Gächter, Simon; Lee, Kyeongtae; Sefton, Martin

Working Paper

The variability of conditional cooperation in sequential prisoner's dilemmas

CeDEx Discussion Paper Series, No. 2022-10

Provided in Cooperation with:
The University of Nottingham, Centre for Decision Research and Experimental Economics (CeDEx)

Suggested Citation: Gächter, Simon; Lee, Kyeongtae; Sefton, Martin (2022) : The variability of conditional cooperation in sequential prisoner's dilemmas, CeDEx Discussion Paper Series, No. 2022-10, The University of Nottingham, Centre for Decision Research and Experimental Economics (CeDEx), Nottingham

This Version is available at:
https://hdl.handle.net/10419/261248

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

University of Nottingham
UK | CHINA | MALAYSIA

Discussion Paper No. 2022-10

Simon Gächter, Kyeongtae Lee and Martin Sefton

The Variability of Conditional Cooperation in Sequential Prisoner's Dilemmas

March 2022

CeDEx Discussion Paper Series
ISSN 1749 - 3293

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The Centre for Decision Research and Experimental Economics was founded in 2000, and is based in the School of Economics at the University of Nottingham.

The focus for the Centre is research into individual and strategic decision-making using a combination of theoretical and experimental methods. On the theory side, members of the Centre investigate individual choice under uncertainty, cooperative and non-cooperative game theory, as well as theories of psychology, bounded rationality and evolutionary game theory. Members of the Centre have applied experimental methods in the fields of public economics, individual choice under risk and uncertainty, strategic interaction, and the performance of auctions, markets and other economic institutions. Much of the Centre's research involves collaborative projects with researchers from other departments in the UK and overseas.

Please visit http://www.nottingham.ac.uk/cedex for more information about the Centre or contact

Suzanne Robey
Centre for Decision Research and Experimental Economics
School of Economics
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: +44 (0)115 95 14763
suzanne.robey@nottingham.ac.uk

The full list of CeDEx Discussion Papers is available at

http://www.nottingham.ac.uk/cedex/publications/discussion-papers/index.aspx

The Variability of Conditional Cooperation in Sequential Prisoner's Dilemmas*

Simon Gächter[1,2,3], Kyeongtae Lee[4] , and Martin Sefton[1]

[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

25 March 2022

Abstract. We examine how conditional cooperation is related to the material payoffs in a Sequential Prisoner's Dilemma experiment. We have subjects play eight SPDs with varying payoffs, systematically varying the material gain to the second-mover and the material loss to the first-mover when the second-mover defects in response to cooperation. We find that few second-movers are conditionally cooperative in all eight games, and most second-movers change their strategies from game to game. Second-movers are less likely to conditionally cooperate when the gain is higher and when the loss is lower. This pattern is consistent with models of distributional preferences.

Keywords: prisoner's dilemma, conditional cooperation,
JEL codes: A13, C91

* 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 Robin Cubitt, Friederike Mengel and Brown Bag audiences at the Centre of Decision Research and Experimental Economics for their comments and suggestions.

1. Introduction

Conditional cooperation is widely observed in social dilemmas. Whereas the pursuit of narrowly defined selfish interests would result in a lack of cooperation, many people are willing to forgo their selfish interests and cooperate, but only as long as other group members also cooperate. This pattern of behavior is particularly clear in controlled experiments investigating contributions to public goods (Brandts and Schram, 2001; Fischbacher, Gächter and Fehr, 2001; Gächter, 2007; Keser and van Winden, 2000; Kocher *et al.*, 2008). These experiments also reveal substantial heterogeneity: for example, in some of these studies some group members are classified as "free riders" (i.e. defecting regardless of the behavior of others), others as "conditional cooperators" (i.e. cooperating as long as others do so), and still others as "unconditional cooperators" (i.e. cooperating independently of the behavior of others). Not much is known, however, about whether such a classification reflects stable personality traits whereby the participant would exhibit similar behavioral patterns in similar situations, or whether the classification applies only to the specific experimental setting and parameters.

In this study, we examine whether the behavioral pattern exhibited by a given participant, such as conditional cooperation, varies across payoff variations. Two previous studies have examined the variability of conditional cooperation across payoff variations. In a meta-study, Thöni and Volk (2018) found that the proportion of conditional cooperators is similar across 17 public goods experimental studies employing different parameters (e.g., marginal per capita return, group size). In contrast, Clark and Sefton (2001), using a between-subject design where different subjects are assigned to different payoff treatments, find that conditional cooperation in sequential prisoner's dilemmas significantly decreases when the payoff to defecting against a cooperator doubles. In contrast to these studies, we examine the individual-level variability of conditional cooperation using a within-subject design: the experiment consists of eight games where payoffs differ, and subjects make decisions in all eight games.

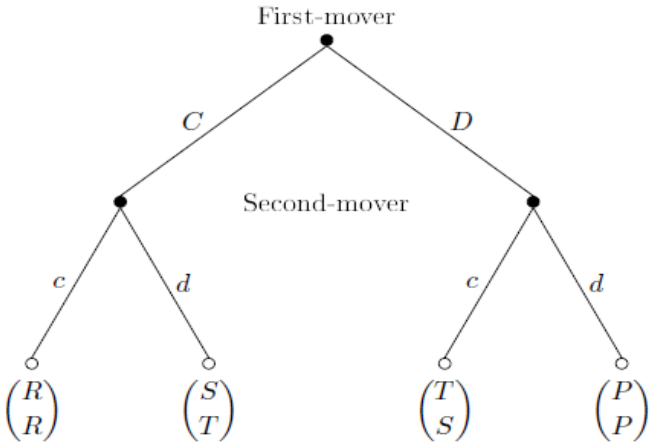
Examining the within-subject variability of conditional cooperation across payoff variations is important for at least two reasons. First, it allows us to understand the nature of conditional cooperation: whether conditional cooperation reflects underlying social preferences, or whether conditional cooperation reflects a desire to reciprocate the cooperation

of others in a way that is robust to changes in material incentives.[1] Social preference models, which define preferences over one's own and other's material payoffs (e.g., Andreoni and Miller, 2002; Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Cox, Friedman and Gjerstad, 2007; Fehr and Schmidt, 1999), are capable of explaining conditional cooperation, but at the same time predict that it will be influenced by material incentives. In contrast, if conditional cooperation reflects a principled stand against free-riding, eschewing material gains in order to reciprocate the cooperation of others, then conditional cooperation is expected to be robust across payoff variations.

Second, the efficacy of interventions to promote cooperation depends on whether conditional cooperation is influenced by payoff variations. For example, leading by example would be an effective mechanism to achieve cooperative outcomes if followers are generally conditionally cooperative (Gächter, Nosenzo, Renner and Sefton, 2012). On the other hand, if conditional cooperation is sensitive to payoffs, then this implies that there would be settings where leading by example is ineffective.

In order to study the within-subject variability of conditional cooperation, we use the sequential prisoner's dilemma. In the sequential prisoner's dilemma (henceforth SPD), there are two players - First-mover and Second-mover - who sequentially choose whether to cooperate or defect. The extensive form of the SPD is shown in Figure 1.

FIGURE 1. The Sequential Prisoner's Dilemma (SPD)



Note: $T > R > P > S$ and $2R > T + S$

[1] As discussed by Fehr and Fischbacher (2004), Gächter *et al.* (2017), and Katuščák and Miklánek (2018), conformity to what is perceived as "socially appropriate" and willingness to sacrifice material payoffs in order to follow such norms could also be a candidate explanation for conditional cooperation.

In this game First-mover chooses to either cooperate (C) or defect (D), and then, after observing this choice, Second-mover chooses to cooperate (c) or defect (d). Subjects' combined earnings are maximized when both cooperate, resulting in each player receiving *R*. However, if First-mover cooperates, Second-mover maximizes own earnings by defecting, in which case First-mover receives *S* and Second-mover receives *T*. If First-mover defects, Second-mover maximizes own earnings by also defecting, so that each player receives *P*. If players are own-earnings maximizers, and this is common knowledge, there is a unique equilibrium where First-mover defects and Second-mover defects regardless of the first-mover's choice.

Previous experiments with the SPD have shown that subjects hardly ever cooperate in response to defection (Clark and Sefton, 2001; Miettinen, Kosfeld, Fehr and Weibull, 2020). In contrast, the response to cooperation is more variable, and there are non-negligible proportions of subjects who cooperate (i.e., conditional cooperators) and defect (i.e., free-riders). There are two factors related to payoffs that could plausibly affect the response to cooperation in the SPD. First, *damage* refers to the cost imposed on First-mover when Second-mover responds by defecting rather than cooperating: *R* - *S*. Second, *gain* refers to the gains to Second-mover from defecting rather than cooperating: *T* - *R*. We define DAMAGE ($\frac{R-S}{R}$) as the percentage loss imposed on First-mover by defecting in response to cooperation, while we define GAIN ($\frac{T-R}{R}$) as the percentage gain from defecting against a cooperator.

Compared to the public goods game which has frequently been used to study conditional cooperation, the SPD is simpler in that it has only two players and binary actions. Moreover, it allows us to separately vary the factors of damage and gain.[2] Using eight SPDs where payoffs vary, we implement a within-subject experiment in which every subject makes decisions in the role of both First-mover and Second-mover for each game. For Second-mover decisions, we ask how the subject would respond to defect and how they would respond to cooperate, with the actual decision being determined by the response to the first-mover's actual choice. That is, we elicit Second-mover's strategy for playing each SPD. To rule out confounding factors, such as belief updating, no feedback on any of the individual games is provided until the end of the experiment. Our experimental design allows us to examine the

[2] In the standard public goods game (see Chaudhuri, 2011), varying the MPCR (Marginal Per Capita Return) affects the levels of GAIN and DAMAGE simultaneously, making it difficult to disentangle the effect of GAIN and DAMAGE on decisions.

following two questions: Is conditional cooperation influenced by payoff changes? If so, How is conditional cooperation influenced by damage and gain?

Our main findings are two-fold. First, we find that conditional cooperation is sensitive to payoff variations. Only 26% of subjects submitted the same Second-mover strategy across all eight games. This is composed of 13% of subjects who were consistently free-riders and 13% who were consistently conditional cooperators. The remaining 74% of subjects changed their Second-mover strategy at least once across games.

Second, conditional cooperation is significantly influenced by DAMAGE and GAIN. Conditional cooperation increases with DAMAGE and decreases with GAIN. Second-movers are more likely to conditionally cooperate when free-riding has a larger negative impact on the first-mover's earnings, and less likely to conditionally cooperate when they gain more from free-riding.

Our finding suggests that classifications of individuals as "conditional cooperators" or "free-riders" should not be generalized to other games with different material payoffs. In other words, a conditional cooperator in one game may be a free-rider in another, and vice versa. The variability of conditional cooperation with GAIN and DAMAGE also suggests that the effectiveness of leading by example in promoting cooperation may be limited when GAIN is high or DAMAGE is low.

This study is also related to research that uses theories of social preferences to explain unselfish behavior in various contexts. The within-subject variation of conditional cooperation with the levels of damage and gain is consistent with the predictions of several distributional preference models (e.g., Fehr and Schmidt, 1999). To this extent, our results support the view that conditional cooperation reflects underlying social preferences.

The remainder of the paper is organized as follows. In Section 2, we review the related literature that examines the variability of conditional cooperation. In Section 3 we present our experimental design and procedures. In Section 4 we present our results, and in Section 5 we conclude.

2. Related literature

In this section, we review related studies that examine the variability of conditional cooperation. A number of studies examine the variability of conditional cooperation over time with mixed results. Brosig *et al.* (2007) conducted SPDs three times within three months using the same subjects and random-matching and found that the rate of conditional cooperation diminished

across repetitions. Andreozzi *et al.* (2020) similarly found conditional cooperation diminished with repetition. Muller *et al.* (2008) elicited subjects' strategy across five repetitions of a public goods game. Although, only 37% of subjects always choose the same strategy across all five games, previous choices were useful predictors of subsequent choices. For example, 69% of subjects who conditionally cooperated in any of the first four games also conditionally cooperated in the fifth game. Volk *et al.* (2012) elicited subjects' strategies in a public goods game three times over the course of five months and observed that conditional cooperation was remarkably stable over time. Half of their subjects chose the same strategy in all three games, and 71% of these conditionally cooperated.

Two further studies examine the variability of conditional cooperation across different contexts by comparing behavior in a public goods game and a SPD. Eichenseer and Moser (2020) and Mullett *et al.* (2020) report that subjects who are conditionally cooperative in a SPD are also conditionally cooperative in a public goods game.

We are not aware of any study examining how within-subject variation of payoffs affects conditional cooperation.[3] In fact, we are only aware of two studies that examine whether payoff variation affects conditional cooperation. Thöni and Volk (2018) found that the proportion of conditional cooperators is quite similar across 17 public goods experiments, which employ different parameters (i.e. marginal per capita return, group size). In contrast, Clark and Sefton (2001), using a between-subjects SPD experiment, found that doubling the temptation payoff, *T*, resulted in a significantly lower rate of conditional cooperation. Our study differs from these two studies in that we ask subjects to make decisions in eight sequential prisoner's dilemmas with systematically varying payoffs.[4] This within-subject design allows us to examine how payoff variations affect conditional cooperation at the individual level.

3. Experimental Design & Procedures

3.1. Experimental design

The experiment is based on a simple sequential two-player prisoner's dilemma game. Each subject must decide whether to cooperate or defect. First-mover decides first and Second-

[3] Several studies examine how decisions in the simultaneous prisoner's dilemmas are influenced by payoff variations (e.g., Ahn *et al.*, 2001; Gächter, Lee and Sefton, 2021; Schmidt *et al.*, 2001; Vlaev and Chater, 2006).
[4] We note two further, potentially important, differences between us and Clark and Sefton. First, in their Double Temptation treatment the payoffs are such that mutual cooperation does not maximize combined earnings. In all our games mutual cooperation maximises combined earnings. Second, within a treatment their subjects play ten repetitions of the game (using a perfect stranger matching protocol), giving feedback at the end of each game. We give no feedback between games.

mover decides after observing First-mover's decision. Subjects had to make decisions in eight such games with varying payoffs.

Table 1 shows the payoff parameterization used in the experiment. Payoffs were chosen to be strictly positive multiples of ten in order to avoid zero or non-rounded payoffs. R (500) is constant across all games while there are two distinct values of P (200, 400). Thus, we study games with two different levels of efficiency. We use EFF = (R – P)/R as a measure of efficiency. There are also two distinct values of T (600, 800) and four distinct values of S (20, 90, 40, 180). This is the same parameterization used in Gächter et al. (2021) for studying cooperation in simultaneous PDs.

TABLE 1. Payoff parameters for Sequential Prisoner's Dilemma Games

Game	R	P	S	T	EFF	DAMAGE	GAIN
G1	500	200	90	600	0.60	0.82	0.20
G2	500	200	20	600	0.60	0.96	0.20
G3	500	200	90	800	0.60	0.82	0.60
G4	500	200	20	800	0.60	0.96	0.60
G5	500	400	180	600	0.20	0.64	0.20
G6	500	400	40	600	0.20	0.92	0.20
G7	500	400	180	800	0.20	0.64	0.60
G8	500	400	40	800	0.20	0.92	0.60

In the sequential PD, we are mainly interested in Second-mover's response to cooperation, and so we focus on the indices of DAMAGE and GAIN. DAMAGE ($\frac{R-S}{R}$) refers to the losses imposed on First-mover when Second-mover defects against cooperation, while GAIN ($\frac{T-R}{R}$) represents the gains to Second-mover from defecting in response to First-mover's cooperation. Note that gains and losses are measured in percentage terms so that, for example, a value 0.82 of DAMAGE implies that if Second-mover defects against a cooperator this will reduce First-mover's monetary payoff by 82% compared to the payoff from mutual cooperation. Similarly, a value 0.20 of GAIN implies that if Second-mover defects in response to cooperation Second-mover accrues a 20% gain in monetary payoff compared to the payoff from mutual cooperation. Note that with this parameterization we study a 2×2 variation in DAMAGE and GAIN for each level of efficiency.

Our main interest lies in exploring how variation in DAMAGE and GAIN affect conditional cooperation. The Nash equilibrium assuming self-interested agents is mutual defection, regardless of the levels of DAMAGE and GAIN. Therefore, the self-interested agent model predicts that conditional cooperation never occurs in SPDs. Distributional preference models on the other hand, which define an individual's preferences over own and other's material earnings, allow for the possibility of rational conditional cooperation for some payoff parameters. These models predict that Second-mover's strategies will depend on the material payoffs and preference parameters. For example, the Fehr and Schmidt (1999) model predicts that Second-mover will free-ride if $\beta < (T - R)/(T - S)$ and conditionally cooperate if $\beta < (T - R)/(T - S)$ where β is Second-mover's marginal disutility from advantageous inequality. Thus, conditionally cooperation is more likely as T decreases, S decreases, or R increases, and so the likelihood of conditional cooperation increases with DAMAGE and decreases with GAIN.

3.2. Experimental procedures

We conducted the online interactive experiment in Spring 2019 using MTurk. Subjects were residents of the United States. We conducted five sessions with a total of 138 participants. None of the subjects participated in more than one session. Each participant was paired with another subject after he/she had read the instructions and passed some control questions.[5] Each pair then played all eight games of Table 1 with no feedback.

For each game, subjects had to answer eight additional control questions about the payoffs before making decisions. These additional control questions were intended to ensure that subjects understood the implications of their decisions and recognized the payoff changes across games. Subjects then made decisions as First-mover and as Second-mover. Both decision tasks were presented on the same screen. In the *First-mover's decision*, they simply chose whether to cooperate or to defect as First-mover. In the *Second-mover's decision*, we asked subjects to decide in the following two situations: i) if First-mover cooperates, and ii) if First-mover defects. Therefore, we elicited Second-mover strategies using the strategy method (Selten, 1967).[6] Rather than use the terms "cooperate" or "defect", we labeled options neutrally as A or B, with labeling randomly chosen at the pair level in each game. To control for potential

[5] The instructions are included in Appendix A.

[6] Regarding potential differences between responses elicited using the strategy method and those using a direct response method, previous studies found no statistical differences in subjects' responses between these two methods (see Brandts and Charness, 2000; 2011 for a review).

order effects, we randomized the sequence of games and the order of tasks (First-mover's decision and Second-mover's decision) at the pair level

Subjects did not receive any feedback on the other's choice or the outcome of each game until the end of the experiment. Once subjects completed the tasks for all games, we asked them to complete a short post-experimental questionnaire eliciting basic demographic information.

We implemented the experiment using the software LIONESS (Giamattei *et al.*, 2020). Subjects were paired with another participant on a real-time basis and they conducted each task at the same time. This implies a subject had to wait until the opponent made a decision to proceed to the next game. As subjects needed to wait until their opponent made a decision, long waiting times could increase the risk of reduced attention. We took the following 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, such an inactive participant was removed from the experiment and the remaining person was able to continue the experiment.

To elicit subjects' responses in an incentive-compatible way, we implement the following payment scheme. At the end of the session, one of eight games was randomly chosen at the pair-level for payment. If both subjects completed the entire experiment, they were paid according to the outcome of this game as follows. One of the pair was randomly chosen to be First-mover, and the other was selected to be Second-mover. Then, subjects were reminded of their decisions and informed about the outcome for this game. For Second-mover's decision we used their conditional response to First-mover's decision. If one of the pair had dropped out during the experiment, the computer randomly selected the payoff-relevant game for the remaining subject. Then the computer randomly selected one out of four monetary outcomes (i.e. T, R, P, or S) of the chosen game for payment to the remaining subject. We explained this payment scheme clearly in the instructions. This payment procedure gives subjects a monetary incentive to take both First-mover decisions and Second-mover decisions seriously in all games as any of these decisions can become payoff-relevant.

As normally occurs in online experiments, there was a non-negligible attrition rate: 32 out of 138 subjects (23%) dropped out during the experiment.[7] For subjects who completed the experiment, the average age was 34 years (between 19 and 65 years) and 37% were female. Subjects' earnings ranged from $1.20 to $9.00, averaging $4.59. On average, the experiment lasted about 30 minutes, including the completion of a post-experimental questionnaire. Subjects were informed of their payment immediately upon completion of the experiment and were paid within 24 hours.

4. Results

Our focus is on Second-mover decisions as these give a direct measure of conditional cooperation. Since Second-mover can condition choices on First-mover's decision, Second-mover has four pure strategies. Conditional cooperation (henceforth CC) involves cooperating if and only if First-mover cooperates. Free-riding (FR) involves defecting regardless of First-mover's choice. Unconditional cooperation (UC) involves cooperating regardless of First-mover's choice. Lastly, mismatching (MM) occurs when Second-mover defects in response to cooperation and cooperates in response to defect.

As in most social dilemma experiments (Fischbacher *et al.*, 2001; Fallucchi *et al.*, 2019; Muller *et al.*, 2008), we find that CC and FR strategies predominate: the average proportion of CC and FR in eight games is 38% and 45%, respectively. In contrast, the average proportion of UC and MM in eight games is only 12% and 5%, respectively. Thus, we focus on CC and FR strategies.

The following analysis is structured to discuss our main research questions. First, does conditional cooperation vary with payoffs? If so, are there systematic patterns between within-subject variations in conditional cooperation and payoff variations? For our analysis, we only include the decisions of subjects who completed the experiment: thus, our data set consists of 848 observations (106 subjects $\times$ eight games).

4.1. Does Conditional Cooperation Vary with Payoffs?

Only 26% of subjects use the same strategy in all eight games: 13% always defect, 13% always conditionally cooperate. None of the subjects always unconditionally cooperate or always choose mismatching strategy. The remaining 74% of subjects change strategies across games:

[7] The dropout rate in our experiment is not too different from that of similar interactive online experiments. For example, Arechar *et al.* (2018) report a 20% dropout rate in their interactive 4-player public goods game, and Gächter *et al.* (2020) reports 24% dropout rate in their interactive eight simultaneous prisoner's dilemma game.

we refer to these subjects as "switchers" for the rest of the analysis. On average, switchers conditionally cooperate three times out of eight games. Similarly, switchers free-ride three times out of eight games, on average.

Aggregating across all eight games, 38% of strategies are conditional cooperation. Table 2 reports the proportion of CC strategies depending on the levels of DAMAGE and GAIN in the *high* and *low* efficiency games.

TABLE 2. Proportions of Conditionally Cooperative Strategies

(a) High Efficiency Games (EFF = 0.6)

		DAMAGE		
		Low (= 0.82)	*High* (= 0.96)	Δ (*H-L*)
GAIN	*Low* (= 0.20)	35.8%	42.5%	+6.7%p
	High (= 0.60)	33.0%	38.7%	+5.7%p
	Δ (*H-L*)	-2.8%p	-3.8%p	

(b) Low Efficiency Games (EFF = 0.2)

		DAMAGE		
		Low (= 0.64)	*High* (= 0.92)	Δ (*H-L*)
GAIN	*Low* (= 0.20)	44.3%	48.1%	+3.8%p
	High (= 0.60)	27.4%	34.9%	+7.5%p
	Δ (*H-L*)	-16.9%p***	-13.2%p***	

Note: EFF = (*R* – *P*)/*R*, DAMAGE = (*R* – *S*)/*R*, GAIN = (*T* – *R*)/*R*.* *p* < *0.1;* ** *p* < *0.05;* *** *p* < *0.01* indicate *p*-values based on McNemar's test.

We find that the proportion of CC strategies increases with DAMAGE and decreases with GAIN. Holding GAIN and efficiency constant conditional cooperation increases by 5.9 percentage points on average when DAMAGE increases (the absolute differences range from 3.8 to 7.5 percentage points). Holding DAMAGE and efficiency constant conditional cooperation decreases by 9.2 percentage points on average when GAIN increases (the absolute differences range from 2.8 to 16.9 percentage points). To examine whether conditional cooperation varies significantly across games, we conduct pairwise comparisons using McNemar's test. We compare the games one by one holding the level of efficiency constant while varying each payoff index (i.e. DAMAGE, GAIN). In the *low* efficiency games (e.g., G5

~ G8), subjects are significantly less likely to conditionally cooperate when GAIN increases (G5 vs G7: McNemar's test: $\chi^2 = 10.80$, $p = 0.001$, G6 vs G8: $\chi^2 = 7.00$, $p = 0.008$). For other cases the differences are not statistically significant.

Overall, while the pairwise comparisons of the effect of DAMAGE are not significant, the consistent pattern of results across games and the statistically significant effect of GAIN in the low efficiency games are suggestive. In the next section, we use regression methods to analyze the relationship between conditional cooperation and payoffs.

4.2. Regression Analysis

In this subsection, we examine how Second-mover's decisions are influenced by variation in payoffs using regression analysis. Table 3 shows the results of panel multinomial logit regressions.

TABLE 3. Determinants of Conditional Cooperation

	(1) All	(2) Switcher	(3) All	(4) Switcher
DAMAGE	0.274**	0.356**	0.276**	0.374**
	(0.113)	(0.147)	(0.113)	(0.150)
GAIN	-0.245***	-0.323***	-0.263***	-0.348***
	(0.067)	(0.087)	(0.067)	(0.087)
EFF	-0.101	-0.133	-0.096	-0.127
	(0.073)	(0.096)	(0.076)	(0.101)
Round	-0.016**	-0.020**	-0.017**	-0.022**
	(0.007)	(0.009)	(0.007)	(0.009)
Controls	*No*	*No*	*Yes*	*Yes*
Observations	848	624	808	600
Log-likelihood	-731.4	-656.0	-676.8	-610.6
BIC	1,604.4	1,447.2	1775.5	1624.2

Notes: Average marginal effects from panel multinomial logit regression with robust standard errors clustered on individuals. DAMAGE: (R – S)/R, GAIN: (T – R)/R, EFF: (R – P)/R. Controls: demographic variables, task characteristics, and session effects.* *p < 0.1; ** p < 0.05; *** p < 0.01*

In Table 3, column (1) examines how conditional cooperation varies with the indices of DAMAGE ($\frac{R-S}{R}$), GAIN ($\frac{T-R}{R}$), and EFF ($\frac{R-P}{R}$). Column (2) provides corresponding regression results for the subsample of switchers. In columns (3) and (4) we include controls for individual characteristics and session effects in the regressions. In all regressions, we include a control for the order in which subjects played the games (Round $\in \{1, 2, ..., 8\}$ is the position of the game in the sequence of eight games).

First, we find that DAMAGE and GAIN have a significant effect on conditional cooperation (Column (1)): conditional cooperation increases with DAMAGE and decreases with GAIN. Second-movers are more likely to conditionally cooperate when free-riding has a larger negative impact on First-mover's earnings: in our experiment, the probability of CC increases by 2.7 percentage points when DAMAGE increases by 0.1 percentage point. For GAIN, Second-movers are less likely to conditionally cooperate when they gain more from free-riding: the probability of CC decreases by 2.5 percentage points when GAIN increases by 0.1 percentage points.

Second, switchers are more sensitive to payoff variations (Column (2)). For switchers, who represent 74% of subjects, the effect sizes of DAMAGE (0.356) and GAIN (-0.323) are approximately 1.3 times greater than the effect sizes based on the complete sample. The overall patterns of conditional cooperation for the subsample of switchers are similar to those patterns based on the complete sample, but the effect sizes of payoff indices become larger for the switchers.

Third, the results are robust to a set of controls for individual characteristics and session effects (Columns (3) and (4)). The significance level of regressors are unchanged, and the effect sizes of damage and gain are similar to those in columns (1) and (2). Except for age, individual characteristics (i.e. gender, political orientation, MTurk experience, employment, education, income level, ethnicity) (Wald test, $\chi^2 = 25.01$, $p = 0.247$) and task characteristics (labeling of cooperative choice as A or B, task order dummies) ($\chi^2 = 8.03$, $p = 0.236$) are all jointly insignificant. Session effects are marginally significant at the 10% level ($\chi^2 = 18.65$, $p = 0.097$).

We do not find a significant relation between the index of efficiency (EFF) and conditional cooperation: in all regressions the effect size of EFF is small and insignificant. Lastly, we also observe a round effect: the probability of CC decreases by 1.6 - 2.2 percentage points from one game to the next.

Note that free-riding is also prevalent in our experiment, representing 45% of Second-mover strategies. Thus, Table 4 reports the determinants of free-riding based on the same panel multinomial logit regressions. We find that free-riding is also significantly influenced by the variations of DAMAGE and GAIN: free-riding decreases with DAMAGE, but increases with GAIN. Second-movers are more likely to free-ride when free-riding has a smaller negative impact on the first-mover's earnings (decreasing DAMAGE). In contrast, second-movers are more likely to free-ride when the gains from free-riding become larger (increasing GAIN). Note that the signs of the marginal effects on FR are exactly opposite to the marginal effects

on CC. This reflects the fact that there are relatively few cases of unconditional cooperation and mismatching, and so subjects tend to switch between conditional cooperation and free-riding as the levels of DAMAGE and GAIN vary.

TABLE 4. Determinants of Free Riding

	(1) All	(2) Switcher	(3) All	(4) Switcher
DAMAGE	-0.260**	-0.337**	-0.227*	-0.308*
	(0.118)	(0.157)	(0.118)	(0.160)
GAIN	0.155**	0.200**	0.160**	0.211**
	(0.069)	(0.092)	(0.068)	(0.093)
EFF	0.008	0.007	-0.012	-0.017
	(0.070)	(0.093)	(0.070)	(0.097)
Round	0.021***	0.028***	0.021***	0.029***
	(0.006)	(0.008)	(0.006)	(0.008)
Controls	*No*	*No*	*Yes*	*Yes*
Observations	848	624	808	600
Log-likelihood	-731.4	-656.0	-676.8	-610.6
BIC	1,604.4	1,447.2	1775.5	1624.2

Notes: Average marginal effects from panel multinomial logit regression with robust standard errors clustered on individuals DAMAGE: (R – S)/R, GAIN: (T – R)/R, EFF: (R – P)/R. Controls: demographic variables, task characteristics, and session effects.* *p < 0.1; ** p < 0.05; *** p < 0.01*

Table 4 reports the following findings, which are similar to the results for the determinants of conditional cooperation (Table 3). First, switchers are more sensitive to variations of DAMAGE and GAIN: the marginal effects of DAMAGE and GAIN for switchers are 1.3 times greater than the marginal effects for all subjects. Second, EFF has an insignificant effect on free-riding. Third, there exists a round effect: the probability of free-riding increases by 2.1 - 2.9 percentage points from one game to the next.

Including controls for individuals and session effects (Columns (3)-(4)) weakens the significance of DAMAGE, but other than this, the overall results are qualitatively similar when individual characteristics and session dummies are included: the significance of GAIN is not affected by the inclusion of control variables, and the effect sizes of DAMAGE and GAIN do not change much with the inclusion of control variables.[8]

[8] We also examine how unconditional cooperation and mismatching are influenced by payoff variations. Neither are significantly affected by DAMAGE, GAIN, or EFF.

5. Conclusion

To our knowledge, our study is the first to empirically examine the within-subject variability of conditional cooperation when payoffs vary. To do this, we have subjects play eight one-shot sequential prisoner's dilemma games with varying payoff parameters. We find that conditional cooperation varies across games, and most subjects change strategies across games. This switching between strategies varies systematically with the distributional consequences of free-riding relative to conditionally cooperating. Subjects conditionally cooperate more often when free-riding imposes larger losses on the first-mover, or when free-riding provides smaller gains for oneself.

These findings provide two important implications. First, the within-subject variation of conditional cooperation with payoffs suggests that conditional cooperation should be viewed as an endogenous behavior arising from interaction between underlying motives and payoff variations, rather than a preference itself (Arifovic and Ledyard, 2012). A majority of subjects change their second-mover strategy when material payoffs change, and so classifications of individuals as conditional cooperators or free-riders should not be generalized to other games with different material payoffs.

Second, these results suggest that conditional cooperation may reflect underlying social preferences. The finding that strategies are sensitive to the cost imposed on the opponent as well as the gain to self suggests that a substantial proportion of subjects care not only their own material payoffs but also the other's material payoffs. Moreover, the way conditional cooperation varies with damage and gain is consistent with the predictions of several distributional preference models (e.g., Fehr and Schmidt, 1999). In further research it would be useful to design experiments that separate the predictions of alternative models.

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.

Andreoni, J., and J. Miller. (2002). 'Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism', *Econometrica*, vol. 70(2), pp. 737–753.

Andreozzi, L., Ploner, M. & Saral, A.S. (2020). 'The stability of conditional cooperation: beliefs alone cannot explain the decline of cooperation in social dilemmas', *Scientific Reports* 10, 13610.

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

Arifovic, J., and J. Ledyard. (2012). 'Individual evolutionary learning, other-regarding preferences, and the voluntary contributions mechanism', *Journal of Public Economics*, vol. 96, pp. 808–823.

Blanco, M., D. Engelmann, A.K. Koch, and H. T. Normann. (2014). 'Preferences and beliefs in a sequential social dilemma: a within-subjects analysis', *Games and Economic Behavior*, vol. 87, pp. 122–135.

Bolton, G.E., and A. Ockenfels. (2000). 'ERC: A Theory of Equity, Reciprocity, and Competition', *American Economic Review.*, vol. 90(1), pp. 166–193.

Brandts, J., and G. Charness. (2000). 'Hot vs. cold: sequential responses in simple experimental games', *Experimental Economics*, vol. 2, pp. 227–238.

Brandts, J., and G. Charness. (2011). 'The strategy versus the direct-response method: a first survey of experimental comparisons', *Experimental Economics*, vol. 14, pp. 375–398.

Brandts, J., and A. Schram. (2001). 'Cooperation and noise in public goods experiments: applying the contribution function approach', *Journal of Public Economics*, vol. 79, pp. 399–427.

Brosig, J., T. Riechmann, and J. Weimann. (2007). 'Selfish in the end? An investigation of consistency and stability of individual behavior', MPRA Paper 2035, University Library of Munich.

Charness, G., and M. Rabin. (2002). 'Understanding Social Preferences with Simple Tests', *The Quarterly Journal of Economics*, vol. 117(3), pp. 817–869.

Chaudhuri, A. (2011). 'Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature', *Experimental Economics*, vol. 14, pp. 47–83.

Clark, K., and M. Sefton. (2001). ' The Sequential Prisoner's Dilemma: Evidence on Reciprocation', *The Economic Journal*, vol. 111, pp. 51–68.

Cox, J.C., D. Friedman, and S. Gjerstad. (2007). ' A Tractable Model of Reciprocity and Fairness', *Games and Economic Behavior,* vol. 59(1), pp. 17–45.

Eichenseer, M., and J. Moser. (2020). 'Conditional cooperation: Type stability across games', *Economics Letters,* vol. 188, forthcoming.

Fallucchi, F., R. A. Luccasen, and T. L. Turocy. (2019). 'Identifying discrete behavioural types: a re-analysis of public goods game contributions by hierarchical clustering', *Journal of the Economic Science Association,* vol. 5, pp. 238–254.

Fehr, E., and K. M. Schmidt. (1999). 'A Theory of Fairness, Competition, and Cooperation', *The Quarterly Journal of Economics,* vol. 114, pp. 817–868.

Fehr, E., and U. Fischbacher. (2004). 'Social norms and human cooperation', *Trends in cognitive sciences,* vol. 8(4), pp. 185–190.

Fischbacher, U., S. Gächter and E. Fehr. (2001). 'Are People Conditionally Cooperative? Evidence from a Public Goods Experiment', *Economics Letters,* vol. 71(3), pp. 397–404.

Gächter, S. (2007). 'Conditional cooperation. Behavioral regularities from the lab and the field and their policy implications ', In B. S. Frey & A. Stutzer (Eds.), Economics and psychology. A promising new cross-disciplinary field. Cambridge: MIT Press.

Gächter, S., L. Gerhards, and D. Nosenzo. (2017). 'The importance of peers for compliance with norms of fair sharing', *European Economic Review,* vol. 97, pp. 72–86.

Gächter, S., K. Lee, M. Sefton, and T. O.Weber (2021). 'Risk, Temptation, and Efficiency in the One-Shot Prisoner's Dilemma', CESifo Working Paper No.9449.

Gächter, S., D. Nosenzo, E. Renner, and M. Sefton (2012). 'Who Makes a Good Leader? Social Preferences and Leading-by-Example', *Economic Inquiry,* vol. 50(4), pp. 953–967.

Giamattei, M., Yahosseini, K. S., Gächter, S. and L. Molleman. (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.

Katuščák, P. & T. Miklánek. (2018). 'What Drives Conditional Cooperation in Public Goods Games?', CERGE-EI Working Paper Series No.631.

Keser, C., and F. van Winden. (2000). 'Conditional Cooperation and Voluntary Contributions to Public Goods', *The Scandinavian Journal of Economics,* vol. 102(1), pp. 23–39.

Kocher, M.G., T. Cherry, S. Kroll, R.J. Netzer, and M. Sutter. (2008). 'Conditional cooperation on three continents', *Economics Letters*, vol. 101(3), pp. 175–178.

Miettinen, T., M. Kosfeld, E. Fehr, and J.W. Weibull. (2020). 'Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions', *Journal of Economic Behavior and Organization,* vol. 173, pp. 1–25.

Muller, L., M. Sefton, R. Steinberg, and L. Vesterlund. (2008). 'Strategic behavior and learning in repeated voluntary contribution experiments', *Journal of Economic Behavior and Organization*, vol. 67, pp. 782–793.

Mullett, T., R. McDonald, and G. Brown. (2020). 'Cooperation in public goods games predicts behavior in incentive-matched binary dilemmas: Evidence for stable prosocialit', *Economic Inquiry*, vol. 58(1), pp. 67–85.

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

Selten, R. (1967). 'Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes', In: Sauermann, H. (Ed.), Beiträge zur experimentellen Wirtschaftsforschung. J.C.B. Mohr (Paul Siebeck), Tűbingen,, pp. 136–168.

Thöni, C., and S. Volk. (2018). 'Conditional cooperation: Review and refinement', *Economics Letters*, vol. 171, pp. 37–40.

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

Volk, S., C. Thöni, and W. Ruigrok. (2012). 'Temporal stability and psychological foundations of cooperation preferences', *Journal of Economic Behavior and Organization*, vol. 81, pp. 664–676.

Appendix A. *Experimental Instructions*

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.

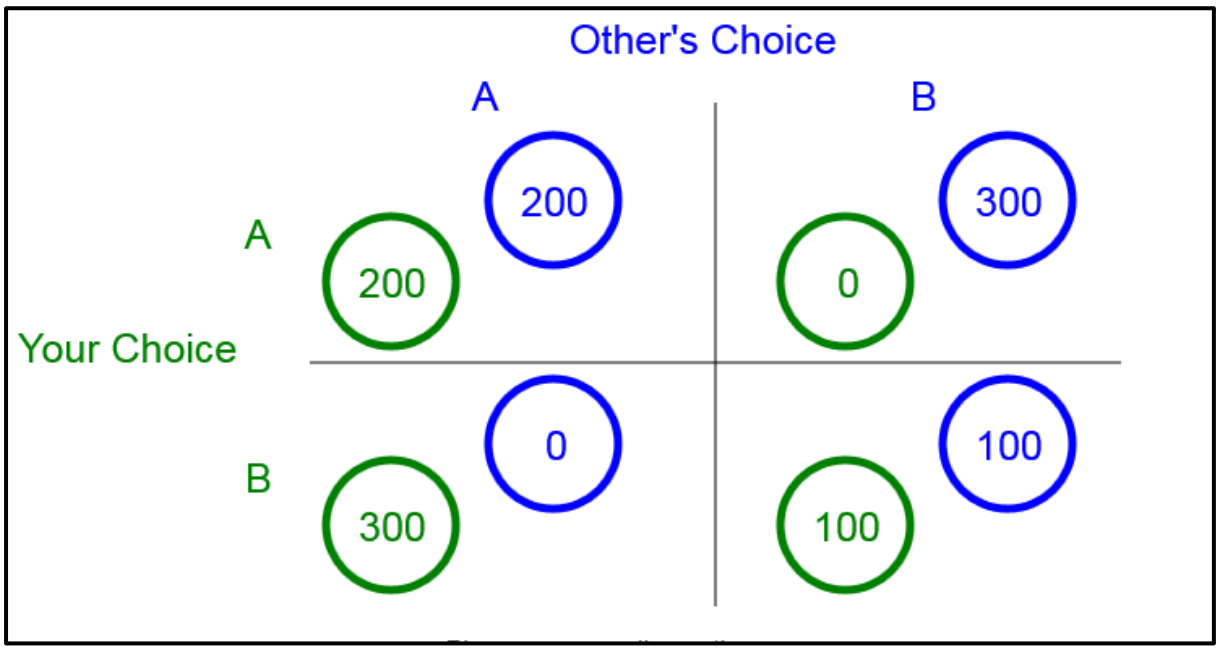
You will receive a code to collect your payment via MTurk upon completion.

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.

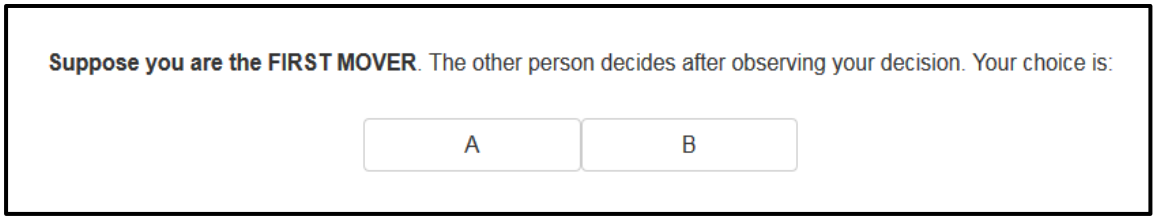
Please note that the values in the table will differ in each decision situation.

Tasks

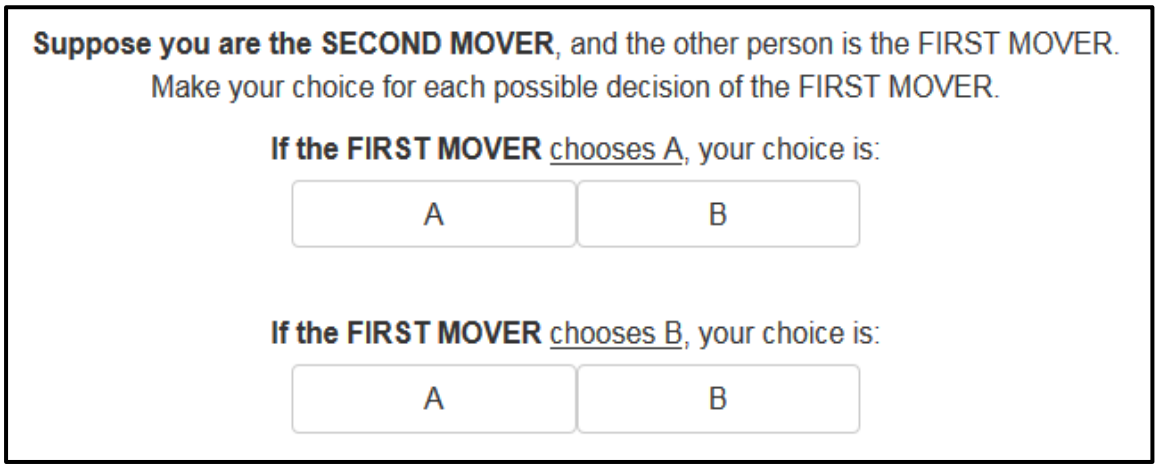
In each decision situation, you must complete **two types** of tasks, which we will refer to below as the "FIRST MOVER's decision" and "SECOND MOVER's decision". The FIRST MOVER decides first whether to choose A or B. The SECOND MOVER is then informed of the FIRST MOVER's decision. The SECOND MOVER then decides whether to choose A or B.

We want to know what you would do in the role of the FIRST MOVER and what would you do in the role of the SECOND MOVER. Thus you will be prompted to make decisions in both roles.

● For the "FIRST MOVER's decision" task, you will see the following screen and you must choose A or B:



● For the "SECOND MOVER's decision" task, You will see the following screen and you must choose A or B in two possible cases: (1) if the FIRST MOVER chooses A (2) if the FIRST MOVER chooses B



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. The computer will randomly choose either you or the other person to be the first-mover. If you are chosen to be the first-mover, your first-mover's decision will be matched with the second-mover's decision of the other person. If the other person is chosen to be the first-mover, your second-mover's decision will be matched with the first-mover's decision of the other person. 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 **the computer randomly selects you to be the first-mover. This implies that your payoff relevant decision will be your first-mover's decision.** Assume that you choose A as the first-mover's decision in the above example screen. Assume that the other person matched with you makes the following second-mover's decisions: he/she chooses A if you choose A, and chooses B if you choose B. As a consequence, you will earn 200 Tokens and the other person will earn 200 Tokens.

Example 2. Assume that **the computer randomly selects the other person to be the first-mover. This implies that your payoff relevant decision will be your second-mover's decision.** Assume that you make the following second-mover's decisions: you choose B if the FIRST MOVER chooses A, and choose B if the FIRST MOVER chooses B in the above example screen. Assume that the other person matched with you chooses A as the first-mover's decision. 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.