T-S}\right)$ on cooperation. These two studies are closely related to our own as greed and fear are alternative measures of temptation and risk (based on a different normalization to those used in the TEMPT and RISK indices). Schmidt et al. (2001) varied the values of R and P across six games while keeping the values of T and S constant and found similar effect sizes of greed and fear on cooperation. Note, however, that an increase in greed could reflect higher temptation or lower efficiency (i.e., TEMPT increases and EFF decreases with greed when T and S are held constant). Similarly, an increase in fear could likewise reflect either an increase in risk or a decrease in efficiency. Ahn et al. (2001) is more closely related to us as they varied the payoffs across four games by using *high* and *low* values of T and S but holding R and P constant. Thus, efficiency is kept constant in their study and variation in T and S results in separate variation in RISK and TEMPT. Ahn et al. (2001) found that greed (or TEMPT) has a greater impact than fear (or RISK) on cooperation. Note that all three studies provided feedback between games during the experiment, and therefore cooperation might be affected by the outcome of previous games as well as by payoff changes. Indeed, all three studies report significant feedback effects. In our experiments, no feedback between games is provided.

Finally, Au et al. (2012) and Ng and Au (2016) study the relative risk of cooperation (henceforth riskiness) which they define as $\left(\frac{R-S}{(R-S)+(T-P)}\right)$, and examine how riskiness and participants' risk attitudes affect cooperation. Au et al. (2012) employed 18, 16, and 28 PDs in three experiments, while Ng and Au (2016) used 24 PDs. No feedback was provided until the end of the experiment in either study. Both studies found that the effect of riskiness of PDs depends on participants' risk attitude: risk-averse participants are more likely to cooperate in a less risky game, while risk-seeking participants are more likely to cooperate in a riskier game. However, the measure of riskiness does not disentangle risk, temptation, and efficiency: riskiness increases as T decreases or R increases. Therefore, increasing cooperation of risk-seeking participants with increasing riskiness might be caused by either decreasing temptation or increasing efficiency or both. The orthogonal variation of payoffs in our Studies 1 and 2 avoids these problems.

2.2. Mengel's meta-analysis

A particularly relevant study for our purposes is Mengel (2018) which examines the relative effect of RISK and TEMPT using data from previously published research supplemented by additional experiments that Mengel conducted either in the lab or on AMT. For the 73 games that were played either as one-shot games or in a random matching protocol, Mengel finds that RISK best explains the variation in cooperation rates, while TEMPT cannot explain this variation after controlling for RISK and EFF.

We report a re-analysis of this dataset, using the same OLS regression specification, in Table 2. The dependent variable is the average cooperation rate. Column 1 reproduces the results reported in Table 3 Column 1 of Mengel (2018). RISK is significantly negatively, and EFF is significantly positively, associated with the average cooperation rate. The coefficient on TEMPT is virtually zero and insignificant.

In some of the games in the full sample $P = S$ and so defecting is only weakly dominant, while in two games $T > R > S > P$ so the game has two strict equilibria. In Column 2, we restrict the sample to games that meet the first PD condition (i.e., $T > R > P > S$). The effect of RISK on cooperation substantially decreases and becomes only marginally significant.

Column 3 further restricts the subsample to games that meet both PD conditions (i.e., $T > R > P > S$ and $2R > T + S$) and shows that neither RISK nor TEMPT are significantly associated with the variation in average cooperation rates, with the caveat that the

Table 2
Average cooperation rate regressed on payoff indices using Mengel's (2018) dataset.

	(1) Full sample	(2) Imposing $T > R > P > S$	(3) Imposing $T > R > P > S$ & $2R > T + S$	(4) Imposing $T > R > P > S$ & $2R \leq T + S$
RISK	−0.255***	−0.142*	−0.045	−0.178
	(0.061)	(0.074)	(0.123)	(0.105)
TEMPT	0.003	0.050	−0.492	−0.165
	(0.080)	(0.084)	(0.305)	(0.179)
EFF	0.291***	0.360***	0.301*	0.443***
	(0.089)	(0.096)	(0.149)	(0.122)
Constant	0.370***	0.218**	0.304**	0.370*
	(0.084)	(0.097)	(0.130)	(0.200)
Adj. R^2	0.35	0.24	0.17	0.42
Obs.	73	66	36	30

Notes: Coefficients of OLS models with standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Model 1 reproduces the estimates of Table 3, Column 1, in Mengel (2018).

Table 3
Determinants of cooperative choice in the preliminary experiment (15 PD games).

Dependent variable: cooperation dummy	
RISK	−0.044
	(0.036)
TEMPT	−0.083
	(0.087)
EFF	0.399***
	(0.060)
Control variables	*Yes*
Constant	0.249
	(0.340)
Within R^2	0.10
Obs. (Clusters)	930 (62)

Notes: Coefficients of a random effects linear probability model with robust standard errors clustered on participants in parentheses. The control variables are round, age, gender, nationality, Business/Economics major, spending, and political attitude. The full results are in Online Appendix B, Table B3. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

sample size is considerably reduced when we restrict attention to games that meet both PD conditions. For comparison we include Column 4 which is based on games meeting the first PD condition but not the second (i.e., $T > R > P > S$ and $2R \leq T + S$). This is an attempt to establish whether the reduced effect of RISK in Column 3 compared to Column 2 is due to a strong association between cooperation and RISK when $2R \leq T + S$, or whether it reflects low power due to the reduced number of observations. In Column 4 the coefficient on RISK is approximately four times that of Column 3, and although insignificant this suggests that the reduced effect of RISK in Column 3 is driven by excluding games where $2R \leq T + S$ where there is a strong association of cooperation with RISK.[3]

It is important to note that the studies included in Mengel's dataset had their own idiosyncratic reasons for selecting their parameters and the variation between the parameterizations is therefore inevitably somewhat unsystematic. In our experiments we design the payoffs explicitly for comparing the effects of payoff indices. Furthermore, the experiments in Mengel's dataset used different instructional materials and framing of tasks: these differences unrelated to payoffs may affect cooperation across experiments. In our experiments, we control these non-payoff factors by holding them constant within our design.

2.3. A preliminary experiment

We conducted our preliminary experiment with 62 participants playing 15 games that meet the two standard PD conditions and vary the RISK, TEMPT and EFF indices over a wide range (see Online Appendix A for the instructions and Online Appendix B for the experimental design details, game parameters, procedures, and additional results). We chose convenient non-negative payoff

[3] One can speculate about why RISK has a strong association with cooperation in games with $T + S > 2R$. It may be that when $T + S > 2R$ cooperation increases when RISK is lower (S is higher) because the asymmetric outcome is more appealing for efficiency reasons (as we will see below, EFF is an important consideration). The difficulty of interpreting RISK and EFF when the asymmetric outcome maximizes the sum of payoffs underscores our focus on games where efficiency requires mutual cooperation.

parameters to vary the RISK, TEMPT and EFF indices over a wide range yielding a low, medium, and high level for each index similar to the studies that entered Mengel's (2018) dataset.

Across the 15 games, cooperation rates varied between 0.37 and 0.77. In Table 3, we report the effect of payoff indices on cooperation using a linear probability model with participant random effects. Robust standard errors are clustered on participants. Using a random effects model allows us to estimate the effects of individual characteristics (i.e., age, gender, nationality, major, spending, and political attitude). The dependent variable is a cooperation dummy, and the explanatory variables are payoff indices (RISK, TEMPT, EFF), with controls for individual characteristics and the round in which the respective game was played.

We find a positive and highly significant coefficient of EFF, whereas neither RISK nor TEMPT have a statistically significant effect on cooperation. An increase in EFF of 0.1 is associated with a 3.99 percentage points higher probability of cooperating. The full model results, robustness checks and additional analyses are in Online Appendix B.

Although the 15 games included in the experiment managed to achieve a large variation in the payoff indices comparable to the studies that entered Mengel's (2018) dataset, this design has the drawback that the induced variation in payoff indices is not fully orthogonal. That is, it gives limited ability to conduct clean non-parametric tests of whether cooperation varies when one index is varied, holding other indices constant. Also, we did not elicit beliefs and so it does not allow us to examine, or control for, the effect of beliefs on choices. Beliefs are interesting because related research in public goods experiments shows that beliefs strongly influence cooperation (e.g., Frey and Meier (2004); Croson (2007); Fischbacher and Gächter (2010); Dufwenberg et al. (2011)). Game parameters in public goods games do causally shift beliefs and cooperation and because many people are conditional cooperators, increased beliefs increase cooperation (e.g., Gächter and Marino-Fages (2023)).

3. Methods

3.1. Experimental design and procedures for within-subjects Study 1

For Study 1, we create different PDs by varying RISK, TEMPT and EFF orthogonally. This allows us to identify the effect of a single payoff index on behavior while holding constant the other two. First, we fix a *low* and *high* level for each of the three payoff indices. We then generated $2^3 = 8$ payoff matrices representing all possible variations of the two levels across the three payoff indices. The payoffs are presented in Table 4. $R = 500$ is constant across all PDs, while our experiment has two distinct values of $T \in \{600, 800\}$ and $P \in \{200, 400\}$, and four distinct values of $S \in \{20, 90, 40, 180\}$. This procedure yields the values 0.55 and 0.90 for RISK, 0.17 and 0.38 for TEMPT, and 0.20 and 0.60 for EFF.

After reading the instructions (see Online Appendix A), participants completed two tasks presented on the same screen for each PD. First, they indicated their decision (cooperate or defect) with decisions neutrally labelled as options 'A' and 'B'. The labels were presented in a random order with randomization at the pair level to control for potential presentation effects (i.e., 'A' was the cooperative decision in some games but not in others).

Second, participants indicated their belief about the other person's decision by selecting the likelihood (between 0 and 100 percent) of the other player choosing option 'A'. We did not incentivize belief elicitation to avoid a potential hedging problem (Blanco et al. (2010)) that may occur when both choice task and belief elicitation are incentivized.[4]

To control for potential order effects, we randomized at the pair level the sequence in which the decision and belief elicitation tasks were displayed. To ensure that participants recognize the payoff changes and fully understand how all potential outcomes depend on decisions, participants had to answer eight game-specific control questions about how decisions affect own and other payoff. These questions had to be correctly answered before decisions and beliefs could be entered.

Participants did not receive feedback on the others' choices or the game outcomes until the end of the session. Once participants completed the tasks for all games, we asked them to complete a short post-experimental questionnaire. At the end of the session, one game was randomly chosen, at the pair level, for payment. Participants were reminded of their decisions and informed about the outcome for the randomly chosen game.

We ran our experiments online with two subject pools: students recruited from a volunteer database at the University of Nottingham (UoN, $n = 162$) and workers recruited on Amazon Mechanical Turk (AMT, $n = 160$). We did this because students are the typical subject pool for the experiments on PDs which inspired our study (see Section 2) and well suited for studying conceptual questions (see Gächter (2010)). However, given that students tend to be less cooperative than older people (e.g., Arechar et al. (2018); Gächter and Herrmann (2011); List (2004); Matsumoto et al. (2016); Praxmarer et al. (2024)), the question of generalizability of results arises: How robust are results on payoff variation for cooperation across subject pools with likely different levels of baseline cooperativeness? We used the same software (LIONESS Lab, Giamattei et al. (2020)) and near-identical instructions for both subject pools.

Because Study 1 was conducted online in both subject pools, we expected a non-negligible attrition rate during gameplay. We used the following procedure to determine payoffs considering potential dropouts. If both participants completed the entire experiment, they were paid according to the outcome of the randomly chosen game. If one of the pair had dropped out during the experiment, the computer randomly selected the payoff-relevant game and randomly selected one of the four monetary outcomes of the chosen game

[4] Another possibility would be to incentivize either the choice task or belief elicitation. This, however, would complicate the instructions making them more difficult to understand. Moreover, Trautmann and van de Kuilen (2015) find that unincentivized and incentivized elicitation perform equally well in terms of accuracy.

Table 4
Payoff parameters for Studies 1 and 2.

Game	T	R	P	S	RISK	TEMPT	EFF	Mean cooperation rates		
								Study 1		Study 2
								UoN	AMT	AMT
G1	600	500	200	90	0.55	0.17	0.60	0.49	0.59	0.61
G2	600	500	200	20	0.90	0.17	0.60	0.45	0.60	0.64
G3	800	500	200	90	0.55	0.38	0.60	0.36	0.47	0.59
G4	800	500	200	20	0.90	0.38	0.60	0.38	0.40	0.53
G5	600	500	400	180	0.55	0.17	0.20	0.38	0.50	0.47
G6	600	500	400	40	0.90	0.17	0.20	0.33	0.48	0.50
G7	800	500	400	180	0.55	0.38	0.20	0.28	0.45	0.56
G8	800	500	400	40	0.90	0.38	0.20	0.28	0.42	0.53

Notes: Payoffs in experimental currency.

for payment to the remaining participant. We explained this payment scheme in the instructions.

As we implemented real-time matching of participants in Study 1, we were concerned that decreasing attention might lead to prolonged waiting times. We took several measures to retain attention and encourage successful completion of the experiment. Before participants entered the experiment, we told them to avoid distractions during the experiment. In addition, participants who were inactive for more than 30 s (i.e., no mouse movement or no keyboard input) got an alert voice message and a blinking text on their browser. If an inactive participant did not respond to the alert message for a further 30 s, they were removed from the session so that the remaining participant could complete the experiment. Three participants (2 %) recruited from UoN and 39 of participants (24 %) recruited via AMT dropped out during the experiment. The relatively high attrition rate amongst participants recruited via AMT is consistent with similar interactive online experiments (Arechar et al. (2018)).

The sessions lasted for approximately 30 min, including the completion of a post-experimental questionnaire. Participants were informed of their payment immediately upon completion of the experiment and were paid within 24 h. Participants recruited at UoN earned on average £4.79 (SD = £2.33); Participants recruited via AMT earned on average $5.00 ($SD$ = $2.43), which amounts to an hourly wage of $10.[5] Further descriptive statistics and comparisons of our subject pools are in Online Appendix C.

3.2. Experimental design and procedures for between-subjects Study 2

For Study 2, we adapt the experimental design of Study 1 to a *between-subjects* design using the eight games of Study 1. The only difference from Study 1 is that each participant plays only one one-shot game randomly selected from G1 to G8 shown in Table 4. This experiment was pre-registered (AEARCTR-0,009,784).[6] The instructions were the same as for Study 1, except for the adaptation to one game play (see Instructions for Study 2 in Online Appendix A).

Based on a power calculation we aimed at recruiting 200 participants per game, that is, a total of 1600 participants.[7] Because these numbers are infeasible in the UoN laboratory and because our results from Study 1 are largely similar between UoN and AMT anyway (apart from higher baseline levels of cooperation in AMT – see Fig. 1) we ran Study 2 on AMT only.

1609 participants completed the experiment. The sessions lasted for approximately 15 min, including the completion of a post-experimental questionnaire. Participants were informed of their payment immediately upon completion of the experiment. Participants earned on average $3.13 ($SD$ = $0.92). Online Appendix C includes the full descriptive statistics.

4. Results

4.1. Results from the within-subject experiment (Study 1)

Cooperation. Across the eight games, cooperation rates vary from 0.28 to 0.49 in UoN and from 0.40 to 0.60 in AMT (see Table 4). On average, UoN participants cooperated in 2.96 of the 8 games, which is significantly lower than AMT participants who cooperated in 3.91 games (Mann-Whitney $Z = -2.86$, $p = 0.004$). This is consistent with previous studies, discussed above, that find lower levels of cooperative behavior across student than non-student subject pools. 67 % of UoN participants (70 % of AMT participants) were switchers, 25 % (17 %) always defected and 8 % (13 %) always cooperated.

The left panel of Fig. 1 illustrates the average cooperation rates in Study 1 in each of the eight PDs separately by payoff index and sample. Panels (*a*) and (*d*) show games connected by a line which only differ in their level of RISK. The line pattern illustrates the

[5] The hourly wage of $10 compares well to the federal minimum wage $7.25 at the time of the experiment. The results of Kocher et al. (2008) (in lab public goods games) and Amir et al. (2012) (in AMT public goods games and trust games) suggests that results are robust to higher stakes.

[6] For the details of preregistration, see https://www.socialscienceregistry.org/trials/9784.

[7] In Study 1, TEMPT emerged as the most important of the three indices in explaining cooperation. The cooperation rate under low TEMPT was 0.4 vs 0.6 under high TEMPT, which turned out to be the biggest effect size. Given this treatment difference, a 5% significance test of the equality of two proportions would have 95% power with a sample size of 160 per treatment. To account for heterogeneity on AMT, we planned to recruit 200 participants for each of the 8 games.

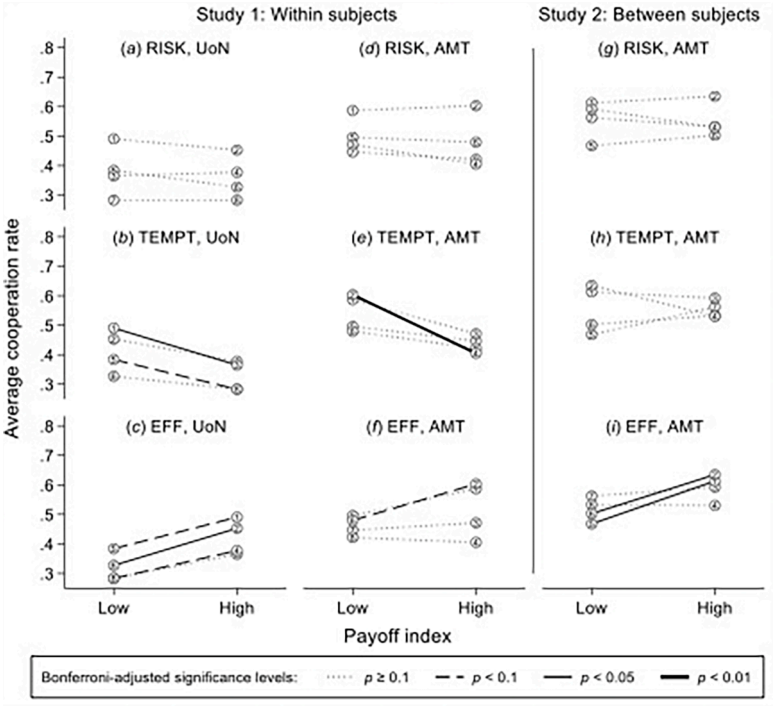

Fig. 1. Average cooperation rates in the eight Prisoner's Dilemma games of Study 1′s UoN sample (Panels *a-c*), Study 1′s AMT sample (Panels *d-f*) and Study 2 (Panels *g-i*). The line patterns indicate the Bonferroni-adjusted significance levels of two-sided McNemar's tests (Study 1) and Fisher's exact tests (Study 2). The game number is shown in the respective marker. See Online Appendix D, Table D1–D2 for the uncorrected *p*-values.

Bonferroni-adjusted significance levels of non-parametric McNemar tests. We find no significant differences in cooperation rates across low- and high-RISK games for any of the four possible pair-wise comparisons possible in either sample.

Panels (*b*) and (*e*) show games that differ only in their level of TEMPT connected by a line. For the UoN sample, we find a significantly lower cooperation frequency as TEMPT increases for two of the four comparisons possible. Similarly, the AMT sample includes one highly significant decrease in the cooperation rates as TEMPT increases. Finally, Panels (*c*), and (*f*) show games that differ only in their level of EFF connected by a line. The UoN sample provides strong evidence for a positive effect of EFF on cooperation as we find that three out of four comparisons show at least a weakly significant increase in the cooperation frequency as EFF increases. The AMT sample shows one weakly significant increase in the cooperation frequency as EFF increases. We will complement these results with a regression analysis reported below, but before we do so, we discuss how payoffs affect beliefs.

Beliefs. As beliefs have been identified as an important driver of cooperative behavior in similar games, such as the public good game (e.g., Croson (2007); Fischbacher and Gächter (2010); Gächter and Renner (2018)) and the sequential PD (e.g., Baader et al., 2024), we now examine how the variation in payoffs affects beliefs. Fig. 2 shows the average expected likelihood that the other player cooperates separately by payoff index and sample. On average, AMT participants held higher average cooperative beliefs than UoN participants (Mann-Whitney $Z = -2.44$, $p = 0.015$) but in terms of belief accuracy (average belief compared to cooperation rate) we find a weakly significantly higher accuracy in the UoN subject pool (for details see Online Appendix E, Table E1).

In Panels (*a*) and (*d*) games that differ only in their level of RISK, but not in TEMPT or EFF, are connected by a line. Beliefs across these two games are directly comparable. No clear effect of a change in RISK on average beliefs emerges, as average beliefs decrease in some games but increase in others. A series of non-parametric Wilcoxon signed-rank tests shows insignificant differences in the average beliefs in both the UoN and the AMT sample. Panels (*b*) and (*e*) illustrate pairs of games that only differ in TEMPT. Beliefs about the other player's cooperativeness decrease as TEMPT increases, but the effect is only marginally significant for one of the four game pairs in the UoN sample. Panels (*c*) and (*f*) show the pairs of games differing in EFF only. We find that an increase in EFF is associated with an increase in the average cooperative belief for almost all pairs of games. The difference between the low- and high-EFF games is highly significant for one game pair and significant for two of the game pairs in the UoN sample. For the AMT sample, we find highly significant differences for one of the four game pairs. The next step in our analysis is a regression analysis that controls for beliefs.

Regression results. In Table 5, we report the effect of payoff indices on cooperation and belief using linear (probability) models with participant random effects and robust standard errors clustered on participants separately for both samples. In all models, we control for the subject pool, individual characteristics, and task characteristics (i.e., the round in which the respective game was played, whether the decision task or belief task appeared at the top of the screen and labelling of cooperative choice as A or B). The full model results are in Online Appendix F, Table F1.

The models in Columns 1–2 show that the effect of RISK on cooperation is small in magnitude and insignificant in both samples.

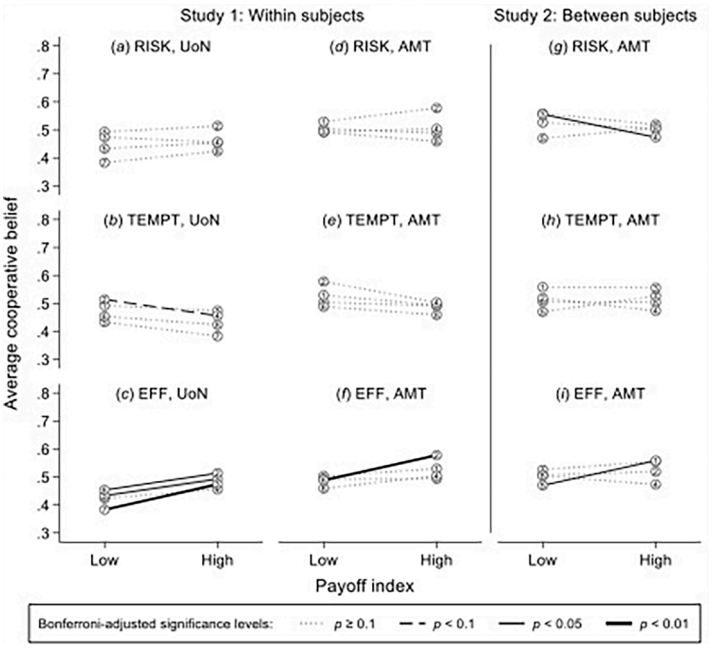

Fig. 2. . Average cooperative *beliefs* in the eight Prisoner's Dilemma games of Study 1′s UoN sample (Panels *a-c*), Study 1′s AMT sample (Panels *d-f*) and Study 2 (Panels *g-i*). The line patterns indicate the Bonferroni-adjusted significance levels of two-sided Wilcoxon signed-rank tests (Study 1) and Mann-Whitney tests (Study 2). The game number is shown in the respective marker. See Online Appendix D, Table D3–D4 for the uncorrected *p*-values.

Table 5
Payoff indices, beliefs, and cooperation in Studies 1 and 2.

	Within-subjects experiment (Study 1, models (1) to (6))						Between-subjects experiment (Study 2, models (7) to (9))		
Dependent variable:	(1) UoN Cooperation	(2) AMT Cooperation	(3) UoN Belief	(4) AMT Belief	(5) UoN Cooperation	(6) AMT Cooperation	(7) AMT Cooperation	(8) AMT Belief	(9) AMT Cooperation
RISK	−0.094	−0.073	0.001	−0.014	−0.094	−0.067	−0.012	−0.042	0.012
	(0.062)	(0.062)	(0.038)	(0.051)	(0.058)	(0.062)	(0.067)	(0.030)	(0.065)
TEMPT	−0.408***	−0.506***	−0.147**	−0.174**	−0.323***	−0.431***	−0.032	−0.029	−0.015
	(0.108)	(0.117)	(0.062)	(0.078)	(0.107)	(0.112)	(0.113)	(0.050)	(0.110)
EFF	0.245***	0.145**	0.155***	0.076*	0.155***	0.113	0.179***	0.048*	0.152***
	(0.058)	(0.073)	(0.033)	(0.042)	(0.059)	(0.072)	(0.059)	(0.026)	(0.057)
Belief					0.582***	0.434***			0.565***
					(0.061)	(0.068)			(0.053)
Control variables	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Constant	0.180	0.575***	0.476***	0.552***	−0.098	0.334**	0.588***	0.771***	0.153
	(0.193)	(0.155)	(0.098)	(0.105)	(0.164)	(0.139)	(0.094)	(0.047)	(0.097)
(Within) R^2	0.06	0.04	0.07	0.13	0.15	0.06	0.11	0.64	0.17
Obs. (Clusters)	1232 (154)	952 (119)	1232 (154)	952 (119)	1232 (154)	952 (119)	1601	1601	1601

Notes: Coefficients of a random effects linear probability model (Cols. 1–2, 5–6) or linear model (Cols. 3–4) with robust standard errors clustered on participants in parentheses. Coefficients from a linear probability model (Cols. 7, 9) or linear model (Col. 8) with robust standard errors in parentheses. The control variables are round, order of tasks, order of choices, age, gender, ethnicity, Business/Economics major (UoN only), spending/income, political attitude, and previous experience in experiments. Full estimation results are in Online Appendix F, Table F1-F2. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

TEMPT appears to be the most influential determinant of cooperation. The coefficients on TEMPT are negative, highly significant, and show a larger effect than EFF and RISK in both samples. An increase in TEMPT of 0.1 is associated with a 4.08 (5.06) percentage points lower probability of cooperating in the UoN (AMT) sample. EFF also appears as an influential determinant of cooperation (although the effect size is smaller than TEMPT). A 0.1 increase in EFF increases cooperation by 2.45 percentage points for UoN participants and 1.45 percentage points for AMT participants.[8]

In Columns 3–4, we estimate the effect of payoff indices on beliefs, which is an important co-variate of our behavioral outcome measure (UoN: $r_s = 0.48$, $p < 0.001$; AMT: $r_s = 0.41$, $p_s < 0.001$; Pooled samples). We find a significantly negative effect of TEMPT on belief across both samples. EFF positively affects beliefs in both samples, with a highly significant coefficient of EFF for UoN and a weakly significant and smaller coefficient for AMT.

The positive correlation of beliefs and cooperation is a common result in the literature on related social dilemma games (e.g., Dufwenberg et al. (2011)). It is consistent with experiments that causally manipulated beliefs (e.g., Frey and Meier (2004)) or held beliefs constant via the strategy method (e.g., Fischbacher and Gächter (2010); Gächter et al. (2022)). On the aggregate level, this can be taken as evidence for conditional cooperation, although this masks a substantial individual-level heterogeneity on the correlation between individual beliefs and behavior (see Online Appendix G for an illustration and discussion).

Columns 5–6 present results from the model which includes the payoff indices and beliefs as explanatory variables. For both samples, the coefficient of TEMPT and EFF are reduced in size when Belief is added to the model. For AMT, the effect of EFF even becomes statistically insignificant. This implies that the total effect of these two payoff indices on cooperation is (partially) mediated through beliefs.

More formally, we can decompose the total effect of the payoff indices into direct and indirect components via the mediator variable Belief using the method proposed by Baron and Kenny (1986).[9] See Online Appendix H for the details. For example, the total effect of TEMPT on cooperation in the UoN sample, comprising direct and indirect effects, is given by the highly significant and negative coefficient in Column 1 ($b = -0.408$, $p < 0.001$). The Baron and Kenny method proposes that the indirect effect through Belief can be approximated by multiplying the direct effect of TEMPT on the mediator Belief ($b = -0.147$, $p = 0.017$; Column 3) with the direct effect of the mediator Belief on cooperation ($b = 0.582$, $p < 0.001$; Column 5), yielding a significant negative indirect effect ($b = -0.086$, $p = 0.011$), which accounts for 21 % of the total effect in the UoN sample. For the AMT sample, we also find a significant indirect effect of TEMPT mediated through Belief ($b = -0.076$, $p = 0.018$), which accounts for 15 % for the total effect. Regarding the indirect effect of EFF mediated through Belief on behavior, we find a (highly) significant indirect effect in both samples (UoN: $b = 0.090$, $p < 0.001$; AMT: $b = 0.033$, $p = 0.043$). In UoN the indirect effect accounts for 37 % of the total effect and in AMT it accounts for 23 %.

All regressions include task characteristics and individual characteristics as controls. The individual characteristics are generally insignificant (see Online Appendix F, Table F1 for details). Round is significantly negative (except in model (6)) despite no feedback between games. This is consistent with "virtual learning" (Weber (2003)) that has also been observed in public goods games (e.g., Neugebauer et al. (2009)).

As a final step in our analysis of Study 1, we take advantage of the within-subject nature of the data and examine the consistency of cooperative behavior across the two different levels of a payoff index. We evaluate consistency using an assumption of (weak) monotonicity: for instance, someone who cooperates under high TEMPT should cooperate under low TEMPT. To count the number of violations in monotonicity, we compare twelve pairs of games: 4 pairs which only differ in RISK, 4 pairs which only differ in TEMPT, and 4 pairs which only differ in EFF. For RISK (TEMPT), cooperating in a higher RISK (TEMPT) game but defecting in a lower RISK (TEMPT) game holding other payoffs indices constant is counted as a violation of monotonicity. For EFF, cooperating in a low EFF game but defecting in a high EFF game holding other payoff indices constant is counted as a violation.

Fig. I1 in Online Appendix I shows the number of violations by each payoff index. In both UoN and AMT samples, participants violate monotonicity assumptions at least once at the following rates:

- RISK: 37 % (43 %) of UoN (AMT) participants.
- TEMPT: 31 % (32 %) of UoN (AMT) participants.
- EFF: 30 % (36 %) of UoN (AMT) participants.

There are no systematic subject pool differences in the degree of monotonicity violations for any of the payoff indices (Fisher's exact tests, all $p \geq 0.304$). These results imply that in the UoN sample the findings of Fig. 1 and Table 5 are not due to systematic and robust index-specific inconsistencies in behavior. In the AMT sample the higher rate of non-monotonic choices in RISK compared to TEMPT and EFF might, however, have contributed to insignificant results in RISK.

[8] We also ran the regressions including a high EFF dummy interacted with RISK and TEMPT to examine whether there was a differential effect of RISK and/or TEMPT across high versus low EFF games. For the AMT sample we find that the effect of TEMPT is stronger in the high EFF games. We find no differential effect of TEMPT in the UoN sample and no differential effect of RISK in either sample. See Online Appendix F, Table F3 for details. We also ran the regressions without individual characteristics and the results are qualitatively unchanged.

[9] While this is a frequently used methodology, it is important to acknowledge that it rests on relatively strong assumptions of linear models and the absence of confounding effects between the mediator and outcome variable (for a discussion of mediation analysis in economics, see for example, Celli (2022)).

4.2. Discussion of Study 1

RISK does not have a significant impact on cooperation in Study 1. In contrast, TEMPT has a highly significant negative effect on cooperation and EFF has a significant positive effect on cooperation. In addition, we find similar effects of TEMPT and EFF on beliefs. Beliefs appear as a partial mediator of the effect of TEMPT and EFF, accounting for a substantial share of the total effect. These results are observed in both subject pools, except for the significant effect of EFF on cooperation in the AMT sample disappearing after controlling for beliefs.

One concern about these results is that they may be sensitive to the within-subject design nature of Study 1. The within-subject payoff variations might have allowed participants to learn through enhanced experience in game play or induced participants to change their decisions either because of a perceived experimenter demand effect ("payoffs changed, so I should change my decisions too") or because changing payoffs makes them more salient (for a discussion of within- vs. between-subjects designs see Charness et al. (2012)). To address these issues, we designed Study 2 where participants played only one game, and games varied between subjects.

4.3. Results from the between-subjects experiment (Study 2)

Results on cooperation. Across the eight games, cooperation rates vary from 0.47 to 0.64. Fig. 1 panels (*g*)-(*i*) illustrate the average cooperation rates in each of the eight PDs by payoff index. We find no significant differences in cooperation rates across low- and high-RISK games for any of the four possible pair-wise comparisons. The same is true when comparing low- and high-TEMPT games. We do find significant differences in cooperation rates across low- and high-EFF games for two pair-wise comparisons. In both pair-wise comparisons, cooperation rates increase as EFF increases.

Results on beliefs. Fig. 2 panels (*g*)-(*i*) shows the average expected likelihood that the other player chooses 'cooperate' separately for each payoff index. Beliefs about other player's cooperativeness decrease as RISK increases: the effect is significant at the 5 % level for one of the four pairs. TEMPT has an ambiguous effect on beliefs as we observe an unclear pattern between beliefs and TEMPT. Increasing EFF strengthens the beliefs about other's cooperativeness: the effect of EFF is significant at the 5 % level for one of the four pairs.

Regression results. Next, in Table 5 Columns 7–9, we report the effect of payoff indices using linear (probability) models. Again, we focus on payoff indices RISK, TEMPT, and EFF, and relegate the full regression results to Online Appendix F, Table F2. The analysis parallels Study 1 but is adapted to the strict one-shot nature of the data.

Column 7 reveals a positive and highly significant effect of EFF on cooperation. The coefficients for RISK and TEMPT are not significantly different from zero. Similarly, Column 8 indicates a positive and weakly significant effect of EFF on Belief, while RISK and TEMPT are not significantly different from zero.[10] Cooperation and beliefs again appear highly significantly correlated ($r_s = 0.38$, $p < 0.001$). Column 9 presents the results of the model that includes the three payoff indices and Belief. The coefficient for EFF is highly significant but smaller compared to that reported in Column 7. The coefficient for Belief is highly significant, positive, and similar in size compared to Study 1. A mediation analysis reveals that the significant total effect of EFF on cooperation comprises a significant indirect effect through Belief ($b = 0.028$, $p = 0.035$), which accounts for 15 % of the total effect.

4.4. Discussion of Study 2

Study 2 tested the role of payoff parameters across eight one-shot PDs in between-subject experiments (with $n \approx 200$ per game). This avoids potential learning through enhanced experience in game play or experimenter demand effects and salience effects that come from within-subject variation in payoffs. As in the preliminary experiment and Study 1, a higher EFF results in a higher cooperation rate, and RISK is insignificant. In Study 2, unlike in Study 1, TEMPT is insignificant. Overall, in our studies of one-shot PDs, EFF robustly influences cooperation in both within- and between-subject designs.

An important question is how sensitive our results are to the specific indices that we use. In the next section, we analyze three related indices. These indices are normalized loss and normalized gain, which are akin to RISK and TEMPT, and the K-index, which resembles EFF.

5. Related payoff indices: normalized loss, normalized gain, and K-index

In this section, we report evidence on related payoff indices that have received attention in the experimental PD literature. Unlike our payoff indices, these are defined on three or four payoff parameters, as will become clear below.

5.1. Normalized loss and normalized gain

A game's payoff matrix can be normalized by subtracting P from all payoffs of the PD payoff matrix (see Table 1) and dividing by $R - P$ (see, e.g., Stahl (1991); Dal Bó and Fréchette (2018); Embrey et al. (2018)). This yields a payoff for mutual cooperation of 1 and a

[10] The relatively high R^2 in this regression model is particularly noteworthy. Online Appendix F, Table F2 reveals that the only highly significant control variable is the labelling of strategies. Taken together, this suggests that most participants in Study 2 expected the other player to choose the strategy that was labelled as A, independent of the variation in payoff parameters.

payoff for mutual defection of 0 in the normalized payoff matrix. Thus, the game's efficiency—defined as the payoff difference between mutual cooperation R and mutual defection P—is set to 1. Normalized loss is given by $\frac{S-P}{R-P} = -l$ and therefore $l = \frac{P-S}{R-P}$, which captures the risk of cooperation against a defector (similar to the RISK index but normalized by $R - P$ instead of P). Normalized gain is defined as $\frac{T-P}{R-P} = 1 + g$ which implies that $g = \frac{T-R}{R-P}$, that is, g measures the gain from defecting against a cooperator (similar to the TEMPT index but normalized by $R - P$ instead of R). Across our eight games, l and g vary orthogonally within the sets of four low/high-EFF games (see Online Appendix J for a summary and illustration).

Table 6 reports regression results focusing on the normalized payoff indices for *low-EFF* games, with the full results including controls reported in Online Appendix J, Table J2-J3. In Study 1, l has no statistically significant effect on cooperation in either the UoN or AMT samples, while g has a highly significant negative effect on cooperation in the UoN sample only (Col. 1–2). Similarly, we do not find a statistically significant effect of l on belief for either sample and we find a significant negative effect of g on belief in UoN only (Col. 3–4). The model that includes belief as an explanatory variable reveals weakly significant negative effects of l and g on cooperation for UoN only, and a highly significant positive effect of Belief for both samples (Col. 5–6). The total effect of g on cooperation shown in Column 1 comprises a 33 % indirect effect mediated through belief which is negative and highly statistically significant ($b = -0.013$, $p = 0.009$). In Study 2, we find no significant effects of l or g on cooperation or beliefs (Col. 7–8). However, the full model including Belief as an explanatory factor shows a highly significant positive effect of belief on cooperation (Col. 9).

Table 7 reports regression results for the *high-EFF* games (see Online Appendix J, Table J4-J5 for the full results). In Study 1, l has no statistically significant effect and g has a highly significant negative effect on cooperation across both samples (Col. 1–2). Again, we do not find a statistically significant effect of l on belief for either the UoN or AMT sample. g has a significant negative effect on belief in the AMT sample only (Col. 3–4). When estimating the model which includes belief as an explanatory variable, we find no statistically significant effects of l, but we do find highly significant negative effects of g on cooperation in the UoN and AMT samples. The coefficient for belief is highly significant and positive for both samples (Col. 5–6). The total effect of g on cooperation shown in Col. 1–2 can be decomposed in direct and indirect effects. For the UoN sample, the indirect effect mediated through beliefs is statistically insignificant ($b = -0.019$, $p = 0.158$). Yet for AMT, the total effect comprises a significant and negative 14 % indirect effect mediated through Belief ($b = -0.034$, $p = 0.019$). In Study 2, we only find a weakly significant negative effect of g on cooperation and no significant effect of the normalized payoff indices on belief (Col. 7–8). The full model, which includes belief, shows no significant effects of l or g but a highly significant positive effect of belief on cooperation (Col. 9). We find no evidence for a significant indirect effect of g on cooperation mediated through belief ($b = -0.007$, $p = 0.300$).[11]

Additional evidence for the role of normalized loss and gain on cooperation can be obtained from our preliminary experiment (see Online Appendix B, Fig. B2 and Table B4). Our regression analysis reveals highly significant negative effects of both, normalized loss and gain on cooperation ($b = -0.011$, $p < 001$; $b = -0.012$, $p < 001$; resp.). Note, however, that the 15 PD games included in the pretest do not provide an orthogonal variation in normalized loss and gain.

5.2. A summary index of PD parameters: The K-index

Recall that the K-index is defined as $\frac{R-P}{T-S}$ (Rapoport (1967)). It is based on all four PD payoff parameters, and it captures the gains from mutual cooperation over mutual defection, $R - P$, relative to the range of payoffs, $T - S$. Because $T > R > P > S$, the K-index $\in (0, 1)$. For the K-index values of our games, see Table J1 in Online Appendix J.

The K-index is an index of cooperation: the higher the K-index, the more beneficial is mutual cooperation, that is, the lower is the conflict of interest between collective benefit and private gain (see also Balliet and Van Lange (2013)). We therefore expect cooperation to increase in the K-index, in line with previous literature (for recent meta-analyses of PD games using the K-index, see, e.g., Balliet and Van Lange (2013); Thielmann et al. (2020); Yuan et al. (2022); and Spadaro et al. (2022)).

The K-index is interesting because it is a summary index of the severity of the cooperation problem. But for our purposes, the K-index analysis that follows below also serves as a robustness check for EFF, which shares the same numerator, $R - P$, with the K-index.

Fig. 3 shows how the K-index of a game is related to the average cooperation rate, separately for each study and subject pool. In line with the previous literature, we see that the K-index and cooperation are positively related in all studies and all subject pools, although with some interesting differences between them.

- In Studies 1 and 2, the K-index is between 0.13 and 0.59. Interestingly, for all eight levels of the K-index, cooperation rates are higher in the AMT subject pool, where cooperation rates range from 0.40 to 0.60, whereas in the UoN subject pool, they range from 0.28 to 0.49. Cooperation rates are positively correlated with the K-index: this correlation is highly significant for UoN participants, whereas it is marginally insignificant for AMT participants (UoN: $r_s = 0.90$, $p = 0.002$; AMT: $r_s = 0.62$, $p = 0.102$).
- In Study 2, which only used AMT participants, cooperation rates range from 0.47 to 0.61, and the correlation of cooperation rates and the K-index is similar to Study 1 for AMT participants: $r_s = 0.59$, $p = 0.120$.

In the pooled dataset, disregarding study, and subject pool, we have 24 distinct average cooperation rates. Here, the Spearman

[11] An alternative way to jointly test the effect of normalized indices and beliefs on behavior is to create a composite index of these factors. Online Appendix J, Table J6 shows that cooperation behavior is jointly affected by the games' incentives as captured by the normalized indices and expected behavior in others.

Table 6
Normalized loss l, normalized gain g, beliefs, and cooperation in low-EFF games.

	Within-subjects experiment (Study 1, models (1) to (6))						Between-subjects experiment (Study 2, models (7) to (9))		
Dependent variable:	(1) Study 1: UoN Cooperation	(2) Study 1: AMT Cooperation	(3) Study 1: UoN Belief	(4) Study 1: AMT Belief	(5) Study 1: UoN Cooperation	(6) Study 1: AMT Cooperation	(7) Study 2 Cooperation	(8) Study 2 Belief	(9) Study 2 Cooperation
Normalized loss l	−0.030	−0.011	0.008	−0.017	−0.034*	−0.002	−0.004	−0.003	−0.002
	(0.020)	(0.022)	(0.013)	(0.018)	(0.019)	(0.023)	(0.024)	(0.010)	(0.023)
Normalized gain g	−0.039***	−0.025	−0.023**	−0.010	−0.026*	−0.020	0.021	−0.002	0.022
	(0.014)	(0.017)	(0.009)	(0.012)	(0.014)	(0.016)	(0.017)	(0.007)	(0.016)
Belief					0.571***	0.529***			0.512***
					(0.082)	(0.083)			(0.077)
Control variables	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Constant	0.178	0.345**	0.464***	0.499***	−0.090	0.074	0.536***	0.753***	0.150
	(0.223)	(0.173)	(0.112)	(0.121)	(0.190)	(0.146)	(0.128)	(0.068)	(0.133)
(Within) R^2	0.05	0.01	0.08	0.10	0.12	0.04	0.12	0.64	0.16
Obs. (Clusters)	616 (154)	476 (119)	616 (154)	476 (119)	616 (154)	476 (119)	802	802	802

Notes: Coefficients of a random effects linear probability model (Cols. 1–2, 5–6) or linear model (Cols. 3–4) with robust standard errors clustered on participants in parentheses. Coefficients from a linear probability model (Cols. 7, 9) or linear model (Col. 8) with robust standard errors in parentheses. The control variables are round, order of tasks, order of choices, age, gender, ethnicity, Business/Economics major (UoN only), spending/income, political attitude, and previous experience in experiments. Full estimation results are in Online Appendix J, Table J2-J3. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 7
Normalized loss l, normalized gain g, beliefs, and cooperation in high-EFF games.

	Within-subjects experiment (Study 1, models (1) to (6))						Between-subjects experiment (Study 2, models (7) to (9))		
Dependent variable:	(1) Study 1: UoN Cooperation	(2) Study 1: AMT Cooperation	(3) Study 1: UoN Belief	(4) Study 1: AMT Belief	(5) Study 1: UoN Cooperation	(6) Study 1: AMT Cooperation	(7) Study 2 Cooperation	(8) Study 2 Belief	(9) Study 2 Cooperation
Normalized loss l	−0.106	−0.163	−0.046	0.052	−0.075	−0.184	−0.010	−0.104	0.052
	(0.134)	(0.140)	(0.083)	(0.089)	(0.123)	(0.145)	(0.142)	(0.064)	(0.138)
Normalized gain g	−0.133***	−0.237***	−0.029	−0.078**	−0.114**	−0.204***	−0.083*	−0.012	−0.076
	(0.047)	(0.056)	(0.028)	(0.033)	(0.047)	(0.054)	(0.050)	(0.022)	(0.048)
Belief					0.668***	0.434***			0.604***
					(0.071)	(0.092)			(0.075)
Control variables	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Constant	0.354*	0.845***	0.601***	0.643***	−0.049	0.567***	0.791***	0.812***	0.301**
	(0.209)	(0.165)	(0.105)	(0.110)	(0.188)	(0.160)	(0.121)	(0.056)	(0.132)
(Within) R^2	0.07	0.09	0.04	0.20	0.18	0.10	0.12	0.65	0.18
Obs. (Clusters)	616 (154)	476 (119)	616 (154)	476 (119)	616 (154)	476 (119)	799	799	799

Notes: Coefficients of a random effects linear probability model (Cols. 1–2, 5–6) or linear model (Col. 3–4) with robust standard errors clustered on participants in parentheses. Coefficients from a linear probability model (Cols. 7, 9) or linear model (Col. 8) with robust standard errors in parentheses. The control variables are round, order of tasks, order of choices, age, gender, ethnicity, Business/Economics major (UoN only), spending/income, political attitude, and previous experience in experiments. Full estimation results are in Online Appendix J, Table J4-J5. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

correlation is $r_s = 0.46$, $p < 0.023$. A simple OLS regression of cooperation rate on K-index returns a coefficient for the K-index of 0.309 (with a 95 % CI of [0.081, 0.536]; $R^2 = 0.26$), which is slightly lower than the estimated coefficient (0.44) of the K-index in the meta-analysis of Yuan et al. (2022) (see their Table 3). Thus, overall, a PD's K-index predicts its average cooperation rate .

Table 8 reports the effect of the K-index on cooperation and beliefs. In Study 1, we find a positive and highly significant effects on cooperation across the UoN and AMT samples (Col. 1–2) as well as positive and highly significant effects on belief for both samples (Col. 3–4). The model that includes belief as an explanatory variable reveals positive and highly significant effects of both, the K-index and belief, for the UoN and AMT samples (Col. 5–6).

Decomposing the total effects of the K-index on cooperation shown in Columns 1–2 reveals that positive and highly significant indirect effects mediated through beliefs account for 30 % of the total effect in the UoN sample and 19 % of the total effect in the AMT sample (UoN: $b = 0.116$, $p < 0.001$; AMT: $b = 0.055$, $p = 0.009$).

Similarly, we find a highly significant positive effect of the K-index on cooperation and belief in Study 2 (Col. 7–8). For the full model, including belief, the coefficient for the K-index is highly significant and positive albeit somewhat reduced in size. The coefficient for belief is also positive and highly significant (Col. 9). The total effect of the K-index on belief comprises a 17 % positive and

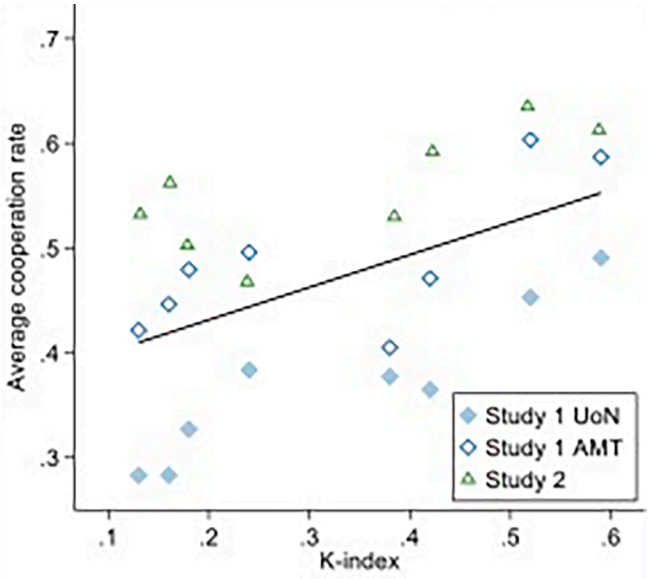

Fig. 3. Average cooperation rates for each level of a game's K-index, by study and subject pool. *Note*: The black line shows the predicted values from a linear regression.

Table 8
K-index, beliefs, and cooperation.

	Within-subjects experiment (Study 1, models (1) to (6))						Between-subjects experiment (Study 2, models (7) to (9))		
Dependent variable:	(1) Study 1: UoN Cooperation	(2) Study 1: AMT Cooperation	(3) Study 1: UoN Belief	(4) Study 1: AMT Belief	(5) Study 1: UoN Cooperation	(6) Study 1: AMT Cooperation	(7) Study 2 Cooperation	(8) Study 2 Belief	(9) Study 2 Cooperation
K-index	0.382***	0.297***	0.199***	0.126***	0.265***	0.242***	0.228***	0.069**	0.189***
	(0.073)	(0.084)	(0.041)	(0.049)	(0.072)	(0.083)	(0.071)	(0.032)	(0.070)
Belief					0.583***	0.440***			0.564***
					(0.061)	(0.068)			(0.053)
Control variables	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Constant	−0.025	0.342**	0.435***	0.482***	−0.279*	0.127	0.570***	0.730***	0.158**
	(0.180)	(0.140)	(0.092)	(0.096)	(0.151)	(0.123)	(0.073)	(0.038)	(0.078)
(Within) R^2	0.06	0.03	0.07	0.13	0.14	0.06	0.11	0.64	0.17
Obs. (Clusters)	1232 (154)	952 (119)	1232 (154)	952 (119)	1232 (154)	952 (119)	1601	1601	1601

Notes: Coefficients of a random effects linear probability model (Cols. 1–2, 5–6) or linear model (Cols. 3–4) with robust standard errors clustered on participants in parentheses. Coefficients from a linear probability model (Col. 7, 9) or linear model (Col. 8) with robust standard errors in parentheses. The control variables are round, order of tasks, order of choices, age, gender, ethnicity, Business/Economics major (UoN only), spending/income, political attitude, and previous experience in experiments. Full estimation results are in Online Appendix J, Table J7-J8. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

significant indirect effect mediated through belief ($b = 0.039$, $p = 0.018$).

Additional support for the role of the K-index in explaining cooperation comes from the analysis of our preliminary experiment (see Online Appendix B, Fig. B3 and Table B4), which reveals a highly significant and positive effect of the K-index on cooperation ($b = 0.442$, $p < 001$).

5.3. Discussion

In this section we have investigated the robustness of our conclusions by looking at three closely related payoff indices: normalized loss, which resembles RISK; normalized gain, which resembles TEMPT; and the K-index, which resembles EFF. These indices all share the same numerator with our respective indices but have different denominators.

The results based on these alternative payoff indices largely confirm the findings based on RISK, TEMPT and EFF. In neither Study 1 nor 2 do we find a significant effect of normalized loss (akin to RISK) on cooperation. We do however find some evidence that normalized gain matters for cooperation, particularly in games with high efficiency. These results should be interpreted with caution as

we did not design the experiments for a controlled variation of the normalized indices and thus the variation in normalized loss and gain is larger in high-efficiency games.

6. Towards an explanation of our results

How can social preferences explain our results on the positive impact for cooperation of EFF (and the K-index) in all experiments? Why does TEMPT matter in the within-subject study but not in the between-subject study? Why does RISK never matter? A full-fledged formal analysis of what theories of social preferences predict in our games is beyond the scope of this paper. Instead, we use the basic psychological motives incorporated in the various theories and their experimental tests as likely sources of psychological considerations that players of prisoner's dilemma games might entertain. Our answers are inevitably somewhat speculative because we did not set up the experiments to test a particular theory (unlike, e.g., the horserace conducted by Miettinen et al. (2020)). Participants in one-shot games are also unlikely to be performing a full-fledged strategic analysis of the games they play but rather employ heuristics based on the comparative attractiveness of various possible outcomes (see, e.g., the approach of Stewart et al. (2016) and Lugrin et al. (2024) who use eye-tracking methods in 2×2 games). In the following we discuss considerations that might guide cooperation decisions in our experiments.

In abstract, anonymous games with monetary payoffs, like in ours, a likely consideration for many people is based on their distributional preferences. Many people are *inequality averse* both when it is to their advantage and when it is to their disadvantage (see Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) for the theoretical arguments and supporting evidence, and Blanco et al. (2011) and Beranek et al. (2015) for empirical estimates on inequality aversion parameters). Inequality aversion renders strategy combinations resulting in equal payoffs $[(R,\ R)$ and $(P,\ P)]$ somewhat more attractive (and thereby 'focal' or 'salient' for inequality averse people) than strategies resulting in unequal payoffs $[(T,S)$ and $(S,T)]$. Combined with the fact that $R > P$, inequality aversion and preferring more money over less money makes mutual cooperation attractive, and this attraction increases the larger $R - P$ (and hence EFF) is. Many people's distributional preferences also contain *preferences for efficiency,* whereby people are willing to incur some cost to maximize payoffs (e.g., Charness and Rabin (2002); Engelmann and Strobel (2004)). Therefore, for efficiency-concerned people, cooperation becomes more likely as EFF increases.

The attractiveness of EFF is further reinforced by a range of well-established motives beyond distributional preferences whose relevance likely also increases as EFF increases. The following motives are also likely to positively influence beliefs about the likelihood of cooperation by a player's opponent thereby further increasing the likelihood of cooperation:

- *Warm glow, altruism*, and *Kantian morality*, according to which people derive some utility from the act of cooperating (Andreoni (1995); Palfrey and Prisbrey (1997)) and cooperating is "the right thing to do" (e.g., Alger and Weibull (2013));
- *Team reasoning,* the idea that players view their opponent as a team member and play the action that maximizes team payoffs, which prescribes mutual cooperation (e.g., Bacharach (2006));
- *Magical thinking,* which is a belief that one's own act of cooperation makes cooperation by the opponent more likely (Shafir and Tversky (1992); Daley and Sadowski (2017));
- *Reciprocity, guilt aversion, and conditional cooperation* by which people are more likely to cooperate if they expect others to cooperate and believe their opponent expects them to cooperate (e.g., Guttman (1986); Sugden (1984); Dufwenberg et al. (2011); Fischbacher and Gächter (2010)).

Why does TEMPT matter in the within-subject experiments of Study 1 but not in the between-subjects experiments of Study 2? A candidate behavioral explanation is related to *salience*. A stimulus is salient if it automatically attracts a decision maker's attention and one source of salience is contrast with surroundings (see Bordalo et al. (2022) for a review of the literature). In the within-subject experiments of Study 1 participants played eight games with changing parameters (T, S, P relative to a fixed R) thereby creating contrasts that made changes in TEMPT salient, whereas in the between-subjects experiments of Study 2 participants only played one game with a given TEMPT parameter (and hence no contrast due to change). This means that the stimulus of TEMPT attracts more attention, and hence is more salient, in Study 1 than in Study 2. Because TEMPT is an appeal to one's self-interest, TEMPT is more likely to enter players' considerations when it is salient, that is, in Study 1 and less likely in Study 2.

RISK has no significant impact on cooperation in any of our three experiments. A likely reason is that for RISK to become relevant, people need to believe that their opponent is likely to defect in which case most people want to defect anyway.

7. Summary

The Prisoner's Dilemma occupies a place of fundamental importance in social science research on cooperation as it represents the simplest setting in which individual and collective interests diverge. An extensive body of experimental research uses money payoffs to generate games where individuals maximize their own earnings by defecting, while combined earnings are maximized by cooperating. This research shows that many individuals cooperate, even in one-shot games, but nevertheless the literature offers an incomplete account of how the money payoffs affect cooperation.

In this paper we examine the separate influences of the unilateral incentives to defect and the efficiency gains from cooperation. Following Mengel (2018), we designed experiments to examine the index of RISK to measure the incentive to defect against a defector, the index of TEMPT to measure the incentive to defect against a cooperator, and the index of EFF to measure the efficiency gains from cooperation. To probe the robustness of our results, we also analyze our data by using three related payoff indices: normalized loss

(closely related to RISK); normalized gain (closely related to TEMPT), and the K-index (closely related to EFF). These related payoff indices share the same numerator with our respective index but have different denominators.

We find that (i) RISK and normalized loss do not influence cooperation systematically in any of our experiments; (ii) TEMPT and normalized gain reduce cooperation in our within-subject experiment of Study 1, but not in our between-subject Study 2; and (iii) cooperation increases significantly with EFF and the K-index. Thus, in conclusion, a robust finding from our experiments is that the gains from mutual cooperation over mutual defection influence cooperation positively in one-shot Prisoner's Dilemmas.

Data availability

Data and analysis code are available at https://doi.org/10.17605/OSF.IO/MPRSC.

Acknowledgements

This work was supported by the European Research Council [grant numbers ERC-AdG 295707 COOPERATION and ERC-AdG 101020453 PRINCIPLES] and the Economic and Social Research Council [grant number ES/K002201/1]. Ethical approval for the experiments was obtained from the Nottingham School of Economics Research Ethics Committee. We are grateful to an Associate Editor and the referees whose comments have greatly improved our paper. We also thank conference participants in Amsterdam, and Colin Camerer, Robin Cubitt, Michalis Drouvelis, Matthew Embrey, José Guinot Saporta, Orestis Kopsacheilis, David K. Levine, Peter Moffatt, Arno Riedl, Chris Starmer, Robert Sugden, Fabio Tufano, Dennie van Dolder and especially Friederike Mengel for helpful comments and provision of her data. S.G. acknowledges the hospitality of briq Bonn while working on this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.euroecorev.2024.104753.

References

Ahn, T.K., Ostrom, E., Schmidt, D., Shupp, R., Walker, J.M., 2001. Cooperation in pd games: fear, greed, and history of play. Public Choice 106, 137–155.
Alger, I., Weibull, J.W., 2013. Homo moralis—Preference evolution under incomplete information and assortative matching. Econometrica 81, 2269–2302.
Amir, O., Rand, D.G., Gal, Y.a.K., 2012. Economic games on the internet: the effect of $1 stakes. PLoS One 7, e31461.
Andreoni, J., 1995. Warm glow versus cold prickle - the effects of positive and negative framing on cooperation in experiments. Q. J. Econ. 110, 1–21.
Arechar, A.A., Gächter, S., Molleman, L., 2018. Conducting interactive experiments online. Exp. Econ. 21, 99–131.
Au, W.T., Lu, S., Leung, H., Yam, P., Fung, J.M.Y., 2012. Risk and prisoner's dilemma: a reinterpretation of coombs' re-parameterization. J. Behav. Decis. Mak. 25, 476–490.
Bacharach, M., 2006. Beyond Individual Choice. Teams and Frames in Game Theory. Princeton University Press, Princeton.
Balliet, D., 2010. Communication and cooperation in social dilemmas: a meta-analytic review. J. Confl. Resolut. 54, 39–57.
Balliet, D., Parks, C., Joireman, J., 2009. Social value orientation and cooperation in social dilemmas: a meta-analysis. Group Process. Intergr. Relat. 12, 533–547.
Balliet, D., Van Lange, P.A.M., 2013. Trust, conflict, and cooperation: a meta-analysis. Psychol. Bull. 139, 1090–1112.
Beranek, B., Cubitt, R., Gächter, S., 2015. Stated and revealed inequality aversion in three subject pools. J. Econ. Sci. Assoc. 1, 43–58.
Blanco, M., Engelmann, D., Koch, A., Normann, H.T., 2010. Belief elicitation in experiments: is there a hedging problem? Exp. Econ. 13, 412–438.
Blanco, M., Engelmann, D., Normann, H.T., 2011. A within-subject analysis of other-regarding preferences. Games Econ. Behav. 72, 321–338.
Bolton, G.E., Ockenfels, A., 2000. ERC: a theory of equity, reciprocity, and competition. Am. Econ. Rev. 90, 166–193.
Bordalo, P., Gennaioli, N., Shleifer, A., 2022. Salience. Ann. Rev. Econ. 14, 521–544.
Celli, V., 2022. Causal mediation analysis in economics: objectives, assumptions, models. J. Econ. Surv. 36, 214–234.
Charness, G., Gneezy, U., Kuhn, M.A., 2012. Experimental methods: between-subject and within-subject design. J. Econ. Behav. Organ. 81, 1–8.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Q. J. Econ. 117, 817–869.
Charness, G., Rigotti, L., Rustichini, A., 2016. Social surplus determines cooperation rates in the one-shot prisoner's dilemma. Games Econ. Behav. 100, 113–124.
Cooper, R., DeJong, D., Forsythe, R., Ross, T., 1996. Cooperation without reputation: experimental evidence from prisoner's dilemma games. Games Econ. Behav. 12, 187–318.
Croson, R., 2007. Theories of commitment, altruism and reciprocity: evidence from linear public goods games. Econ. Inq. 45, 199–216.
Dal Bó, P., Fréchette, G.R., 2018. On the determinants of cooperation in infinitely repeated games: a survey. J. Econ. Lit. 56, 60–114.
Daley, B., Sadowski, P., 2017. Magical thinking: a representation result. Theor. Econ. 12, 909–956.
Dufwenberg, M., Gächter, S., Hennig-Schmidt, H., 2011. The framing of games and the psychology of play. Games Econ. Behav. 73, 459–478.
Embrey, M., Fréchette, G.R., Yuksel, S., 2018. Cooperation in the finitely repeated prisoner's dilemma. Q. J. Econ. 133, 509–551.
Engel, C., Zhurakhovska, L., 2016. When is the risk of cooperation worth taking? The prisoner's dilemma as a game of multiple motives. Appl. Econ. Lett. 23, 1157–1161.
Engelmann, D., Strobel, M., 2004. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. Am. Econ. Rev. 94, 857–869.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Q. J. Econ. 114, 817–868.
Fischbacher, U., Gächter, S., 2010. Social preferences, beliefs, and the dynamics of free riding in public good experiments. Am. Econ. Rev. 100, 541–556.
Flood, M.M., 1958. Some experimental games. Manage Sci. 5, 5–26.
Frank, R.H., Gilovich, T., Regan, D.T., 1993. Does studying economics inhibit cooperation? J. Econ. Perspect. 7, 159–171.
Frey, B.S., Meier, S., 2004. Social comparisons and pro-social behavior. Testing 'conditional cooperation' in a field experiment. Am. Econ. Rev. 94, 1717–1722.
Gächter, S., 2010. (Dis)advantages of student subjects: what is your research question? Behav. Brain Sci. 33, 92–93.
Gächter, S., Herrmann, B., 2011. The limits of self-governance when cooperators get punished: experimental evidence from urban and rural russia. Eur. Econ. Rev. 55, 193–210.
Gächter, S., Kölle, F., Quercia, S., 2022. Preferences and perceptions in provision and maintenance public goods. Games Econ. Behav. 135, 338–355.
Baader, M., Gächter, S., Lee, K., Sefton, M., 2024. Social preferences and the variability of conditional cooperation. CeDEx Discussion Paper 2024-04.

Gächter, S., Marino-Fages, D., 2023. Using the strategy method and elicited beliefs to explain group size and MPCR effects in public good experiments. IZA Discussion Paper 16605.
Gächter, S., Renner, E., 2018. Leaders as role models and 'belief managers' in social dilemmas. J. Econ. Behav. Organ. 154, 321–334.
Giamattei, M., Yahosseini, K.S., Gächter, S., Molleman, L., 2020. Lioness lab: a free web-based platform for conducting interactive experiments online. J. Econ. Sci. Assoc. 6, 95–111.
Guttman, J., 1986. Matching behavior and collective action. Some experimental evidence. J. Econ. Behav. Organ. 7, 171–198.
Kocher, M.G., Martinsson, P., Visser, M., 2008. Does stake size matter for cooperation and punishment? Econ. Lett. 99, 508–511.
List, J.A., 2004. Young, selfish and male: field evidence of social preferences. Econ. J. 114, 121–149.
Lugrin, C., Konovalov, A., Ruff, C.C., 2024. Facilitating cooperation by manipulating attention. PsyArXiv. 10.31234/osf.io/m62qp.
Matsumoto, Y., Yamagishi, T., Li, Y., Kiyonari, T., 2016. Prosocial behavior increases with age across five economic games. PLoS One 11, e0158671.
Mengel, F., 2018. Risk and temptation: a meta-study on prisoner's dilemma games. Econ. J. 128, 3182–3209.
Miettinen, T., Kosfeld, M., Fehr, E., Weibull, J., 2020. Revealed preferences in a sequential prisoners' dilemma: a horse-race between six utility functions. J. Econ. Behav. Organ. 173, 1–25.
Moisan, F., ten Brincke, R., Murphy, R.O., Gonzalez, C., 2018. Not all prisoner's dilemma games are equal: incentives, social preferences, and cooperation. Decision 5, 306–322.
Murnighan, J.K., Roth, A.E., 1983. Expecting continued play in prisoner's dilemma games:a test of several models. J. Confl. Resolut. 27, 279–300.
Neugebauer, T., Perote, J., Schmidt, U., Loos, M., 2009. Self-biased conditional cooperation: on the decline of cooperation in repeated public goods experiments. J. Econ. Psychol. 30, 52–60.
Ng, G.T.T., Au, W.T., 2016. Expectation and cooperation in prisoner's dilemmas: the moderating role of game riskiness. Psychon. Bull. Rev. 23, 353–360.
Palfrey, T.R., Prisbrey, J.E., 1997. Anomalous behavior in public goods experiments: how much and why? Am. Econ. Rev. 87, 829–846.
Praxmarer, M., Rockenbach, B., Sutter, M., 2024. Cooperation and norm enforcement differ strongly across adult generations. Eur. Econ. Rev. 162, 104659.
Rapoport, A., 1967. A note on the "index of cooperation" for prisoner's dilemma. J. Conflict. Resolut. 11, 100–103.
Rapoport, A., Chammah, A.M., 1965. Prisoners' Dilemma. A Study in Conflict and Cooperation. The University of Michigan Press, Ann Arbor.
Sally, D., 1995. Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992. Ration. Soc. 7, 58–92.
Schmidt, D., Shupp, R., Walker, J., Ahn, T.K., Ostrom, E., 2001. Dilemma games: game parameters and matching protocols. J. Econ. Behav. Organ. 46, 357–377.
Shafir, E., Tversky, A., 1992. Thinking through uncertainty: nonconsequential reasoning and choice. Cogn. Psychol. 24, 449–474.
Snowberg, E., Yariv, L., 2021. Testing the waters: behavior across participant pools. Am. Econ. Rev. 111, 687–719.
Spadaro, G., Graf, C., Jin, S., Arai, S., Inoue, Y., Lieberman, E., Rinderu, M.I., Yuan, M., Van Lissa, C.J., Balliet, D., 2022. Cross-cultural variation in cooperation: a meta-analysis. J. Pers. Soc. Psychol. 123, 1024–1088.
Stahl, D.O., 1991. The graph of prisoners' dilemma supergame payoffs as a function of the discount factor. Games Econ. Behav. 3, 368–384.
Stewart, N., Gächter, S., Noguchi, T., Mullett, T.L., 2016. Eye movements in strategic choice. J. Behav. Decis. Mak. 29, 137–156.
Sugden, R., 1984. Reciprocity: the supply of public goods through voluntary contributions. J. Econ. 94, 772–787.
Thielmann, I., Spadaro, G., Balliet, D., 2020. Personality and prosocial behavior: a theoretical framework and meta-analysis. Psychol. Bull. 146, 30–90.
Trautmann, S.T., van de Kuilen, G., 2015. Belief elicitation: a horse race among truth serums. Econ. J. 125, 2116–2135.
Van Lange, P.A.M., Balliet, D., Parks, C.D., Van Vugt, M., 2014. Social Dilemmas. The Psychology of Human Cooperation. Oxford University Press, Oxford.
Vlaev, I., Chater, N., 2006. Game relativity: how context influences strategic decision making. J. Exp. Psychol. Learn. Mem. Cogn. 32, 131–149.
Weber, R.A., 2003. Learning' with no feedback in a competitive guessing game. Games Econ. Behav. 44, 134–144.
Yuan, M., Spadaro, G., Jin, S., Wu, J., Kou, Y., Van Lange, P.A.M., Balliet, D., 2022. Did cooperation among strangers decline in the united states? A cross-temporal meta-analysis of social dilemmas (1956–2017). Psychol. Bull. 148, 129–157.