**Institute for Economic Studies, Keio University**

**Keio-IES Discussion Paper Series**

# 制度選択における個人･チーム不連続性効果―自発的公共財供給実験からの事実

**亀井 憲樹、ケイティ タベロ**

**2022 年 11 月 10 日**
**DP2022-015**
**https://ies.keio.ac.jp/publications/21393/**


Institute for Economic Studies, Keio University
2-15-45 Mita, Minato-ku, Tokyo 108-8345, Japan
ies-office@adst.keio.ac.jp
10 November, 2022

# 制度選択における個人･チーム不連続性効果―自発的公共財供給実験からの事実

亀井 憲樹、ケイティ タベロ

IES Keio DP2022-015
2022 年 11 月 10 日
JEL Classification: C92; D72; H41
キーワード: 制度;公共財;経済実験;罰則;不連続性効果

【要旨】

本研究の目的は、制度選択における意思決定主体としてのチームと個人の行動比較を、実験室内実験の手法を用いて考察することである。被験者は、有限回繰り返し公共財ゲームにおいて、正式な罰則とインフォーマルな罰則の選択を投票で行った。グループに正式な罰則スキームが遂行された場合には、チームは個人よりも非協力行動に対して強い抑止力を有する罰則制度を構築し、高い協力規範を実現した。グループがインフォーマルな罰則スキームを導入した場合には、チームは個人よりも非協力者に罰則を集中的に科し（反社会的な罰則行動を抑制し）、それにより高い貢献行動を持続した。このように検出された行動パターンは、熟議と学習を通じた行動経済学におけるいわゆる「真実は勝つ」効果（“truth wins”）で説明される。本実験結果は、ただ乗りなどモラルハザード問題への対処法として、組織内に意思決定主体としてチームを持つことの有効性を示唆している。

亀井 憲樹
慶應義塾大学経済学部
〒108-8345
東京都港区三田2-15-45
kenju.kamei@keio.jp

ケイティ タベロ
ダラム大学ビジネススクール
〒DH1 3LB
英国ダラム郡ミルヒルレーン

# The Individual-Team Discontinuity Effect on Institutional Choices: Experimental Evidence in Voluntary Public Goods Provision

This version: November 2022

Kenju Kamei[#1] and Katy Tabero[#2]

[#1] Faculty of Economics, Keio University, 2-15-45, Mita, Minato-ku, Tokyo 108-8345, Japan. Email: kenju.kamei@keio.jp.

[#2] Department of Economics and Finance, Durham University, Mill Hill Lane, Durham, DH1 3LB, United Kingdom. Email: katy.tabero@durham.ac.uk.

**Abstract:** A laboratory experiment is used to show that teams as a decision-making unit behave more efficiently than individuals in an institutional setting. Subjects make voting choices over formal versus informal (peer to peer) sanctions in a finitely repeated public goods dilemma. When a formal sanction scheme is selected in their groups, teams vote for deterrent sanction rates much more frequently than individuals. When an informal sanction scheme is selected, teams inflict costly punishment more frequently on low contributors than individuals, thereby reducing the relative frequency of "misdirected" punishment among teams. As such, teams sustain cooperation surprisingly better than individuals regardless of which scheme is enacted. These behavioral patterns are consistent with the idea of "truth wins" which proposes that teams achieve better choices than individuals through deliberation and learning. The results underscore the effectiveness of having teams as a decision-making unit in organizations in combating a moral hazard problem, such as free riding.

*Keywords:* institution, public goods, experiment, punishment, discontinuity effect

*JEL classification codes:* C92, D02, D72, H41

Acknowledgement: This project was supported by a grant-in-aid from the Japan Center for Economic Research. This project was also supported by Northern Ireland and North East Doctoral Training Partnership. The authors thank John Hey for his hospitality when they conducted the experiment at the University of York, and for Louis Putterman, Stefan Penczynski, Marie Claire Villeval, Daniel Li and the audience in the research seminar at the Gate Lab in Lyon and 2022 European ESA Meeting in Bologna for helpful comments. The authors also thank Mark Wilson (an IT manager at the University of York) for support in managing the computers and the setup of the z-Tree software in the experimental sessions.

## 1. Introduction

Teams have seen increasing popularity as a decision-making unit within organizations in the last half a century; this applies to both the public and private sector, and across a breadth of industries (see Lawler *et al.* 1992, 1995; Devine *et al.* 1999; Kersley *et al.* 2005). For example, Eurofound (2020) found that just under 70% of workers in the EU27 claimed to work as part of a team, and in only the transport industry did this fall to a low of 60%. Teams also form the basis of many decision-making units in the public sphere, ranging from the domestic context, such as councils (and also political factions), committees, and cabinets (ministries and agencies), to international relations, such as in international organizations like the United Nations, in which each country operates as a decision-making unit that summarizes their citizens' views and casts a single vote in making an organizational decision. The use of teams and team-based structures in an organization, especially those that offer more autonomy in terms of decision-making and problem-solving, has been linked to improved productivity and profitability under certain conditions (e.g., see Pfeffer 1998; Guzzo and Dickson 1996; Cohen and Bailey 1997, and Delarue 2008, for reviews and examples). Despite its importance, however, teams' institutional formation and their behaviors under endogenously constructed rules have not received attention in the literature on institutions.

Scholars studying workers' performances and interactions have actively used experimental games and human subjects in controlled laboratory settings for the last several decades. In such a setup, each worker subject is assigned to a group, given a fixed endowment, and simultaneously decides how to use the endowment (exert costly effort) for the group. Theoretically, optimal effort provision cannot be achieved in typical environments due to workers' free riding, whereby they pursue their own self-interest. A large number of experiments have been conducted in the social sciences (such as economics and political science) and in psychology to study worker behaviors in such voluntary provision of public goods when *individuals* are the decision-making unit in a group (see, e.g., Ledyard [1995] and Chaudhuri [2011] for a survey). It is now known that, without any institution to assist collaboration, while some individuals initially attempt to cooperate with their peers, cooperation cannot be sustained at a high level as they learn of their peers' opportunistic behaviors with repetition (e.g., Fischbacher and Gächter 2010). However, groups can sustain cooperation when the members can voluntarily monitor their peers' contribution behaviors (e.g., Grosse *et al.* 2011; Nicklisch *et al.* 2021), inflict costly punishment peer to peer (e.g., Fehr and Gächter 2000, 2002), or introduce a centralized incentive scheme regarding punishment and rewards (e.g., Falkinger *et al.* 2000). In particular, scholars have advanced the field during the last 15 years by exploring individuals' ability to *construct* and *operate* centralized governance by voting, finding that without any guidance, groups can achieve high efficiency through such endogenous institution formation, despite taking some time to learn better institutional formation (e.g., Gürerk *et al.* 2006; Kosfeld *et al.* 2009; Sutter *et al.* 2010; Ertan *et al.* 2009; Kamei *et al.* 2015; Fehr and Williams 2018). However, surprisingly, no attention has been paid to self-governance capacity and institutional formation when teams, as a decision-making unit (voter), constitute a group.

Theoretical modeling for decision-making by teams is usually based on the same assumptions made of the rational, self-interested individual. Hence, the neglect of teams' self-governance possibility is natural, and the use of *individuals* in a laboratory can be thought of as a simplification for experimentation in the literature. However, this assumption may not be correct according to the findings from another, but substantial, literature on group or team decision-making. This research area proposes the so-called "individual-team discontinuity effect" (simply "discontinuity effect," hereafter): teams may behave more efficiently than individuals (see, e.g., Charness and Sutter [2012], Kugler *et al.* [2012] and Kerr *et al.* [2004] for a survey). Such discontinuity effects have been detected in various setups, for example, in beauty contest games (e.g., Kocher and Sutter 2005), ultimatum games (e.g., Robert and Carnevale 1997; Bornstein and Yaniv 1998), signaling games (e.g., Cooper and Kagel 2005), centipede games (e.g., Bornstein *et al.* 2004), trust games (e.g., Kugler *et al.* 2007), coordination games (e.g., Feri *et al.* 2010), and monetary policy decisions (e.g., Blinder and Morgan 2005). It is possible that teams construct institutions differently from individuals in the voluntary provision of public goods.

This paper provides the first experimental study by utilizing a repeated linear public goods game and letting teams (decision-making units) govern their assigned group through building sanctioning institutions by voting. Members of each team communicate with one another to make joint voting and contribution decisions. The institutional formation and their behaviors under constructed institutions are compared against the case where the units are individuals to study the following specific questions:

***Question 1*:** Do teams utilize sanctioning institutions differently by voting from individuals?

***Question 2:*** If yes, how does the efficiency differ between individuals and teams? For example, do teams sustain cooperation more easily than individuals in an institutional setting?

There are two possible key hypotheses to these two questions. The first mechanism is the so-called Condorcet's (1785) jury theorem and behavioral public choice theorem (e.g., Ertan *et al.*, 2009). This states that if the probability of an individual being correct is larger (smaller) than ½, then the probability of the majority in a team choosing for the correct answer is larger (smaller) than that in which each individual votes correctly. This hypothesis is valid when members do not influence each other in deciding on a team decision (the "independence" condition). The other mechanism is the so-called "truth wins." This hypothesis states that teams can achieve more efficient outcomes through communication, learning and deliberation (Laughlin, 2015).

The experiment results are well consistent with the "truth wins" idea. First, remarkably, teams achieve much higher efficiency than individuals thanks to the former's effective use of the sanctioning institutions. In particular, given an opportunity to construct a formal sanction scheme, individuals vote for inefficient, non-deterrent sanction rates much more than 50% of the time. By sharp contrast, teams vote for deterrent sanction rates, i.e., the rates that make free riding materially unprofitable, more than 50% of the time. This pattern is inconsistent with the Condorcet's jury theorem, which underscores the role of influence and deliberation in the team decision-making procedure. When informal (peer-to-peer)

punishment is collectively enacted in a group, its teams punish low contributors more frequently than individuals, which helps reduce the relative frequency of "misdirected" punishment, i.e., punishment of high contributors. Moral hazard in groups is a central issue in organizations as it can hurt productivity (e.g., Holmstrom 1982). While recent experiments suggest that it can endogenously be resolved by allowing agents to construct institutions (e.g., Gürerk *et al.* 2006; Kosfeld *et al.* 2009; Sutter *et al.* 2010; Ertan *et al.* 2009; Kamei *et al.* 2015; Fehr and Williams 2018), the finding of the present study underlines the clear role of organizational structure in strengthening a group's ability to govern themselves, whether under formal or informal schemes. This would open up a new research direction in the field concerning the shape of efficient organizations.

The present paper is related to the large literature on the theory of the firm. Here, team decision-making is treated as a coordination problem in which the same processes involved in individual decision-making are used, but feature additional complexities relating to imperfect information, monitoring, and agency costs (e.g., Alchian and Demsetz 1972; Marschak and Radner 1972). Marschak and Radner (1972), for example, build a model using teams of individuals that have homogenous preferences (that align with the common goal), but heterogeneous information. It focuses on ways in which team members eliminate the intra-team information gap so as to face the same situation that an individual decision-maker faces. However, teams usually have difficulties in doing so, due to the costs of gathering information and mixed incentives of sharing information (see also Gibbons *et al.* [2013] for an overview).[1] By contrast, teams may be modeled as superior decision-makers to individuals when individuals are assumed to be bounded rational, due to the teams' increased ability to store and process information, e.g., through shared memory (Bainbridge 2002). Unlike the assumption of these prior studies, all team members in the present experiment have the *same* information described in the experiment instructions. The discontinuity effect detected in this study therefore suggests a need to bolster existing theoretical models, perhaps explicitly incorporating the beneficial communication, deliberation, and influence process with even symmetric information.

Further, this paper also contributes to empirical literature on management, organizational economics, and personnel economics that studies team decision-making and team production. First, prior research in management argues that managerial decision-making via top management *teams* can lead to better organizational outcomes, such as performance and innovation (e.g., Carmeli *et al.* 2008; Aboramadan 2020; Certo *et al.* 2008). The superiority of management teams is especially strong when the teams have great heterogeneity in terms of, say, age, education, and background (e.g., Aboramadan 2020; Certo *et al.* 2008). Nevertheless, it is difficult to draw causal inferences from these studies for various reasons, for example, because there is possible selection bias in the management team formation, and many studies rely on self-assessed/reported questionnaires. Second, team production (such as that in production

[1] Prior research in management has thus explored effective ways to coordinate and share information held by workers in organizations (e.g., Grant 1996).

sites) is also shown to lead to better work performance than individual production in the empirical research (e.g., Ichniowski *et al.* 1997), especially when teams have a greater spread in abilities across workers (e.g., Hamilton *et al.* 2003). However, the human resource practices in teams vary multiple dimensions simultaneously, making it difficult to identify the role of the team decision process in isolation. In addition, team decision-making per se is not the prior research's focus, and hence, scholars have not attempted to identify its treatment effects in the past.

The rest of the paper proceeds as follows: Section 2 describes the experimental design, Section 3 discusses hypotheses, and Section 4 reports experimental results. Sections 5 and 6 report results from finite mixture modeling and communication dialogues, respectively. Section 7 concludes. Appendix A in supplementary materials provides a summary of the related literature.

## 2. Experiment Design

The experiment is built on a linear public goods game. Subjects play the games under one treatment condition (between-subjects design).[2] Six treatments are constructed by varying two dimensions (Table 1). The first dimension is the decision-making unit, either an individual or a three-person team. The second dimension is the institutional environment; either there are no sanctioning institutions, or units can use sanction schemes. Two different strengths of punishment are considered because the efficacy of sanctioning mechanism can differ by its strength. The treatments are named as "I-No (Individual, No Voting)," "I-Voting (Individual, Voting)," "I-Voting-ST (Individual, Voting, Strong)," "T-No (Team, No Voting)," "T-Voting (Team, Voting)," and "T-Voting-ST (Team, Voting, Strong)."

The sanction scheme is designed based on Kamei *et al.* (2015). Each decision-making unit votes whether to execute a formal or informal sanction scheme in their group. A novel part of the design is that unlike all prior experimental studies on institutions (e.g., Kamei *et al.* 2015; Kosfeld *et al.* 2009; Traulsen

**Table 1:** *Summary of Treatments*

| Treatment name | Decision-making unit | Voting | Cost ratio in punishment[#1] | Number of groups (sessions) | Number of subjects |
|---|---|---|---|---|---|
| I-No | individuals | No | n.a. | 12 (2) | 36 |
| I-Voting | individuals | Yes | 1:3 | 11 (2) | 33 |
| I-Voting-ST | individuals | Yes | 1:5.5 | 11 (2) | 33 |
| T-No | teams | No | n.a. | 12 (7) | 108 |
| T-Voting | teams | Yes | 1:3 | 11 (6) | 99 |
| T-Voting-ST | teams | Yes | 1:5.5 | 11 (6) | 99 |
| Total | | | | 68 (25) | 408 |

*Notes:* [#1] The ratio (1: $x$) means that for each point a punisher spends in reducing another's payoff, $x$ points are deducted from the payoff of the punished. The ratio of 1:3 (1:5.5) means $x = 3$ and $y = 5$ ($x = 5.5$, and $y = 10$) – see Subsection 2.2.

[2] A between-subjects design is more appropriate than a within-subjects design to avoid possible democratic spill-over (e.g., Kamei 2016) or behavioral spill-over effects (e.g., Bednar *et al.* 2012; Cason *et al.* 2012).

*et al.* 2012, Zhang *et al.* 2014; Kamei 2019a; Fehr and Williams 2018), the present study is the first to explore endogenous institutional choices when the decision-making units are *teams*. The treatments with individuals being as the decision-making units will act as a control treatment.

### *2.1. Common Features in All Treatments*

A partner matching protocol is used in all six treatments. At the onset of the experiment, decision-making units are randomly assigned to a group whose size is three (three individuals or three teams, dependent on the treatment), and the group composition stays the same throughout the entire experiment. The number of periods is set at 24 to allow for the evolution of institutional choice and cooperation behavior over time. The periods are grouped into six phases of four periods each (Figure 1). The number of periods is common knowledge to the subjects. Subject identity is kept anonymous in the experiment.

In each period, every decision-making unit will be assigned an endowment of 20 points (62.5 points = 1 pound sterling), and then simultaneously decide how many points to allocate between their private and public accounts. Contribution amounts must be non-negative integers and not exceed 20. A marginal per-capita return (MPCR) is set at 0.6. In other words, when decision-making unit *i* contributes $c_{i,t}$ to the public account, she receives the following payoff $\pi_{i,t}$ in that period:

$$\pi_{i,t} = (20 - c_{i,t}) + 0.6 \sum_{j=1}^{3} c_{j,t}. \quad (1)$$

In the three treatments with teams, each member in a team *i* receives the team's payoff to make the payoff consequence the same for team members in the team treatments and individuals in the individual treatments.[3] At the end of a given period, each unit is informed of (i) their own payoff and (ii) the amounts contributed to the public account by two other units in their own group in a random order.

The structure of Phase 1 (also called "Part 1") is the same for all six treatments. Subjects repeat the public goods game without any sanctioning opportunities (No Sanction [NS] scheme, hereafter) four times with the same group membership, thereby helping subjects learn the basic structure of the PGG game and the dynamic free riding problem. Phases 2 to 6 (collectively "Part 2," hereafter) differ by whether they can use sanction schemes, as summarized in Sections 2.2 and 2.3. Panels A and B of Figure 1 summarize the schematic diagrams.

### *2.2. The Individual Treatments*

Each phase of Part 2 in the I-Voting and I-Voting-ST treatments begins with each decision-making unit voting on the *formal* versus *informal* sanction scheme (FS and IS hereafter).[4] Voting is mandatory and does not cost subjects. As discussed below, theory predictions based on the selfishness of players are different between FS and IS. At the beginning of each phase, the decision-making units decide

[3] The same per-subject payoff consequences for individuals and teams are usually used in the design of prior related studies on team decision-making (e.g., Cason and Mui 1997; Kamei 2019b).
[4] The formal (informal) sanctioning scheme is called group determined fines (individual reduction decisions) in the experiment. The same wording was used in the experiment sessions of Kamei *et al.* (2015).

on which scheme they would prefer to use (Figure 1.B). Whichever scheme receives the majority of votes (i.e., more than or equal to two votes) will be enacted in that group for all four periods of the phase. Sections 2.2.1 and 2.2.2 summarize the details of the IS and FS schemes, respectively.

In the sanction-free I-No treatment, subjects play the PGG under the NS scheme for all five phases in Part 2 (Figure 1.A). There is a 40-second pause between the adjacent phases to control for the restart effects (Andreoni 1998; Kamei *et al.* 2015) that may be present in the voting treatments.

*2.2.1. Informal Sanction Scheme*

If IS is chosen, a punishment stage follows the allocation stage in each period of the phase. In the punishment stage, a decision-making unit $i$ can reduce the payoff of each of the other units ($j$) in their group by assigning punishment points $p_{i\to j} \in \{0, 1, 2, \ldots, 10\}$ at a private cost. While each punishment point costs the recipient $x$ points ($x > 1$), it costs the punisher one point. Following a prior experimental framework (e.g., Fehr and Gächter 2000, 2002), the punishment points allocated by others cannot make the recipients' earnings for that period negative. However, each decision-making unit always incurs the cost of imposing punishments. The payoff for unit $i$ in period $t$ playing IS can be expressed as follows:

$$\pi_{i,t} = \max\left\{\left(20 - c_{i,t}\right) + 0.6\sum_{j=1}^{3} c_{j,t} - x\sum_{j\neq i} p_{j\to i}, 0\right\} - \sum_{j\neq i} p_{i\to j}. \quad (2)$$

To limit delayed revengeful punishment among members, contribution decisions of the other two decision-making units appear in a random order in the punishment stage (e.g., Fehr and Gächter 2000, 2002; Denant-Boemont *et al.* 2007; Kamei *et al.* 2015).

At the end of the punishment stage, subjects are informed of (i) the total payoff reductions due to punishment points imposed by the other two group members (in total, not broken down by member), (ii) the total cost spent imposing punishment on other members, and (iii) their own payoffs.

*2.2.2. Formal Sanction Scheme*

When FS is in place, the allocation stage is followed by an automatic punishment stage in each period. Specifically, groups in the FS scheme determine by voting in advance at what rate allocations to the *private* account are penalized. Voting is mandatory and cost-free. The available sanction rates (*SR*, hereafter) are 0.0, 0.2, 0.4, 0.6, 0.8, 1.0, and 1.2. A median voting rule is used. Participants vote four times, once at the onset of each period for that phase (a new rate can be selected in each period).

Imposing sanctions is costly. First, there is a fixed cost (administrative cost) imposed when using the FS of 4 points, which is applied to each unit's payoff for that period (See the 6-C treatment of Kamei *et al.* [2015]). Second, when a member is fined, the group incurs a variable cost of imposing the sanctions, i.e., $3/y$ of the amount deducted from the member's payoff. The cost is equally shared among the three units in the group, meaning that each unit pays $(1/3)(3/y) = 1/y$ of the sanctions.[5]

[5] To mirror the cost of informal punishment, the FS scheme features a proportional cost. However, unlike the IS mechanism the variable cost will be borne by the whole group.

To parallel the IS scheme, the deductions resulting from formal punishment cannot result in a negative payoff, but the cost of implementing those sanctions and the administrative cost can. Specifically, the payoff of decision-making unit *i* for that period is calculated first using Equation (1), and then the sanction rate is applied to the amount that *i* held in the private account. If applying the sanction rate results in a negative payoff, then it will be set at 0 (otherwise it will not be changed). The shared cost of imposing the sanctions and the administrative cost are then deducted from that period's earnings. In short, the payoff for decision-making unit *i* in period *t* under the FS scheme is calculated as follows:

$$\pi_{i,t} = \max\left\{\left(20 - c_{i,t}\right) + 0.6\sum_{j=1}^{3} c_{j,t} - SR_t(20 - c_{i,t}),0\right\} - \frac{1}{y}\sum_{j=1}^{3} SR_t\left(20 - c_{j,t}\right) - f, \quad (3)$$

where $f = 4$ (administrative cost). Should the group select a sanction rate of 0.0, their payoffs would remain effectively unchanged from that without the FS scheme (however, they still incur the fixed administrative cost of 4 points per period). If *i* receives a negative payoff due to strong punishment, then it will be deducted from their accumulated payoffs from other periods.

Equation (3) suggests that for each sanction imposed, the cost ratio between the punished and the punishers is 1 + 1/y : 2/*y* (punished decision-making unit: two other units). To further make the FS scheme comparable to the IS scheme, the cost ratio is set to be the same as the IS scheme, namely, $1 + 1/y: 2/y = x: 1$. This reduces to the following condition for *x* and *y*.

$$y = 2x - 1. \quad (4)$$

Modest punishment intensity, $x = 3$ and $y = 5$, is used for the I-Voting and T-Voting, while strong punishment intensity, $x = 5.5$ and $y = 10$, is used for the I-Voting-ST and T-Voting-ST treatments.

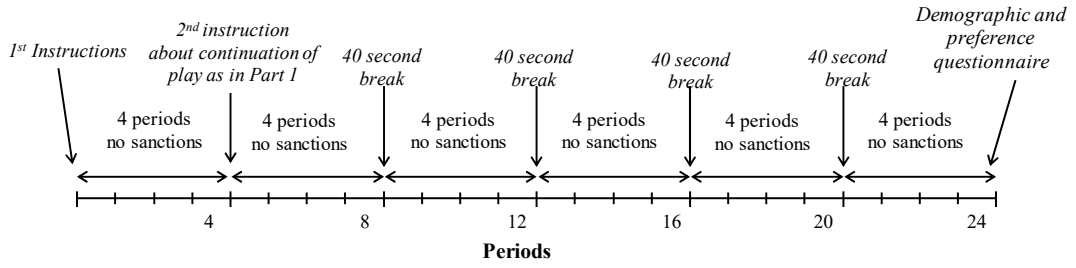
At the end of each period, they are informed of (i) the two other units' allocation decisions in a random order, (ii) their own payoffs before reductions, (iii) their final payoffs in the period, and (iv) a breakdown of reductions due to fines, the cost of administering fines, and the fixed administration cost.

*2.3. The Team Treatments*

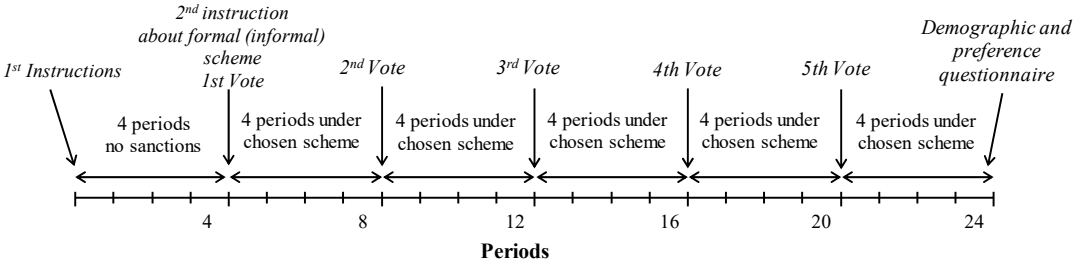
The T-No, T-Voting and T-Voting-ST treatments are identical to, respectively, the I-No, I-Voting and I-Voting-ST treatments (Figure 1), except that the decision-making units are *three-person teams*, not individuals. Three subjects playing as a team will jointly make a single decision as a decision-making unit. At the onset of the experiments, subjects are randomly assigned to a team of three, and the team composition does not change throughout the entire experiment. The teams are then randomly assigned to a group of three teams (thus each group consists of nine subjects) before the experiment commences.

The team's joint decision-making follows Kamei (2019b, 2021). Three members in a team communicate with each other for 60 seconds using a computer chat screen before making each team decision. This enables us to perform transcript analysis post-experiment. Members are not allowed to communicate verbally, eliminating the risk of contamination of the experiment which may occur if players were able to overhear another team's discussions. The members are only able to communicate

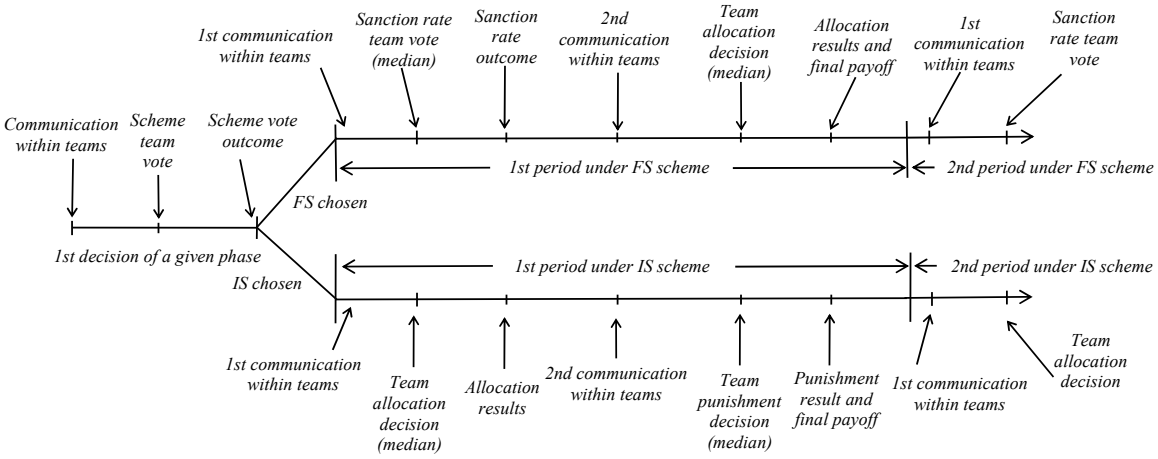
**Figure 1:** *Schematic Diagram*




(A) I-No and T-No treatments




(B) I-Voting, I-Voting-ST, T-Voting and T-Voting-ST treatments




(C) Phase Structure in the T-Voting and T-Voting-ST treatments

with other members of their own team.[6] Anonymity is kept preserved, such that the subjects are identified by fixed Player IDs in the chat screen, and they are instructed that disclosing any information that may

[6] The use of electronic chat windows is one of the most common procedures (e.g., Charness and Sutter 2012; Kugler *et al.* 2012). While some studies set the duration of each communication stage much more than 60 seconds, prior papers such as Kagel (2018) and Kamei (2019b) set the duration of each communication stage to 60 seconds or less. The authors read all the teams' communication logs and counted the number of *explicit* agreement cases (treated a communication log as a disagreement if there was no communication, only irrelevant communication, or one of three team members did not communicate, unless a pre-agreed strategy was still in play, agreements were considered implicit, or teams disagree or do not try to reach consensus). Even with such strict judgment, at least 80.5% of team decisions were classified as agreed. This suggests that the 60-seconds duration was sufficient for the communication. A detailed analysis of communication logs will be executed in Section 6.

identify themselves or using offensive language in communication is prohibited.[7]

A team's three joint decisions are determined using the median voting rule – see Figure 1.C. This includes the allocation decisions in the public goods game (all team treatments), and punishment decisions under the IS scheme and sanction rate votes under the FS scheme (T-Voting and T-Voting-ST treatments). The specific procedure is as follows: The three members in a team first discuss strategies and decisions with their team. After the communication stage, each member privately and simultaneously submits their preferred decision (e.g., an amount they wish to contribute as the team's joint contribution decision).[8] The median of three submissions becomes the team's decision. Each team member is informed of the submissions of their two other team members, anonymously and in a random order.

A team's joint scheme choice (FS or IS) is based on a majority rule. As in the other team decision-making, each team member votes on which scheme they prefer after communication, with the team's majority choice (an option with at least two votes) being the team's joint voting decision.[9]

## 3. Hypothesis

Standard theory based on the assumption of agents' self-interest and common knowledge of rationality is straightforward when sanctioning schemes are absent (I-No and T-No treatments, and Phase 1 of the four voting treatments). Equation (1) implies that $\partial\pi_{i,t}/\partial c_{i,t} = -0.4 < 0$, which means that contributing nothing to the public account is the strictly dominant strategy for every decision-making unit. Thus, mutual free riding characterizes the unique Nash Equilibrium of this game. Repetition does not alter this prediction with the logic of backward induction.

The standard theory assumption also predicts that having IS does not alter equilibrium play from that in the NS scheme because punishment activities are costly. From Equation (2), $\partial\pi_{i,t}/\partial p_{i\to j} = -1 < 0$ for all *i*. Thus, it is materially beneficial for each unit to not punish one another ($p_{i\to j} = 0$), in which case their payoff would be unaffected when compared to the payoff in the allocation stage (Equation (1)).

In contrast, standard theory prediction based on pure selfishness is different in the FS scheme from that in the NS or IS scheme (e.g., Falkinger *et al.* 2000; Kamei *et al.* 2015). $\partial\pi_{i,t}/\partial c_{i,t}$ is calculated from Equation (3) as: $-0.4 + SR_t(1 + 1/y)$. This means that units contribute nothing when the enacted *SR* is 0.0 or 0.2 as then $\partial\pi_{i,t}/\partial c_{i,t} < 0$, but they contribute the full endowment amount when $SR \geq 0.4$. Each unit obtains a payoff of 32 points (= 0.6 × 60 – 4) when a deterrent sanction rate is enacted, while

[7] A subject receives a fine of 10 pounds with an apparent violation of this rule. No one disclosed any identifiable information, and only seven out of 306 subjects (2.28%) had to pay the fine with the rule of offensive language.
[8] Where the team members agree on a decision, they can submit that decision. If they do not agree on a decision as a team, however, they can submit whatever decision they prefer. Three team members submitted the same decisions in almost all cases in the team treatments (2,049 out of 2,448 team allocation decisions, 581 out of 672 team sanction rate votes, and 1,176 out of 1,296 team informal punishment decisions).
[9] All three team members agreed how to vote in almost all cases in the experiment (they submitted the same vote in 278 out of 330 cases).

they obtain a payoff of 16 points (= 20 + 0.6 × 0 – 4) when a non-deterrent sanction rate is enacted instead. Hence, the theory predicts that groups would choose FS rather than IS, and then vote for a deterrent sanction rate, after which each unit contributes the full endowment amount to the public account (trembling-hand perfect equilibrium).

However, the superiority of the FS over the IS scheme changes under certain conditions, once people's other-regarding preferences are considered (see, e.g., Fehr and Schmidt 2006 and Sobel 2005 for a survey). Prior experimental research has shown that real human subjects can sustain contributions at high levels under certain conditions when the IS scheme is available (e.g., Fehr and Gächter 2000, 2002; Anderson and Putterman, 2006; Nikiforakis and Normann, 2008) due to peer-to-peer punishment inflicted driven by such non-selfish preferences. The costly punishment activities and the maintenance of contributions can be rationalized successfully by, for example, the inequity-averse preference model, as has been proven in Fehr and Schmidt (1999). Thus, people's institutional choices between the IS and FS schemes are not obvious for real human subjects as they can achieve high cooperation regardless of which scheme is selected, and they may prefer IS to FS to avoid a fixed administrative cost. Empirically, such choices are known to be affected by which scheme is more materially beneficial, as shown by many prior experiments (e.g., Gürerk *et al.* 2006; Kosfeld *et al.* 2009; Sutter *et al.* 2010; Ertan *et al.* 2009; Kamei *et al.* 2015; Fehr and Williams 2018).

The main aim of this paper is to investigate how teams utilize sanctioning institutions differently from individuals, and as a result how teams make contribution decisions differently from individuals under a given sanction scheme. As described above, a theoretical analysis that does not incorporate the internal aspects of team decision-making suggests the same behaviors for teams and individuals. However, such analysis misses an important dimension for teams, namely, intra-team dynamics or the preference aggregation process.

A model that may then be applicable to the present setup is that of household bargaining behavior (see, e.g., Chiappori and Mazzocco [2017] for a survey). The models in this area explicitly include intra-household members' utilities to explain the household's behavior as a whole when only observable behavior is aggregated for the household. It is possible to treat household members and the whole household as, respectively, team members and the team in the model if it is applied to the context of the present study. The model, however, does not predict different behaviors for individuals and teams in the present experiment, and its predictions are the same as those without considering the intra-team dynamics just discussed, as explained below.

The model of household behavior considers two cases: one in which preferences are other-regarding among household members (i.e., the utility of each household member is affected by other household members' utilities), and the other in which they are egotistic (i.e., the utility is not affected by other household members'). It is obvious that if team members are assumed to have egotistic preferences, then there will be no effects due to the difference in the decision-making units, because the utilities of

subjects, whether individuals in the individual treatments or teams or team members in the team treatments, are expressed by the same forms, for example, by Equations (1), (2) or (3) dependent on the treatment condition. If it is instead assumed that people have other-regarding preferences, the utility of member $i$ in a team $k$ is expressed as: $U_i(u_i, u_j, u_l)$, where $j$ and $l$ are the two other team members of $i$ in team $k$. With $s$ being the utility weight of regard for others outcomes, the utility is further expressed as: $U_i = u_i + s \cdot u_j + s \cdot u_l = (1 + 2s) \cdot u_i$, which is just a linear transformation of one's own utility, because the three members of team $k$ have the same payoff consequences. This expression clearly demonstrates that there are no effects relating to the decision-making unit in this setup.

***Hypothesis 1*** *(theory without considering the internal aspects of team decision-making, or the model of household bargaining behavior): The difference in the decision-making format, team- or individual decision-making, does not have any impact in the experiment.*

The approach just discussed does not consider the preference or information aggregation process when three members in a team make a joint decision. There are two other different theories that may explain differing behaviors between teams and individuals. The first one relates to the notion of 'wisdom of the crowds' and is captured by theories of preference aggregation such as the so-called Condorcet jury theorem, or what Ertan *et al.* (2009) call the "behavioral public choice theorem" (also see Hauser *et al.* 2014). Condorcet's (1785) jury theorem suggests that, assuming the event of individuals in a given population choosing the correct answer is independent (unconditional independence), the probability of the majority of individuals voting for a correct answer hinges on the individual correctness probability. That is, if the probability of an individual being correct exceeds ½, and that this is the same for all individuals in the population (unconditional competence), then the probability of the majority in a group voting for the correct answer is larger than that in which each individual votes correctly. Similarly, if general competence is below ½, then the probability of a majority voting for an incorrect answer is worse in a group than by individuals and increases with group size. In terms of this experiment, it can be interpreted that the behavior of individuals in the individual treatments would indicate the individual correctness probability. Should individuals be able to select the strategically correct answer more than 50% of the time, then it follows that teams of three formed of similar individuals will have a greater probability of voting for the strategically correct answer, whether a deterrent sanction rate or lack of punishment. This tendency is summarized as Hypothesis 2 below:

***Hypothesis 2*** *(Condorcet jury theorem, and the so-called behavioral public choice theorem)****:*** *(a) If the majority of individuals vote for deterrent (non-deterrent) sanction rates, then teams are more likely to vote for deterrent (non-deterrent) sanction rates. As a result, teams achieve higher (lower) levels of contributions under the FS scheme. (b) If pro-social punishment is more (less) prevalent than anti-social punishment among individuals, then teams inflict punishment more (less) pro-socially, thereby achieving higher (lower) levels of contributions under the IS scheme, than individuals.*

Hypothesis 2 does explain voting outcomes in prior experimental research where each individual votes independently as the decision-making unit (Ertan *et al.*, 2009; Hauser *et al.*, 2014). It should be worth noting, however, that Hypothesis 2 is valid only when the independence of the probability of individuals being correct holds. This independence assumption is unlikely to hold for the present experiment as the three team members have intra-team communication and deliberation before making each team decision. In this sense, team decision-making may be more than the aggregation of three members' preferences, and as such one can expect that Hypothesis 2 will not hold perfectly.

The second mechanism similarly explores preference combination, but unlike the models discussed above, does *not* assume independence. Instead focusing on influence and learning, this mechanism is more closely related to the notion of "truth wins." The generalization by Friedkin and Bullo (2017) of the DeGroot (1974) learning model takes the initial set of preferences or judgements for those within a team and allows for teammates to influence each other over time using a weighted averaging mechanism. This mechanism takes into account attachment to one's own judgement as well as the influence of the other members' judgements (which may be 0, as in the case of independence or individual decision-making). As Friedkin and Bullo (2017) note, it is possible for teams to converge to both correct and incorrect conclusions depending on the distribution of initial judgements and the calculative logic adopted.

In the context of the present experiment, correct answers can be considered the options that lead to the highest utilities, since, as already discussed, voting patterns revealed in prior experiments on institutions showed the strong behavioral effects of material outcomes in driving units' choices of sanctioning institutions (e.g., Gürerk *et al.* 2006; Kosfeld *et al.* 2009; Sutter *et al.* 2010; Ertan *et al.* 2009; Kamei *et al.* 2015; Fehr and Williams 2018). Thus, the learning model allows for one or more team members to persuade their teammates of the correct logic and so lead them to a better decision, i.e., selecting deterrent sanction rates when the FS scheme is in effect, and punishing low rather than high contributors when the IS scheme is in effect.[10] Recent experiments on problem-solving suggest asymmetry regarding influence and persuasion. For instance, He *et al.* (2022) found that more cognitively able and knowledgeable members can influence less knowledgeable members more strongly if they work together, thereby making it easier for the latter to find correct answers while the former is little affected by incorrect suggestions made by the less able member. Similarly, Schulze and Newell (2016) found that group members claimed that their group made decisions by implementing the most effective strategy proposed by one of its members when confronting probability matching bias, enabling them to outperform most individuals and match their best individuals. Further, Bonner *et al.* (2002) report that groups utilize ranking information to assign more weight to the suggestions of high performers when completing

[10] Well-targeted peer-to-peer punishment plays a vital role in achieving a high payoff for people in the IS scheme (e.g., Hermann *et al.*, 2008).

moderately difficult problem-solving tasks. Prior experimental research on team decision-making in a different context also supports the role of deliberation and learning as a mechanism of improved decision-making. For instance, Cooper and Kagel (2022) find that teams learn to outperform individuals over time through careful consideration and experimentation with strategies when no singular optimal strategy is available. In sum, while we may still see examples of teams that converge to poor strategies when every member, or a particularly dominant member, is incorrect, it is expected on average that deliberation and learning will allow teams to discover the optimal strategy more quickly than individuals.

***Hypothesis 3*** *(truth wins)**: (a) Under the FS scheme, teams will enact deterrent sanction rates more frequently than individuals, and as a result, the former contribute larger amounts than the latter. (b) Under the IS scheme, teams will inflict punishment more selectively on low, rather than high, contributors, thereby achieving higher contribution norms, than individuals.*

## 4. Experimental Results

The experiment was conducted at the University of York (see Appendix E.1 for the implementation). Section 4.1 provides an overview of the decision-making units' average behaviors and examines treatment differences in contributions and payoffs. Section 4.2 investigates scheme voting behavior, while Section 4.3 provides a comparison between units in utilizing the sanctioning institutions.

### *4.1. Treatment Differences in Contributions and Payoffs*

Groups experienced typical free riding dynamics when sanctioning schemes were unavailable (Figure 2). The average contribution of individuals in the I-No treatment began at 62% of the endowment and gradually decreased over time. In line with the literature, mild restart effects were seen in Phases 4 to 6 (Andreoni 1988), and end-game defection was evident in period 24. The average contribution across all periods was 10.19 points (50.9% of the endowment) in the I-No treatment. The average contribution of teams was also modest, 10.57 points (52.9% of the endowment), in the T-No treatment, and the dynamics followed a declining contribution trend, similar to that of individuals in the I-No treatment.[11] No difference by the decision-making unit is unsurprising because of the floor effect, typical to the serious free-riding dynamics in repeated PGG.

A key comparison in this study is units' decisions to contribute under voting. Contribution trends differ drastically between individuals and teams when they voted on the sanctioning schemes. This is clearly at odds with Hypothesis 1. The difference was especially large under the mild punishment intensity (Figure 2.A). On the one hand, teams in the T-Voting treatment learned to cooperate gradually from phase to phase. Remarkably their average contribution amounts were more than 80% of the endowment in the final three phases. On the other hand, individuals in the I-Voting treatment did not

[11] Unlike this trend, teams cooperated much more strongly than individuals in Kamei (2019b) where the group size was two. The difference between Kamei (2019b) and the T-No treatment, however, resonates with the idea that cooperation is more difficult to evolve when the group size is three, rather than two (e.g., Cox and Stoddard 2018).

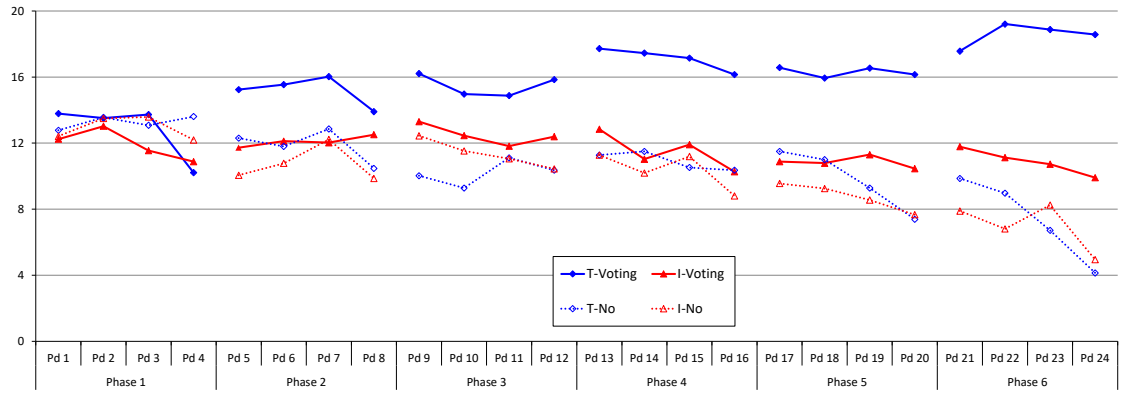
follow as strong a learning pattern, although they did not learn to free ride either. The individuals' average contribution amounts hovered between 10 and 12 points. The clear difference between the T-Voting and I-Voting treatments is consistent with the discontinuity-effect hypothesis, demonstrating its application in an institutional choice setting.

When the punishment intensity was strong, cooperation evolved at a further higher level among teams – see Figure 2.B. The average contribution in the T-Voting-ST treatment was close to the full contribution level in each phase of Part 2. With strong punishment, individuals (in the I-Voting-ST treatment) were also able to gradually learn to cooperate over time. The difference between the I-Voting-ST and I-Voting treatments suggests that individuals' contribution behaviors are sensitive to the punishment effectiveness, as has been shown by Anderson and Putterman (2006) and Nikiforakis and Normann (2008). Having said this, the difference in the average contribution was consistently large between individuals and teams even under the strong punishment intensity.

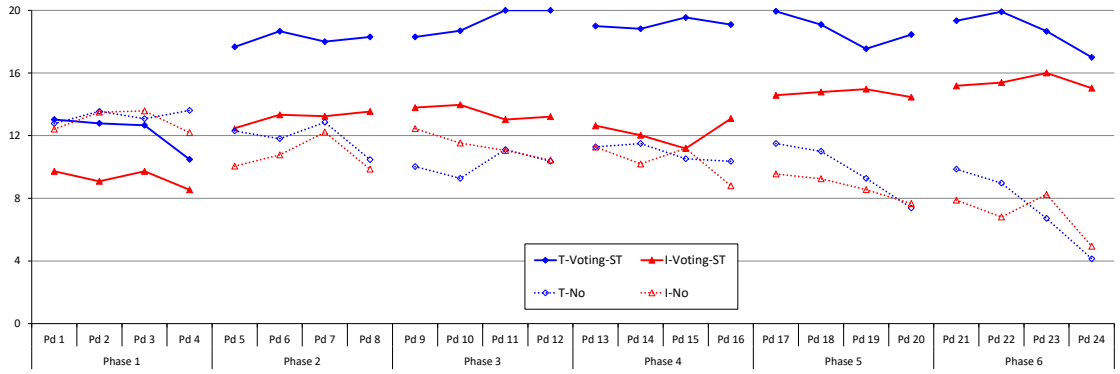
So, why did teams perform better than individuals when they voted on sanctioning schemes? An answer may be either by the mere structure that teams are composed of three members and aggregate the preferences when deciding on a team decision (Hypothesis 2), or by the tendency that teams use punishment opportunities more efficiently than individuals driven by the former's communication, deliberation and learning (Hypothesis 3). An analysis in Section 4.3 will reveal the validity of Hypothesis 3.

Figure 3 reports the trends of average payoffs. It shows first that individuals persistently incurred large losses due to punishment when its intensity was modest, consistent with the idea that individuals' failure to learn to cooperate seen in Figure 2.A triggers negative emotional responses from their peers (e.g., Casari and Luini 2009; Gächter *et al.* 2008). As a result, individuals received lower payoffs in the I-Voting than in the I-No treatment in all phases except Phase 6 (Figure 3.A). Second, teams also experienced such negative welfare losses under the modest punishment intensity (Figure 3.A). However, the negative impact in the T-Voting treatment was limited to Phases 2 and 3. Instead, the teams achieved *higher* payoffs in Phases 4 to 6, relative to the T-No treatment. Considering the teams' increasing

**Figure 2:** *Average Contribution Period by Period*




(A) Treatments with Modest Punishment Intensity




(B) Treatments with Strong Punishment Intensity

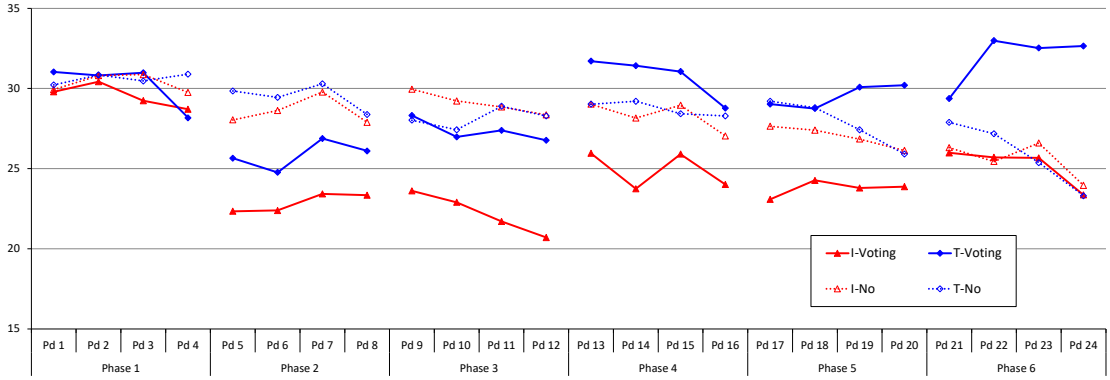
*Note*: The unit of the vertical axis in each panel is points.

contribution trend, this implies that, in later phases, teams did not need to discipline their group members through costly punishment, because the group successfully cooperated then (Figure 2.A).

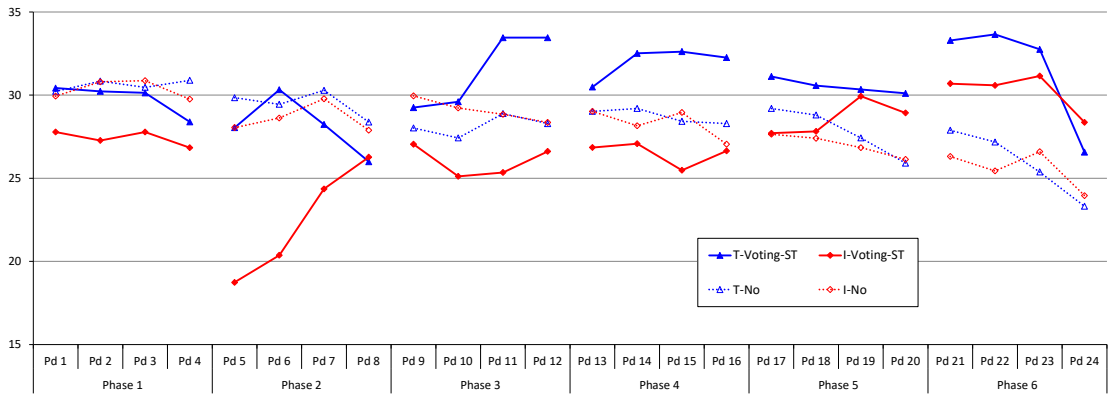
Third, likewise, when the punishment intensity was strong, having the sanctioning schemes led to similar negative welfare consequences in groups. However, the duration in which groups suffered from losses was shorter relative to the treatments with modest punishment (Figure 3.B). In other words, the availability of strong punishment induced the members to learn to cooperate smoothly, thereby helping reduce the welfare loss due to punishment activities.

A series of non-parametric tests were performed to judge treatment differences statistically (Table 2), which confirms most of the patterns seen in Figures 2 and 3. First, without the sanctioning schemes, units (whether individuals or teams) had a significantly lower level of contribution in Part 2 (Phases 2 to 6) than in Part 1 (Phase 1) of the experiment. Second, in both the T-Voting and T-Voting-ST treatments, teams' contribution behaviors were significantly stronger in Part 2 than in Part 1. As a result, the teams did not experience a drop in payoffs after Part 1, unlike in the T-No treatment. An across-treatment comparison in Part 2 further demonstrates that teams contributed larger amounts when the sanctioning schemes were

**Figure 3:** *Average Payoff Period by Period*




(A) Treatments with Modest Punishment Intensity




(B) Treatments with Strong Punishment Intensity

*Notes*: The unit of the vertical axis in each panel is points.

available than otherwise (see $H_0$: (c) = (d) in Table 2).[12] Third, individuals earned significantly less in Part 2 than in Part 1 of the experiment in the I-Voting treatment, but not in the I-Voting-ST treatment.[13]

Lastly, a closer look at the data by scheme uncovers three further patterns. First, in the I-Voting and I-Voting-ST treatments, while cooperation did not evolve when FS was in place, individuals maintained strong cooperation norms when IS was instead in effect (panels (a) and (b) of Appendix Figure C.1). Hence, the individuals' overall failure to learn cooperation (Figure 2) is partly attributable to their selection of sanction rates and/or contribution behaviors under the FS scheme. Second, due the low cooperation norms and administrative cost payments, the individuals persistently earned much less under the FS scheme, relative to the I-No treatment (panels (a) and (b) of Appendix Figure C.2), and the difference is significant (Table 2). Under the IS scheme, individuals in the I-Voting (I-Voting-ST) treatment successfully cooperated with each other in Phase 6 (from Phase 4),[14] but they received lower payoffs than those in the I-No treatment in Phases 2 to 5 (Phase 2 to 3). The low payoffs in the earlier phases were due to losses from intensive punishment activities. Hence, learning to cooperate with informal punishment requires a sufficiently long length of interactions in the experiment, as Gächter *et al.* (2008) demonstrated.

[12] The same positive effect can be found even if the two team treatments are not pooled (Panel C of Appendix B).
[13] A regression was also performed as a supplementary analysis to analyze the contribution trend in Part 2 (Appendix Table C.1). It confirms that when the sanctioning schemes were unavailable, units, whether individuals or teams, decreased contribution amounts significantly over time. By contrast, teams increased contribution amounts significantly from phase to phase in both the T-Voting and T-Voting-ST treatments. A regression also confirms that the contribution trend differs by punishment intensity when the units are individuals: an increasing (somewhat decreasing) contribution trend in the I-Voting-ST (I-Voting) treatment. It further shows that the payoff trend is similar to the contribution trend: declining trends for the I-No and T-No treatments *versus* an increasing trend in the T-Voting treatment (the maintenance of high payoff in the T-Voting-ST treatment) – see Appendix Table C.2.
[14] A group-level Mann-Whitney test finds that the average contribution in Phase 6 (Phases 4-6) under the IS scheme in the I-Voting (I-Voting-ST) treatment is different from that in the I-No treatment at two-sided $p = 0.0709$ (0.0196). Likewise, the average payoff in Phase 6 (Phases 4-6) under the IS scheme in the I-Voting (I-Voting-ST) treatment is different from that in the I-No treatment at two-sided $p = 0.0709$ (0.0245). Note that there were only three groups playing the IS scheme in Phase 6 for the I-Voting treatment, making statistical significance difficult to obtain.

## **Table 2:** *Average Contribution and Payoff*

I. Contribution

| | Avg. contribution based on all data | | | Avg. contribution under a given sanction scheme in Phases 2-6 | | | | |
|---|---|---|---|---|---|---|---|---|
| | (i) Phase 1 | (ii) Phases 2-6 | *p*-value for $H_0$: (i) = (ii) | (iii) FS | *p*-value for $H_0$: (i) = (iii) | (iv) IS | *p*-value for $H_0$: (i) = (iv) | *p*-value for $H_0$: (iii) = (iv)[#1] |
| [Individual treatments:] | | | | | | | | |
| (a) **I-No** | 12.92 | 9.64 | 0.0414** | --- | --- | --- | --- | --- |
| (b) **Indiv Voting** (I-Voting, I-Voting-ST) | 10.60 | 12.68 | 0.2914 | 10.24 | 0.8313 | 15.04 | 0.2790 | 0.2330 |
| (b1) I-Voting | 11.92 | 11.57 | 0.9292 | 9.69 | 0.4838 | 13.66 | 0.9594 | 0.7353 |
| (b2) I-Voting-ST | 9.27 | 13.80 | 0.1549 | 10.88 | 0.8590 | 16.23 | 0.2026 | 0.1614 |
| [Team treatments:] | | | | | | | | |
| (c) **T-No** | 13.26 | 10.04 | 0.0096*** | --- | --- | --- | --- | --- |
| (d) **Team Voting** (T-Voting, T-Voting-ST) | 12.53 | 17.67 | 0.0001*** | 18.02 | 0.0002*** | 17.30 | 0.0166** | 0.1054 |
| (d1) T-Voting | 12.81 | 16.53 | 0.0128** | 16.87 | 0.0209** | 16.28 | 0.0827* | 0.0966* |
| (d2) T-Voting-ST | 12.24 | 18.80 | 0.0033*** | 18.81 | 0.0051*** | 18.78 | 0.1282 | 0.7532 |
| [Across-treatment comparisons:] | | | | | | | | |
| *p* for $H_0$: (a) = (b) | 0.1882 | 0.2273 | --- | --- | --- | --- | --- | --- |
| *p* for $H_0$: (c) = (d) | 0.7051 | 0.0000*** | --- | --- | --- | --- | --- | --- |
| *p* for $H_0$: (a) = (c) | 0.9310 | 0.6861 | --- | --- | --- | --- | --- | --- |
| *p* for $H_0$: (b) = (d) | 0.2007 | 0.0074*** | --- | 0.0003*** | --- | 0.0554* | --- | --- |

II. Payoff

| | Avg. payoff based on all data | | | Avg. payoff under a given sanction scheme in Phases 2-6 | | | | |
|---|---|---|---|---|---|---|---|---|
| | (i) Phase 1 | (ii) Phases 2-6 | *p*-value for $H_0$: (i) = (ii) | (iii) FS | *p*-value for $H_0$: (i) = (iii) | (iv) IS | *p*-value for $H_0$: (i) = (iv) | *p*-value for $H_0$: (iii) = (iv)[#1] |
| [Individual treatments:] | | | | | | | | |
| (a) **I-No** | 30.34 | 27.71 | 0.0414** | --- | --- | --- | --- | --- |
| (b) **Indiv Voting** (I-Voting, I-Voting-ST) | 28.48 | 25.27 | 0.0575* | 23.38 | 0.0086*** | 27.09 | 0.0304** | 0.1252 |
| (b1) I-Voting | 29.54 | 23.79 | 0.0208** | 22.88 | 0.0357** | 24.81 | 0.0218** | 0.0280** |
| (b2) I-Voting-ST | 27.42 | 26.75 | 0.7897 | 23.97 | 0.1731 | 29.07 | 0.5076 | 0.8886 |
| [Team treatments:] | | | | | | | | |
| (c) **T-No** | 30.61 | 28.03 | 0.0096*** | --- | --- | --- | --- | --- |
| (d) **Team Voting** (T-Voting, T-Voting-ST) | 30.02 | 29.90 | 0.9353 | 29.71 | 0.8092 | 30.10 | 0.1701 | 0.1252 |
| (d1)T-Voting | 30.25 | 29.07 | 0.5337 | 28.38 | 0.3743 | 29.56 | 0.1823 | 0.1386 |
| (d2) T-Voting-ST | 29.79 | 30.73 | 0.4236 | 30.63 | 0.5076 | 30.89 | 0.7353 | 0.9165 |
| [Across-treatment comparisons:] | | | | | | | | |
| *p* for $H_0$: (a) = (b) | --- | 0.2343 | --- | --- | --- | --- | --- | --- |
| *p* for $H_0$: (c) = (d) | --- | 0.0661* | --- | --- | --- | --- | --- | --- |
| *p* for $H_0$: (a) = (c) | --- | 0.6861 | --- | --- | --- | --- | --- | --- |
| *p* for $H_0$: (b) = (d) | --- | 0.0514* | --- | 0.0004*** | --- | 0.3061 | --- | --- |

*Notes*: All *p*-values are based on two-sided tests. Group-level Wilcoxon signed rank (Mann-Whitney) tests were conducted for within(across)-treatments comparisons. See Panel A of Appendix B for the standard errors. See Panel C of Appendix B for more detailed across-treatment comparisons. "Indiv Voting" includes the I-Voting and I-Voting-ST treatments. "Team Voting" includes the T-Voting and T-Voting-ST treatments. To supplement the nonparametric tests (Table 2) and the regression analyses (Appendix Tables C.1 and C.2), additional regression was conducted using group-level average contribution or payoff as the dependent variable (Appendix Table C.3). The results consistently confirm strong discontinuity effects under voting. [#1] Only groups that had experienced both the FS and IS schemes in Part 2 were used. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

Third, the picture is markedly different in the team treatments. Whether in the FS or IS scheme, cooperation was sustained at high levels (panels (c) and (d) of Appendix Figure C.1). Regardless of which scheme was in effect, the teams' contribution amounts with the sanction schemes were significantly larger, relative to the T-No treatment (Table 2). Teams also quickly responded to the informal punishment received from their peers. Although payoff losses due to punishment were large in Phases 2 and 3 (in Phase 2) with the IS scheme in the T-Voting (T-Voting-ST) treatment, they achieved high payoffs after these phases. Despite administrative cost payments, teams in the T-Voting-ST treatment did earn more than those in the T-No treatment across *all* phases (panels (c) and (d) of Appendix Figure C.2). The clearly better outcomes among teams than individuals reject Hypothesis 1, and underscore the effects that the internal aspects of team decision-making (intra-team dynamics, and/or the preference aggregation process in team decision-making) have under the sanctioning institutions.

**Result 1**: *(a) Decision-making units (whether individuals or teams) reduced their contributions over time when sanctioning schemes were unavailable. (b) With the sanctioning schemes, individuals in the I-Voting treatment sustained their initial level of cooperation, and individuals in the I-Voting-ST gradually increased cooperation further. (c) Inconsistent with Hypothesis 1, the impact of voting was much stronger for teams: Under each punishment intensity, teams increased their cooperation more strongly and quickly than individuals regardless of which sanction scheme was chosen.*

### *4.2. Scheme Choice*

The strong efficiency under the IS scheme was not driven by a small number of groups. Despite standard theory based on selfishness of players predicting the superiority of the FS scheme, on average 47.3%, 63.0%, 53.3%, and 46.1% of decision-making units voted for the IS scheme in the I-Voting, T-Voting, I-Voting-ST, and T-Voting-ST treatments, respectively (Table 3.I). As a result of majority voting, groups adopted the IS scheme similar percentages of the time, i.e., 47.3%, 58.2%, 54.6%, and 40.0% of the time in the corresponding treatments (Table 3.II). Group-level Wilcoxon signed rank tests confirm that units' voting for the IS scheme and the vote outcomes were not the result of error. Group-level Mann-Whitney tests also indicate that scheme choice behaviors did not differ between individuals and teams (Panel K of Appendix B). A look at the across-phase trend also indicates that the popularity of the IS scheme was quite stable.

Realized relative effectiveness of FS and IS schemes affected voting. Seven, nine, eight, and six groups experienced both the FS and IS schemes at least once as a result of voting. Using these groups, Figure 4 demonstrates that decision-making units were more likely to vote for the scheme under which they had previously experienced higher payoffs on average. This resonates with the idea that people's institutional choices are guided by material outcomes (e.g., Ertan *et al.* 2009; Kamei *et al.* 2015),[15] and it

[15] To supplement this finding, a regression analysis was also conducted regarding how decision-making units' voting in Phase 6 (the final phase) may be influenced by relative payoff ratios they experienced before that phase. As shown in Appendix Table C.4, the relative payoff ratio is a significantly positive predictor for their selection of

may be a general phenomenon as the role of realized payoffs has been demonstrated in another setup, e.g., voting on leadership in PGG (e.g., Güth *et al.* 2007).

**Table 3:** *Scheme Choice and Voting Outcome*

I. Percentages of Times that Decision-Making Units Voted for the IS Scheme

| | Phase 2 | Phase 3 | Phase 4 | Phase 5 | Phase 6 | Overall | *p*-value for Wilcoxon signed rank tests[#1] |
|---|---|---|---|---|---|---|---|
| I-Voting | 48.5% | 63.6% | 42.4% | 51.5% | 30.3% | 47.3% | 0.0022*** |
| I-Voting-ST | 54.5% | 48.5% | 45.5% | 60.6% | 57.6% | 53.3% | 0.0017*** |
| T-Voting | 48.5% | 84.8% | 63.6% | 66.7% | 51.5% | 63.0% | 0.0016*** |
| T-Voting-ST | 33.3% | 48.5% | 48.5% | 54.5% | 45.5% | 46.1% | 0.0017*** |
| Average | 46.2% | 61.4% | 50.0% | 58.3% | 46.2% | 52.4% | 0.0000*** |

II. Percentages of Times that the IS Scheme was Selected in Groups

| | Phase 2 | Phase 3 | Phase 4 | Phase 5 | Phase 6 | Overall | *p*-value for Wilcoxon signed rank tests[#1] |
|---|---|---|---|---|---|---|---|
| I-Voting | 54.5% | 63.6% | 36.4% | 54.5% | 27.3% | 47.3% | 0.0021*** |
| I-Voting-ST | 54.5% | 54.5% | 36.4% | 63.6% | 63.6% | 54.5% | 0.0021*** |
| T-Voting | 45.5% | 81.8% | 54.5% | 63.6% | 45.5% | 58.2% | 0.0015*** |
| T-Voting-ST | 27.3% | 36.4% | 36.4% | 54.5% | 45.5% | 40.0% | 0.0197** |
| Average | 45.7% | 59.2% | 41.1% | 59.2% | 45.7% | 50.0% | 0.0000*** |

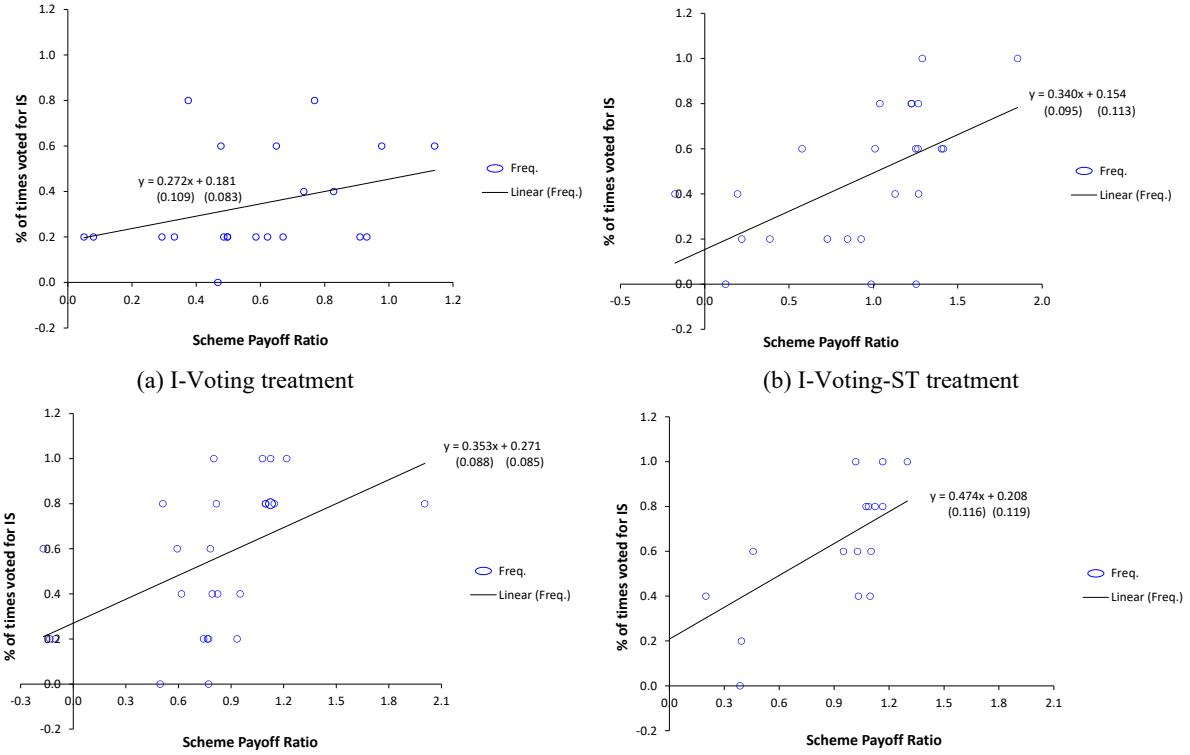
*Notes*: [#1] *p*-values here are one-sided as the theory predicts a specific direction. The null hypothesis is that the percentage of the time that units or groups select the IS scheme is less than or equal to 5%, assuming that errors happen with a 5% probability. In order to perform Wilcoxon signed rank tests, the overall percentage of decision-making units that voted for IS was calculated for each group in panel I (the percentage of times when IS was enacted was calculated for each group in panel II). After that, using the group-average observations Wilcoxon signed rank tests were performed. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

Figure 4 also indicates two more interesting patterns. First, there is a large variation for decision-making units' voting behaviors: the correlations between units' scheme votes and experienced relative payoff ratios are far from perfect. This implies strong heterogeneity in subjects' cooperation and punishment tendencies (e.g., Fischbacher *et al.* 2001; Fischbacher and Gächter 2010; Kamei 2014). Second, the punishment intensity influenced the relative effectiveness of informal sanctions. Under the modest punishment intensity, a considerable majority of the units – i.e., 95.24% and 66.66% of the units in the I-Voting and T-Voting treatments, respectively, had lower payoffs on average under the IS than the FS scheme due to punishment losses. However, the informal punishment became more effective under the strong punishment intensity. The percentages of the units who on average earned less under the IS than the FS scheme are a minority, i.e., 41.67% and 38.89% in the I-Voting-ST and T-Voting-ST treatments, respectively. Around 32% of groups exclusively selected one of the schemes across the five phases in Part 2. Except for one group in the I-Voting treatment, the groups' persistence in one scheme can be explained

the IS scheme both in the individual voting and team voting treatments (when data are pooled irrespective of the punishment intensity). The role of experience is also supported by an analysis of communication logs (Section 6).

by their success in cooperation under that scheme. The average contributions of groups that always selected IS were 19.93 and 19.21 points in the I-Voting-ST and T-Voting-ST treatments, respectively.[16] The average contributions of groups that always selected FS were 15.16, 19.54, 18.29, and 19.67 points in the I-Voting, I-Voting-ST, T-Voting, and T-Voting-ST treatments, respectively.[17]

**Figure 4:** *Scheme Choice and Relative Payoff Ratio*




(a) I-Voting treatment

(b) I-Voting-ST treatment

(c) T-Voting treatment

(d) T-Voting-ST treatment

*Notes:* The figures were depicted based on the data from the groups that experienced both the FS and IS schemes in Part 2. The horizontal axis (x-axis) is calculated by a given unit's average payoff under the IS scheme divided by their average payoff under the FS scheme across all periods. The vertical axis (y-axis) is the percentage of times the unit voted for IS and takes a value between 0 and 1. The size of each point indicates its frequency. The numbers in parentheses in the linear

[16] The numbers of groups that selected the IS (FS) scheme for all phases were 1(3), 1(2), 0(2), and 4(1) in the I-Voting, I-Voting-ST, T-Voting, and T-Voting-ST treatments, respectively. The average contribution of the group that exclusively selected IS in the I-Voting treatment was 11.2 points.

[17] Factors other than realized payoffs also affected the groups' scheme choice outcomes (Appendix Table C.5). First, subjects' preferences for fairness drove their scheme choices. Subjects provided their views on the fairness of each scheme in the post-experiment questionnaire. A calculation found that the fairer its members on average perceived the IS scheme relative to the FS scheme, the more likely a group was to implement the IS scheme. Second, subjects' levels of trust in others also drove their selection of the IS scheme. Specifically, the more strongly group members believed that people were trustworthy, the more frequently a group implemented the IS scheme. Third, cognitive ability affected voting, especially when the punishment strength was modest. A calculation shows that a more mathematically able group was more likely to select the IS scheme in the I-Voting and T-Voting treatments. Recall that sustaining cooperation with informal punishment was difficult when punishment strength was modest. However, more cognitively able groups might have attempted to build cooperative relationships without relying on the alternative centralized mechanism since the FS scheme entailed an administrative cost.

equation (OLS) in each panel are robust standard errors clustered by group ID. The slopes in the linear lines in panels a, b, c, and d are significantly positive at two-sided $p = 0.046$, 0.009, 0.004, and 0.009, respectively.

**Result 2**: *(a) Despite standard theory based on selfish preferences predicting the superiority of the FS scheme, around half of the groups adopted the IS scheme. (b) Decision-making units voted for the scheme under which they had previously experienced higher payoffs. (c) Almost all groups that selected one scheme (FS or IS) for all phases achieved successful cooperation in that scheme.*

### *4.3. Discontinuity Effects in Utilizing the Sanctioning Institutions*

While the analysis in Section 4.2 usefully uncovered the motives behind the units' scheme choices (FS or IS), it does not help to determine which hypothesis, Hypothesis 2 or Hypothesis 3, drove Result 1. Section 4.3 will explain that the higher efficiency of teams relative to individuals was driven by the "truth wins" mechanism: in line with Hypothesis 3, there were clear differences in the ways in which decision-making units utilize the sanctioning institutions.

#### *4.3.1. Voting and Contribution Behaviors in the FS Scheme*

Units' decisions to contribute under the FS scheme were strongly influenced by their group's sanction rate. A regression analysis finds that units were significantly more likely to contribute large amounts, the higher the sanction rate their group had implemented (Appendix Table C.7). Having a deterrent sanction rate effectively improves units' decisions to contribute (again see Appendix Table C.7). The larger impact of having stronger punishment is consistent with prior research on formal sanctioning institutions (e.g., Falkinger *et al.* 2000; Kamei *et al.* 2015), suggesting that a centralized solution of the free riding problem is to enforce an incentive mechanism in a society or organization.

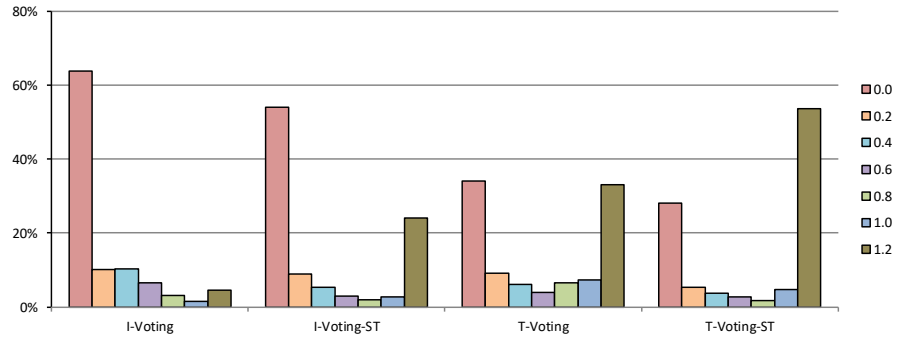
Result 1 was driven by the difference in voting. As shown in Figure 5, the popularity of sanction rates differs markedly between individuals and teams. First, the sanction rate of 0.0 was the focal point among the individuals. Strikingly, individuals in the I-Voting and I-Voting-ST treatments voted for the zero sanction rate on 63.79% and 54.00% of the occasions, respectively (Figure 5.A). As a result of the majority rule applied, the regime without any sanctions, the same regime as in Phase 1, was implemented on 70.69% and 57.00% of the occasions, respectively, in these two treatments (Figure 5.B).[18,19]

Given this revealed voting patterns of individuals, Hypothesis 2 predicts that if team decision-making were a mere aggregation process of three members' preferences, then teams would vote for non-deterrent sanction rates, and therefore collectively enact non-deterrent formal schemes in their groups, more frequently than individuals. However, clearly contrary to this hypothesis, teams voted for the zero

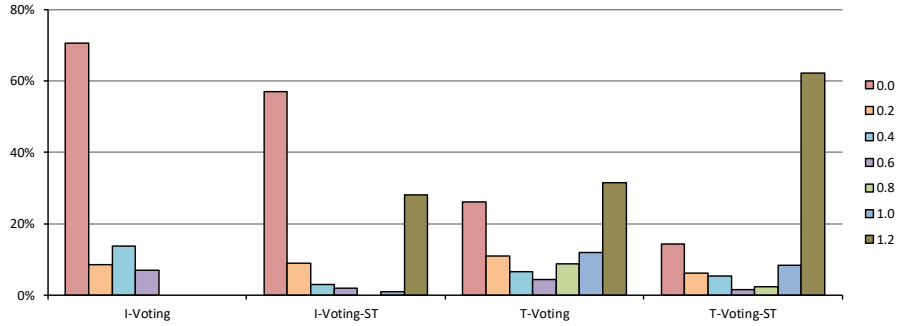
[18] In Kamei *et al.* (2015), almost all individuals successfully constructed deterrent schemes. The difference between this study and Kamei *et al.* could be due to the difference in the research site: the USA versus England. Alternatively it could be due to the difference in the group size – three in this study versus five in Kamei *et al.* (2015).
[19] The outcome of the zero sanction rate is somewhat larger than the percentage of the voters who preferred it (e.g., 70.69% > 63.79%). This is due to the majority voting system because it tends to outnumber the preferences of minorities – a phenomenon called the Condorcet's (1785) jury theorem or the behavioral public choice theorem (e.g., Ertan *et al.* 2009; Hauser *et al.* 2014).

sanction rate only 34.06% and 28.03% of the time in the T-Voting and T-Voting-ST treatments, respectively. Instead, consistent with Hypothesis 3, teams used voting much more efficiently than individuals: teams voted for deterrent sanction rates (0.4 or above) on 56.88% and 66.67% of the occasions in the T-Voting and T-Voting-ST treatments, respectively. In particular, teams' preferences for the highest sanction rate – 1.2 per point allocated to the private account – were strikingly strong (Figure 5.A). In the T-Voting-ST treatment, teams voted for the highest rate on 53.54% of the occasions. With the majority rule, 31.52% (26.09%) and 62.12% (14.39%) of the vote outcomes were the highest (zero) sanction rate in the T-Voting and T-Voting-ST treatments, respectively.[20] The average realized group sanction rates were 0.64 and 0.89, both of which are deterrent, in the T-Voting and T-Voting-ST treatments, respectively. However, average realized sanction rates were much smaller in the individual treatments, i.e., 0.11 and 0.39 in the I-Voting and I-Voting-ST treatments, respectively.[21,22] The difference in the severity of selected sanction

**Figure 5:** *Voting on Sanction Rates and Vote Outcome*




(A) Distributions of Decision-Making Units' Voting




(B) Distributions of Vote Outcomes

[20] The percentages of cases in which a group selected the zero (highest) sanction rate in Phases 2 to 6 are significantly different between individual and team voting at two-sided $p = 0.0080$ ($p = 0.0319$), according to a group-level Mann-Whitney test, when pooled data are used – see Panel F of online Appendix B.
[21] The average realized sanction rates are significantly different at two-sided $p = 0.0116$ between individual versus team voting when pooled data are used (see Panel F of Appendix B).
[22] Figure C.3 reports the popularity of sanction rates, period by period. It indicates that teams' strong preferences for deterrent sanction rates were stable across all periods, while individuals' preferences for non-deterrent sanction rates were strong from earlier periods and became even stronger gradually as the experiment progressed.

**Table 4:** *Average Contribution by Sanction Rate under the FS scheme*

| Sanction rate | (a) Individual Voting | | | (b) Team Voting | | | (c) Mann-Whitney tests[#1] | | |
|---|---|---|---|---|---|---|---|---|---|
| | (i) All data | (ii) I-Voting | (iii) I-Voting-ST | (i) All data | (ii) T-Voting | (iii) T-Voting-ST | (i) $H_0$: a.i = b.i | (ii) $H_0$: a.ii = b.ii | (iii) $H_0$: a.iii = b.iii |
| 0.0 or 0.2 (non-deterrent) | 7.52 (4.08) | 7.86 (4.66) | 7.04 (3.59) | 15.12 (5.87) | 14.09 (6.93) | 16.42 (4.99) | 0.0110** | 0.1415 | 0.0274** |
| 0.4 or above (deterrent) | 17.67 (5.38) | 16.72 (6.29) | 18.34 (4.27) | 19.10 (1.51) | 18.50 (2.14) | 19.42 (0.72) | 0.0268** | 0.1467 | 0.0949* |

*Notes*: Numbers in parenthesis are standard errors based on group averages. [#1] Two-sided *p* for group-level Mann-Whitney tests. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

rates well explains the stronger contribution behaviors of teams in the voting treatments (Figure 2, Appendix Figure C.1), and supports Hypothesis 3 – the truth wins mechanism.[23]

Result 1 was also partly affected by decision-making units' decisions to contribute in the FS scheme, which differs between individuals and teams even when the same sanction rates prevailed. Strikingly, on average, teams contributed significantly more than individuals, whether sanctions were deterrent or not (see columns a.i, b.i, and c.i of Table 4). The difference was especially large under non-deterrent sanction rates (i.e., rates of 0.0 or 0.2). This difference cannot be explained by a selectivity bias. Notice that more cooperative groups can be assumed to select stronger sanction rates, making mutual cooperation easier (Appendix Table C.7). If this assumption is correct, the least cooperative units would be overrepresented in groups that enacted non-deterrent sanction rates for the T-Voting (T-Voting-ST) rather than the I-Voting (I-Voting-ST) treatment, because such weak sanction rates were realized only in a small fraction of groups in the team treatments (Figure 5, Table 4).

As the maintenance of group cooperation norms leads to large long-term payoffs, the teams' stronger behavioral responses to sanction rates suggest that, with the FS being enacted, teams may be more far-sighted and less myopic loss averse than individuals (Sutter 2007, 2009; Bougheas *et al.* 2013), and these responses are again consistent with the positive effects of deliberation and learning that the truth win mechanism proposes.

**Result 3**: *(a) While individuals voted for the zero sanction rate more than 50% of the time in the two individual treatments, teams did so much less than 50% of the time in the two team treatments, inconsistent with Hypothesis 2. Instead, teams voted for deterrent sanction rates more than 50% of the time, and this more efficient voting by teams is consistent with Hypothesis 3. As a result, (c) teams enacted significantly stronger sanction rates than individuals. Specifically, the average sanction rate in the T-Voting (T-Voting-ST) treatment was 0.64 (0.89), while the average sanction rate in the I-Voting (I-*

[23] As was done for the groups' scheme choices in Section 4.2, bivariate correlations between a group's average sanction rate and their attribute variables were calculated to explore what non-material factors might have affected their selection of sanction rates (Appendix Table C.6). The calculation reveals (a) female subjects' possible dislike of using punishment, (b) economics students' rational voting, i.e., voting for strong sanction rates, (c) a positive impact of perceived fairness under the FS scheme on voting, and (d) subjects' intention to encourage others' contributions through centralized punishment, for both individuals and teams.

*Voting-ST) treatment was 0.11 (0.39). (b) Teams contributed significantly more than individuals for given sanction rates.*

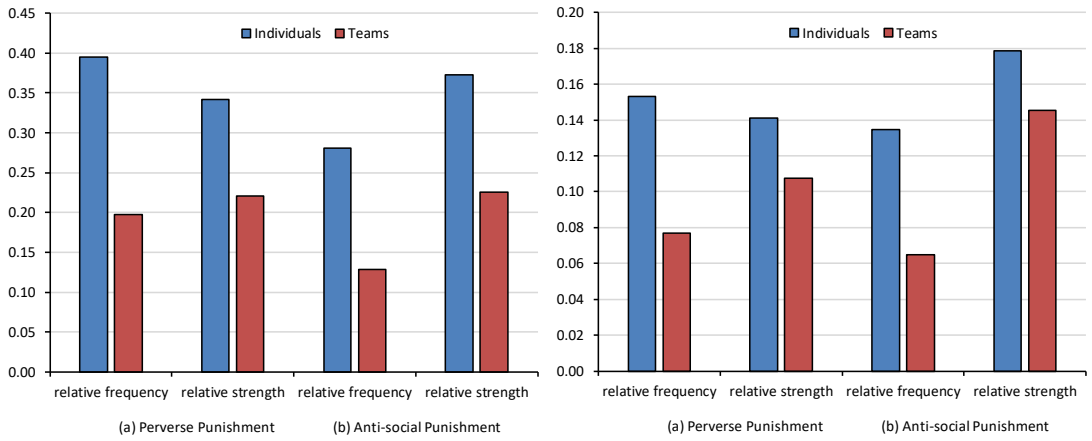
*4.3.2. Contribution and Punishment Behaviors in the IS Scheme*

Decision-making units inflicted costly punishment based on the distribution of contributions in their group (Appendix Table C.8). First, the smaller the amount a decision-making unit *j* contributed to the public account relative to *i*, the more strongly *i* punished *j*. Second, contributing more than another member also attracted punishment by that member to some degree, but such anti-social punishment is significantly weaker than pro-social punishment. These two patterns, which hold for all treatments, are in line with the prior research (e.g., Fehr and Gächter 2000; Kamei and Putterman 2015). But then, what may explain the difference in efficiency between individuals and teams in the IS scheme? The difference in the overall punishment distribution provides the answer.

Figure 6 reports the relative strength and frequency of anti-social (perverse) punishment to pro-social (non-perverse) punishment. It reveals that (a) pro-social (non-perverse) punishment was more prevalent than anti-social (perverse) punishment among individuals, and that (b) pro-social (non-perverse) punishment was even more dominant among teams than individuals.[24] The teams' better targeted punishment behaviors are consistent with both Hypothesis 2 and Hypothesis 3.

In addition, behavioral responses to punishment received differ by the decision-making unit

**Figure 6:** *Relative Strength and Frequency of Perverse/Anti-Social Punishment*




(A) I-Voting versus T-Voting treatments (B) I-Voting-ST versus T-Voting-ST treatments

*Notes*: Following Herrmann *et al.* (2008), (i) punishment from *i* to *j* in period *t* is defined as anti-social if *j* contributed more than *i* or when both *i* and *j* are 20-contributors in that period, and (ii) punishment that is not anti-social is called pro-social. Following Cinyabuguma *et al.* (2006), (iii) punishment from *i* to *j* in period *t* is defined as perverse if *j* contributed more than their group average or when all in their group contributed the full endowment amount in that period, and (iv) punishment that is not perverse is called non-perverse.

[24] Due to the small sample size, the difference is only significant at $p = 0.0544$ for the relative frequency if a one-sided group-level Mann-Whitney test is used for pooled data. However, a finite mixture modeling analysis in Section 5 reveals significant different punishment strategies between individuals and teams.

(see again Table C.8). Individuals were insensitive to punishment received: Pro-social punishment did not encourage individuals to contribute larger amounts in the following periods. Anti-social punishment also did not significantly discourage the recipients' subsequent cooperative behaviors. Instead, individuals tended to conform to group norms, forming contribution decisions based on their group's last-period contribution behavior. By contrast, teams strongly responded to peers' anti-social punishment: the larger the anti-social punishment teams received, the more strongly they reduced their own contributions in the following periods. In addition, pro-social punishment helped teams boost cooperation for the T-Voting-ST treatment. These differences in the behavioral responses are more in line with the "truth wins" hypothesis. Notice that simple preference aggregation predicts that teams would be more insensitive than individuals to punishments received, considering that individuals tended not to respond to punishment. Unlike this prediction, teams responded more strongly than individuals to punishments received so as to make punishment better targeted and to sustain contributions at high levels.

## 5. Structural Estimations of Punishment Types under the IS Scheme

The main experimental finding of the previous section was that (a) teams are able to sustain cooperation at a higher level than individuals when they can vote on sanctioning institutions, and (b) the teams' high efficiency is driven by their effective use of punishment. This is consistent with the "truth wins" mechanism summarized as Hypothesis 3.

In the previous analysis, while the percentages of types as a voter, i.e., rational or irrational, were precisely compared among units in the FS scheme (Figure 5), it is still unclear what percentages of units punished anti-socially or pro-socially in the IS scheme. Section 5 analyzes the differences in punishment type in the IS scheme more accurately by using finite mixture modeling.[25,26]

Finite mixture modeling assumes a set of possible behavioral types in advance and then assigns a probability measure over the types to each subject so that the likelihood is maximized (McLachlan and Peel 2000; Moffatt 2016). Table 5 reports the estimation results.[27] Two models were estimated by assuming different sets of three punishment types, as there are two approaches to define punishment patterns (Herrmann *et al.* 2008; Cinyabuguma *et al.* 2006). The first model assumes the pro-social punisher, the anti-social punisher, and the selfish type (Herrmann *et al.* 2008), while the second model

[25] Some units voted on sanction rates in an indecisive manner (e.g., voted for deterrent rates in some periods and for non-deterrent rates in the other periods). To explain their behavior, finite mixture modeling analysis was conducted for these units by assuming possible types (e.g., a type who votes based on their punishment received in the last period). However, almost all models were unable to be estimated (failed to converge) due to a small sample size.
[26] As an additional analysis, units' contribution types were structurally estimated using the finite mixture modeling approach. The result uncovered that the distribution of contribution types under the FS scheme differs largely from those under the IS scheme, as the decision-making units decided contribution amounts in response to the sanction rate collectively enforced in the current period under the FS scheme (Appendix Table C.9).
[27] Typical to a maximum likelihood method, estimation results may depend on what starting values are assumed. In each model of Table 5, starting values were chosen to achieve the highest log likelihood. The selected starting values in some models coincide with the starting values based on the method suggested by Moffatt (2016).

assumes the non-perverse punisher, the perverse punisher, and the selfish type (Cinyabuguma *et al.* 2006). The pro-social and anti-social punishers, and the perverse and non-perverse punishers, are defined the same as in Section 4.3.2. The selfish type is defined as a player who does not inflict punishment throughout.

Consider, first, Models A.i and B.i to see behavioral differences between individuals and teams with a larger dataset. The results show that a larger percentage of teams, relative to individuals, inflicted punishment on low contributors (60.0% versus 49.4% in panel I, and 65.6% versus 48.2% in panel II). The difference in the classified type is especially large in panel II: According to a two-sided Kolmogorov-Smirnov test, the percentage of non-perverse punishers is significantly larger among teams than individuals at $p = 0.025$. On the other hand, types that engage in "misdirected" punishment are regularly present regardless of the decision-making format.[28] This implies that the issue of misdirected punishment is ubiquitous whether among individuals or teams.

The estimation results by the respective treatment provide further nuanced explanations for the discontinuity effects detected in Section 4. First, strikingly, the percentage of anti-social (perverse) punishers is only 9.1% (12.4%) in the T-Voting treatment, which is less than one fourth (a half) of the percentage in the I-Voting treatment, while the percentages of pro-social (non-perverse) punishers do not differ much between the two treatments.[29] Hence, under weak punishment intensity, there is strong evidence that team decision-making effectively prevents units from engaging in misdirected punishment.

**Result 4:** *(a) On average, a significantly larger percentage of teams, relative to individuals, inflicted punishment on low contributors, while "misdirected" punishment was observed both for individuals and teams. (b) Under the weak punishment intensity, team decision-making effectively prevented decision-making units from engaging in misdirected punishment.*

Stronger punishment intensity makes individuals reluctant to anti-socially punish members (compare Models A.ii and A.iii of Table 5), perhaps being afraid of inviting blind revenge in the following periods (e.g., Ostrom *et al.*, 1992). This is consistent with the higher efficiency of the I-Voting-ST relative to the I-Voting treatment seen in Result 1.b and Figures 2 and 3. Reflecting this, teams cannot be judged superior to individuals for their punishment type choices under the strong punishment intensity. In particular, as seen in Models A.iii and B.iii, the differences of the estimated percentages of pro-social versus anti-social punishers (non-perverse versus perverse punishers) are large for both the individuals and teams: 44.4% (44.5%) in the T-Voting-ST treatment, and 27.8% (38.8%) in the I-Voting-ST treatment. So, why did the T-Voting-ST treatment perform much better, compared with the I-Voting-ST

[28] Regarding misdirected punishment, no consistent patterns were seen between unit types across definitions: "anti-social" ("perverse") punishment was less (more) frequent among teams than individuals.

[29] A two-sided Kolmogorov-Smirnov test finds that the percentages of anti-social (perverse) punishers are significantly different between the T-Voting and I-Voting treatments at $p < 0.001$ ($p = 0.018$), while the percentages of pro-social (non-perverse) punishers are not significantly different between the T-Voting and I-Voting treatments at $p = 0.391$ ($p = 0.148$).

treatment, at sustaining cooperation (Figures 2 and 3)? The answer to this may be due to the difference in the percentages of pro-social or non-perverse punishers per group. Notice that, remarkably, around 67-68% of teams, i.e., on average *two out of three* units per group, are classified as pro-social/non-perverse punishers in the T-Voting-ST treatment (Model B.iii). The corresponding percentage is around 45-52% in the I-Voting-ST treatment (Model A.iii), which means that there is often *only one* pro-social/non-perverse punisher per group. It might have been challenging for *a single member* to discipline two members of her group in the I-Voting-ST treatment as punishment is privately costly.

**Table 5:** *Estimated Percentages of Punishment Types in the IS Scheme*

| Treatment: | A. Individual Voting | | | B. Team Voting | | |
|---|---|---|---|---|---|---|
| | (i) All data | (ii) I-Voting | (iii) I-Voting-ST | (i) All data | (ii) T-Voting | (iii) T-Voting-ST |
| **I. Pro-social versus Anti-social punishment** | | | | | | |
| Classified types [%] | | | | | | |
| *Pro-social* | 49.4% (7.7)*** | 44.2% (10.1)*** | 52.9% (10.9)*** | 60.0% (8.2)*** | 48.3% (12.9)*** | 67.9% (11.6)*** |
| *Anti-social* | 25.5% (6.0)*** | 41.5% (9.3)*** | 25.1% (8.2)*** | 19.1% (5.7)*** | 9.1% (5.0)* | 23.5% (9.3)** |
| *Selfish* | 25.1% (7.0)*** | 14.3% (7.8)* | 22.0% (9.4)** | 20.9% (7.7)*** | 42.6% (12.9)*** | 8.6% (8.1) |
| # of obs. | 1,344 | 624 | 720 | 1,296 | 768 | 528 |
| Wald $\chi^2$ | 118.77 | 103.89 | 43.70 | 162.02 | 128.48 | 41.56 |
| Prob > Wald $\chi^2$ | < 0.001 | < 0.001 | < 0.001 | < 0.001 | < 0.001 | < 0.001 |
| **II. Perverse versus Non-perverse punishment** | | | | | | |
| Classified types [%] | | | | | | |
| *Non-perverse* | 48.2% (7.1)*** | 49.8% (10.3)*** | 46.1% (10.4)*** | 65.6% (8.1)*** | 60.3% (10.3)*** | 67.4% (11.6)*** |
| *Perverse* | 16.7% (5.3)*** | 26.8% (9.7)*** | 26.5% (8.1)*** | 20.3% (5.5)*** | 12.4% (5.8)** | 23.9% (9.4)** |
| *Selfish* | 35.2% (6.6)*** | 23.4% (8.9)*** | 27.4% (9.6)*** | 14.2% (6.8)** | 27.3% (9.7)*** | 8.8% (8.2) |
| # of obs. | 1,344 | 624 | 720 | 1,296 | 768 | 528 |
| Wald $\chi^2$ | 120.31 | 61.92 | 14.56 | 123.22 | 70.73 | 39.78 |
| Prob > Wald $\chi^2$ | < 0.001 | < 0.001 | < 0.001 | < 0.001 | < 0.001 | < 0.001 |

*Notes*: The numbers in parentheses are standard errors. All models were estimated by having a tremble term. Estimation results in each model occasionally varied dependent on their starting values, due to multiple local equilibria of the likelihood function. As such, starting values were initially set based on the method suggested by Moffatt (2016), and then systematically varied to achieve the global maximum log likelihood. The selected starting values coincide with the starting value based on the method suggested by Moffatt (2016) for models A.i, A.ii and B.ii of panel I and models A.i, A.ii, and A.iii of panel II. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

## 6. Team Communication Dialogues

Teams' communication dialogues contain richer information that may explain the reasoning behind team decisions, although a look at the communication dialogues do not help explain the behavioral differences between individuals and teams as such contents are unavailable for individuals that could be used for comparison. As discussed in Cooper and Kagel (2022), analysis using behavioral data is based on assumptions imposed by each estimation model. Human reasoning and responses are usually complex and can differ from those assumed in economic models; as such, it is still meaningful to investigate teams'

communication dialogues.[30] As a final supplementary analysis, communication dialogues were analyzed following the standard coding procedure in the experimental literature (e.g., Cooper and Kagel 2005, 2022; Cason and Mui 2015; Kagel and McGee 2016; Leibbrandt and Sääksvuori 2012).

Two research assistants (RAs) were hired as independent coders. They were not informed about any substance of the research, such as the research aim, to avoid demand effects. They were simply provided with the instructions, communication logs, and the list of codes, and were then asked to assign as many relevant codes as possible to each log. The full list of codes is available in online Appendix D.2. Once the two RAs finished coding, the researchers checked for discrepancies between the two coders' classifications and highlighted any differences. After that, each coder was given the other coder's assigned codes and could reconsider their own coding, with the knowledge that the other coder would independently follow the same reconsideration process. This *reconsideration* process was first used, and confirmed effective in catching any errors in initial coding, by Elten and Penczynski (2020). Appendix D.1 includes the detail of the coding procedure adopted in the present paper.

Cohen's Kappa (Cohen, 1960) is the most popular form of agreement analysis and is hence used to judge the reliability of coding (e.g., Cason and Mui 2015; Leibbrandt and Sääksvuori 2012). The final Kappas after reconsideration were 0.88, 0.90, and 0.87 in the T-No, T-Voting, and T-Voting-ST treatments, respectively. Appendix D.3 includes the Kappas for individual codes, showing that almost all codes have high Kappa values. Regression analyses below use codes whose Kappa is above 0.4. 0.4 is often used as a criterion for the reliability of codes (e.g., Landis and Koch 1977, Bougheas *et al.* 2013, Cason *et al.* 2012).

### *6.1. Voting on Sanction Rates in the FS Scheme*

As discussed in Section 4.3.1, a large fraction of decision-making units, even teams (34.06% and 28.03% of occasions in the T-Voting and T-Voting-ST treatments, respectively), voted for the zero sanction rate. Two codes were considered to capture this inefficient voting behavior:

C1: Suggests 0.0 sanction rate/desire to have effectively no fine due to ideological reasons (e.g., dislike of coercive measures) or simply due to their tastes against the cost.

C2: Suggests 0.0 sanction rate/desire to have effectively no fine due to confusion of the incentive structure.

The earlier analysis in Section 4.3.1 at the same time found that teams selected stronger sanction rates much more frequently than individuals (Figure 5). Thus, two additional codes were also considered to explain possible sources for this efficient voting behavior as follows:

C5: Discusses rate based on deterrence, i.e., deterrent (non-deterrent) if it is $\geq$ ($<$) 0.4.

C6: Discusses effects of a strong sanction rate, other than deterrence (e.g., why 1.2 is preferred to 0.8).

The key difference between C5 and C6 is whether members recognize the relationship between

[30] Cooper and Kagel (2022) demonstrated that the distributions of subjects' strategy choices in an IRPD are estimated significantly differently between coding exercises from communication logs and the Strategy Frequency Estimation Method, because subjects often employ a strategy which is incompletely specified at the outset.

sanction rates and material incentives in the game. The sanction rate should be set equal to or greater than 0.4 to induce other units to contribute fully to the public account. A rational decision-making unit would be indifferent between the sanction rates of, for example, 0.4 and 0.8. The two coders assigned Codes C1, C2, C5, and C6 at least once for 28.1%, 43.9%, 63.2%, and 26.3% of the teams playing FS, respectively.

Table 6.A reports key estimation results of a regression where the dependent variable is team voting on a sanction rate in the FS scheme. The results first indicate that C1 and C2 are both significantly negative predictors for units' sanction rate preferences. This confirms that some subjects' dislike of using centralized punishment and/or confusion does harm efficient institutional formation. Second, C6 is a significant positive predictor for their preferred sanction rates. C5 has also a significant and positive coefficient for the T-Voting-ST treatment, but not when all data are used (column (1)). A close look by the authors at the coding results for Code C6 and the teams' communication logs indicate that teams often had negative reactions and intolerance towards low contributions, and therefore had preferences for the maximum sanction rate to punish such acts. This result collaborates with the fact that the sanction rate of 1.2 was the most popular among the deterrent sanction rates (Figure 5).

In summary, it can be concluded that voting for deterrent sanction rates was driven by their negative reactions and intolerance towards low contributions and their learning about its impact.

*6.2. Informal Punishment Decisions in the IS Scheme*

Both individuals and teams inflicted punishment not only pro-socially but also anti-socially (Section 4.3.2). Four codes are considered to investigate motives behind these punitive behaviors:

F1: Suggests punishment for a contribution higher than their own (anti-social).
F2: Suggests no punishment for a contribution higher than their own (pro-social).
F3: Suggests punishment for a contribution lower than their own (pro-social).
F4: Suggests no punishment for a contribution lower than their own.

Codes F1 to F4 are defined using the anti- or pro-social punishment classification (Hermann *et al.* 2008). As in the earlier analyses, four more codes (F5 to F9) are also considered based on the perverse or non-perverse punishment definition (Cinyabuguma *et al.* 2006). The result shown in Section 6.2 is based on Codes F1 to F4, and is similar when Codes F5 to F9 are instead used (Appendix D.4.b).

In order to control for factors related to confusion, errors, and mistakes evident in the communication, Code F19 is also considered:

F19: Confusion, errors, mistakes (e.g., failing to understand the punishment cost).

Table 6.B reports key regression results. It first shows that Code F19 is a positive predictor for units' punishment decisions. Thus, some units' punishment activities are due to their low cognitive ability. However, even after controlling for Code F19, Codes F1 and F3 are positive predictors for units' decisions to punish (and the size of the absolute values of the coefficient estimates are much larger than for F2 and F4, respectively). This means that punishment motives are heterogeneous (Kamei 2014), and

units have clear intentions to punish pro-socially, or anti-socially, parallel to the observations from the decision data.

The results reveal three further patterns. First, emotion (Code F16: Suggests punishment as an emotional response) drives punishment. Second, some units inflict punishment on those whose contribution is less than a certain threshold (Code F9: Suggests punishment based on absolute contribution e.g. below or above a specific number). Third, costs for punishment (Code F11: Expresses desire to avoid punishment regardless of contribution due to the cost in imposing punishment) and the fear of retaliation (Code F13: Expresses desire to avoid punishment to prevent retaliation) discourage punishment.

**Table 6:** *Reasoning behind Units' Use of Punishment*

A. Team votes on a sanction rate in the FS scheme

Dependent variable: a sanction rate voted by team *i* in period *t*

| | (1) Pooled data | | (2) T-Voting | | (3) T-Voting-ST | |
|---|---|---|---|---|---|---|
| | Coefficient | S.E. | Coefficient | S.E. | Coefficient | S.E. |
| C1 dummy | -1.475*** | 0.270 | -1.319*** | 0.284 | -1.557*** | 0.525 |
| C2 dummy | -1.565*** | 0.229 | -1.041*** | 0.256 | -1.862*** | 0.391 |
| C5 dummy | 0.226 | 0.182 | -0.164 | 0.215 | 0.991*** | 0.314 |
| C6 dummy | 1.161*** | 0.339 | 0.747* | 0.403 | 1.299** | 0.651 |
| # of observations | 672 | --- | 276 | --- | 396 | --- |
| Wald $\chi^2$ | 136.33 | --- | 75.53 | --- | 78.49 | --- |
| Prob > Wald $\chi^2$ | 0.000 | --- | 0.000 | --- | 0.000 | --- |

*Notes:* Tobit regressions. Decision-making unit random effects were included to control for the panel structure. The regression includes all C codes and G codes with Kappa being above 0.4, phase dummies, and the Period within phases variable as independent variables. The full estimation result can be found in online Appendix Section D.4.a. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

B. Team informal punishment decisions in the IS scheme

Dependent variable: total punishment points assigned from team *i* to the other two teams in *i*'s group in period *t*

| | (1) Pooled data | | (2) T-Voting | | (3) T-Voting-ST | |
|---|---|---|---|---|---|---|
| | Coefficient | S.E. | Coefficient | S.E. | Coefficient | S.E. |
| F1 dummy | 6.349*** | 1.284 | 6.747*** | 1.506 | 3.946*** | 1.473 |
| F2 dummy | -3.610*** | 1.308 | -3.916** | 1.538 | -3.409** | 1.529 |
| F3 dummy | 8.835*** | 1.101 | 10.352*** | 1.381 | 7.524*** | 1.116 |
| F4 dummy | -2.909** | 1.233 | -6.107*** | 1.497 | 1.621 | 1.074 |
| F9 dummy | 3.576*** | 1.116 | 4.569*** | 1.368 | 9.646*** | 1.773 |
| F11 dummy | -3.259** | 1.443 | -0.019 | 1.990 | -8.875*** | 1.507 |
| F13 dummy | -3.736** | 1.592 | -5.046* | 2.741 | 0.775 | 1.076 |
| F16 dummy | 6.103*** | 2.100 | -5.132 | 3.805 | 9.854*** | 1.519 |
| F19 dummy | 7.964*** | 1.741 | 6.153*** | 2.065 | 12.468*** | 2.736 |
| # of observations | 648 | --- | 384 | --- | 264 | --- |
| Wald $\chi^2$ | 172.55 | --- | 150.91 | --- | n.a. | --- |
| Prob > Wald $\chi^2$ | 0.000 | --- | 0.000 | --- | n.a. | --- |

*Notes:* Tobit regressions with decision-making unit random effects. Codes associated with the definition of anti-social/pro-social punishment were used. The regression includes all F and G codes (except F5 to F8) with Kappa being above 0.4, phase dummies, and the Period within phases variable as independent variables. The full estimation result can be found in online Appendix Section D.4.b. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

### *6.3. Contribution Decisions*

Coding analyses, summarized in Table 7, suggest qualitatively similar, and important motives for contributions for all team treatments. First, units with unconditional willingness to cooperate contributed large amounts (variable i). Apart from such altruistic motives, some units aimed to encourage other units to cooperate, or to avoid discouraging already cooperative teams, through contributing large amounts (variable ii). Second, however, there were some units who discussed unconditional free riding in the communication stage, and did so as their team contribution decisions (variable iii), consistent with the prevalence of such free rider types in public goods dilemmas (e.g., Fischbacher *et al.* 2001; Fischbacher and Gächter 2010). Those who had inclinations to cooperate tended to decrease contributions out of distrust for the other teams or safety (variable iv).

There is one interesting pattern for the units' reasoning under the FS scheme. The result shows that units' desire to avoid receiving fines (code D9), rather than material calculations in the public goods game (D10), drove their strong contribution behaviors. This implies that positive effects of formal sanctioning institutions widely documented in prior research, such as in Falkinger *et al.* (2000) and Kamei *et al.* (2015), may emerge merely from people's dislikes of receiving formal punishment, regardless of their cognitive ability or understanding of the material incentives within the game.

**Table 7:** *Reasoning behind Units' Contribution Decisions*

Dependent variable: contribution amount of team *i* in period *t*

| Codes included in the regression: | (1) No scheme | | (2) Under FS scheme | | (3) Under IS scheme | |
|---|---|---|---|---|---|---|
| | Coefficient | S.E. | Coefficient | S.E. | Coefficient | S.E. |
| i. Contribute high always (Codes A2, D1, E1 dummies) | 4.954*** | 0.617 | 8.916*** | 2.049 | 3.015*** | 1.201 |
| ii. Contribute high to encourage others to cooperate (Codes A3, D3, E3 dummies) | 5.428*** | 0.612 | 2.380 | 2.594 | 5.558*** | 1.570 |
| iii. Contribute low always (Codes A4, D2, E2 dummies) | -4.098*** | 0.625 | -6.524*** | 2.107 | -7.058*** | 1.143 |
| iv. Contribute low out of distrust (Codes A5, D4, E4 dummies) | -4.429*** | 0.714 | -12.415*** | 2.700 | -7.788*** | 1.608 |
| vi. Contribute to avoid fines (Code D9 dummy) | --- | --- | 9.203*** | 2.388 | --- | --- |
| vii. Contribute based on material payoff maximization (Code D10 dummy) | --- | --- | -2.624 | 2.046 | --- | --- |
| # of observations | 1,128 | --- | 672 | --- | 648 | --- |
| Wald $\chi^2$ | 749.45 | --- | 212.5 | --- | 254.42 | --- |
| Prob > Wald $\chi^2$ | 0.000 | --- | 0.000 | --- | 0.000 | --- |

*Notes:* Tobit regressions with decision-making unit random effects. The regressions include all relevant codes (all A codes, D codes, and E codes in columns (1), (2) and (3), respectively) and G codes with Kappa being above 0.4, phase dummies, and the Period within phases variable as independent variables. The full estimation results can be found in online Appendix Section D.4.c. *, **, and *** indicate significance at the 0.10 level, at the 0.05 level, and at the 0.01 level, respectively.

### *6.4. Scheme Choice*

The remaining analysis is on communication dialogues related to scheme choices. The same kind of regression analysis using classification codes was performed. However, a relatively large number of the codes were omitted in the analysis due to collinearity. Nevertheless, four patterns are worth

mentioning. First, units' support for the FS scheme is driven by their dislike of the unpredictable/variable nature of the IS scheme (Code B2). Second, however, some units voted for the FS scheme in the experiment with a clear intention to construct the NS by selecting the zero sanction rate (Code B3). Third, some units voted against the FS scheme to avoid the fixed administrative charge of operating the scheme (Code B4). Lastly, consistent with the results summarized in Figure 4, members discussed prior experiences/contributions/behaviors under IS and FS schemes in order to decide which sanctioning scheme to vote for (Code B11). Online Appendix D.4.d includes the detail of the estimation results.

## 7. Conclusion

Team decision-making is ubiquitous whether in the public or private sphere. The literature in the theory of the firm has so far assumed that team decision-making is inferior to individual decision-making due to imperfect information, monitoring issues, and agency costs. In their theoretical context, team decision-making is just identical to individual decision-making when complexities in teams are resolved (e.g., Alchian and Demsetz 1972; Marschak and Radner 1972). Furthermore, team decision-making has received no attention in the experimental literature in an institutional setting to date either. While during the last two decades numerous scholars have studied members' institutional choices and self-governance possibilities by letting them vote in experiments (e.g., Gürerk *et al.* 2006; Kosfeld *et al.* 2009; Sutter *et al.* 2010; Ertan *et al.* 2009; Kamei *et al.* 2015; Fehr and Williams 2018), no studies used teams as the decision-making unit (voter). Using individuals as the decision-making unit could be a nice simplification if the following assumption is correct: teams make the same institutional choices as individuals on the condition that the former hold the same information and face the same incentive structure as the latter. However, to the authors' knowledge, there is no research to compare institutional formation and behaviors under the selected institutions between individuals and teams. Moreover, little research has been conducted to study the role of team decision-making in the empirical literature in management and organizations.

This paper demonstrated, for the first time, that teams may be more able than individuals to form efficient institutions by voting and therefore overcome free riding in groups more effectively. In the experiment, decision-making units, teams or individuals, were given a voting opportunity to either construct a formal sanction scheme or to use informal punishment in a public goods dilemma. The results showed that teams achieved surprisingly higher levels of group contributions than individuals in the public goods game. The strong effects of team decision-making were driven by teams' effective use of the sanctioning institutions. When the formal scheme was selected, teams voted to enact deterrent sanction rates much more frequently than individuals. The difference in voting is remarkable: while the majority of individuals in the individual treatments voted for the zero sanction rate, teams voted for deterrent sanction rates more than 50% of the time. When peer-to-peer punishment was instead selected, teams inflicted costly punishment more frequently on low contributors than individuals. These results are consistent with

the “truth wins” hypothesis. This hypothesis explains that teams achieve better choices than individuals through deliberation and learning.

While the results obtained from the present experiment are sufficiently clear, this study is only the first step in researching the individual-team discontinuity effect on institutional choices in dilemma situations. There are many directions for further research. For example, this study set both the team size and group size to three. The sizes of teams and/or groups could be much larger in real organizations, however. The design setup chosen in this study was necessary because with larger team and group sizes the experiment would have been too costly in terms of payment size and the difficulty in implementing the experiment. However, it would be a useful robustness check to study the same research questions by changing the group size and/or team size. For another example, the three team members communicated with each other anonymously, i.e., without being allowed to disclose their identifiable information, to jointly make a single decision in the experiment. This design setup is standard in the current experimental literature (e.g., Charness and Sutter [2012], Kugler *et al.* [2012] and Kerr *et al.* [2004]) and is useful to identify the effects of team decision-making in isolation while controlling for any effect of team composition. In the typical workplace environment (excluding some anonymous online work), however, members of a team are fully or partially aware of the identity of each other. It would therefore be worthwhile studying how the discontinuity-effect phenomenon differs by the anonymity condition within teams. Another important direction of further research is to study possible discontinuity effects when conflicts among team members prevail (e.g., Glaetzle-Ruetzler *et al.*, 2021). The present study assumes for simplicity that the three members in a team received the same payoffs, but there are many real-world situations where team members receive different payoffs. Lastly, of course, the finding of this research also opens up further avenues for theoretical research, for example, in the theory of the firm, as according to the finding of the present experiment, teams, as decision-making units, make different choices through deliberation and learning compared with individuals, even if they face the same incentive structure.

## References


Aboramadan M (2020) Top Management Teams Characteristics and Firms Performance: Literature Review and Avenues for Future Research. *Int. J. Organ. Anal.* 29(3):603-628.

Ahn T.K., Ostrom E, Schmidt D, Shupp R, Walker J (2001) Cooperation in PD games: Fear, greed, and history of play. *Public Choice* 106(1/2): 137-55.

Alchian A, Demsetz H (1972) Production, Information Costs, and Economic Organization. *Am. Econ. Rev.* 62(5): 777-795.

Anderson C, Putterman L (2006) Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games Econ. Behav.* 54(1):1-24.

Andreoni J (1988) Why Free Ride? Strategies and Learning in Public Goods Experiments. *J. Pub. Econ.* 37:291-304.

Appelbaum E, Batt R, 1994. *The New American Workplace: Transforming Work Systems in the United States*, ILR Press, Ithaca, NY.

Bainbridge S (2002) Why a Board? Group Decisionmaking in Corporate Governance. *Vanderbilt Law Rev.* 55(1).

Bednar J, Chen Y, Liu T, Page S (2012) Behavioral Spillovers and Cognitive Load in Multiple Games: an Experimental Study. *Games Econ. Behav.* 74:12-31.

Blinder A, Morgan J (2005) Are Two Heads Better than One? Monetary Policy by Committee. *J. Money Credit Bank.* 37(5):789-811.

Bornstein G, Kugler T, Ziegelmeyer A (2004) Individual and Group Decisions in the Centipede Game: Are Groups More ‘‘Rational’’ Players? *J. Exp. Soc. Psychol.* 40:599-605.

Bornstein G, Yaniv I (1998) Individual and Group Behavior in the Ultimatum Games: Are Groups More “Rational” Players? *Exp. Econ.* 1:101-108.

Bougheas S, Nieboer J, Sefton M (2013) Risk-taking in Social Settings: Group and Peer Effects. *J. Econ. Behav. Organ.* 92:273-283.

Carmeli A, Sheaffer Z, Halevi M Y (2009) Does Participatory Decision-Making in Top Management Teams Enhance Decision Effectiveness and Firm Performance? *Pers. Rev.* 38(6):696-714.

Casari M, Luini L (2009) Cooperation under Alternative Punishment Institutions: An Experiment. *J. Econ. Behav. Organ.* 71(2):273-282.

Cason T, Mui V (1997) A Laboratory Study of Group Polarisation in the Team Dictator Game. *Econ. J.* 107:1465-83.

Cason T, Mui V (2015) Rich communication, social motivations, and coordinated resistance against divide-and-conquer: A laboratory investigation. *Euro. J. Polit. Econ.* 37:146-159.

Cason T, Savikhin A, Sheremeta R (2012) Behavioral Spillovers in Coordination Games. *Eur. Econ. Rev.* 56:233-245.

Cason T, Sheremeta R, Zhang J (2012) Communication and efficiency in competitive coordination games. *Games Econ. Behav.* 76:26-43.

Certo T, Lester R, Dalton C, Dalton D (2006) Top Management Teams, Strategy and Financial Performance: A Meta-Analytic Examination. *J. Manag. Stud.* 43(4):813-839.

Charness G, Sutter M (2012) Groups Make Better Self-Interested Decisions. *J. Econ. Perspect.* 26:157-76.

Chaudhuri A (2011) Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature. *Exp. Econ.* 14(1):47-83.

Chiappori P-A, Mazzocco M (2017) Static and Intertemporal Household Decisions. *J. Econ. Lit.* 55(3): 985-1045.

Cinyabuguma M, Page T, Putterman L (2006) Can second-order punishment deter perverse punishment? *Exp. Econ.* 9:265-279.

Cohen J (1960) A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* 20:37-46.

Cohen S, Bailey D (1997) What Makes Teams Work: Group Effectiveness Research from the Shop Floor to the Executive Suite. *J. Manag.* 23(3): 230-290.

Condorcet M (1785) Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix, *de l'Impr. royalede l'Impr. royale*, Paris.

Cooney R (2004) Empowered self-management and the design of work teams. *Pers. Rev.* 33(6):677-692.

Cooper D, Kagel J (2005) Are two heads better than one? Team versus individual play in signaling games. *Am. Econ. Rev.* 95(3):477-509.

Cooper D, Kagel J (2022) Using Team Discussions to Understand Behavior in Indefinitely Repeated Prisoner's Dilemma Games. Working Paper.

Cox C, Stoddard B (2018) Strategic thinking in public goods games with teams. *J. Pub. Econ.* 161:31-43.

DeGroot M (1974) Reaching a Consensus. *Journal of the American Statistical Association.* 69(345):118-121.

Denant-Boemont L, Masclet D, Noussair C (2007) Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Econ. Theory* 33:154-167.

Delarue A, Van Hootegem G, Procter S, Burridge M (2007) Teamworking and organizational performance: A review of survey-based research. *Int. J. Manag. Rev.* 10(2):127-148.

Devine D, Clayton L, Philips J, Dunford B, Melner S (1999) Teams in Organizations: Prevalence, Characteristics, and Effectiveness. *Small Group Res.* 30(6):678-711.

Elten J, Penczynski S (2020) Coordination games with asymmetric payoffs: An experimental study with intra-group communication. *J. Econ. Behav. Organ.* 169:158-188.

Ertan A, Page T, Putterman L (2009) Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *Eur. Econ. Rev.* 53:495-511.

Eurofound, Cedefop (2020). European Company Survey 2019: Workplace practices unlocking employee potential. *European Company Survey 2019*, Publications Office of the European Union, Luxembourg.

Falkinger J, Fehr E, Gächter S, Winter-Ebmer R (2000) A Simple Mechanism for the Efficient Provision of Public Goods: Experimental Evidence. *Am. Econ. Rev.* 90(1):247-264.

Fehr E, Gächter S (2000) Cooperation and Punishment in Public Goods Experiments. *Am. Econ. Rev.* 90(4):980-994.

Fehr E, Gächter S (2002) Altruistic punishment in humans. *Nature* 415:137-140.

Fehr E, Williams T (2018) Social Norms, Endogenous Sorting and the Culture of Cooperation. University of Zurich Department of Economics Working paper No. 267.

Feri F, Irlenbusch B, Sutter M (2010) Efficiency Gains from Team-based Coordination—Large-Scale Experimental Evidence. *Am. Econ. Rev.* 100:1892-912.

Fischbacher U, Gächter S, Fehr E (2001) Are People Conditionally Cooperative? Evidence from a Public Goods Experiment. *Econ. Lett.* 71(3):397-404.

Fischbacher U, Gächter S (2010) Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments. *Am. Econ. Rev.* 100(1):541-56.

Gächter S, Renner E, Sefton M (2008) The Long-Run Benefits of Punishment. *Science* 322(5907):1510.

Gibbons R, Matouschek N, Roberts J (2013) Decisions in Organizations (Ch. 10) in *The Handbook of Organizational Economics* (ed. by R. Gibbons and J. Roberts), 373-431, Princeton University Press.

Gillet J, Schram A, Sonnemans J (2009) The Tragedy of the Commons Revisited: the Importance of Group Decision-Making. *J. Pub. Econ.* 93:785-797.

Glätzle-Rützler D, Lergetporer P, Sutter M, 2021. Collective intertemporal decisions and heterogeneity in groups. *Games Econ. Behav.* 130:131-147.

Grant R (1996) Toward a Knowledge-based Theory of the Firm. *Strateg. Manag. J.* 17:109-122.

Grosse S, Putterman L, Rockenbach B (2011) Monitoring in Teams: Using Laboratory Experiments to Study a Theory of the Firm. *J. Eur. Econ. Assoc.* 9(4):785-816.

Gunnthorsdottir A, Houser D, McCabe K (2007) Disposition, History and Contributions in Public Goods Experiments. *J. Econ. Behav. Organ.* 62(2):304-15.

Gürerk O, Irlenbusch B, Rockenbach B (2006) The Competitive Advantage of Sanctioning Institutions. *Science* 312(5770):108-111.

Güth W, Levatia V, Sutter M, der Heijden E, 2007. Leading by Example with and without Exclusion Power in Voluntary Contribution Experiments, *J. Pub. Econ.* 91:1023-1042.

Guzzo R, Dickson M (1996) Teams in Organizations: Recent Research on Performance and Effectiveness. *Annu. Rev. Psychol.* 47:307-338.

Hamilton B, Nickerson J, Owan H (2003) Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation. *J. Polit. Econ.* 111(3):465-497.

Hauser O, Rand D, Peysakhovich A, Nowak M (2014) Cooperating with the future. *Nature* 511:220-223.

He Y, Lien J, Zheng J (2022) Making Use of the Wisdom of Crowds: Stuck in the Majority Rule. SSRN Working Paper.

Herrmann B, Thöni C, Gächter S (2008) Antisocial Punishment across Societies. *Science* 319:1362-1367.

Holmstrom B (1982) Moral Hazard in Teams. *Bell J. Econ.* 13(2):324-340.

Ichniowski C, Shaw K, Prennushi G (1997) The Effects of Human Resource Management Practices on Productivity: A Study of Steel Finishing Lines. *Am. Econ. Rev.* 87:291-313.

Isenberg D (1986) Group Polarization: A Critical Review and Meta-Analysis. *J. Pers. Soc. Psychol.* 50(6): 1141-1151.

Kagel J (2018) Cooperation through Communication: Teams and Individuals in Finitely Repeated Prisoners' Dilemma Games. *J. Econ. Behav. Organ.* 146:55-64.

Kagel J, McGee P (2016) Team versus Individual Play in Finitely Repeated Prisoner Dilemma Games. *Am. Econ. J. Micro* 8(2):253-76.

Kamei K (2014) Conditional Punishment. *Econ. Lett.* 124(2):199-202.

Kamei K (2016) Democracy and Resilient Pro-social Behavioral Change: An Experimental Study. *Soc. Choice Welf.* 47(2):359-378.

Kamei K (2019a) Cooperation and Endogenous Repetition in an Infinitely Repeated Social Dilemma. *Int. J. Game Theory* 48(3):797-834.

Kamei K (2019b) The Power of Joint Decision-Making in a Finitely-Repeated Dilemma. *Oxford Econ. Pap.* 71(3):600-622.

Kamei K (2021) Teams do Inflict Costly Third Party Punishment as Individuals do: Experimental Evidence. *Games* 12(1):22.

Kamei K, Putterman L (2015) In Broad Daylight: Fuller Information and Higher-Order Punishment Opportunities can Promote Cooperation. *J. Econ. Behav. Organ.* 120:145-159.

Kamei K, Putterman L, Tyran J-R (2015) State or Nature? Endogenous Formal versus Informal Sanctions in the Voluntary Provision of Public Goods. *Exp. Econ.* 18:38-65.

Kerr N, Tindale S (2004) Group Performance and Decision Making. *Annu. Rev. Psychol.* 55:623-655.

Kersley B, Alpin C, Forth J, Bryson A, Bewley H, Dix G, Oxenbridge S (2005) Inside the Workplace: Findings from the 2004 Workplace Employment Relations Survey. WERS.

Kocher M, Sutter M (2005) The Decision Maker Matters: Individual versus Group Behavior in Experimental Beauty-Contest Games. *Econ. J.* 115:200-223.

Kosfeld M, Okada A, Riedl A (2009) Institution Formation in Public Goods Games. *Am. Econ. Rev.* 99(4):1335-55.

Kugler T, Kausel E, Kocher M (2012) Are Groups More Rational than Individuals? A Review of Interactive Decision Making in Groups. *WIREs Cognitive Science* 3:471-482.

Kugler, T, Bornstein G, Kocher M, Sutter M (2007) Trust between Individuals and Groups: Groups are Less Trusting than Individuals but Just as Trustworthy. *J. Econ. Psych.* 28:646-57.

Landis R, Koch G (1977) An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. *Biometrics* 33(2):363-74.

Laughlin, PR, (2015) Social Combination Processes of Cooperative Problem-Solving Groups. *Progress in social psychology*. Psychology Press.

Lawler E, Mohrman S A, Ledford G (1992) *Employee Involvement and Total Quality Management: Practices and Results in Fortune 1000 Companies*. San Francisco: Jossey-Bass.

Lawler E, Mohrman S A, Ledford G (1995) *Creating High Performance Organizations: Impact of Employee Involvement and Total Quality Management*. San Francisco: Jossey-Bass.

Ledyard J (1995) Public Goods: A Survey of Experimental Research, pages 111-194 in J. Kagel and A. Roth (eds.), *Handbook of Experimental Economics*. Princeton University Press.

Leibbrandt A, Sääksvuori L (2012) Communication in Intergroup Conflicts. *Eur. Econ. Rev.* 56(6):1136-47.

Marschak J, Radner R (1972) *Economic Theory of Teams*. New Haven: Yale University Press.

McLachlan G, Peel D (2000) *Finite Mixture Models*. New York: Wiley.

Moffatt P (2016) *Experimetrics: Econometrics for experimental economics*, Macmillan International Higher Education.

Müller W, Tan F (2013) Who Acts More Like a Game Theorist? Group and Individual Play in a Sequential Market Game and the Effect of the Time Horizon. *Games Econ. Behav.* 82:658-74.

Nicklisch A, Putterman L, Thöni C (2021) Trigger-Happy or Precisionist? On Demand for Monitoring in Peer-based Public Goods Provision. *J. Pub. Econ.* 200:104429.

Nikiforakis N, Normann H-T (2008) A Comparative Statics Analysis of Punishment in Public-Good Experiments. *Exp. Econ.* 11:358-369.

Ostrom E, Walker J, Gardner R (1992) Covenants With and Without a Sword: Self-Governance is Possible. *Am. Polit. Sci. Rev.* 86(2):404-417.

Pfeffer J (1998) Seven Practices of Successful Organizations. *Calif. Manage. Rev.* 40(2):96-124.

Robert C, Carnevale P (1997) Group Choice in Ultimatum Bargaining. *Organ. Behav. Hum. Decis. Process.* 72:256-279.

Sunstein C (2007) *Republic.com 2.0*, Princeton University Press.

Sutter M (2007) Are Teams prone to myopic loss aversion? An Experimental Study on Individual Versus Team Investment Behavior. *Econ. Lett.* 97:128-132.

Sutter M (2009) Individual Behavior and Group Membership: Comment. *Am. Econ. Rev.* 99:2247-2257.

Sutter M, Haigner S, Kocher M (2010) Choosing the Carrot or the Stick? – Endogenous Institutional Choice in Social Dilemma Situations. *Rev. Econ. Stud.* 77(4):1540-1566.

Traulsen A, Röhl T, Milinskim M (2012) An Economic Experiment Reveals that Humans Prefer Pool Punishment to Maintain the Commons. *Proc. R. Soc. B: Biol. Sci.* 279:3716-3721.

Tyran J-R, Feld L (2006) Achieving Compliance when Legal Sanctions are Non-deterrent. *Scan. J. Econ.* 108(1):135-156.

Zhang B, Li C, Silva H, Bednarik P, Sigmund K (2014) The Evolution of Sanctioning Institutions: an Experimental Approach to the Social Contract. *Exp. Econ.* 17(2):285-303.