I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14895

Risk, Temptation, and Efficiency in the One-Shot Prisoner's Dilemma

Simon Gächter
Kyeongtae Lee
Martin Sefton
Till O. Weber

NOVEMBER 2021

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14895

Risk, Temptation, and Efficiency in the One-Shot Prisoner's Dilemma

Simon Gächter
University of Nottingham, IZA and CESifo Munich

Kyeongtae Lee
Bank of Korea

Martin Sefton
University of Nottingham

Till O. Weber
Newcastle University

NOVEMBER 2021

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Risk, Temptation, and Efficiency in the One-Shot Prisoner's Dilemma*

The prisoner's dilemma (PD) is arguably the most important model of social dilemmas, but our knowledge about how a PD's material payoff structure affects cooperation is incomplete. In this paper we investigate the effect of variation in material payoffs on cooperation, focussing on one-shot PD games where efficiency requires mutual cooperation. Following Mengel (2018) we vary three payoff indices. Indices of *risk* and *temptation* capture the unilateral incentives to defect against defectors and co-operators respectively, while an index of *efficiency* captures the gains from cooperation. We conduct two studies: first, varying the payoff indices over a large range and, second, in a novel orthogonal design that allows us to measure the effect of one payoff index while holding the others constant. In the second study we also compare a student and non-student subject pool, which allows us to assess generalizability of results. In both studies we find that temptation reduces cooperation. In neither study, nor in either subject pool of our second study, do we find a significant effect of risk.

JEL Classification: A13, C91
Keywords: prisoner's dilemma, cooperation, temptation, risk, efficiency

Corresponding author:
Simon Gächter
School of Economics
University of Nottingham
Sir Clive Granger Building
University Park
Nottingham NG7 2RD
United Kingdom
E-mail: simon.gaechter@nottingham.ac.uk

* 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 Colin Camerer, Robin Cubitt, Matthew Embrey, José Guinot Saporta, Orestis Kopsacheilis, David K. Levine, Peter Moffatt, Chris Starmer, Robert Sugden, Fabio Tufano, Dennie van Dolder and especially Friederike Mengel for helpful comments. We would also like to thank participants at the 2019 CCC meeting in Amsterdam.

1. Introduction

In many naturally occurring economic and social environments there is a conflict between individual and collective interests. The canonical model to represent such a conflict is the Prisoner's Dilemma (PD) and so it plays an important role in social science research and is the topic of a vast literature in economics, sociology, political science, and social psychology. A large experimental literature has shown 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)).[1] This literature has studied a wide variety of factors that affect cooperation (for surveys see, e.g., Balliet *et al.,* (2009); Van Lange *et al.,* (2014)), but perhaps from an economics perspective the most fundamental factor to consider is the material payoff structure. As we discuss in detail in Section 2, a surprisingly small literature has studied the effect of *variation* in the payoff matrix. In this paper, we provide, across two studies, a systematic analysis of the role of the material payoff structure for cooperation in one-shot PD games.

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.

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.

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 ($T > R$). This condition

[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, Mengel (2018) and Dal Bó and Fréchette (2018), respectively.

ensures that the dominant strategy for money maximising participants is to defect. Rapoport and Chammah (1965) impose a second condition, $2R > T + S$, to ensure that mutual cooperation maximises combined earnings. The remainder of the paper focuses on one-shot PDs that satisfy both conditions.

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 condenses a game's incentives into a single index based on all four elements of the payoff matrix. Following this, other indices have been developed that focus on different elements of the payoff matrix, such as the unilateral incentives to defect against co-operators ($T - R$) or defectors ($P - S$). Most recently, Mengel (2018) proposed new indices of RISK $\equiv \frac{P-S}{P}$, measuring the percentage loss from cooperating against a defector, and TEMPT $\equiv \frac{T-R}{T}$, measuring the percentage gain from defecting against a cooperator. These indices allow for an intuitive interpretation of the incentives to defect in terms of percentages and eliminate potential collinearity between the indices. In our paper we examine the role of material incentives on cooperation by focussing on the relationship between Mengel's indices and cooperation.

Our experiments are motivated by several observations about the previous literature (which we will discuss further 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. In our experiments we have participants play several games with different payoff

matrices, but we control for learning effects by giving no feedback between periods. We also control for order effects by randomising the order in which games are presented to participants.

Most relevant to our research is the recent paper by Mengel (2018). While few experiments examine controlled variation in payoffs, payoffs do differ considerably across studies and she takes advantage of this variation to conduct a meta-analysis of the roles of RISK and TEMPT, controlling for a third index of *efficiency*, EFF $\equiv \frac{R-P}{R}$. 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, as we show in Section 2, this result does not hold in a restricted sample of games in which mutual cooperation maximises combined earnings (i.e., imposing $2R > T + S$). In the restricted sample neither RISK nor TEMPT has a significant effect on cooperation after controlling for efficiency. 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 two new experimental studies. In Study 1, we run a lab experiment in which participants played 15 one-shot games that meet our two PD conditions while aiming for large variation in the RISK, TEMPT and EFF indices. Despite wide variation in these payoff indices, we find no evidence that cooperation is systematically related to RISK. In contrast, we find that cooperation is significantly higher when EFF is higher, and we also find some suggestive evidence that cooperation decreases with TEMPT. However, this design includes only a few instances where one index varies while the other two indices are held constant.

In Study 2, 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. 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 (Gächter and

Herrmann (2011); List (2004)). Study 2 allows us to test whether results based on student samples are transferable to a more diverse 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.

Our two studies together suggest that, in one-shot PDs where efficiency requires mutual cooperation, variation in TEMPT has a larger impact than RISK on cooperation. The remainder of this study is organised as follows. Section 2 reviews the related literature. Section 3 presents the experimental design, procedures and the results of Study 1. Section 4 details the same for Study 2. Section 5 concludes.

2. Related literature and our contribution

There is a vast experimental literature on PDs (for more recent 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 behaviour, but a complicated one in which to study the role of incentives. Embrey et al. (2018) and Mengel (2018) discuss the effect of payoffs on cooperation in finitely repeated PDs. The role of 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), which surveys 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 her 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 prisoner's dilemmas. 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 prisoner's dilemma games 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' ($\frac{T-R}{T-S}$) and 'fear' ($\frac{P-S}{T-S}$) 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 normalisation 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 ($\frac{R-S}{(R-S)+(T-P)}$), 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 payoff indices in our main study 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 she 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 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 Mengel (2018), Table 3, Col. 1. 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 Column 2, we restrict the sample to the 36 games in Mengel's dataset that meet *both* PD conditions ($T > R > P > S$ and $2R > T + S$). For the restricted sample, the effect of RISK on the average cooperation rate becomes smaller and statistically insignificant. The estimated effect of TEMPT is larger but still insignificant. The effect of EFF is positive and weakly significant. Thus, the meta-analytic results on the importance of RISK in explaining cooperation are sensitive to whether payoff parameters satisfy $2R > T + S$.

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 not entirely systematic. In our experiments we design the payoffs explicitly for comparing the effects of payoff indices. Furthermore, in Mengel's dataset experiments vary in numerous other respects not related to payoffs. For example, instructional materials and framing of the task vary. In our experiments we control these other factors that may affect cooperation by holding them constant within our design.

TABLE 2. Average cooperation rate regressed on payoff indices.

	(1) All games (Mengel 2018; Table 3, Col. 1)	(2) $T > R > P > S$ & $2R > T + S$
RISK	-0.255*** (0.060)	-0.045 (0.123)
TEMPT	0.003 (0.080)	-0.492 (0.305)
EFF	0.291*** (0.089)	0.301* (0.149)
Constant	0.370*** (0.084)	0.304** (0.130)
Adj. R^2	0.350	0.167
Obs.	73	36

Notes: Data from stranger matching and one-shot games included in Mengel (2018). All columns show OLS coefficients with standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

3. Study 1

3.1. Experimental design and procedures

For Study 1 and inspired by Simpson (2003) and Mengel (2018), we devised 15 games that meet the two standard PD conditions ($T > R > P > S$ and $2R > T + S$). We chose convenient non-negative payoff parameters to vary the RISK, TEMPT and EFF indices over a wide range yielding a low, medium and high level for each index.[2] Table 3 presents the payoff parameters. The two standard PD conditions and non-negative payoff parameters restrict the theoretically possible variation of the payoff indices such that RISK $\in$ (0, 1], TEMPT $\in$ (0, 0.5) and EFF $\in$ (0, 1). The implemented payoff parameters cover almost the entire possible range, with RISK varying from 0.04 to 1, TEMPT from 0.1 to 0.49 and EFF from 0.04 to 0.98. The K-index ranges from 0.02 to 0.88 across games. The design also includes several sets of games across which only one payoff parameter changes while holding the others constant. Games 1, 4 and 7 vary only in RISK. Games 2, 5 and 8 constitute a second set varying only in RISK. Games 10 and 11 vary only in TEMPT. Three sets of games vary only in EFF: Games 7, 10 and 13; Games 8, 11, 14; Games 9 and 12.

[2] The sessions included two further games that violate the standard PD conditions and are thus excluded from the analysis.

After reading the instructions (see Appendix A), participants were presented with a game's payoff matrix on the computer screen and indicated their decision (cooperate or defect). The decisions were neutrally labelled as options 'A' and 'B'. To control for potential order effects, we randomised the sequence of games at the pair level. Only after making their decisions for all games did participants learn the outcome of each PD. Participants then completed a short post-experimental questionnaire, and one of the games was randomly chosen, at the pair level, for payoff.

We ran our experiments with student participants at the University of Nottingham (two sessions, $n = 62$). The experiment was computerised and conducted with z-Tree (Fischbacher (2007)). Participants were recruited using ORSEE (Greiner (2015)). None of the participants took part in more than one session. The sessions lasted for approximately one hour and the average earnings (including a £3 show-up fee) were £11.86 (SD = £3.32). Participants were paid in cash at the end of the session.

TABLE 3. Payoff parameters for Study 1.

Game	T	R	P	S	RISK	TEMPT	EFF	K-index	Cooperation Rate
G1	12	10.8	4.8	4.6	0.04	0.10	0.56	0.81	0.65
G2	12	8.4	4.8	4.6	0.04	0.30	0.43	0.49	0.53
G3	9.8	5.6	3.2	3	0.06	0.33	0.52	0.50	0.60
G4	12	10.8	4.8	2.4	0.50	0.10	0.56	0.63	0.68
G5	12	8.4	4.8	2.4	0.50	0.30	0.43	0.38	0.56
G6	9.8	6.2	4.8	2.4	0.50	0.37	0.23	0.19	0.37
G7	12	10.8	4.8	0	1.00	0.10	0.56	0.50	0.55
G8	12	8.4	4.8	0	1.00	0.30	0.43	0.30	0.60
G9	9.8	5	4.8	0	1.00	0.49	0.04	0.02	0.37
G10	12	10.8	0.2	0	1.00	0.10	0.98	0.88	0.77
G11	12	8.4	0.2	0	1.00	0.30	0.98	0.68	0.66
G12	9.8	5	0.2	0	1.00	0.49	0.96	0.49	0.76
G13	12	10.8	8	0	1.00	0.10	0.26	0.23	0.44
G14	12	8.4	8	0	1.00	0.30	0.05	0.03	0.37
G15	8	6	4	2	0.50	0.25	0.33	0.33	0.56

Notes: Payoffs in £. Cooperation Rate is the average cooperation rate we find in our experiments. For a discussion see Section 3.2.

3.2. Results

Across the 15 games, cooperation rates vary from 0.37 to 0.77 (see Table 3). As observed in previous experiments, cooperation rates are positively correlated with the K-index (Spearman rank correlation, $r_s = 0.82$, $p < 0.001$). Of the participants, 81% were 'switchers' who altered their behaviour at least once over the 15 games, 8% always chose 'defect', and 11% always chose 'cooperate'. On average, participants chose 'cooperate' in 8.47 of the 15 games. This suggests that the large variation in payoff indices implemented over the 15 games induced substantial variation in game play, and thus the impact of each payoff index warrants further investigation.

Figure 1 plots the average cooperation rate in each of the 15 games against the respective RISK, TEMPT and EFF index. We find no significant association between the average cooperation rate and RISK ($r_s = -0.00$, $p = 0.992$) or TEMPT ($r_s = -0.27$, $p = 0.333$). However, the average cooperation rate is strongly positively and highly significantly correlated with EFF ($r_s = 0.92$, $p < 0.001$).

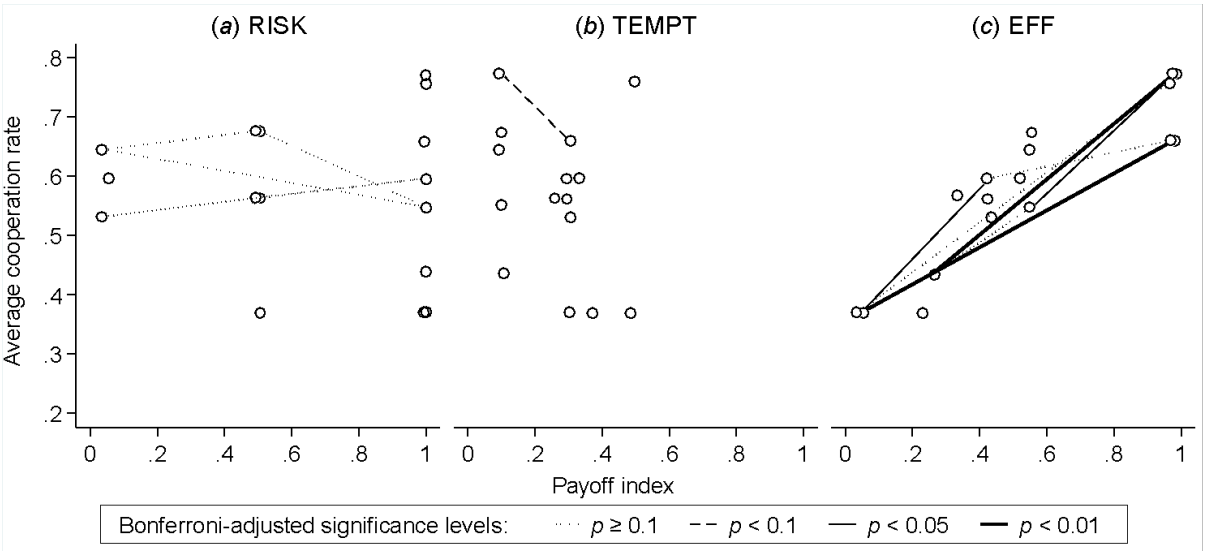


FIGURE 1. Average cooperation rates and payoff indices of the 15 Prisoner's Dilemma games in Study 1. The line patterns indicate the Bonferroni-adjusted significance levels of a two-sided McNemar's test in pairs of games that vary only in a single payoff index.

In Figure 1, pairs of games are connected by a line if one payoff index changes while the other two remain constant. The line pattern illustrates the Bonferroni-adjusted significance levels of a non-parametric McNemar's test for differences in the cooperation rates across a particular pair of games (see Appendix B for details). Panel (*a*) shows six pairs of games in which only RISK varies. We cannot reject the null hypothesis of equal cooperation rates across

any of the pairs. Panel (*b*) shows one pair of games that varies only in TEMPT, and we find a weakly significantly lower cooperation rate associated with higher TEMPT. Panel (*c*) indicates seven pairs of games differing in EFF only, and we find substantial evidence of an effect of this index on behaviour: we can (strongly) reject the null hypothesis of equal cooperation rates for five of the seven pair-wise comparisons possible.

Next, in Table 4, we report the effect of payoff indices on cooperation using a linear probability model with participant fixed effects. Robust standard errors are clustered on participants. The dependent variable is a cooperation dummy, and the explanatory variables are payoff indices (K-index, RISK, TEMPT, EFF) and the round in which the respective game was played.

TABLE 4. Determinants of cooperative choice in Study 1.

Dependent variable: cooperation dummy	(1)	(2)
K-index	0.442*** (0.065)	
RISK		-0.044 (0.036)
TEMPT		-0.083 (0.087)
EFF		0.399*** (0.060)
Round	-0.001 (0.003)	-0.001 (0.003)
Constant	0.386*** (0.031)	0.432*** (0.042)
BIC	768.0	767.9
Within R^2	0.084	0.097
Obs. (Clusters)	930 (62)	930 (62)

Notes: All columns show coefficients from a linear probability model with participant fixed effects. Robust standard errors clustered on participants in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

In Column 1, we find that the probability of cooperation increases with the K-index, which is consistent with previous experimental PD studies (e.g., Rapoport and Chammah (1965), Vlaev and Chater (2006)). A 0.1 point increase in K-index is associated with a 4.4 percentage points higher probability of choosing 'cooperate'. In Column 2, we examine the

effects of RISK, TEMPT, and EFF on cooperation. 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 4.0 percentage points higher probability of choosing 'cooperate'. We do not observe a significant effect of the round on cooperation in either Column 1 or 2.

Although the 15 games included in Study 1 managed to achieve a large variation in the payoff indices, this design has the drawback that the induced payoff variation 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. Our Study 2 addresses this limitation.

4. Study 2

4.1. Experimental design and procedures

For Study 2, we create different PDs by varying RISK, TEMPT and EFF orthogonally. This allows us to identify the effect of a single payoff index on behaviour while holding constant the other two. First, we fix a *low* and *high* level for each of three payoff indices. We then generated $2^3 = 8$ payoff matrices representing all possible variation of the two levels across the three payoff indices.

TABLE 5. Payoff parameters for Study 2.

Game	T	R	P	S	RISK	TEMPT	EFF	K-index	Cooperation Rate	
									UoN	AMT
G1	600	500	200	90	0.55	0.17	0.60	0.59	0.49	0.59
G2	600	500	200	20	0.90	0.17	0.60	0.52	0.45	0.60
G3	800	500	200	90	0.55	0.38	0.60	0.42	0.36	0.47
G4	800	500	200	20	0.90	0.38	0.60	0.38	0.38	0.40
G5	600	500	400	180	0.55	0.17	0.20	0.24	0.38	0.50
G6	600	500	400	40	0.90	0.17	0.20	0.18	0.33	0.48
G7	800	500	400	180	0.55	0.38	0.20	0.16	0.28	0.45
G8	800	500	400	40	0.90	0.38	0.20	0.13	0.28	0.42

Notes: Payoffs in experimental currency. Cooperation Rate is the average cooperation rate we find in our experiments. For a discussion of our results see Section 4.2.

For each payoff index, we fix the low value approximately at the 25th and the high value approximately at the 75th percentile from the 34 PDs included in the meta-analysis by Mengel (2018) that meet the standard PD conditions ($T > R > P > S$ and $2R > T + S$). Therefore, the low and high values reflect typical values used in previous studies (see Appendix C).[3] The payoffs are presented in Table 5. $R = 500$ is constant across all PDs, while our experiment has two distinct values of $T = \{600, 800\}$ and $P = \{200, 400\}$, and four distinct values of $S = \{20, 90, 40, 180\}$. 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. The K-index ranges from 0.13 to 0.59.

After reading the instructions (see Appendix D), participants completed two tasks presented on the same screen for each PD. First, they indicated their decision (cooperate or defect) with decision neutrally labelled as options 'A' and 'B'. The labels were presented in a random order with randomisation at the pair level to control for potential presentation effects (i.e., 'A' was the cooperative decisions for some games but not in others). Second, participants indicated their belief about the other person's decision by selecting their expected probability (between 0 and 100 percent) of the other player choosing option 'A'. To control for potential order effects, we randomised at the pair level the sequence in which the decision and belief elicitation tasks were displayed. To ensure that participants recognise 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 any 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 at the University of Nottingham (UoN, $n = 162$) and workers 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 theoretical 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)), the question of generalisability of results

[3] Slight variations from these quartiles occurred because we wanted to use payoffs that were strictly positive multiples of ten to give participants convenient round numbers.

arises: How robust are results on payoff variation for cooperation across subject pools with likely different levels of baseline cooperativeness?

We ran our experiments using the same software LIONESS Lab (Giamattei *et al.,* (2020)) and identical instructions for both subject pools. Because Study 2 was conducted online in both subject pools, we expected a non-negligible attrition rate. 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 for payment to the remaining participant. We explained this payment scheme clearly in the instructions.

As we implemented real-time matching of participants in Study 2, 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 seconds (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 seconds, 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 minutes, 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 hours. Participants recruited at UoN earned on average £4.79 (*SD* = £2.33); Participants recruited via AMT earned on average $5.00 (*SD* = $2.43). Further descriptive statistics about our subject pools are in Appendix E.

4.2. Results

Results on cooperation. Across the 8 games, cooperation rates vary from 0.28 to 0.49 in UoN and from 0.40 to 0.60 in AMT (see Table 5). Again, cooperation rates are positively correlated with the K-index (UoN: $r_s = 0.90$, $p = 0.002$; AMT: $r_s = 0.62$, $p = 0.102$). On average, UoN participants chose 'cooperate' in 2.96 of the 8 games, which is significantly lower compared to 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 behaviour across student than non-student subject pools. 67% of UoN participants (70% of AMT participants) were switchers, 25% (17%) always chose 'defect' and 8% (13%) always chose 'cooperate'.

Figure 2 illustrates the average cooperation rates 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 Bonferroni-adjusted significance levels of non-parametric McNemar tests (see Appendix F Table F1 for details). 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.

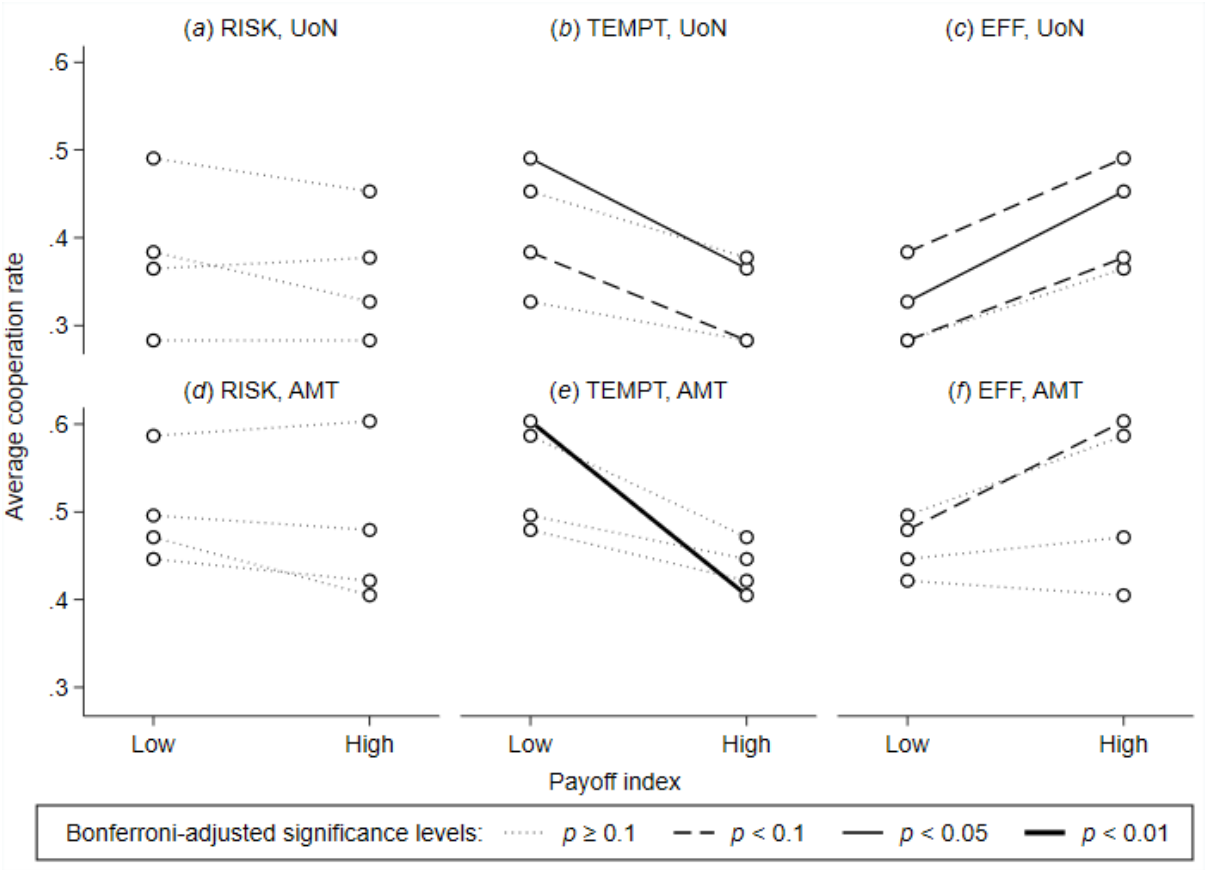


FIGURE 2. Average *cooperation rates* in the eight Prisoner's Dilemma games of Study 2. The line patterns indicate the Bonferroni-adjusted significance levels of two-sided McNemar's tests.

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.

Next, in Table 6, we report the effect of payoff indices on cooperation using a linear probability model with participant fixed effects separately for both samples. Robust standard errors are clustered on participants. The dependent variable is a cooperation dummy, and the explanatory variables are payoff indices (K-index, RISK, TEMPT, EFF), the round in which the respective game was played, and the task characteristics (i.e., whether the decision task or belief task appeared at the top of the screen, and labelling of cooperative choice as A or B).

TABLE 6. Determinants of *cooperative choices* in Study 2.

	(1) UoN	(2) AMT	(3) UoN	(4) AMT
K-index	0.385***	0.287***		
	(0.071)	(0.082)		
RISK			-0.077	-0.072
			(0.061)	(0.061)
TEMPT			-0.401***	-0.516***
			(0.105)	(0.115)
EFF			0.250***	0.131*
			(0.057)	(0.072)
Round	-0.021***	-0.012**	-0.021***	-0.014**
	(0.006)	(0.006)	(0.006)	(0.006)
Constant	0.321***	0.481***	0.510***	0.719***
	(0.037)	(0.040)	(0.068)	(0.077)
BIC	919.2	794.1	928.6	796.5
Within R^2	0.054	0.027	0.058	0.039
Obs. (Clusters)	1,272 (159)	968 (121)	1,272 (159)	968 (121)

Notes: All columns show coefficients from a linear probability model with participant fixed effects. Robust standard errors clustered on participants in parentheses. The order of tasks (randomly determined in each round, 1 if the belief elicitation is placed in the upper section and decision task in the lower section of the screen, 0 otherwise), and labelling dummy (randomly determined in each round, 1 if cooperation is labelled to option B, 0 if cooperation is labelled as option A) are included in the regressions. ** $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$*

Similar to the results in Study 1, we find that the probability of cooperation increases with the K-index in both samples (Cols. 1-2). UoN participants are more sensitive to the payoff variations than AMT participants: an increase in the K-index of 0.1 is associated with 3.85

percentage points (2.87 percentage points) higher probability of choosing 'cooperate' for UoN (AMT) participants.

Next, we estimate the effects of RISK, TEMPT and EFF on cooperation (Cols. 3-4). The effect of RISK is small in magnitude and insignificant in both subject pools. In contrast, 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 4.0 (5.2) percentage points higher probability of choosing 'cooperate' 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 percent increase in EFF increases cooperation by 2.5 (1.3) percentage points for UoN (AMT) participants.

Results on beliefs. As beliefs have been identified as an important driver of cooperative behaviour in similar games, such as the public good game (e.g., Croson (2007); Fischbacher and Gächter (2010); Gächter and Renner (2018)) we now examine how the variation in payoff indices affects beliefs. Figure 3 shows the average expected likelihood that the other player chooses 'cooperate' separately by payoff index and sample. On average, AMT participants held higher average cooperative beliefs than UoN participants (Mann-Whitney $Z = 2.44$, $p = 0.015$).

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 (see Appendix F Table F2 for details). 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.

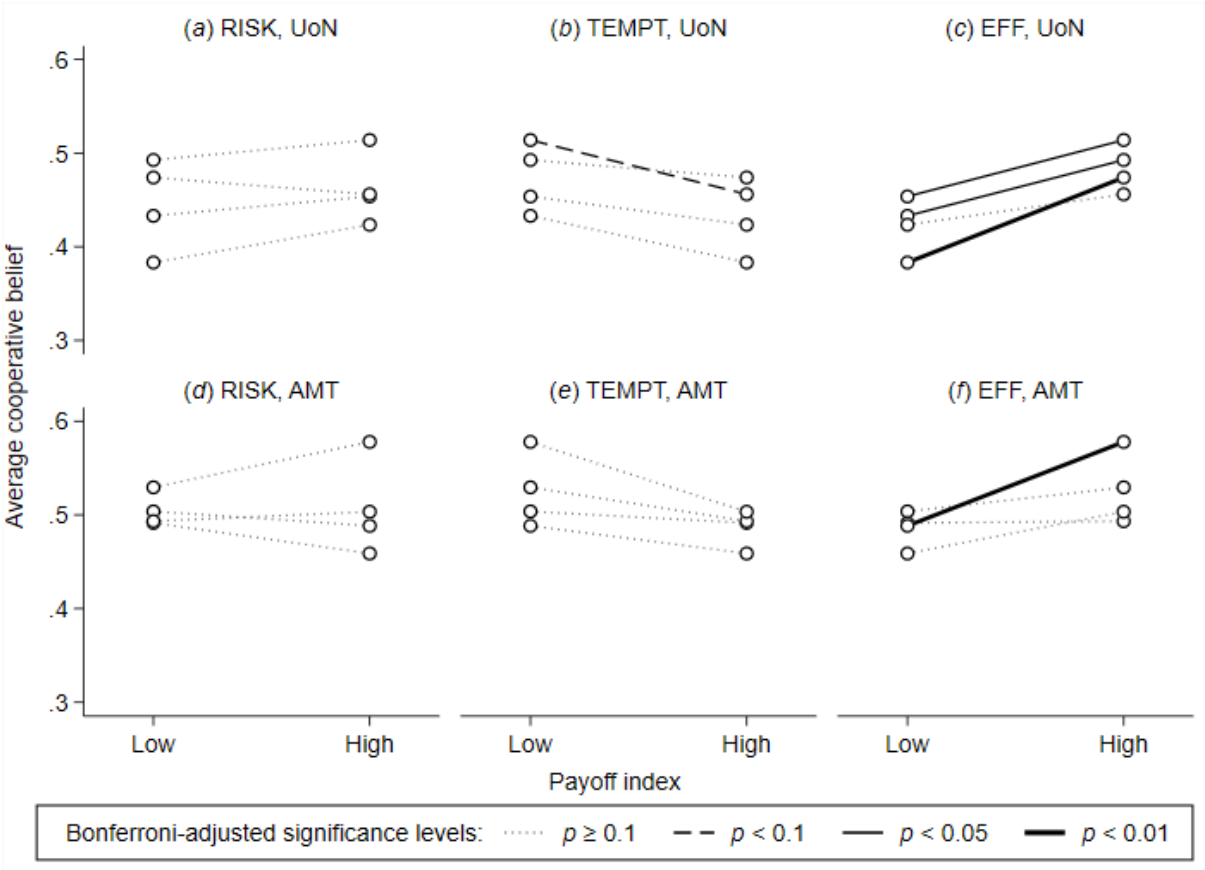


FIGURE 3. Average cooperative *beliefs* in the eight Prisoner's Dilemma games of Study 2. The The line patterns indicate the Bonferroni-adjusted significance levels of Wilcoxon signed-rank tests.

Next, in Table 7 we report the effects of the payoff indices on beliefs using linear regression with participant fixed effects for each sample separately. The analysis parallels that of Table 6. In Columns 1-2, we find that participants are more likely to expect that their opponents would cooperate as the K-index increases. An increase in the K-index of 0.1 is associated with increasing cooperative beliefs of 2.0 (1.1) percentage points for UoN (AMT) participants. Like the results for cooperative decisions (Table 6), the beliefs of UoN participants are more sensitive to the payoff variations than AMT participants.

Looking at the effect of our three payoff indices, TEMPT has the largest effect size in both samples (Cols 3-4). Moreover, the effect sizes of TEMPT are similar in both samples: A 0.1 increase in TEMPT is associated with decreasing cooperative beliefs of around 1.6 percentage points. EFF has a significant impact on beliefs only for the UoN sample: a 0.1 increase in EFF is associated with a 1.5 percentage points increase in belief (Col. 3), whereas for AMT the effect size is 0.7 percentage points (and insignificant). Lastly, the effect of RISK is negligible and insignificant in both samples.

Overall, when a payoff index changes, the effect on beliefs appears to follow a similar pattern to the effect on cooperative decisions. However, the variation in cooperation rates is larger than the variation in beliefs.

TABLE 7. Determinants of *beliefs* in Study 2

	(1) UoN	(2) AMT	(3) UoN	(4) AMT
K-index	0.200***	0.114**		
	(0.040)	(0.048)		
RISK			0.017	-0.017
			(0.038)	(0.050)
TEMPT			-0.161***	-0.160**
			(0.061)	(0.077)
EFF			0.154***	0.066
			(0.033)	(0.041)
Round	-0.015***	-0.011***	-0.015***	-0.012***
	(0.003)	(0.004)	(0.003)	(0.004)
Constant	0.491***	0.602***	0.523***	0.669***
	(0.024)	(0.030)	(0.038)	(0.055)
BIC	-319.6	-154.6	-310.1	-143.1
Within R^2	0.067	0.136	0.070	0.138
Obs. (Clusters)	1,272 (159)	968 (121)	1,272 (159)	968 (121)

Notes: All columns show coefficients from a linear regression with participant fixed effects. Robust standard errors clustered on participants in parentheses. The order of tasks (randomly determined in each round, 1 if the belief elicitation is placed in the upper section and decision task in the lower section of the screen, 0 otherwise), and labelling dummy (randomly determined in each round, 1 if cooperation is labelled to option B, 0 if cooperation is labelled as option A) are included in the regressions. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

5. Conclusion

The PD occupies a place of fundamental importance in social science research 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 maximise their own earnings by defecting, while combined earnings are maximised 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 present two studies examining the separate influences of the unilateral incentives to defect and the efficiency gains from cooperation. Following Mengel (2018) we

use her index of RISK to measure the incentive to defect against a defector, her index of TEMPT to measure the incentive to defect against a co-operator, and her index of EFF to measure the efficiency gains from cooperation. In Study 1, participants play 15 one-shot prisoner's dilemma games with a wide variation in these payoff indices. The results show that the likelihood of cooperation increases in the game's efficiency index (EFF) but is not significantly affected by the risk index (RISK). There is some weak evidence that cooperation rates decrease with the temptation index (TEMPT).

In Study 2, participants played eight prisoner's dilemma games which varied the payoff indices orthogonally. In this controlled design, the likelihood of cooperation decreases when TEMPT increases. We also find a positive, but weaker, effect of EFF on cooperation. As in Study 1 we find no significant effect of RISK. Our elicited (unincentivised) beliefs also indicate that participants expect opponents to be less likely to cooperate when TEMPT is higher and more likely to cooperate when EFF is higher. However, participants' beliefs do not vary significantly with RISK. Note that we observe these common findings across two different subject pools: Amazon Mechanical Turk workers and University of Nottingham students.

Our comparison of students with a socio-demographically more diverse AMT sample is an attempt to learn about the generalizability of our findings. The question of generalizability is interesting because we expected – and observed – cooperation levels to be higher among non-students than students. The fact that variation of payoff indices has similar effects on cooperation (and beliefs) across subject pools with different socio-demographic characteristics and different levels of baseline cooperativeness is therefore reassuring for the robustness of our findings.

Taken together, the evidence from both studies suggests that variation in TEMPT has a greater influence than RISK on cooperation in prisoner's dilemma games when efficiency requires mutual cooperation. In fact, in neither study, nor in either subject pool in Study 2, do we observe a significant effect of RISK.

Our results are consistent with previous research that examines other indices based on the incentives to defect. Ahn et al. (2001) discuss two distinct pressures to defect. One, that they term 'greed', is based on the gain from defecting against a co-operator, while the other, termed 'fear', is based on the loss from cooperating against a defector. These indices are similar to the TEMPT and RISK indices that we analyse, and their result that greed has a larger impact on behaviour than fear mirrors our result that TEMPT matters more than RISK. We emphasise

that our result is obtained in a setting where participants cannot learn from previous encounters and where we randomise the order in which participants experience games.

At first glance, our results appear to contradict the findings of the meta-analysis by Mengel (2018), who concluded that RISK explains the cooperation levels in one-shot Prisoner's Dilemma games while TEMPT best accounts for behaviour in repeated games. Our results and her's can be resolved by noting the importance of the parameter restriction $2R > T + S$, which implies that mutual cooperation maximises efficiency. Imposing this restriction reduces the influence of RISK in Mengel's study. Understanding why this is so would be an interesting topic for future research.

To conclude, as in many previous studies, we find evidence of cooperation, even in carefully controlled anonymous one-shot games. But we emphasise that this cooperation is not random, it varies systematically with the material payoffs of the game. These material payoffs define the incentives to defect and the mutual gains from cooperating. The orthogonal design of our second study identifies clearly which of these incentives matter most for cooperation. We find that cooperation is higher when the gains from cooperation are higher. However, the incentive to free ride on co-operators has an even stronger, negative, impact on cooperation. This implies that if a choice architect could influence the material payoffs of a prisoner's dilemma, they would promote mutually beneficial cooperation most effectively by limiting the temptation to free ride on co-operators.

References

Ahn, T. K., Ostrom, E., Schmidt, D., Shupp, R. and Walker, J. (2001). 'Cooperation in PD games: Fear, greed, and history of play', *Public Choice,* vol. 106(1-2), pp. 137-155.

Arechar, A. A., Gächter, S. and Molleman, L. (2018). 'Conducting interactive experiments online', *Experimental Economics,* vol. 21(1), pp. 99-131.

Au, W. T., Lu, S., Leung, H., Yam, P. and Fung, J. M. Y. (2012). 'Risk and prisoner's dilemma: A reinterpretation of coombs' re-parameterization', *Journal of Behavioral Decision Making,* vol. 25(5), pp. 476-490.

Balliet, D., Parks, C. and Joireman, J. (2009). 'Social value orientation and cooperation in social dilemmas: A meta-analysis', *Group Processes & Intergroup Relations,* vol. 12(4), pp. 533-547.

Charness, G., Rigotti, L. and Rustichini, A. (2016). 'Social surplus determines cooperation rates in the one-shot prisoner's dilemma', *Games and Economic Behavior,* vol. 100, pp. 113-124.

Cooper, R., DeJong, D. V., Forsythe, R. and Ross, T. W. (1996). 'Cooperation without reputation: Experimental evidence from prisoner's dilemma games', *Games and Economic Behavior,* vol. 12(2), pp. 187-218.

Croson, R. (2007). 'Theories of commitment, altruism and reciprocity: Evidence from linear public goods games', *Economic Inquiry,* vol. 45(2), pp. 199-216.

Dal Bó, P. and Fréchette, G. R. (2018). 'On the determinants of cooperation in infinitely repeated games: A survey', *Journal of Economic Literature,* vol. 56(1), pp. 60-114.

Embrey, M., Fréchette, G. R. and Yuksel, S. (2018). 'Cooperation in the finitely repeated prisoner's dilemma', *The Quarterly Journal of Economics,* vol. 133(1), pp. 509-551.

Engel, C. and Zhurakhovska, L. (2016). 'When is the risk of cooperation worth taking? The prisoner's dilemma as a game of multiple motives', *Applied Economics Letters,* vol. 23(16), pp. 1157-1161.

Fischbacher, U. (2007). 'Z-tree: Zurich toolbox for ready-made economic experiments', *Experimental Economics,* vol. 10(2), pp. 171-178.

Fischbacher, U. and Gächter, S. (2010). 'Social preferences, beliefs, and the dynamics of free riding in public good experiments', *American Economic Review.,* vol. 100(1), pp. 541–556.

Flood, M. (1958). 'Some Experimental Games', *Management Science*, vol. 5(1), pp. 5-26.

Frank, R. H., Gilovich, T. and Regan, D. T. (1993). 'Does studying economics inhibit cooperation', *Journal of Economic Perspectives,* vol. 7(2), pp. 159-171.

Gächter, S. (2010). '(Dis)advantages of student subjects: What is your research question?', *Behavioral and Brain Sciences,* vol. 33(2-3), pp. 92-93.

Gächter, S. and Herrmann, B. (2011). 'The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia', *European Economic Review,* vol. 55(2), pp. 193-210.

Gächter, S. and Renner, E. (2018). 'Leaders as role models and 'belief managers' in social dilemmas', *Journal of Economic Behavior & Organization,* vol. 154, pp. 321-334.

Giamattei, M., Yahosseini, K. S., Gächter, S. and Molleman, L. (2020). 'Lioness lab: A free web-based platform for conducting interactive experiments online', *Journal of the Economic Science Association,* vol. 6(1), pp. 95-111.

Greiner, B. (2015). 'Subject pool recruitment procedures: Organizing experiments with orsee', *Journal of the Economic Science Association,* vol. 1(1), pp. 114-125.

List, J. A. (2004). 'Young, selfish and male: Field evidence of social preferences', *Economic Journal,* vol. 114(492), pp. 121-149.

Mengel, F. (2018). 'Risk and temptation: A meta-study on prisoner's dilemma games', *Economic Journal,* vol. 128(616), pp. 3182-3209.

Murnighan, J. K. and Roth, A. E. (1983). 'Expecting continued play in prisoner's dilemma games: A test of several models', *Journal of Conflict Resolution,* vol. 27(2), pp. 279-300.

Ng, G. T. T. and Au, W. T. (2016). 'Expectation and cooperation in prisoner's dilemmas: The moderating role of game riskiness', *Psychonomic Bulletin & Review,* vol. 23(2), pp. 353-360.

Rapoport, A. (1967). 'A note on the" index of cooperation" for prisoner's dilemma', *Journal of Conflict Resolution,* vol. 11(1), pp. 100-103.

Rapoport, A. and Chammah, A. M. (1965). *Prisoners' dilemma. A study in conflict and cooperation*, Ann Arbor: The University of Michigan Press.

Sally, D. (1995). 'Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992', *Rationality and Society,* vol. 7(1), pp. 58-92.

Schmidt, D., Shupp, R., Walker, J., Ahn, T. K. and Ostrom, E. (2001). 'Dilemma games: Game parameters and matching protocols', *Journal of Economic Behavior & Organization,* vol. 46(4), pp. 357-377.

Snowberg, E. and Yariv, L. (2021). 'Testing the waters: Behavior across participant pools', *American Economic Review,* vol. 111(2), pp. 687-719.

Van Lange, P. A. M., Balliet, D., Parks, C. D. and Van Vugt, M. (2014). *Social dilemmas. The psychology of human cooperation*, Oxford: Oxford University Press.

Vlaev, I. and Chater, N. (2006). 'Game relativity: How context influences strategic decision making', *Journal of Experimental Psychology-Learning Memory and Cognition,* vol. 32(1), pp. 131-149.

Online Appendix to

Risk, Temptation, and Efficiency in the One-Shot Prisoner's Dilemma

Simon Gächter[1,2,3,†], Kyeongtae Lee[4] , Martin Sefton[1] and Till O. Weber[5]

[1] Centre for Decision Research and Experimental Economics (CeDEx), University of Nottingham, UK

[2] IZA Bonn, Germany [3] CESifo Munich, Germany

[4] Economic Research Institute, Bank of Korea, Korea

[5] Newcastle University Business School, Newcastle University, UK

[†] Corresponding author: simon.gaechter@nottingham.ac.uk

26 November 2021

Contents

Appendix A. *Instructions used in Study 1*

You are now taking part in an economic experiment. Depending on the decisions made by you and other participants, you can earn a considerable amount of money. It is therefore very important that you read these instructions with care.

These instructions are solely for your private use. **It is prohibited to communicate with other participants during the experiment.** If you have any questions, please raise your hand. A member of the experiment team will come and answer them in private. If you violate this rule, you will be dismissed from the experiment and you will forfeit all payments.

You will solve several tasks during this experimental session. After this experimental session, one task will be randomly selected for payoff.

Additionally, you will receive a show-up fee of £3. Your earnings will be paid to you privately in cash at the end of the session.

At the end of the session, you will be asked to fill in a questionnaire. The answers you provide in this questionnaire are completely anonymous. They will not be revealed to anyone either during the experiment or after it. Furthermore, your responses to the questionnaires will not affect your earnings during the experiment.

You will be randomly matched with another participant. **You will not learn who the other person, who you are matched with, at any point during or after the experiment.**

The experiment

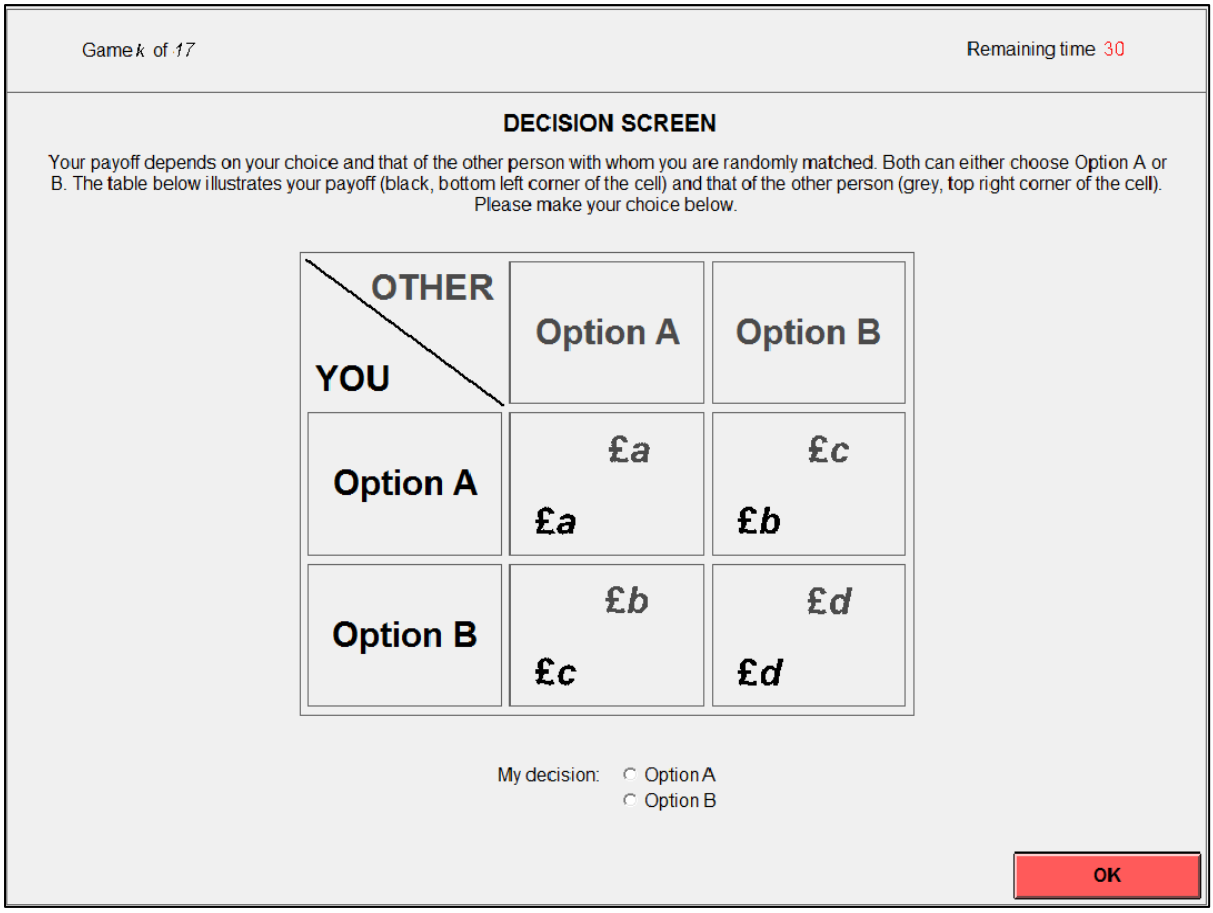
The experiment consists of 17 games and is separated into two stages: the decision stage and the results stage.

At the decision stage, you will have to make a decision for each of the 17 games. The other person, with whom you are randomly matched, will also make a decision for each of the 17 games. During the decision stage, you will not receive any feedback on the choices of the other person and the outcome of the games.

At the results stage, you will receive feedback on the decision taken by you, the other player's decision, as well as the resulting payoffs from these choices.

The decision stage

At the decision stage, you will see the following screen for each game:



In the table shown on the decision screen, your actions and resulting payoffs are given in black (bottom left corner) and the other person's actions and payoffs are given in grey (top right corner). The payoffs shown will be paid to you in case this game is randomly selected at the end of the session. The table is read as follows (black payoffs):

- If you choose Option A and the other participant chooses Option A, you receive £*a*.
- If you choose Option A and the other participant chooses Option B, you receive £*b*.
- If you choose Option B and the other participant chooses Option A, you receive £*c*.
- If you choose Option B and the other participant chooses Option B, you receive £*d*.

Note that the other participant (grey payoffs) is in the same situation as you are. The other participant will receive the following payoff, if this game is randomly selected at the end of the session:

- If the other participant chooses Option A and you choose Option A, the other participant receives £*a*.
- If the other participant chooses Option A you choose Option B, the other participant receives £*b*.
- If the other participant chooses Option B and you choose Option A, the other participant receives £*c*.

- If the other participant chooses Option B and you choose Option B, the other participant receives £*d*.

Keep in mind that you will not receive any feedback on the other person's choices and the other person's payoffs during the decision stage.

The results stage

The results stage starts after all participants have made their decisions for each of the 17 games. At the results stage you will learn the outcomes of each of the 17 games, starting with the first game. First, you will see the payoff table, with **your own choice** highlighted for several seconds. Afterwards, you will see the **other participant's choice and the resulting payoffs** for several seconds.

If you have any questions, please raise your hand and a member of the experiment team will come and answer them in private.

Appendix B. *Non-parametric test results for Study 1*

TABLE B1. McNemar's tests for differences in cooperation across games.

	Games	Variation	Indices held constant	*p*-value
RISK:	G1 vs G4	0.04 vs 0.50	TEMPT = 0.10, EFF = 0.56	0.774
	G2 vs G5	0.04 vs 0.50	TEMPT = 0.30, EFF = 0.43	0.815
	G4 vs G7	0.50 vs 1.00	TEMPT = 0.10, EFF = 0.56	0.057
	G5 vs G8	0.50 vs 1.00	TEMPT = 0.30, EFF = 0.43	0.804
	G1 vs G7	0.04 vs 1.00	TEMPT = 0.10, EFF = 0.56	0.210
	G2 vs G8	0.04 vs 1.00	TEMPT = 0.30, EFF = 0.43	0.524
TEMPT:	G10 vs G11	0.10 vs 0.30	RISK = 1.00, EFF = 0.98	0.092*
EFF:	G7 vs G10	0.56 vs 0.98	RISK = 1.00, TEMPT = 0.10	0.004**
	G10 vs G13	0.26 vs 0.98	RISK = 1.00, TEMPT = 0.10	< 0.001***
	G7 vs G13	0.26 vs 0.56	RISK = 1.00, TEMPT = 0.10	0.092
	G8 vs G11	0.43 vs 0.98	RISK = 1.00, TEMPT = 0.30	0.481
	G11 vs G14	0.05 vs 0.98	RISK = 1.00, TEMPT = 0.30	< 0.001***
	G8 vs G14	0.05 vs 0.43	RISK = 1.00, TEMPT = 0.30	0.003**
	G9 vs G12	0.04 vs 0.96	RISK = 1.00, TEMPT = 0.49	< 0.001***

Notes: To correct for multiple testing, we use Bonferroni-adjusted significance levels. RISK: * $p < 0.017$; ** $p < 0.008$; *** $p < 0.002$. TEMPT: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. EFF: * $p < 0.014$; ** $p < 0.007$; *** $p < 0.001$.

Appendix C. *Payoff indices in previous studies*

TABLE C1. *Descriptive statistics of payoff indices from Mengel (2018)'s data set*

	Average	Q_1	Q_3	SD	Min	Max	N
RISK	0.746	0.667	0.975	0.265	0.167	0.999	34
TEMPT	0.271	0.182	0.333	0.105	0.059	0.444	34
EFF	0.509	0.429	0.571	0.180	0.200	1.000	34

Notes: Q_1 and Q_3 denote the 25th and 75th percentile.

Appendix D. *Instructions used in Study 2*

Note: *These are the instructions used on Amazon Mechanical Turk. The instructions for the sessions conducted at the University of Nottingham used an exchange rate of 100 tokens = £1. Additionally, on the welcome screen, the term 'HIT' was replaced with 'experiment'. Otherwise, the instructions were identical.*

Welcome

Thank you for accepting this HIT. To complete this HIT, you must make some decisions. Including the time for reading these instructions, the HIT will take about 30 minutes to complete. If you are using a desktop or laptop to complete this HIT, we recommend that you maximize your browser screen (press F11) before you start.

It is important that you complete this HIT without interruptions. During the HIT, please **do not close this window or get distracted from the task.** If you close your browser or leave the task, you will not be able to re-enter and we will not be able to pay you.

In this HIT, you will be matched with one other participant. Each of you will make decisions for 8 decision situations. In each situation, each of you will earn Tokens depending on your decisions.

At the end of the HIT, one of the decision situations will be randomly chosen. Your earnings from this situation will be converted from Tokens to Dollars at a rate of 1**00 Tokens = $ 1.** This will be added to **your participation fee of $1.00.** Depending on your decisions, you may make up to $8.00 more in addition to the $1.00 participation fee. In the same way, Tokens earned by the person matched with you in that same situation will also be converted to Dollars at a rate of 100 Tokens = $ 1.

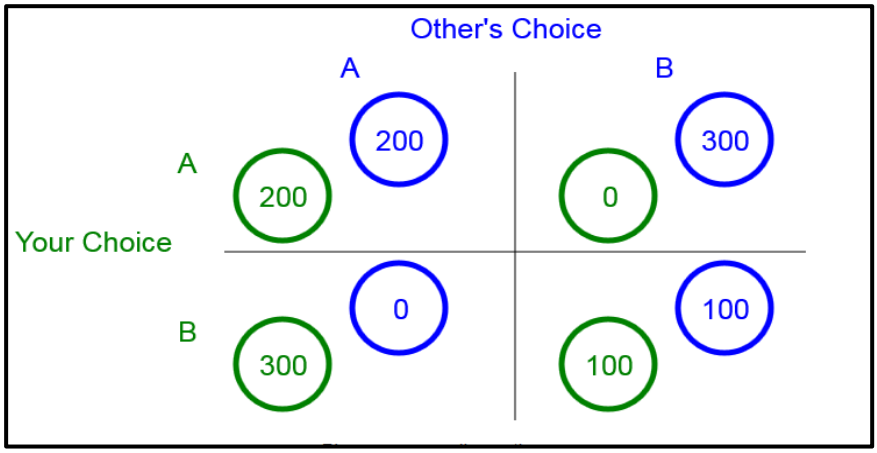
You will receive a code to collect your payment via MTurk upon completion.

Please click "Continue" to start the HIT.

Instructions

The HIT consists of 8 decision situations.

Each decision situation will be presented on a screen like the **example screen** below.



You and the other person will be making choices between **A** and **B**. Your earnings are the values in the green circle, and the other person's earnings are the values in the blue circle. The table is read as follows:

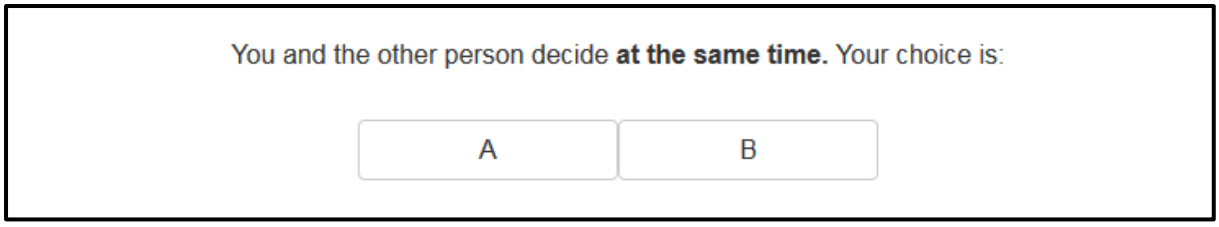
- If you choose A and the other person chooses A, you will earn 200 Tokens and the other person will earn 200 Tokens.
- If you choose A and the other person chooses B, you will earn 0 Tokens and the other person will earn 300 Tokens.
- If you choose B and the other person chooses A, you will earn 300 Tokens and the other person will earn 0 Tokens.
- If you choose B and the other person chooses B, you will earn 100 Tokens and the other person will earn 100 Tokens.

Please note that the values in the table will differ in each decision situation.

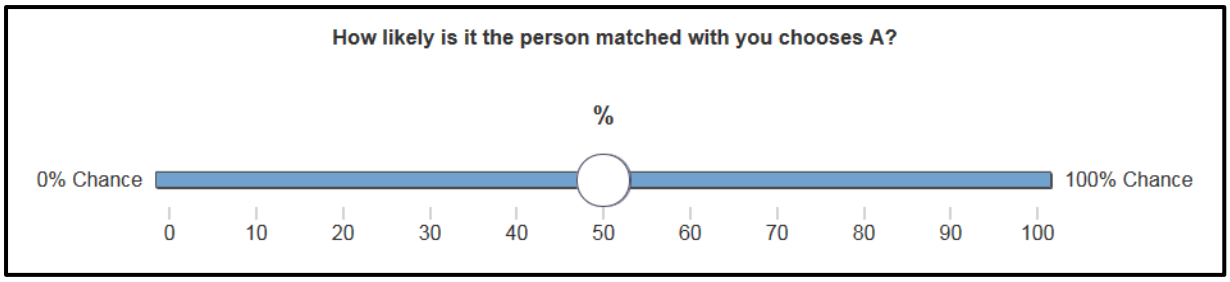
Tasks

In each decision situation, you must complete **two types** of tasks, which we will refer to below as the "decision" and "prediction".

- For the "decision" task, you will see the following screen and you must choose A or B:



● For the “prediction” task, you will see the following screen and you must indicate how likely you think it is that the other person will choose A:



During the HIT, you will not receive any feedback on the other person's choice or the outcomes of the decision situations.

Your dollar earnings

On completion of the HIT, you will be paid your participation fee of $ 1.

In addition, one of the decision situations will be randomly chosen for your additional dollar earnings. Your earnings and the other person’s earnings will be determined depending on choices of you and the other person in that situation. Two examples should make this clear.

Example 1. Assume that you choose A and the other person matched with you chooses A in the above example screen. As a consequence, you will earn 200 Tokens and the other person will earn 200 Tokens.

Example 2. Assume that you choose B and the other person matched with you chooses A in the above example screen. As a consequence, you will earn 300 Tokens and the other person will earn 0 Tokens.

At the end of the HIT

On completion of the HIT, one of the decision situations will be randomly chosen as explained above. You will be informed of your choices and earnings for that decision situation, and you will be paid these earnings in addition to your participation fee.

Note that we will not be able to pay you if you do not complete the HIT. If the person you are matched with does not complete the HIT, the computer will randomly select one of the four possible earnings in the randomly chosen decision situation, and you will be paid these earnings in addition to your participation fee.

Your participation fee and the additional earnings will be paid to you within two working days.

Appendix E. *Descriptive statistics for participants in Study 2*

TABLE E1. Descriptive statistics for UoN and AMT participants.

Variable	UoN	AMT
Age (years)	22.27 (4.33)	33.56 (9.76)
Female (%)	60.38	49.59
Ethnicity (%)		
Asian	27.04	11.57
Black	6.92	7.44
White	48.43	71.90
Latin	1.26	6.61
Annual household pre-tax income (%)		
Less than $70,000		56.20
$70,000 or more		41.32
Average spending per month (%)		
Less than £400	67.30	
£400 or more	26.42	
University graduate (%)		61.16
Full-time worker		56.19
Experience in similar studies (%)	49.06	38.01

Notes: *SD* in parenthesis. AMT Participants were asked to choose one of the categories regarding their household pre-tax income: Less than $30,000, $30,000-$49,999, $50,000-$69,999, $70,000-$89,999, $90,000 or more, prefer not to say. UoN participants were asked to choose one of the categories for reporting their average spending per month excluding a rent: Less than £200, £200-£399, £400-£599, £600-£799, £800-£999, £1,000 or more, prefer not to say. "Experience in similar studies" is defined as participants who participated in similar studies more than once or twice.

Appendix F. *Non-parametric test results for Study 2*

TABLE F1. McNemar's tests for differences in cooperation across games.

		UoN *p*-value	AMT *p*-value
Low vs High RISK	Low TEMPT, Low EFF	0.188	0.860
	Low TEMPT, High EFF	0.489	0.851
	High TEMPT, Low EFF	1.000	0.743
	High TEMPT, High EFF	0.885	0.215
Low vs High TEMPT	Low RISK, Low EFF	0.014*	0.441
	Low RISK, High EFF	0.012**	0.034
	High RISK, Low EFF	0.296	0.296
	High RISK, High EFF	0.104	< 0.001***
Low vs High EFF	Low RISK, Low TEMPT	0.016*	0.061
	Low RISK, High TEMPT	0.092	0.775
	High RISK, Low TEMPT	0.006**	0.024*
	High RISK, High TEMPT	0.020*	0.864

Notes: To correct for multiple testing, we use Bonferroni-adjusted significance levels. * $p < 0.025$; ** $p < 0.013$; *** $p < 0.003$.

TABLE F2. Wilcoxon signed-rank tests for differences in cooperative beliefs across games.

		UoN *p*-value	AMT *p*-value
Low vs High RISK	Low TEMPT, Low EFF	0.338	0.717
	Low TEMPT, High EFF	0.842	0.060
	High TEMPT, Low EFF	0.175	0.374
	High TEMPT, High EFF	0.658	0.775
Low vs High TEMPT	Low RISK, Low EFF	0.100	0.792
	Low RISK, High EFF	0.151	0.280
	High RISK, Low EFF	0.251	0.275
	High RISK, High EFF	0.016*	0.041
Low vs High EFF	Low RISK, Low TEMPT	0.007**	0.236
	Low RISK, High TEMPT	< 0.001***	0.718
	High RISK, Low TEMPT	0.006**	< 0.001***
	High RISK, High TEMPT	0.034	0.066

Notes: To correct for multiple testing, we use Bonferroni-adjusted significance levels. * $p < 0.025$; ** $p < 0.013$; *** $p < 0.003$.